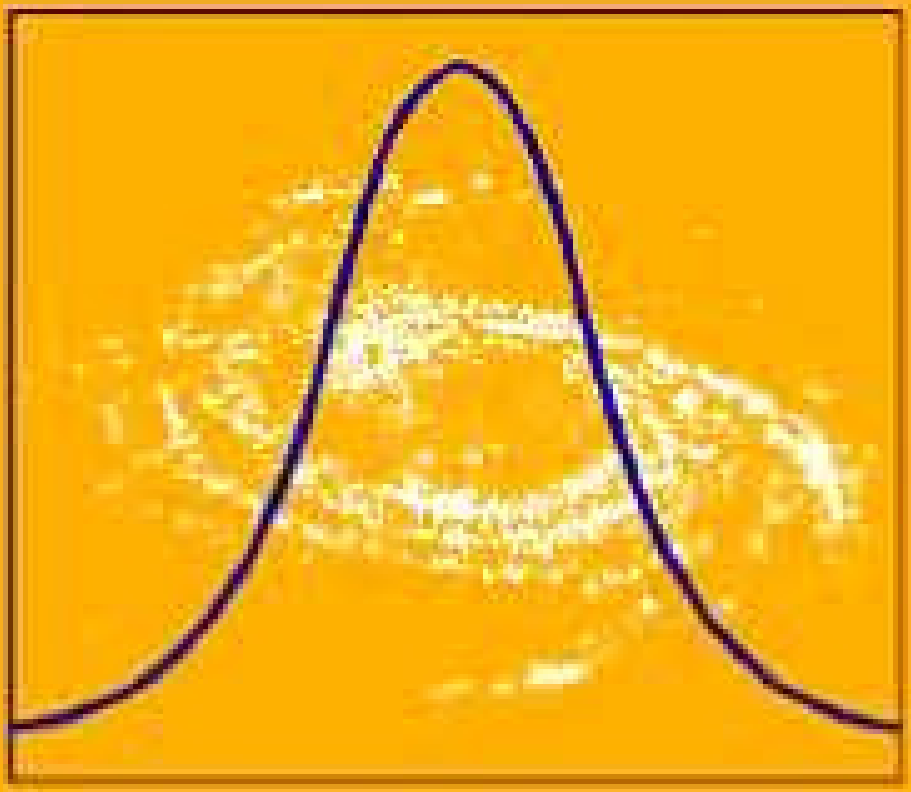
Eric D. Feigelson / G. Jogesh Babu

Statistical Challenges in Astronomy



Springer

Statistical Challenges in Astronomy

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

Eric D. Feigelson G. Jogesh Babu
Editors

Statistical Challenges in Astronomy

With 104 Illustrations

Springer

Eric D. Feigelson
Department of Astronomy
and Astrophysics
Pennsylvania State University
University Park, PA 16802
USA
edf@astro.psu.edu

G. Jogesh Babu
Department of Statistics
Pennsylvania State University
University Park, PA, 16802
USA
babu@stat.psu.edu

Cover art: Conference logo of the cross-disciplinary conference, "Statistical Challenges in Modern Astronomy," held on August 11–14, 1991, at the University Park campus of Pennsylvania State University.

PACS: 95.75/MSC: 62P35

Library of Congress Cataloging-in-Publication Data
Statistical challenges in astronomy / editors, Eric D. Feigelson, G. Jogesh Babu.
p. cm.
Includes bibliographical references and index.
ISBN 0-387-95546-1 (alk. paper)
1. Statistical astronomy—Congresses. I. Title: Statisical challenges in astronomy.
II. Feigelson, Eric D. III. Babu, Gutti Jogesh, 1949–
QB149 .S75 2002
520′.7′2—dc21 2002026661

ISBN 0-387-95546-1 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10886440

Typesetting: Pages created by the authors using a Springer TEX macro package.

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Preface

The third *Statistical Challenges in Modern Astronomy* (SCMA III) conference was held at Penn State University during July 18-21 2001. The SCMA conferences are intended to bring together scholars in two communities that have much in common yet relatively little contact with each other. Astronomers are acquiring enormous (terabyte or larger) datasets that require sophisticated processing and modeling to arrive at important astrophysical conclusions. Great advances have similarly occurred in the development of statistical methodologies in recent decades. The vibrant atmosphere of the SCMA III conference supports our belief that powerful, mutually beneficial synergisms can emerge when astronomers and statisticians get together do discuss astrostatistical problems and approaches. SCMA conferences are designed to foster cross-disciplinary interaction – talks by scholars in one field are followed by commentaries by scholars in the other field. We are extremely grateful to the invited speakers for preparing their talks in advance of the conference to facilitate this valuable cross-talk.

The conference was kicked off by an historical overview by Virginia Trimble and four extremely useful interactive tutorials: Robin Ciardullo and Joe Bredekamp introduced statisticians to basic cosmology and NASA accomplishments, while Steve Arnold and Alanna Connors introduced astronomers to the principles and practice of Bayesian statistics.

The first research session continued with Bayesian strategies for astrophysical modeling astronomical data. Eric Kolaczyk provided a valuable overview of Bayesian methods for Poissonian data, Tom Loredo showed how to plan astronomical observations with optimal efficiency, David van Dyk explained sophisticated nested models to deal with instrumental and Poissonian effects, and Jim Berger provided a convincing analysis of a nonlinear modeling problem.

The rapid growth of astronomical data sets and archives were presented by Joe Bredekamp. George Djorgovski presented plans for the federation of such databases into a vast Virtual Observatory during the next decade. An early glimpse at this database-rich future was provided by Michael Strauss' talk on the Sloan Digital Sky Survey.

The conference then delved into its principal theme: statistical methodologies for modeling fundamental characteristics of the Universe on its largest scales. The first of these cosmological issues is the large-scale structure (LSS) in the Universe; the nonlinear, anisotropic clustering of galaxies in 3-dimenaional space. Vicent Martínez set the stage on the rapid progress

in this field, and specific issues were then developed by two speakers: Alex Szalay on the exciting new results from the Sloan Digital Sky Survey using the Karhunen-Loeve transform; and Rien van de Weygaert on an statistical approach involving tessellations. The second major cosmological issue is the modeling of fluctuations of the cosmic microwave background (CMB). Bayesian, frequentist and nonparametric approaches to CMB studies were presented by Andrew Jaffe, Chad Schafer and Larry Wasserman, respectively.

The next session investigated statistical methodologies for studying the clustering of points in p-dimensional space. This could either be galaxy clustering in 3-space, or any multivariate study of a population in multidimensional parameter space. Three distinguished statisticians introduced astronomers to recent advances in this area: Leo Breiman on decision tree methods, Adrian Raftery on Bayesian clustering methods, and Fionn Murtagh on very-high-dimensionality problems. Dianne Cook showed astronomers developments in data visualization tools, and Bob Nichol presented new computational tools for clustering very large datasets.

After this deep immersion in cosmology, the conferees turned to some practical issues in the daily challenges of astronomical data analysis. Jeff Scargle provided a profound perspective on Bayesian signal detection in both image and time series analysis. Larry Bretthorst placed a major tool, the Lomb-Scargle periodogram for unevenly spaced data, upon a general mathematical footing. We heard from Jean-Luc Starck, Iain Johnstone and Peter Freeman on advances in wavelet analysis, methods that simultaneously treat structure on many scales.

The conference was closed with thoughtful comments by two distinguished leaders, Berkeley statistician John Rice and Oxford astrophysicist Joe Silk. A strong feeling that such astro-statistical interactions are necessary and fruitful for the enrichment of the two fields.

In addition to the invited speakers and discussants, several dozen scientists from many countries presented contributed papers. Many of these are briefly summarized in the final portion of this volume. We thank all participants for their labor on this cross-disciplinary frontier.

Acknowledgments

The success of the SCMA III conference rests on many shoulders. First, we thank the Scientific Organizing Committee for its thoughtful work many months earlier in formulating the goals and recommending speakers for the conference. S.O.C. members consisted of four statisticians – G. Jogesh Babu, Fionn Murtagh, John Rice, and David van Dyk – and four astronomers – Eric D. Feigelson, Alanna Connors, Robert C. Nichol, and Jeffrey Scargle.

Penn State conference planner Rachel Graham and her assistant Fawn Hosterman made all logistical arrangements. The staff of the Penn Stater Conference Hotel also did well. Statistics graduate students James McDermott and Hyunsook Lee provided able assistance in many respects. Tom von Förster from Springer-Verlag warmly supported publications of these Proceedings. The Institute of Mathematical Statistics cosponsored the conference and promoted it in the *IMS Bulletin*.

We particularly would like to thank the several organizations that provided financial support for the conference: the National Science Foundation (Division of Mathematical Sciences), National Aeronautical & Space Administration (Applied Information Systems Research Program), and several branches of the Pennsylvania State University (Department of Statistics, Department of Astronomy & Astrophysics, Eberly College of Science, and Division of Continuing Education), and the Pennsylvania Space Grant Consortium. Their funds assisted many participants, from distinguished professors to graduate students in both fields, with the costs of attending the conference.

Eric D. Feigelson
G. Jogesh Babu

Pennsylvania State University
October 2002

This page intentionally left blank

Contents

This page intentionally left blank

Contributors

Arnaud, Dr. Keith A., Code 662, Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, USA MD 20771, USA

Arnold, Prof. Steven F., Department of Statistics, 313 Thomas Building, Pennsylvania State University, University Park, PA 16802, USA

Babu, Prof. G. Jogesh, Department of Statistics, 319 Thomas Building, Pennsylvania State University, University Park, PA 16802, USA

Barrera, Ruth S., Observatorio Astronmico Nacional, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá D.C., Colombia

Bazell, Dr. David, Eureka Scientific, Inc., 6509 Evensong Mews, Columbia, MD 21044-6064, USA

Berger, Prof. James O., Institute of Statistics and Decision Sciences, Old Chemistry Building, Duke University, Durham, NC 27708-0251, USA

Bernstein, Joseph, Department of Astronomy, University of Michigan, P.O. Box 4121, Ann Arbor, MI 48106, USA

Bhavsar, Prof. Suketu P., Department of Physics and Astronomy, University of Kentucky, 277 Chem/Phys Building, Lexington, KY 40506-0055, USA

Bissantz, Dr. Nicolai B., Astronomisches Institut der Universitt Basel, Venusstrasse 7, CH-4102 Binningen, Switzerland

Bolton, Prof. Charles T., David Dunlap Observatory, University of Toronto, P.O. Box 360, Richmond Hill Ontario L4C 4Y6 Canada

Bredekamp, Dr. Joseph H., Code SS, Office of Space Science, NASA Headquarters, 300 E Street SW, Washington D.C., 20546, USA

Breiman, Prof. Leo, Department of Statistics, 367 Evans Hall, University of California at Berkeley, Berkeley, CA 94720-3860, USA

Bretthorst, Dr. G. Larry, School of Medicine, Campus Box 8227, Washington University, St. Louis, MO 63110, USA

Chu, Prof. Yaoquan, Center for Astrophysics, University of Science and Technology of China, 96 Jinshai Road, Hefei Anhui 230026 China

Ciardullo, Prof. Robin, Department of Astronomy and Astrophysics, Davey Laboratory, Pennsylvania State University, University Park, PA 16802, USA

Connors, Dr. Alanna, Eureka Scientific Inc., 46 Park Street, Arlington, MA 02474, USA

Cook, Prof. Dianne H., Department of Statistics, 325 Snedecor Hall, Iowa State University, Ames, IA 50014, USA

Costello, Dr. Dawn M., Rutgers University, Bristol-Myers Squibb, One Squibb Drive, New Brunswick, NJ 08903, USA

Datta, Dr. Srabani, Department of Applied Mathematics, University of Calcutta, 92 A.P.C. Road, Calcutta, West Bengal 700009, India

de Diego, Dr. Jose A., Instituto de Astronomia, Universidad Nacional Autónoma de Mxico, Apdo Postal 70-264, Mexico D.F., 04510, Mxico

Del Popolo, Dr. Antonio, Catania Astrophysical Observatory, Via S. Sofia 78, I-95125 Catania, Italy

Derriere, Dr. Sebastien, Observatoire Astronomique de Strasbourg, UMR 7550, 11 rue de l'Universit, F-67000 Strasbourg, France

Djorgovski, Prof. S. George, Department of Astronomy, California Institute of Technology, MS 105-24, Pasadena, CA 91125, USA

Doroshkevich, Dr. Andrei, Theoretical Astrophysics Center, Juliane Maries Vej 30, DK-2100 Copenhagen 0, Denmark

Esch, David N., Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Eyheramendy, Susana, Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA

Feigelson, Prof. Eric D., Department of Astronomy and Astrophysics, Pennsylvania State University, 525 Davey Lab, University Park, PA 16802, USA

Fioc, Dr. Michel, Institut d'astrophysique de Paris 98 bis boulevard Arago, F-75014 Paris, France

Freeman, Dr. Peter E., Harvard-Smithsonian Center for Astroastrophysics, MS 81, 60 Garden Street, Cambridge, MA 02138, USA

Galloway, Dr. Duncan K., Center for Space Research, Massachusetts Institute of Technology, 37-571, Cambridge, MA 02139, USA

Genovese, Prof. Christopher, Department of Statistics, Carnegie Mellon University, Baker Hall 132, Pittsburgh, PA 15213, USA

Greenhough, John, University of Warwick, Tocil Flat 9, Coventry CV47AL United Kingdom

Gregory, Prof. Philip C., Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver B.C. V6T 1Z1 Canada

Hans, Christopher M., Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Harel, Ofer, Department of Statistics, 326 Thomas Building, Pennsylvania State University, University Park, PA 16802, USA

Hatziminaoglou, Dr. Evanthia, European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching b München, Germany

Hawkins, Dr. George W., American Association of Variable Star Observers, 25 Birch Street, Cambridge, MA 02138, USA

Heacox, Prof. William D., University of Hawaii at Hilo, 200 W. Kawili Street, Hilo, HI 96720-4091, USA

Ho, Dr. Tin Kam, Bell Laboratories, Lucent Technologies, 700 Mountain Avenue, 2C425, Murray Hill, NJ 07974, USA

Horvath, Dr. Istvan, Department of Physics, BJKMF, Box 12, H-1456 Budapest, Hungary

Indradjaja, Baju, Department of Astronomy, Institute of Technology Bandung, JL Ganesha 10, Bandung West Java 40132, Indonesia

Iskander, Ed M., Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

Jaffe, Dr. Andrew H., Department of Astronomy, University of California at Berkeley, 601 Campbell Hall, Berkeley, CA 94720, USA

Jang, Woncheol, Department of Statistics, Carnegie Mellon University, 229J Baker Hall, Pittsburgh, PA 15213, USA

Jefferys, Prof. William H., Department of Astronomy, University of Texas, Austin, TX 78703, USA

Johnstone, Prof. Iain M., Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305, USA

Jordan, Andres, Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, NJ 08854-8019, USA

Kang, Hosung, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Kashyap, Dr. Vinay L., Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA

Kaspi, Dr. Shai, Department of Astronomy and Astrophysics, 525 Davey Laboratory, Pennsylvania State University, University Park, PA 16802, USA

Kestens, Dr. Elke, Universitair Centrum voor Statistiek, Katholieke Universiteit Leuven, Celestijnenlaan 200 B, B-3001 Leuven, Belgium

Kolaczyk, Prof. Eric D., Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, USA

Lacerda, Pedro, Leiden Observatory, University of Leiden, Post Bvs 9513, Leiden 2300RA, Netherlands

Laidler, Dr. Victoria G., Computer Sciences Corporation, Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

Lee, Hyunsook, Department of Statistics, 326 Thomas Building, Pennsylvania State University, University Park, PA 16802, USA

Lee, Dr. Jeongin, United States Naval Observatory, 3450 Massachusetts Avenue NW, Washington, DC 20392, USA

Lehto, Prof. Harry J., Tuorla Observatory, University of Turku, Vaisalantie 20, FIN-21500 Piikkio, Finland

Lindler, Vice Pres. Don J., Advanced Compuer Concepts, Inc., 11518 Gainsborough Road, Potomac, MD 20854, USA

Loh, Ji Meng, Department of Statistics, Columbia University, 618 Mathematics Building, 2990 Broadway, New York, NY 10027, USA

Loredo, Dr. Thomas, Department of Astronomy, Cornell University, Space Sciences Building, Ithaca, NY 14853-6801, USA

Madore, Dr. Barry F., Infrared Processing and Analysis Center, California Institute of Technology, 770 S. Wilson, Pasadena, CA 91125, USA

Martínez, Prof. Vicent J., Observatori Astronòmic, Universitat de València, Avda Vicent Andres Estelles, E-46100 Burjassot, València, Spain

Martínez-Gonzalez, Dr. Enrique, Instituto de Fisica de Cantabria, Facultad de Ciencias, Avinida de los Castros s/n, 39005 Santander, Spain

Matsubara, Prof. Takahiko, Department of Physics and Astrophysics, Nagoya University, Furo-cho Chikusa-ku, Nagoya 464-8602, Japan

Mattei, Director Janet A., American Association of Variable Star Observers, 25 Birch Street, Cambridge, MA 02138, USA

McDermott, James P., Department of Statistics, 326 Thomas Building, Pennsylvania State University, University Park, PA 16802, USA

Miller, Christopher, Department of Physics, Carnegie Mellon University, 7325 Wean Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA

Mohanty, Dr. Soumya D., Albert Einstein Institute, Max-Planck-Institut fuer Gravitationsphysik, AM Muehlenberg 1, Golm D-14476, Germany

Morrison, Prof. Nancy D., Department of Physics and Astronomy, University of Toledo, 2801 West Bancroft Street, Toledo, OH 43606, USA

Mukherjee, Dr. Soma, Albert Einstein Institute, Max-Planck-Institut fuer Gravitationsphysik, AM Muehlenberg 1, Golm D-14476, Germany

Murtagh, Prof. Fionn D., School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, Ireland

Nichol, Prof. Robert C., Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA

Ombao, Prof. Hernando C., Department of Statistics, University of Pittsburgh, 2706 Cathedral of Learning, Pittsburgh, PA 15260, USA

Peterson, John R., Department of Astronomy, Columbia University, 550 W. 120th Street, New York, NY 10032, USA

Philip, Prof. Ninan, Department of Physics, Cochin University of Science and Technology, Cochin, Kerala 682 022, India

Phillips, Dr. Nicholas G., Code 685, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

Pons-Bordería, Dr. Maria J., Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, Pasco Alfonso XIII 52, Cartagena 30203, Spain

Raftery, Prof. Adrian E., Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA

Rahman, M. Nurur, Department of Physics and Astronomy, University of Kansas, Lawrence, KS 66045-2151, USA

Räth, Dr. Christoph C.W., Max-Planck-Institut für Extraterrestrische Physik, Giessenbachstrasse 1, Garching D-85748, Germany

Rice, Prof. John A., Department of Statistics, University of California at Berkeley, 367 Evans Hall, Berkeley, CA 94720, USA

Scargle, Dr. Jeffrey D., NASA Ames Research Center, Mail Stop 245-3, Moffett Field, CA 94035-1000, USA

Schaap, Willem E., Kapteyn Astronomical Institute, University of Groningen, P. O. Box 800, Groningen 9700, Netherlands

Schafer, Chad M., Department of Statistics, University of California at Berkeley, 367 Evans Hall, Berkeley, CA 94720, USA

Schmidt, Dr. Jason D., Starfire Optical Range, Air Force Research Lab, 3550 Aberdeen Avenue, Kirtland AFB, NM 87117, USA

Silk, Prof. Joseph, Astrophysics, Oxford University Jubilee Terrace, Oxford 0X1 4LM, United Kingdom

Souchay, Dr. Jean, Observatoire de Paris, 61 Avenue de l'Observatoire, Paris 75014, France

Sourlas, Epaminondas V., Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Spencer, Kate A., School of Physics and Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom

Starck, Dr. Jean-Luc, CES-Saclay, Service d'Astrophysique, 91191 Gif-sur-Yvette, France

Strauss, Prof. Michael A., Princeton University Observatory, Princeton University, Peyton Hall, Ivy Lane, Princeton, NJ 08544, USA

Szalay, Prof. Alex, Department of Physics and Astronomy, Johns Hopkins University, 3701 San Martin Drive, Baltimore, MD 21218, USA

Tenorio, Prof. Luis, Mathematical and Computer Sciences, Colorado School of Mines, Golden C) 80401

Trimble, Prof. Virginia L., Department of Physics and Astronomy, University of California at Irvine, 4129 Reines Hall, Irvine, CA 92697-4575, USA

Uribe-Botero, Jose-Antonio, Observatorio Astronomico Nacional, Calle 43 No. 24-27, Apto. 201, Bogotá, Colombia

Valtchanov, Ivan A., Service d'Astrophysique, Centre d'Études de Saclay, Orme des Merisiers, Batiment 709, Gif-sur-Yvette 91191, France

van de Weygaert, Dr. Rien, Kapteyn Institute of Astronomy, University of Groningen, P.O. Box 800, Groningen 9700AV, Netherlands

van Dyk, Prof. David A., Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Vavrek, Roland, Konkoly Observatory, P.O. Box 67, Budapest H 1525, Hungary

Verde, Dr. Licia, Princeton University Observatory, Princeton University, Peyton Hall, Ivy Lane, Princeton, NJ 08544, USA

Wasserman, Prof. Larry A., Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Watts, Dr. Peter I. R., School of Physics and Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom

Wells, Dr. Donald C., National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, VA 22903-2475, USA

Yasuda, Naoki, National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitoka Tokyo 181-8588, Japan

Young, Dr. C Alex, NASA Goddard Space Flight Center, Code 682.3, Greenbelt, MD 20771, USA

Zhu, Dr. Xingfen, Centre for Astrophysics, University of Sciences and Technology of China, 96 Jinshai Road, Hefei Anhui 230026, China

1

Statistical Challenge in Medieval (and Later) Astronomy

Virginia Trimble[1]

ABSTRACT Portions of the history of the interaction between astronomy and statistics are told in the form of short case studies of a number of people who appear (or should appear) in books about both. These should be regarded as notes for a serious discussion of the subject, not the discussion itself.

In memory of Peter August Georg Scheuer from whom I (and many others) first heard that $N^{\frac{1}{2}}$ is sometimes signal rather than noise.

1.1 A demographic introduction

If one is going to explore the contributions of astronomers to statistics and of statisticians to astronomy, one ought perhaps to begin by deciding what is meant by as astronomer, a statistician, and statistics. I will not do so, and merely call attention to the cases, first, of Roger Boscovich of Dubrovnik, who rates a whole section in Hald (1998) for extending the method of least absolute deviations beyond where it had been left by Galileo for application to astronomical observations of latitude but is known only to the subsets of astronomers who collect foreign paper money or speak Serbo-Croatian (in which his name is spelled - and pronounced - quite differently) and, second, of John Michell, who appears in lots of astronomy treatises for inventing the concept of black holes, and, occasionally, for the discovery of binary stars, but does not make it into the statistics histories of Hald (1998), Stigler (1986), or Pearson (1976), despite his binary task having been accomplished by a method that most of us would call both statistical (certainly probabilistic) and innovative.

How large is the overlap between the two communities? Of the 76 astronomers indexed in Abell (1982) who flourished from ancient times up to

[1]Department of Physics and Astronomy, University of California, Irvine, CA 92697, and Astronomy Department, University of Maryland, College Park, MD 20742

about 1850, 49 (from Airy to Zach) appear in one or more of the statistical histories by Stigler (1986, 1999), Hald (1990, 1998), Pearson (1976) and Franklin (2001). They, in turn, mention another 27 astronomers (Arago to Thomas Young) who did not make the Abell cut but who are mentioned in Russell, Dugan, and Stewart (1926), in Hoskin (1999) or some other reasonable place. RDS was the primary introduction astronomy text in English for about 20 years. George Abell wrote the first of the now-ubiquitous books for non-science major courses, with the 4th, 1982, edition the last over which he had control. And Hoskin's volume is the most recent attempt to put the entire history of astronomy between two covers.

Closer to the present, scientists become more and more specialized but in the period from 1850 to 1950 at least the following can reasonably be described as having contributed to the astronomy/statistics interface: Simon Newcomb (1886), Arthur Eddington (1914), and Harold Jeffreys (1939), noted by Hampel (1998), who also regards the work of Cannon, Fleming, and Leavitt as statistical in nature, Jacobus Kapteyn (1922, ending his 40+ years of work on the topic), Jerzy Neyman and Elizabeth Scott (1956), W.M. Smart (...), S. Chandrasekhar (1939), Gunnar Malmquist (1920, 1924), Col. Frank J.M. Stratton, and Robert Trumpler and Harold Weaver (1953).

The sign of the contribution is not always clear. Consider the case of Stratton, who was the last person to have participated officially in every general assembly of the International Astronomical Union and who was one of the officers who held the Union together during the very difficult 1939-1945 period, but whose astronomical work most of us would be hard-pressed to recall. He was also the Cambridge tutor of Ronald Fisher (of the F-distribution and much else), and I cannot resist quoting the following from Hald (1998):

> *The astronomer F.J.M. Stratton (1881-1960), who was Fisher's tutor, lectured occasionally on the theory of errors. We do not know precisely the contents of his course, but in the preface to a book by D. Brunt (1917), the author thanks Stratton, "to whose University lectures I owe most of my knowledge of the subjects discussed in this book, and upon whose notes I have drawn freely." There is nothing original in this book.*

Not knowing Hald, I cannot be sure whether he means this to be as mirth-provoking as it is. Stigler (1999), on the other hand, clearly means to amuse as well as to enlighten when he includes in a section called "Questions to Discovery" a chapter entitled "Who discovered Bayes' Theorem?", one called "Daniel Bernoulli, Leonhard Euler, and Maximum Likelihood" (to which a local wit responded, "Oh, yeah. Old Max. He used to drink a lot."), and one called "Gauss and the invention of least squares." The issue of which items in astro-statistics and statistico-astronomy should be called

discoveries and which inventions is another issue that I will not resolve here. Indeed I will say nothing about Gauss and least squares, since his contributions, the antecedents, and descendants were so well explored by Rao (1997) in SCMA II.

What will appear in the rest of this paper is a series of case studies, of what strike me as fruitful interactions between the fields. None is precisely medieval (how sure are we that the number of cardinal sins falls between 4.65 and 9.35?), though some archeoastronomy items appear at the end. Just how many of the tales get told will depend on the editor, who will remove as many as necessary to get below the assigned page limit.

1.2 Giants in the Land

These stories concern scientists of enormous reputation over a range of disciplines, and I have not consulted the original literature, but retell from Franklin (2001), Hald (9190, 1998), Hoskin (9199), Stigler (1986), Pearson (1976), and other sources read too long ago to be honorably recalled.

1.2.1 Galileo and Least Absolute Deviations

In the simplified version of history we hand our students while they are getting settled into their seats at the beginning of a lecture, the Aristotelian-Aquinian principle of "the immutability of the heavens" was overthrown by Tycho Brahe (1546-1601), who set an upper limit to the geocentric parallax of his nova stellar of 1572 (and also the comet of 1577) placing them beyond the sphere of the moon. But, not surprisingly, he was not the only astronomer of the time to look for this parallax. Incidentally, seeing the geometry of it is rather tricky for modern eyes, but it is a true statement that the new star, if it is close to the earth and turning in the diurnal motion about the pole, will show itself more distant from the pole when it is below the pole on the meridian than when above it (roughly Galileo's words). A certain Scipione Chiaramonti (1565-1652) combined some of Tycho's observations with those of 11 other astronomers to conclude that what we now call NS 1572 was at most 32 earth radii away, with similar conclusions for SB 1604 and the nova stella of 1600 (actually Mira).

This provoked Galileo (1564-1642) in his 1632 **Dialogo** to look again at all the reported measurements of upper and lower culmination altitudes of the 1572 star made by astronomers at latitudes from 38.5 to 56^o north. That is, he is looking for geocentric parallax over a fairly small baseline rather than for earth rotation parallax which can be measured by a single observer and, for circumpolar stars (as SN 1572 was for Tycho) has a baseline of 2 R_e cos (latitude).

Galileo then compared the sums of the absolute values of the errors of the

observations implied if the distance was 32 R_e vs. sufficiently large to yield no parallax. Of the more than 100 pairings of the data points available, Galileo picked 10 most favorable to Chiaramonti's hypothesis and 10 most favorable to his (with no overlap). The sums of the absolute errors in the two cases were 756.9 arc minutes and 83.7 arc minutes respectively. Small being good in this context, he regarded the result as being strong evidence for a translunar location for the event. And so do we.

The method then languished until 1755, when Boscovich applied it to the determination of the lengths of arcs of latitude at various locations (in connection with the problem of determining the shape of the earth - prolate had been claimed). Galileo was also the first to figure out the odds of getting various outcomes from the case of three dice. I checked his numbers by writing down all the combinations, which is presumably how he did it. He got it right, and it is left as an exercise for the Gosset to figure out how the results would change in the case of fermionic or bosonic, rather than distinguishable, dice.

1.2.2 Edmond Halley and survival rates

Halley (1656-1742) is known to astronomers best for his prediction of the return of the period comet now bearing his name. On the astronomical side, he also discovered proper motions of the stars and secular acceleration of the moon, accurately predicted the path of the eclipse of 1715 over England, and served as Astronomer Royal from 1720 until his death (succeeding Flamsteed, who was first).

But he also wrote, in 1694, "... on... degrees of mortality... and prices of annuities." The end of the title makes clear why men of practical bent were concerned with human survival and death rates as a function of age. His work in this area is an interesting illustration of what our grandmothers called "making do with what you have." Since it was English annuities for which he was trying to set a fair price (or anyhow one that people would pay an that would not bankrupt the issuers), he would obviously have liked to have rates of the deaths of English persons (not just men, since annuities were often purchases for widows) as a function of age. But the methods of recording births and deaths in England, mostly in parish registers, did not provide the numbers needed, so he used tables of numbers of births an deaths and total population for Breslau.

According to Pearson (1976), Halley was probably also the author of a 1699 piece in **Philosophical Transactions of the Royal Society** (the first scientific periodical in any language, in case you wondered) called "a calculation of the credibility of human testimony." This is also phrased in the language of how much you should be willing to pay for things. For instance, of someone who is 90% reliable tells you that he has seen your cargo ship safely into the harbor and unloaded without damage, then you should be prepared to pay (only) 10% of the value of the cargo to insure against

the loss of the whole. The paper does not address how you determine the reliability quotient of your informant, which is the aspect of the problem on which we most often stumble even today, whether the issue is astronomical or financial.

1.2.3 Tobias Mayer and the libration of the Moon

Mayer (1723-1762) tackled a problem whose geometry is even more difficult to see than that of geocentric parallax and solved it, using a method (called Mayer's or, more often, the method of averages) that would elude Euler working on the mathematically rather similar problem of mutual perturbations in the orbital motion of Jupiter and Saturn. Mayer's goal (connected with the use of lunar motion for longitude determination) was to find three angles: the one between the true rotation axis of the Moon and the poles of the circumference parallel to the ecliptic, the ecliptic longitude of the node at which the plane of the lunar rotation equator crosses the ecliptic, and the true latitude of the crater (Manilius, a suitable choice in several ways) he had observed. The observations were 27 pairs of angular positions of the crater parallel and perpendicular to (changing) apparent equator of the moon (the circumference parallel to the ecliptic), gathered by him over a couple of years.

Thus he had 27 equations in three unknown. His solution was to group these in three sets, with large, medium, and small (negative) coefficients of the first angle mentioned above, which he regarded as the most important. He then added up the groups (he could alternatively have averaged them) and solved the resulting triple, concluding that the result would be more accurate than that from any three data pairs alone (true) by a factor nine (false; it is at best three if only random errors in the observations are important). He apparently invented $\pm$ as well.

Euler, writing in 1749 (the year before Mayer) was faced with 75 sets of observations of Saturn and Jupiter, gathered over 163 years, from which to extract eight unknown describing the orbits and their interactions. He pulled out the two that were not periodic in the 59-year synodic period of the two planets, and then ground to a halt, when various combinations of the equations let to wildly inconsistent results, saying that the errors had multiplied themselves through combining of observations. Nevertheless, most of us have heard of Euler, and few of Mayer. Indeed, Stigler (1986) notes that the method of averaging (or summing) equations discovered by Mayer is often attributed to Euler. His section heading is, of course, **Saturn, Jupiter, and Euler**.

1.3 Three careful clerics

James Bradley (1693-1762, third Astronomer Royal) and John Michell (c. 1724-1793) turn up in historical astronomy discussions with the words statistics or statistical attached to their persons. Bradley is known to Stigler, but not to Hald, and Michell to neither. These are the two stories for which I returned to the original literature and both remain high on my list of favorites, even after reading many pages in which *f* is pronounced *s*.

1.3.1 Bradley and the aberration of starlight

Bradley set out to find (as many others before him, and after, did) heliocentric parallax as the definitive demonstration of Copernican cosmography. He focused initially on Gamma Draconis, chosen by Robert Hooke for the same purpose, because it comes very close to the London zenith, thus minimizing both atmospheric refraction and flexure of the observing apparatus. By great good luck, the star is also very close to the ecliptic pole. The critical papers are Bradley (1728 on nutation). Hirshfeld recounts many more details than there is space for here.

Aberration is the apparent shift in positions of all stars (independent of distance) caused by earth's orbital motion. The maximum displacement is the ratio of orbit speed of light (10^{-4} or a half-angle close to 20 arseconds), and the standard analogy is walking forward into falling rain and needing to tip your umbrella to keep the drops from hitting you. Bradley seems to have found geometry easy and does not sketch the situation. Incidentally, he is able to report observations taken right through the year. You cannot see stars by daylight from the bottom of a well, but you can with a suitable (preferably long local-length) telescope.

Aberration shows in a year (or less) of data as our direction of motion through space changes and a star near the ecliptic pole seems to move north and south in declination at transit. Bradley continued to follow Gamma Dra over the years at the same time as he moved on to other stars, seeking to confirm the effect. After 20 years, it became clear that there was a systematic residual, with period about 19 years, which we now call nutation and attribute to lunar tides. His second paper makes use of (at least) the following ideas that are statistical in nature:

(a) mean values for the rate of precession of the equinoxes and obliquity of the ecliptic (rather than a favorite, or the most recent, or the oldest);

(b) a weighted mean for the maximum value of aberration for a star exactly at the ecliptic pole, which takes into account data on about 10 stars, giving largest weight to Gamma Dra, which has the longest data string, the smallest polar and ecliptic polar distances, and the brightest apparent magnitude; and

(c) an examination of the distribution of residuals.

He says that, in the comparison between observed declinations (or altitudes at upper culmination) and ones calculated from his final model, 11 of 300 values differ by 2-3" and none by more than 3".

Bradley ends by noting that he suspects that some physically meaningful effect remains to be found (e.g. a secular decrease in the obliquity). In modern terms, the fact that the distribution of errors is flatter than a Gaussian with a standard deviation of 1 arc second is confirmation of his suspicion. He displays a number of tables of observed and model declinations, one of whose implications is that, in 1748 in England, the autumnal equinox came about September 9th.

1.3.2 John Michell and binary stars

Michell is also part of the quest for parallax, because his demonstration that pairs of stars close on the sky are generally bound systems rather than chance superpositions undid the hopes of William Herschel and others to use such pairs for parallax measurement, on the assumption that the fainter star would always be more distant. He also, of course, thereby demonstrated that not all stars have the same absolute brightness, enormously complication "star gauging" or "the" problem of statistical astronomy (next section).

Michell appears in various contexts as:

(a) the inventor of black holes ("all light emitted from such a body would be made to return to it, by its own proper gravity." Michell, 1783);

(b) designer of the Cavendish balance (Cavendish was his executor),

(c) propounder of the idea that earthquake energy travels in waves (based on times at which Lisbon 1755 shook up other European cities); and

(d) the discoverer of binary stars (though it took Herschel's measurement of the first bit of an orbit before all were persuaded).

Michell (1767) began by asking for the probability that any one particular star should happen to be within a certain distance (as for example one degree) of any other given star and finding that it is $(60)^2/(6875.5)^2$ or 1/13131. And the probability that it is not is 13130/13131. He then extends to the probability that no one of whole number of stars n would be within one degree from the proposed star, and its complement, $1-(13130/13131)^n$ that there is one, and so onward to the probability that no one star should be within a distance r of any other star, with n to choose from,

$$P(\text{not}) = \left(1 - \frac{r^2}{(6875.5)^2}\right)^{n\times n}$$

and its complement, the probability that one is.

He makes fairly heavy going of the arithmetic, ending up with a style that resembles that of a modern student whose calculator doesn't have quite enough significant figures in its chips. Apparently $(1+n)^x = 1 + nx + \ldots$ was not part of the standard tool kit, but he gets the right answer, finding, for instance, that for Beta Cap ($n = 230$, $r = 3.3$ *arc min*) the chances are 80:1 against its being a chance alignment. For the six brightest Pleiads, the odds are $496,000 : 1$ against a chance grouping.

If this sort of arithmetic rings a bell, it is probably because you have met it before as the question of how many non-twins must you have in the room before it becomes more likely than not that two of them have the same birthday. The number (about 22) is smaller for Moslems because their year is shorter. I have no idea whether Michell or his predecessors knew about the birthday problem or other events described by the same calculation, but he does seem to have been first to apply it in astronomy.

1.3.3 Nevil Maskelyne and the personal equation

Maskelyne (1732-1811), the fifth Astronomer Royal, like Bradley and Michell, held orders in the Anglican church and is the member of the trio one finds it hardest to associate with the concept of charity, perhaps because he figures as something of a villain in the story of the quest to determine longitude at see. He was indeed a supporter of the method using the motion of the Moon (Maskelyne, 1762), mentioned in connection with Mayer's work. He was also in some sense the discoverer of the first recognized systematic error in astronomy, generally known as the personal equation.

Back in 1796, when the right ascensions of stars were determined from their times of meridian transit, Maskelyne noticed that his assistant, David Kinnebrook, whose work had formerly been consistent with his own, was now recording transit times that were systematically 0.8 sec later than his own. This corresponds to 12 arc seconds or as much as 0.2 miles at sea, and this 68 years after Bradley had measured the polar distances of stars to 1 arc second or better. Rather than rejoicing in the discovery that systematic errors could be much larger than random ones (and that Bradley had been wise to measure altitudes rather than hour angles), Maskelyne waxed wroth and fired Kinnebrook. Twenty-some years later, Bessel (who eventually found the long-sought parallax) looked again at hour angles measured not only by Maskelyne and Kinnebrook but also ones of his own and from Struve (another parallax discoverer), Argelander, Walbeck, and Knorre and found systematic differences up to a second (of time) and more which could vary from year to year.

His way of writing these, as, for instance, $B - S = -0,799$ sec. appears to have given rise to the name "personal equation" (Stigler, 1986). The magnitudes and variations were the sort normally associated with human reaction times, as per the story of Galileo's attempt to measure the speed of light

with dark lanterns on the seven hills of Rome. The name personal equation became customary and the numerical values dropped only with the adoption of automatic and electrified chronographs. The very large difference in **systematic** accuracies of right ascension (with personal equation) and declination data (without it) propagated through astronomy in the form of separate analyzes of the two components for many purposes, statistical and others.

The term was sometimes used for systematic errors of other sorts, for instance by Russell et al. (1926) to describe the tendency of some Mars observers to draw thin, straight lines between the dots and others to avoid this at all costs. Sherlock Holmes uses the phrase to mean something like general intelligence, remarking at one point that he "need not allow for what astronomers call the personal equation" since a particular foe is of first-rate intelligence (like himself, of course).

Any astronomer will be able to come up with other examples of unrecognized systematic errors utterly swamping the recognized random ones. Stigler's (1999) Table 20.1 shows 15 successive published values for the length of the astronomical unit in miles. Only two fall within the error bars of the previous value, and only two have error bars that take in the present official number. This is known to 9 significant figures in metric units (from radar travel times), but only about 6 in miles (owing to disagreement about the conversion factor). My own favorite is the Hubble constant, which has declined from 536 km/sec/Mpc (Mpc = megaparsecs) according to Hubble's initial, 1929, calibration, down to about 65, with 10% error bars at every stage (Trimble, 1996).

Maskelyne also makes a cameo appearance at the beginning of our next story, because he provided some of the key proper motion measurements from which Herschel first charted the motion of the sun relative to its neighbors. Other numbers came from Tobias Mayer, whom you have now also met.

1.4 "THE" problem of stellar statistics

Newton thought of, Michell (1767) and undoubtedly many others developed, and William Herschel is generally given credit for applying the method of determination of the distances and distributions of the stars in space based on the assumption that they are as bright as the sun (see Hoskin, 1963, for details of this story). Herschel called the method star gauging (gaging in his spelling) and by 1785 had put the sun near the center of a flattened system having sharp edges, a uniform density of stars, and an extent of a couple of kiloparsecs, stretching furthest in the directions where we see the most and faintest stars ("the Milky Way"). Even the Cygnus rift is there.

From Herschel's time down to the present, the key problem marching under the banner statistical astronomy has been to turn counts of numbers of stars as a function of apparent brightness into, in historical order:

(a) the size and shape of the system;

(b) the real distribution of stellar brightnesses (after Michell et al. showed that they were not all the same); and

(c) the distribution of the velocities of the stars (as a function of location, brightness, and so forth) after proper motion data and, later, radial velocity measurements showed that the system is not a static one.

Trumpler and Weaver (1953) mark the high-point of this endeavor as a core subject in astronomy.

Why is it a statistical problem? The number of stars you count as a function of apparent magnitude, $A(m)$, is given by

$$A(m) = \omega \int_0^\infty \varphi(M) D(r) r^2 dr$$

where ω is the solid angle you are examining, $\varphi(M)$ is the luminosity distribution, and $D(r)$ is the density of stars as a function of distance in that direction. The implied assumption that $\varphi(M)$ and $D(r)$ are separable functions is already a fatal error if you propose to look more than about one kpc in the galactic plane or 100 pc perpendicular to it. Built in is the relation between apparent and absolute magnitude, $M = m + 5 - 5 \log r - a(r)$, where $a(r)$ is the absorption in magnitudes and constitutes another unknown function. Kapteyn (1922) was the last to do this for $a(r) = 0$ everywhere (though he had earlier suggested values of 0.3 and 2.0 mag/kpc in the galactic plane), and even in this case, one clearly has to go over to sums rather than integrals, leading to a Mayer- or Euler-like problem of many equations in many (but fewer) unknowns and the potential for ending up with nonsense through what Euler called the multiplying of errors (both Gaussian and Poissonian in this case).

McCuskey (1965) and van Rhijn (1965, Kapteyn's colleague and successor) summarize the additional computational difficulties introduced when $a(r) \neq 0$ and make it clear when the confirmation of spiral arms in the Milky Way was left for the radio astronomers (for whom $a(r)$ really is 0 most of the time).

Now try to do the dynamical (stellar population) problem, where the goal is to extract, for instance, $N(M, V)$ from observations of $A_1(m, \mu)$ and $A_2(m, V_r)$ in various directions in the sky, subject to the same unknown $a(r, \theta, \emptyset)$ and the non-separability of the luminosity function, the density distribution, and the kinematic properties. Apart from everything else, one simply must have the counts, apparent magnitudes, proper motions, and radial velocities for the same stars in the same directions in the

sky. Kapteyn's (1906) **Plan of Selected Areas** sought to address this problem. The IAU Commission (32) on Selected Areas eventually voted itself out of existence, but this is the one context in which Kapteyn's name is remembered today in a positive tone of voice. Binney and Tremaine (1987), the relatively modern authority, mention neither Kapteyn nor his star streams, but do make contact with his period via the velocity ellipsoid of Karl Schwarzschild (which has, among other things, the shape of a Gaussian normal in two or three dimensions).

"Data products" from the traditional endeavor called statistical astronomy include:

(a) the luminosity distribution(s) of stars (which we now immediately try to turn into the mass distribution;

(b) the solar motion (first found by Herschel, using proper motions from Mayer and Maskelyne); and

(c) galactic dynamics.

The local distribution of stellar motions was described by Kapteyn as two star streams and by Schwarzschild as an ellipsoid. Neither means quite what you might guess, and I recommend Russell et al. (1926) or their references, Campbell (1913) and Eddington (1914) for clearer expositions than found in more modern references. All wrote before Trumpler (1929) forced galactic absorption upon us. Even so, the problem, in the words of RDS,

> *The problem of stellar statistics is to deduce from the apparent distribution of the stars in the heavens with respect to magnitude, proper motion, radial velocity, parallax, galactic concentration, etc. ... what is the true distribution of the stars in space ... in terms of three statistical functions: the density function, which gives the total number of stars per unit volume. ... the luminosity function, which shows what proportions of these stars have absolute magnitudes lying in successive equal intervals; and the velocity function, which defines the similar distribution of their velocities in space.*

must be sung as "to invert the impossible matrix".

Against this background, the discovery of galactic rotation by Bertil Lindblad and Jan Oort might seem nothing less than miraculous. They did however, have the rotation of M81 (Max Wolf) and M31 (Vesto Melvin Slipher) to guide them.

1.5 A smattering of archeoastronomy

Archeoastronomy includes (at least) two territories - the use of ancient observations to shed light on current questions (Chinese and other sightings of comets and supernovae are the classic examples) and the use of modern astronomy to shed light on ancient cultures (the classic example is the Star of Bethlehem, which I shall not mention at all, statistical considerations not often being important for single events, whether or not miraculous, but this is perhaps as good a place as any to record my prejudice that Bayesian methods, while excellent for changing your mind by a small amount, are much less useful on the road to Damascus).

Was Ptolemy to be trusted? Two aspects of this question have a "goodness-of-fit" answer. First, it seems that some of his observations are "predicted" so well by his model that they must have been back-calculated. This "excessive goodness-of-fit" result is an old one (Newton, 1977). Second, very recently, Schaeffer (2001) has asked whether Ptolemy borrowed his catalog from Hipparchus, and, if so, did it leave a statistical trail. Because the two lived at different latitudes and in different centuries (with precession of the equinoxes), different stars skimmed their horizons with differing degrees of visibility (hence opportunities for accurate observations of position and apparent brightness). The conclusion is that his fourth-quadrant stars are borrowed, the first three new observations.

Alignments of pre-literate and peri-literate monuments have been scrutinized for astronomical significance from the time of Locker to the present (Krupp, 1988, is a good source.). Conclusion range from, "you can see the whole of positional astronomy, including precession and changes in the obliquity at Stonehenge" to "yeah, the door is on the north side." I have dabbled in the now very densely populated part of this territory occupied by the pyramids of Giza (Trimble, 1964). Objectively, one can say things like

(a) the inclination of the shafts from the King's chamber of Cheops' pyramid point (to the accuracy within which they can be determined) to the north celestial pole (where there was a star when the pyramids were built) and to the upper culmination of the middle star in Orion's belt;

(b) the main exit of the Great Temple of Amon-Ra at Karnak points northwest, but misses the direction of sunset at summer solstice by more than the accuracy of the measurement (1.0^o at the time temple was built, Krupp, 1988); and

(c) main axes of 38 other temples built during the Empire period point in 38 other directions, 7 close to the cardinal directions and 6 (NW), 7 (NE), 13 (SE), and 5 (SW) in each of the quadrants (Badawy, 1968, p.184).

You could ask a statistical question about how likely this is to be a chance distribution (and answer it by frequentist or Bayesian methods). But if the answer is to be a contribution to Archeoastronomy, then you must decide what hypothesis you are testing. The choices include perpendicular to the nearby riverbank or to the cliffs behind as well as astronomical orientations.

The next step is supposed to be to test the hypothesis against a new, independent data set, or, failing that, to attempt to multiply the chance probability you find (which is always very small or you wouldn't be doing this sort of thing) by the number of other hypotheses that would be equally interesting. In the Empire Temple case, there is no comparable sample, but lots of hypotheses, and you are left with the usual result, "well, maybe there is something there."

Section 3.3 carried the moral that systematic errors are nearly always larger than random ones. The lesson here is that you must choose a testable hypothesis and stick with it. "Part of Ptolemy's catalogue is more consistent with observations made from Hipparchus' 4-dimensional location than with observations from Ptolemy's own 4-location" is such a hypothesis. "The Egyptians deliberately lined up their temples and pyramids to incorporate astronomical information" is not. Investigations of non-cosmological redshifts (which are now more than 35 years old) seem to me also to suffer from a surfeit of shifting and untestable hypotheses.

1.6 Ancient statistics in modern astronomy

Recent forays of astronomers into statistical territory come sometimes perilously close to reinventing the wheel and making it square. Nevertheless, I think each of the following issues is still a live one and still on the interface.

Density of matter (including dark matter) in the galactic plane This belongs to the tradition territory because the key equation is

$$\left[\frac{d}{dz}\ln\frac{N(z)}{N(z_0)}\right] < V_z^2 >= -4\pi G\sigma_0$$

where $< V_z^2 >$ is the component of the velocity ellipsoid perpendicular to the galactic place and the logarithmic gradient is that of the density of stars perpendicular to the plane. The desired density is σ_0, and the error made if you choose to take $\pi = 3$ will be smaller than other that are unavoidable. The equation and its application go back to Kapteyn and Jeans, though Oort often gets credit, and forward into modern models of the galaxy from Bahcall and Soneira, Kuijken and Gilmore, and others. The main errors are now recognized as systematic rather than random (though the latter are not small), because star populations change systematically away from the galactic plane, rendering color-based parallaxes too large (distance too small) because the more distant stars will be of lower metallicity, loser

mass, and more advanced evolutionary stage. Kapteyn and Jeans actually bracketed modern results, with $\sigma_0 = 0.099$ to 0.143 M_0/pc^3, and we remain uncertain about whether there is a separate disk dark matter component.

Closely related is the attempt to estimate the contribution of very faint, low mass stars to the total density. Small scale surveys (like those from the Hubble Space Telescope) yield a handful of brown and old white dwarfs (random errors win), and large scale ones suffer calibration errors (one of which the late Willem Luyten ungenerously dubbed the Weistrop Watergate).

Malmquist bias and the Scott effect Wherever two or three cosmologists are gathered together, one will say that the others do not understand these: their essence, the difference between them, or how to correct for them. Adriaan Blaauw even objects to the term Malmquist bias, on the ground that the concepts are all to be found in earlier papers by Kapteyn.

$\log N - \log S$ This is a cumulative distribution of source numbers vs. apparent flux. Errors due to binning are thereby removed, but others introduced. Early applications in radio astronomy suffered from confusion (meaning two or more faint sources getting counted as a single brighter one), though the conclusion that there are more distance radio sources than nearby ones stands. Giacconi (1972) used it at a time when very few X-ray sources were yet known or identified to predict that the X-ray background would eventually be resolvable into many distant sources. He too was right.

$P(D)$ **and** $N^{\frac{1}{2}}/N$ The concept that Poissonian fluctuations in numbers of sources within your beam will translate into apparent fluctuations in background surface brightness has been rediscovered at every wavelength. Scheuer (1957) used it to add a few points to the $\log N - \log S$ curve from the Third Cambridge Catalogue (rousing the wrath of the then-powerful steady-state community). Applied to optical observations of elliptical galaxies, it is one of the newer subrungs on the distance ladder (because you can pull out the brightness of the individual brightest stars contribution, declare then to be on the red giant tip, and get a spectroscopic parallax). Applied to the X-ray background, the calculation shows that the number of sources needed is just about what you would get from a $\log N - \log S$ extrapolation, if the background is to be neither more ragged nor smoother than what we see (these sources have now been resolved by Chandra and other missions).

V/V_m was Maarten Schmidt's way of taking into account that he had a flux limited sample with both radio and optical flux limits so that he could use measured redshifts of a very small number to conclude that quasars were commoner in the past. He has said that the basic ideas can again be found in Kapteyn's work (Schmidt, 2000). Recently he has suggested (Schmidt, 2001, personal communication) that the same methodology applied to gamma ray bursters implies that those of short duration are closer (and less beamed) than those of long duration (which optical redshifts now exist).

The **Lutz-Kelker correction** is needed when you look at groups of measured parallaxes encumbered with measurement errors, which are intrinsically asymmetric (since no real parallax can be negative; Chiaramonti had trouble with this!).

Kaplan-Meier survival curves This is my own particular square wheel, honed when I was trying to figure out how to show (or anyhow display) data concerning the long-term publication records of astronomers starting out with Ph.D's from high and low prestige graduate schools. The principle end point was, therefore, ceasing to publish. But it seemed to me (Trimble, 1991) that posthumous publication was an unreasonable expectation (not true - Lundmark was co-author of a 1999 paper), and it removed the deceased from the set of those at risk, so that the curves could turn back up if more people in a cohort died than stopped publishing for other reasons.

Properties of binary star populations There are at least two issues. First, how do you allow for unresolved binaries when counting stars as a function of apparent brightness (part of "the" problem of stellar statistics). This cannot be dealt with until you know the answer to the second issue, what are the real distributions of binary periods, separations, mass ratios, eccentricities and all as a function of age, chemical composition, and whatever else matters. These all fold into various attempts to understand chemical, luminosity, and other evolution tracks for galaxies or their separate stellar populations. Much ink has been expended since Kuiper (1935) interpolated (correctly) and extrapolated (I thing incorrectly) from the handful each of visual and spectroscopic binary orbits available to him. I abandoned the fray in 1990, with the parting shot that the answer you get will depend on the sample you choose to look at. This remains true. Complete information could be obtained only by working to sharp limits in apparent magnitude, magnitude difference and separation (for visual binaries), velocity amplitude and period (for spectroscopic binaries), and light amplitude and period (for eclipsing binaries) and then carrying out the equivalent of V/V_m in about six-dimensional space to get a volume limited sample. This is (marginally) possible for nearby clusters. "all the F V stars in the Yale Bright Star Catalog" or a few other narrowly circumscribed classes, but otherwise impossible.

Can we derive any particular lesson from these more complicated cases? I think so (and it is one that spectroscopists working on stellar structure and evolution were forced, kicking and screaming, to accept a couple of decades ago). It is that, when comparing hypothesis and data, it is better to transform your model into the observed quantities rather than try to put the data into theoretical bins (star color and effective temperature are a characteristic pair). For complex situations, a Monte Carlo simulation is often (not always) the best way to do this - assume a model and calculate what the observers should see. There will, then, in effect be error bars on your theory as well as your observations, but this cannot be helped.

1.7 Conclusions

Statisticians and astronomers have been trespassing on each other's territories for as long as the territories have existed. In addition to the discovery of particular methods and concepts, we can find in this history several lessons. It is easier to analyze data you have taken yourself than other peoples (Mayer vs. Euler). Systematic errors generally exceed random ones (Maskelyne and many more recent examples). It is important to decide which hypothesis you are testing before you do the arithmetic, ideally even before you collect the data (archeo-astronomy and non-cosmological redshifts). And, finally, if as is nearly always the case, there is not a precise correspondence between the quantities you can measure and the ones in your hypothesis, it is best to transform theorists's units into observers' units, rather than the converse.

And the most important lesson is that the story is never completely told. Despite all these pages, I have not mentioned

(1) Oscar Sheynin (1996 and many prior papers), who is the real expert on early astronomical statistics'

(2) the early recognition of interstellar absorption by King (1914, working, as usual on "the" problem of stellar statistics);

(3) all the good things that Simon Newcomb did (despite his role as Whitman's "learned astronomer" and opposition to astrophysics), many of them statistical (corrections of coordinates for refraction, fluctuations of the solar cycle, recognition of the background light of the night sky as not being due to faint stars); or

(4) Lambert (of the reflector), who despite Stigler's (1999) discussion of Bernoulli, Euler, and Old Max, arguably invented Maximum Likelihood (but did not use it for anything).

Acknowledgements: Special thanks to organizers Babu and Feigelson for finding something for me to do at the meeting and suggesting some references. Brenda Corbin of US Naval Observatory was enormously generous in finding and sending copies of papers not just older than UC Irvine but older than the present author.

1.8 References

[1] Abell, G.O. 1982. *Exploration of the Universe*, 4th Ed., Saunders, Holt Rinehart.

[2] Badawy, 1968. *A History of Egyptian Architecture: The Empire*, Univ. of California.

[3] Binney, J. and Tremaine, S.D., 1987. *Galactic Dynamics*, Princeton Univ. Press.

[4] Bradley, J., 1728. *Phil. Trans. Roy. Soc.*, 34, 637-661.

[5] Bradley, J., 1748. *Phil. Trans. Roy. Soc.*, 45, 1-45.
[6] Brunt, D. 1917. *The Combination of Observations*, Cambridge Univ. Press.
[7] Campbell, W.W. 1913. *Stellar Motions*, Yale Univ. Press.
[8] Chandrasekhar, S. 1942. *Principles of Stellar Dynamics*, Chicago; Dover, 1960.
[9] Eddington, A.S. 1914. *Stellar Movement and the Structure of the Universe*, London: McMillan.
[10] Franklin, J. 2001. *The Science of Conjecture*, John Hopkins Univ. Press.
[11] Giacconi, R. 1972. *Ann. NY Acad, Sci.*, 224, 149.
[12] Hald, A. 1990. *A History of Probability and Statistics and their Applications Before 1750*, Wiley/Interscience.
[13] Hald, A. 1998. *A History of Mathematical Statistics from 1750 to 1930*, Wiley.
[14] Hampel, F. 1998. *Can. J. Stat.*, 26, 497.
[15] Hoskin, M.A. 1963. *William Herschel and the Construction of the Heavens*, NY: Science History Publications.
[16] Hoskin, M.A., Ed. 1999. *The Cambridge Concise History of Astronomy*, Cambridge Univ. Press.
[17] Jeans, J. 1922. *Mon. Not. Royal Astro. Soc.*, 82, 122.
[18] Jeffreys, H. 1930. *Theory of Probability*, Oxford: Clarendon Press.
[19] Kapteyn, J.C. 1909. *Astrophys. J.*, 30, 284.
[20] Kapteyn, J.C. 1922. *Astrophys. J.*, 55, 302.
[21] King, E.S. 1914. *Harvard Obs. Annals*, 76, 1.
[22] Krupp, E.C. 1988. In C. Ruggles, Ed., *Records in Stone*, Cambridge Univ. Press.
[23] Kuiper, G.P. 1935. *Publ. Astro. Soc. Pacific*, 47, 38.
[24] Malmquist, G. 1920. *Lund Medd.*, Ser. 2, No.22.
[25] Malmquist, G. 1924. *Lund Medd.*, Ser. 2, No.32.
[26] Maskelyne, N. 1762. *Phil Trans. Roy. Soc.*, 52, 558.
[27] McCuskey, S.W. 1965. In A. Blaauw & M. Schmidt, Eds., *Galactic Structure*, Univ. Chicago Press, 1.
[28] Michell, J. 1767. *Phil. Trans. Roy. Soc.*, 57, 234-264.
[29] Michell, J. 1787. *Phil. Trans. Roy. Soc.*, 74, 35-57.
[30] Newcomb, S. 1886. *Amer. J. Math.*, 8, 343.
[31] Newton, R. 1977. *The Crime of Claudius Ptolemy*, Baltimore.
[32] Oort, J. 1932. *Bull. Astron. Neth.*, 6, 249.
[33] Pearson, K. 1976. *The History of Statistics in the 17th and 18th Centuries*, Griffin & Co., Ed. E.S. Pearson.
[34] Rao, C.R. 1997. In E. Feigelson & J. Babu *Statistical Challenges in Modern Astronomy II*, Springer Verlag.
[35] Russell, H.N., Dugan, R.S. and Stewart, J.O. 1926. *Astronomy*, Boston: Ginn & Co.

[36] Schaeffer, B.E. 2001. *J. Hist. Astron.*, 32. 1.
[37] Scheuer, P.A.G.S. 1957. *Proc. Cam. Phil. Soc.*, 53, 764.
[38] Schmidt, M. 2000. In P.C. van der Kruit & K. van Berkel, Eds. *The Legacy of J.C. Kapteyn*, Kluwer.
[39] Scott, E.L. 1956. *Astron. J.*, 61, 190.
[40] Sheynin, O.B. 1996. *The History of the Theory of Errors*, Egelsbach.
[41] Stigler, S. 1986. *The History of Statistics*, Harvard Univ. Press.
[42] Stigler, S. 1999. *Statistics on the Table*, Harvard Univ. Press.
[43] Trimble, V. 1964. *Mitt. Inst. Orientforschung*, 10, 183.
[44] Trimble, V. 1990. *Mon. Not. Royal Astro. Soc.*, 242, 79.
[45] Trimble, V. 1992. *Scientometrics*, 20, 71.
[46] Trimble, V. 1996. *Publ. Astro. Soc. Pacific*, 108, 1073.
[47] Trumpler, R.J. 1929. *Lowell Observatory Bulletin*, 14, 154 (No. 420).
[48] Trumpler, R.J. and Weaver, H.F. 1953. *Statistical Astronomy.*
[49] van der Kruit, P.C. & K. van Berkel, Eds. 2000. *The Legacy of J.C. Kapteyn*, Kluwer.
[50] van Rhijn, P.J. 1965. In A. Blaauw & M. Schmidt Eds. *Galactic Structure*, Univ. of Chicago Press, 27.

2

Power from Understanding the Shape of Measurement: Progress in Bayesian Inference for Astrophysics

A. Connors[1]

ABSTRACT After a review of the historical context of Bayesian inference in astronomy, we work a tutorial problem involving the search for pulsars in gamma-ray astronomical data; i.e. the detection of periodicities in a Poisson point process. We develop a model called Sparse Bayes Blocks for this purpose. These methods are also effective for estimation of the pulsar period.
This paper is followed by a commentary by statistician Eric D. Kolaczyk.

2.1 Introduction

2.1.1 WHY: Historical Context (A Personal View)

Overview

From antiquity to modern times (early 1900's), fundamental advances in astronomy and statistics had been intertwined (see [4, 2, 5]; and references therein). In modern times, this was not so: the fields had separated. Hence for the past few generations progress in astronomical data analysis proceeded piecemeal, in isolated spurts. Typically, first, one faced a class of problems which could not be solved, or to which one got silly or inconsistent answers using previously standard astrophysics methods. Second, one had access to a lively subcommunity with greater statistical knowledge. Then, an solitary solution to this specific class of problems was proposed, and diffused outwards in the astrophysics community. Well–known frequentist examples include: proper use of confidence intervals and likelihood ratios for parameter estimation and hypothesis comparison [4, 5, 6, 7]; and clarifying linear regressions, esp. in the presence of error bars [8]. Each greatly improved understanding, but often carried with them only bits and pieces

[1]Eureka Scientific

of the wider framework from which they were derived (e.g. [9]).

Bayesian progress followed similar pattern, save that in contrast to "black–box" methods, the formalism carries with it an overarching probabilistic/likelihood framework. Learning this was perceived as a barrier; as was the often high high computation cost. There seemed to be roughly four reservoirs of Bayesian statistical knowledge: statisticians applying their speciality to astronomical problems; physical scientists from other countries with more robust statistics education (Spain, South America, Eastern Block countries); people with some overlap in the engineering/radar/signal processing community; and people who discovered the work of Ed Jaynes [10]. Interestingly, the first and last seemed by far the most influential.

Modern pioneers: 1970's-80's

Independent thinkerA. Bijaoui [11, 12] was probably first to use Bayes in modern astrophysics. (He was the first to try many methods; at this conference his influence is represented in the multi–scale sessions.) The second modern application was the introduction of the EM algorithm (independently) by statisticians Richardson [13] and Lucy[14] for "image deconvolution" in astrophysics. This garnered a sudden increase in popularity after a high–profile data–analysis problem (Hubble Space telescope's mirror being out of focus) could not be fixed by the usual method of building a bigger telescope. (See [15, 16] for a modern EM view.) The third application was the (ill-conditioned) 'deconvolving' of spatial images from radio interferometry (cf. [17]; history in [18]). The researchers discovered Maximum Entropy; then the work of Ed Jaynes; and became 'Evangelical Bayesians'. One of the more spectacular examples was the use of Bayes methods for COBE limits on fluctuations in the CBR. N. Kaiser (private communication) writes that for the CBR data, there was much heated discussion of the silly range of the fluctuations – clearly not supported by common sense, or by likelihood analyses. For example, one analysis [19] happened to have an anomalously low χ^2; using the then standard [20] approach gave an unreasonable low "upper limit" on the fluctuations. At the same time, some of the collaborators (A. Lasenby) had offices near the "evangelical" Maximum Entropy group, and so were introduced to lively discussions on how to form the best likelihood ratio, and what were appropriate priors (see [21, 22, 23], and references in [24]). Last, independent thinker T. Loredo [25] also discovered the work of Ed Jaynes, and in turn influenced those at U. Chicago, LANL, and beyond (e.g. [26, 27, 28]).

1990's: SCMA I and II

At the first SCMA, broadly speaking, the focus tended to be on the Bayesian equivalent of standard problems and related ones that could not readily be handled by frequentist methods [4]. Many of these were physical, parametric models; and one did the marginalization either analytically, or via

simple numerical integration methods (Laplace, quadrature). Some of these solutions are only now coming into standard use (XSPEC, CIAO), despite their simplicity. At the second SCMA, [2], there was more focus on computer methods (esp. MCMC), and more formation of "working groups" and more serious collaboration with statisticians (Duke, CMU, Purdue, Chicago, Harvard, BU, ...). As Feigelson and Babu had predicted, it took a great deal of work to "translate" between astrophysics jargon and culture and that of the statisticians. However this brought not only access to more advanced computational techniques but also to broader perspectives.

SCMA III: 2001

Now we have reached SCMA III, where we can see fruits of those collaborations many built in response to those challenges. What are the new challenges that will be defined here? Is it time for an era of fundamentally rethinking how we do measurements? "Data Analysis" (not even "Statistics") used to be an afterthought. Perhaps now it can be part of how one phrases the scientific question or designs the experiment.

2.1.2 WHY: Astrophysicists now

Likelihood methods in general and Bayesian methods in particular are especially well–suited to modern astronomy and astrophysics. First, astrophysics is unusual in being able to derive reliable quantitative estimates of our errors and uncertainties, and of the entire measurement process [29]. Second, we have both (literally) millenia's worth of prior observations — much of it detailed, quantitative measurements [29]; plus astoundingly precise quantitative predictive theories, from quantum mechanics to relativity and beyond. Third, unfortunately, unlike engineers, accountants, biologists, economists, and others, we — generally speaking — do not have a formal background in probability and statistics beyond what is in 'cookbook' texts [30, 31]. This isolation often invites misapplication, both from unfamiliarity with the larger probability framework, and from missing out on crucial advances (e.g. it was clarified in the 1970's that in certain cases the F-test does not work; only slowly is this being brought into the astrophysics community [9]). On the other hand, together these give us an opportunity, now, to build a more fundamental understanding.

The framework for approaching problems that I advocate here is a 'likelihoodist' perspective [32]. The method for deriving these is Bayesian.

2.1.3 WHY: Bayesian Inference

A properly constructed Bayesian likelihood ratio is always the best measure of all the information in one's data given competing null and interesting hypotheses, plus one's prior information. This is not in dispute by either

classical or Bayesian statisticians. The difficulties lie in: 1) properly incorporating prior information into a clear well-specified prior probability; 2) doing the marginalization (often numerical integration); and 3) parametrizing the model so the first two steps are easier. Recall the Bayes prescription requires four parts:

1. A set of *hypotheses* (class of model + parameters) $\mathcal{H}_0 ... \mathcal{H}n; \theta\mathbf{n}$;

2. An appropriate *sampling statistic* for the data, given the (class of) model plus parameters $p(D \mid \mathcal{H}_n, \theta_\mathbf{n}, I)$; and

3. Previous information I (instrument calibration, background measurements, quantum mechanics, ...), appropriately quantified in *prior probabilities* $\pi(\mathcal{H}_n \mid I), \pi(\theta\mathbf{n} \mid I)$. Combining these via Bayes's Theorem gives the *posterior probability* $p(\theta_n \mid D, I\mathcal{H})$ of the parameters given the data D, model or hypothesis $\mathcal{H}$, and prior information I, or $p(\mathcal{H} \mid D, I)$ of the model or hypothesis given the data and prior information:

$$p(\theta_\mathbf{n}|D, \mathcal{H}_n, I) = \frac{\pi(\theta_\mathbf{n}|I)}{p(|I)} p(D|\ \theta_\mathbf{n} I), \quad \text{or}$$

$$p(\mathcal{H}_n|D, I) = \frac{\pi(\mathcal{H}_n|I)}{p(D|I)} p(D|\mathcal{H}n, I).$$

4. The Bayesian formalism then allows another step: *Marginalizing* over or averaging over uninteresting (or "nuisance") parameters, such as an unknown background rate, unknown continuum flux, pulsar phase, etc etc.

Hence, a statistic can fail to be the best in four ways. The first two are perhaps the most common:

1. Wrong sampling statistic (e.g. χ^2 or Gauss-Normal when the distribution is skewed, or multipeaked; using a periodogram or even traditional wavelets for Poisson data; and so forth.

2. Not the best underlying model class (e.g. using a FT when one does not have a single stationary sinusoid; or — as in our upcoming example! — one expects arbitrarily sharp peaks).

3. More subtle: not the best use of prior information (of which astronomers have lots). It is also is this structure that prescribes how to successively make use of more complexity:

   - imaging / PSF and background information in timing or energy analysis;
   - systematic uncertainty in detection process (deadtime/pile–up; calibration uncertainties; etc);

- previous but uncertain measurements from, say, non-simultaneous observations (*e.g.,* non-contemporaneous radio ephemerides).

4. More subtle yet: not the best handling of "nuisance" parameters (background rates, etc).; i.e how best to summarize information on multidimensional models. This may be the most powerful piece of Bayesian methods. It defines how to reduce the dimension of one's statistic to only the interesting parameters, while still retaining *all* of the information about them contained in the data. By contrast, finding the maximum likelihood or letting some parameters "float free" while interesting ones are fixed may be more familiar and a quicker approximation, but it will only work under special circumstances. These include relying on underlying Gaussian assumptions about the probability–space, which may in fact be highly skewed or even multiply-peaked; and may often be unexpectedly uncalibrated (e.g. [9]).

BUT this also means: for many astrophysical problems, it can be obvious how to write down a more correct statistic! This is part of the opportunity that comes from astrophysicists having ignored basic statistics for so long.

2.1.4 Paper overview

In the next two sections, we work through a single problem (due to [55]). We show the nuts,bolts and struts of building the best (likelihood) measure for any problem. The problem is simple enough to be given as an undergraduate excersise; yet has real potential as a new method (e.g. detection of unknown pulsars at high energies). We work the examples backwards: hardest example first (Bayes model selection) then more familiar parameter estimation. Finally, we briefly widen our perspective to place it into context. We conclude with references to more complicated works, and some challenges.

2.2 Bayes Model Selection: detecting 1D structure

Based on [55], we work through a simple but powerful example of a new method for detecting structure in a signal: 1D, no instrumental response. It illustrates both pitfalls and insights relevant for more complex treatments. For example the process is (almost!) identical to that used for: finding structure in an energy spectrum (complex solar flare line spectra, [33]); structure in spatial imaging data (PET images [34]; [36] new Chandra new work); even higher dimension problems; or even non-periodic timing (detecting a flare or burst on a variable background: [37].

2.2.1 WHAT: CGRO sources and γ–ray pulsars

High energy pulsars

The extremely coherent periodic signal characterizing pulsars was first detected in the radio. Since, pulsars have been detected at all wavelengths, with periods ranging from ms to seconds. Pulsars are thought to be rapidly rotating neutron stars (about the mass of the sun collapsed into $\sim$10 km) with extreme magnetic fields ($10^{12} - 10^{14}$ times that of Earth). These funnel the radiation into a beam something like that of a rotating beacon from a lighthouse. The grand sweep of the massive magnetic fields of the youngest pulsars is thought to power not only the the population of highest energy particles (cosmic rays) bathing out Galaxy, but also drive the generation of the highest energy photons (γ-rays). At lower (optical or X–ray) energies one sees rounded, roughly sinusoidal pulse profiles, or light–curves (brightness as a function of time or phase) suggestive of a larger rounded hot spot on the pulsar itself rotating in and out of the line of site [38]. At high energies, these pulse–profiles can be very sharply peaked — more like the edge of a tightly focused cone of emission swinging in and out of the line of sight ([39, 40]; and references therein).

Previous detections

As yet only six have been detected in γ-rays, most by the Compton Gamma-Ray Observatory (CGRO). They are usually detected first in radio energies (Crab; Vela; etc). However one of the brightest γ-ray sources in the sky turned out to be a nearby pulsar (Geminga; [13, 14]) and has yet to be found in radio. One strongly suspects that many of the unidentified CGRO/EGRET sources are in fact similar γ–ray pulsar neighbors (e.g. [9, 15]). But detecting them is difficult. In part this is due to the intrinsically low number of very high energy photons detected from celestial sources: a decade's worth of surveying the sky can result in only a few hundred photons from one source. In part, could it be that previous carefully crafted and studied pulsar–detecting algorithms aren't optimal for the low counts and sharp–edged pulse profiles at the highest energies?

Models for Detection of structure

Broadly, for the most sensitive *detection*, one wants a model that distills the essential shape of the structure in the lowest number of parameters. Of course the best would be a tightly–constrained physical theory (e.g. in planet detection, modeling the Keplerian orbits; [45]). When that is not practical, one uses the first few terms of flexible non–parametric multi–scale models (Fourier components; wavelets; simple binning; etc).

Previous methods have been either Fourier transform–based (Z_n^2; [2]) or binning–and–χ^2–based (epoch–folding; [17]). In the Z_n^2, one takes the Fourier transform of the pulsar phases indicated by the photon arrival

times. The sum of the squares of its components (e.g. Rayleigh statistic), plus those of its $n-1$ harmonics, are then tested against flatness using a χ^2 distribution. One can show this is similar to assuming the pulsar light–curve can be represented by an exponentiated Fourier transform (c.f. [48, 4, 6] and references therein). This is reasonable at lower energies, where more rounded pulse–profiles may better lend themselves to the standard Fourier transform based methods (e.g. [38]). However at higher energies the pulse profile is often expected to be sharply peaked (see [39]).

In classical epoch–folding, the data arrival times are folded on the known or trial period, then binned into evenly spaced bins (weighted by exposure). One tests against flatness in the resulting histogram via χ^2 [17]. However some difficulties remain: 1) How does one proceed when χ^2 is not appropriate (few counts per bin)? 2) How does one choose the bin size, balancing fine detail (many small bins) versus good χ^2 approximation (larger bins for more counts/bin)? For a typical unidentified CGRO/EGRET source the total counts can be low (a few hundred). Hence neither simple FFT or epoch–folding/χ^2 based methods are the best.

In a seminal "how–to" paper on Bayes in astrophysics, P. Gregory and T. Loredo [18] derived a fully Poisson equivalent of epoch-folding — no χ^2 required. They also used the Bayesian technique of *marginalizing* (or averaging, in parameter–probability–space) over unknown parameters to address the question of the proper number of bins. Still, there is an implicit penalty in using too many model parameters for detecting a feature. (See [51] or [9] for the "Ockham's Razor" that is built into Bayes odds ratios). So if one could use the minimum number of parameters in one's model, yet still capture very sharp features, one could in theory do a better job of detecting pulsars with very narrow peaks.

2.2.2 HOW: Sparse Bayes Blocks

Overview

For high–energy pulsars, one sees that previous methods may not have been the best because either the model (Reason 1) or the sampling statistic (Reason 2) were not the best. Hence, to derive a better statistical tool for catching sharply–peaked pulsars, we shall use both the correct sampling statistic (Poisson), and an improved model. With these we step through the Bayes formalism: specify hypotheses; priors; and sampling statistic; marginalize; and compare.

For our class of interesting hypotheses, we propose to use an extremely simple model called "Sparse Bayes Blocks" [55]. It is a distillation of useful points from epoch–folding ([18] and references therein) and the "Bayes Blocks" *changepoints* approach [3]. Long used in other fields, changepoint models assumes the process to be composed of relatively smooth segments delineated by discontinuous jumps at the "changepoints". These smooth

segments can be constant (as in a histogram), exponential, or any smooth function. The size of each segment (i.e. the width of each bin) is determined by the data, rather than assumed to be evenly spaced. This allows a light–curve with (say) one (or two) very narrow sharp peaks to be described by only four (or eight) parameters: the positions of the two changepoints per peak, and the expected average rate in each segment. This is in contrast to standard epoch–folding binned models, which for our example would require $\sim 10^3$ bins to properly model a single sharp spike. The "sparseness" comes from using the fewest possible model parameters. We are interested in detecting our pulsars first; later we may characterize them with more complex changepoint models [34],[3].

Simplest Interesting hypothesis $\mathcal{H}_2$: two changepoints

The model is piecewise constant, with three segments (0, 1, 2), delineated by two changepoints (ϕ_0, ϕ_1). The rate at the end of the phase cycle (phase = 1.) is required to match that at the beginning of the cycle (phase = 0.); hence the rate of segment 2 is the same as that of segment 0. The model rate μ_i above is r_1 if that phase bin is between the two changepoints (i.e. within the peak); and r_{02} (i.e. background) otherwise.

Simplest Null hypothesis $\mathcal{H}_0$: no changepoints

The model is a single constant segment, with model rate μ_T.

On Priors in Astrophysics, Part I

This is more subtle, and requires some thought. The problem can be expressed *either* as: we are looking for evidence of a new piece of structure, a peak with unknown flux, on top of an existing background; *or* we already know a source of approximately this flux exists, and want evidence that some fraction of this total flux comes from a periodic peak. The first corresponds to looking for an extra component in a multi–component model such as an emission line in an energy spectrum, or a new source in spatial data. This is in general a hard problem; see [9]. The second is simpler and corresponds to thinking of structure as fractional "shape parameters" (e.g. [18]). We will illustrate both methods here.

First, we will work through the first, slightly more conservative assumption: there may be one (or more) extra component(s) on top of a flat background. This leads to separate priors for the unknown rates in each segment. If there were no previous measurements of any kind, at any wavelength, of any similar types of objects; and also absolutely no theoretical predictions (beyond that they cannot be larger than the maximum the instrument can detect), one might choose priors based on invariance arguments. For rates that both must be non–zero and for which one's lack of prior knowledge is unchanged, no matter in what units it is expressed (scale invariance) one can use a log–prior; for rates that can be zero or negative one might use a

constant prior (See [4] for examples and references). However this is astrophysics; there is almost always some kind of previous information —- even if it is only a measurement or prediction of an average rate. In this case, following [22], we can use an exponential prior on the *average rate*. It is a flexible, physical, conjugate prior. With low scale parameter it resembles the log–prior, and with high average rates it resembles the constant prior. Ed Jaynes [54] pointed out this was the Maximum Entropy prior when one knows only a scale for the average rate before the measurement. Also, [37] found that it worked well for catching bursts. It is informative, yet does not strongly bias the outcome.

Therefore, for the average rates r_1, r_{02} on each segment n, the prior π is

$$\pi(r_n \mid I)\mathrm{d}r_n \;=\; \mathrm{e}^{(-\beta_n r_n)}\beta_n \mathrm{d}r_n$$

with scale β given by the inverse of the average from prior measurements.

For the changepoints $\{\phi_n\}$, we used a prior π that is constant in phase (that is, one that is invariant with respect to translations in phase):

$$\pi(\phi_n \mid I)\mathrm{d}\phi_n \;=\; \mathrm{d}\phi_n.$$

Data and sampling statistic

The data are intrinsically Poisson: lists of times (plus energies, positions, data quality indicators) measured by the instrument as photons (or background particles) arrive from a distant source. For pulsar (i.e. period) searches, these arrival times are corrected for the (varying) travel times between the pulsar and moving instrument (Bari-center corrections [2], then folded on the (known or trial) period. The data are then in a the form of a (Poisson–distributed) list of photon (BVC-corrected) arrival times and associated phases (plus energies, positions, etc..) These X-ray or $\gamma-$ray photons can arrive with mean spacings of seconds to weeks — i.e. many pulsar revolutions between each detection. Hence high precision is necessary for the calculations. Sometimes these events are binned onboard the satellite (along with the associated livetime per bin) before being sent to ground for processing.

If μ_i is the average rate on the instrument during phase bin i, δt_i its livetime, and y_i the number of counts measured, then the sampling statistic is:

$$p(y_i \mid \mu_i \delta t_i) \;=\; \mathrm{e}^{(-\mu_i \delta t_i)} \frac{(\mu_i \delta t_i)^{y_i}}{y_I!}.$$

(This assumes the data are binned counts, but in the limit of very small bins it takes the usual Poisson Point Process form.)

Turn the crank: Apply Bayes Theorem to get Posterior

With the model and priors specified, we write down their product (divided by a normalization term) to form the posterior probability. Let Y_{02} be

counts in the background piece, with total livetime T_{02}; Y_1 be counts in the peak, with total livetime T_1; and Y_{TOT} be the total counts in the observation, with total livetime T_{TOT}. Then, with the (conservative) choice of the scale factor β being the same for both, the posterior can be written:

$$p(r_1, r_{02}, \phi_0, \phi_1 \mid \{y_i\}, I, \mathcal{H}_2)\mathrm{d}\phi_0\mathrm{d}\phi_1\mathrm{d}r_{02}\mathrm{d}r_1 \;=$$
$$\mathrm{d}\phi_0\mathrm{d}\phi_1 \times \beta\mathrm{d}r_{02}\mathrm{e}^{-(\beta+T_{02})r_{02}}(r_{02})^{Y_{02}} \times \beta\mathrm{d}r_1\mathrm{e}^{-(\beta+T_1)r_1}(r_1)^{Y_1}$$
$$\times \prod_{i\,=\,0}^{N_{TOT}} \frac{(\delta t_i)^{y_i}}{y_i!} / p(\{y_j\} \mid I).$$

After analytically integrating over the rates r_{02}, r_1 one obtains the marginalized posterior for the changepoints ϕ_0, ϕ_1:

$$\lambda_2(\phi_0, \phi_1 \;\mid\; \mathcal{H}_2, I, \{y_i\})\mathrm{d}\phi_0\mathrm{d}\phi_1 \;=\; \beta^2 \frac{\Gamma[Y_{02}+1]}{(\beta+T_{02})^{(Y_{02}+1)}} \frac{\Gamma[Y_1+1]}{(\beta+T_1)^{(Y_1+1)}}$$
$$\times \mathrm{d}\phi_0\mathrm{d}\phi_1 \times \prod_{k\,=\,1}^{N_{TOT}} \Big(\frac{(\delta t_k)^{y_k}}{y_k!}\Big) / \Big(p(\{y_k\} \mid I)\Big).$$

Null hypothesis, $\mathcal{H}_0$

The model is simply one constant segment, with no changepoints. From above, one can see the marginalized posterior will have the form:

$$\lambda_0 \;=\; \beta \frac{\Gamma[Y_{TOT}+1]}{(\beta+T_{TOT})^{(Y_{TOT}+1)}} \times \prod_{k\,=\,1}^{N_{TOT}} \Big(\frac{(\delta t_k)^{y_k}}{y_k!}\Big) / \Big(p(\{y_k\} \mid I)\Big).$$

Likelihood Ratios

Finally, the payoff: to find the Bayes likelihood ratio as a function of changepoints (ϕ_0, ϕ_1), one divides the likelihood of null hypothesis $\mathcal{H}_0$ into that of the interesting hypothesis $\mathcal{H}_2$:

$$\Lambda_2(\phi_0, \phi_1 \;\mid\; \mathcal{H}_2, \mathcal{H}_0, I, \{y_i\})\mathrm{d}\phi_0\mathrm{d}\phi_1 \;=\; \mathrm{d}\phi_0\mathrm{d}\phi_1$$
$$\beta \frac{\Gamma[Y_1+1]\Gamma[Y_{02}+1]}{\Gamma[Y_{TOT}+1]} \frac{(\beta+T_{TOT})^{(Y_{TOT}+1)}}{(\beta+T_1)^{(Y_1+1)}(\beta+T_{02})^{(Y_{02}+1)}}.$$

This maps out the most probable changepoints. To find global (or total) odds O, or Bayes factor, of $\mathcal{H}_2$ (one peak, two changepoints) versus $\mathcal{H}_0$ (flat, no changepoints) we marginalize (i.e. numerically integrate) the expression above over all changepoints (ϕ_0, ϕ_1):

$$\mathcal{O}(\{y_i\}) \;\mid\; \mathcal{H}_2, \mathcal{H}_0, I) \;=\; \int \mathrm{d}\phi_1\mathrm{d}\phi_2 \Lambda_2(\phi_0, \phi_1 \;\mid\; \mathcal{H}_2, \mathcal{H}_0, I, \{y_i\}).$$

More complicated models are similar in form: for example, one exponential piece for the peak versus one constant piece (only background) looks similar, save for an extra term due to marginalizing over the exponential slope.

2.2.3 On Priors in Astrophysics, Part II

Notice the result above has a dependence on the prior parameter β. Although its effect on the likelihoods for the positions of the changepoints is almost negligible, it has a stronger effect on the Bayes evidence, (i.e. global odds ratio) comparing the null and interesting hypotheses. This is not the case when the problem can be formulated as a question of unknown fractional shapes rather than an unknown extra component: i.e. one knows that a source exists and tests the hypothesis that it is a pulsar, rather than testing for the existence of a source at the same time.

Rephrasing the interesting hypothesis $\mathcal{H}_2$:

Let the total rate be r_T. The fraction of the total counts in the peak is f_1, while the fraction outside the peak is f_{02}, with constraint $f_1 + f_{02} = 1$. The expected number of counts in each time (or phase) bin δt_i is then:

$$r_i = r_T T_{TOT} \delta t_i \big(f_1/T_1\big), \text{ for } \phi_i \in (\phi_0, \phi_1];$$

$$r_i = r_T T_{TOT} \delta t_i \big(f_{02}/T_{02}\big) \text{ otherwise.}$$

As before T_1 and T_{02} represent the livetimes accumulated in the peak and background sections, respectively.

Rephrasing the priors

The prior on the total rate has the same form as before:

$$\pi(r_T \mid I)\mathrm{d}r_T = \mathrm{e}^{(-\beta_T r_T)} \beta_T \mathrm{d}r_T.$$

However the prior on the fractional rates is new. It is uniform on $[0, 1]$ with the constraint that both sum to unity:

$$\pi(f_1 \mid I)\mathrm{d}f_1 = \mathrm{d}f_1, \;\; \pi(f_1 \mid I)\mathrm{d}f_1 = \mathrm{d}f_1; \;\; \text{with constraint } \; f_1 + f_{02} = 1.$$

Alternate posteriors

With these changes, the posterior for the interesting hypotheses becomes:

$$p(r_T, f_1, f_{02}, \phi_0, \phi_1 \mid \{y_i\}, I, \mathcal{H}_2)\mathrm{d}\phi_0\mathrm{d}\phi_1\mathrm{d}r_T\mathrm{d}f_1\mathrm{d}f_{02} = \mathrm{d}\phi_0\mathrm{d}\phi_1 \times$$

$$\beta \mathrm{d}r_T \mathrm{e}^{-(\beta + T_{TOT}) r_{TOT}} (r_{TOT} T_{TOT})^{Y_{TOT}} \times$$

$$\mathrm{d}f_{02} \big(\frac{f_{02}}{T_{02}}\big)^{Y_{02}} \times \mathrm{d}f_1 \big(\frac{f_1}{T_1}\big)^{Y_1} \times \delta(f_1 + f_{02} - 1)$$

$$\times \prod_{i\,=\,0}^{N_{TOT}} \frac{(\delta t_i)^{y_i}}{y_i!} / p(\{y_j\} \mid I).$$

The marginalized posterior becomes:

$$\lambda_2(\phi_0, \phi_1 \;\mid\; \mathcal{H}_2, I, \{y_i\})\mathrm{d}\phi_0\mathrm{d}\phi_1 = \mathrm{d}\phi_0\mathrm{d}\phi_1 \times$$

Preliminary Monte Carlo Results from Connors and Carramiñana 2001

Monte Carlo		CLASSIC — Z_6^2		BAYES — "Sparse BB"				GL92
Model	Cts	n=6	$-\log_{10}$ *Prob*	π_{scale}	$\log_{10}$ $\mathcal{O}_{2,E}$	$\log_{10}$ $\mathcal{O}_{2,GL}$	$\log_{10}$ $\mathcal{O}_{3,GL}$	$\log_{10}$ *Odds*
spike	134	393.7	76.1	140.	96.6	174.	171.	32.1
spike	74	195.8	34.6	70.	41.2	93.1	91.7	14.2
spike	32	102.4	15.7	35.	12.9	39.7	38.6	6.53
spike	13	32.3	2.91	10.	0.44	9.19	8.5	0.65
Vela	561	467.5	91.8	500.	52.4	52.2	70.0	77.6
Vela	277	279.3	52.0	300.	29.1	28.9	41.1	44.3
Vela	138	165.5	28.4	140.	16.2	16.1	26.6	23.0
Vela	72	73.8	10.21	70.	5.65	5.42	7.4	7.40
flat	538	14.9	0.61	500.	-0.431	-0.91	-0.06	-1.92
flat	258	13.4	0.47	300.	-0.453	-0.91	-0.3	-1.79
flat	136	9.4	0.17	140.	-0.445	-0.89	-0.5	-1.78
flat	71	10.1	0.22	70.	-0.0046	-0.40	0.04	-0.95

TABLE 2.1. Note: "CLASSIC" is *classical probability (frequency of occurrence) of the null hypothesis*, rather than a *ratio* of the probabilities of the null and interesting hypotheses, as are the others. 'E' in $\mathcal{O}_{2,E}$ stands for our first choice of parametrization, with an exponential prior on each separate segment. 'GL' stands for the second parametrization, similar to that from GL92. The number tells the number of changepoints used in the model (two or three). GL92 Calculations provided by P. Freeman, private communication; calculated for up to m = 12 bins. "Vela" means CGRO/EGRET 100 MeV - 10 GeV Obs 00 data used as "template" for source shape.

$$\beta \frac{\Gamma[Y_{TOT}+1]}{(\beta+T_{TOT})^{(Y_{TOT}+1)}} \frac{{T_{TOT}}^{Y_{TOT}}}{{T_1}^{Y_1}{T_{02}}^{Y_{02}}} \frac{\Gamma[Y_1+1]\Gamma[Y_{02}+1]}{\Gamma[Y_{TOT}+2]}$$

$$\times \prod_{k\,=\,1}^{N_{TOT}} \Big(\frac{(\delta t_k)^{y_k}}{y_k!} \Big) / \Big(p(\{y_k\} \mid I) \Big).$$

The marginalized posterior for the null hypothesis remains the same — the first four terms from above, plus the last normalization term. The final likelihood ratio for the changepoints then becomes:

$$\Lambda_2(\phi_0, \phi_1 \;\mid\; \mathcal{H}_2, I, \{y_i\}) \;=\; \frac{{T_{TOT}}^{Y_{TOT}}}{{T_1}^{Y_1}{T_{02}}^{Y_{02}}} \frac{\Gamma[Y_1+1]\Gamma[Y_{02}+1]}{\Gamma[Y_{TOT}+2]}.$$

Notice any dependence on the scale parameter β for the prior on the flux has cancelled out. Notice, too, how similar this is to the form in [18] (henceforth GL92), save that the bins can now have arbitrary width and placement.

One can derive the equivalent marginalized likelihood ratio for three changepoints (and higher; see [55]):

$$\Lambda_3(\phi_0, \phi_1 \mid \mathcal{H}_3, I, \{y_i\}) \;=\; \frac{{T_{TOT}}^{Y_{TOT}}}{{T_1}^{Y_1}{T_2}^{Y_2}{T_{03}}^{Y_{03}}} \frac{\Gamma[Y_1+1]\Gamma[Y_2+1]\Gamma[Y_{03}+1]}{\Gamma[Y_{TOT}+3]}.$$

2.2.4 Comparison Tests

In table 1, we list some of the results of Monte Carlo tests from [55]. They simulated three kinds of data: 1) flat background; 2) a Vela pulsar–shaped light–curve, with CGRO/EGRET 100 MeV - 10 GeV Obs 00 data used as a template; and 3) a spike in a single 5×10^{-4} wide bin. Each of these was analyzed with three methods: 1) the current high energy standard, Z_n^2 with $n = 6$; 2) The Bayesian epoch—folding method of GL92; and 3) our new statistic based on two or three "Bayes Blocks". We note that the GL92 method would have performed better had we used a much larger cutoff for number of bins, rather than stopping at the default number $m = 12$.

The preliminary test of the concept is very encouraging. Notice that both "One BB' methods outperformed the classical method on the "single spike" pulse–profiles. Note further that parametrizing the model with an overall rate and shape parameters improved the $\log Odds$ throughout.

2.3 Bayes for Parameter Estimation

Both [56] and [18] give excellent tutorials in Bayesian parameter estimation. For astrophysicists, "parameter estimation" means either "confidence intervals" (the % of data that would fall within contours of constant $\Delta \log(\textit{maximum likelihood})$; [4], [5]); or "credible regions" (the % volume of parameter space contained in contours of constant $\Delta \log(\textit{marginal likelihood})$; [18]). (Occasionally, an astrophysicist might use something simpler such as adding error bars in quadrature, but seldom for serious problems.)

Bayesian parameter estimation looks much the same as its classical counterpart, but with predictable differences. First, one uses the posterior rather than the sampling distribution. Second, one *marginalizes* over uninteresting parameters, and to reduce the dimension of the interesting likelihood statistic (rather than taking the maximum). Third, one does not use look–up tables (c.f. [31, 4]). Instead one steps through a grid of the parameters of interest, directly calculating the appropriate $\Delta \log(\textit{likelihood})$ for each % volume of probability space (e.g. 67.23%,95.45%,99.73%, etc).

BENEFITS

1. No problems dealing with multiple peaks, skewed distributions, or any other non–Gauss–Normal case (no CLT required).

2. Nice summary of highly dimensional models.

3. One can explicitly see the effect of each part (priors, model parameters, model choice) on the final outcome.

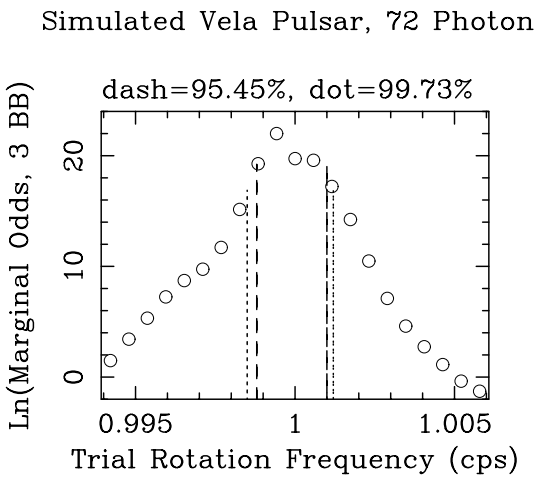


FIGURE 2.1. Parameter limits on the unknown rotation period: Credible regions delineating 95.45% and 99.73% (equivalent to 2 and 3 σ) of the volume of posterior probability space.

COSTS

1. Especially when the integrals (marginalization) cannot be done analytically, it can cost one significant computation time.

2. Although Bayesian parameter estimation is simpler than Bayesian model comparison, as results do not depend as strongly on the choice of prior, for weak data constraints it can still be important.

With this in mind, let us work through our brief example. Suppose one of the parameters that we have assumed to be fixed, in the previous example, is in fact unknown. This might be the true source position; the pulsar rotation frequency (and its derivatives); or any other physical quantity —- the formalism remains the same. (This also lets us demonstrate the modular 'hierarchical form' one might use for taking into account instrumental uncertainties; e.g. [5, 6] demonstrate including imaging information.)

Here, we will work through the example of the unknown rotation frequency. As our test data–set, we will use a Monte Carlo one based on the EGRET $> 5GeV$ histogram of [39]. We assume a rotation period of 1.00 cycles/sec; and period derivative of zero. The simulated observation lasts for 2 days — enough to accumulate 72 photons.

First, we will form the marginalized posterior likelihood as a function of the pulsar rotation frequency ν, via the procedure we have done before.

Prior

Following [18] and [56], we assign a log prior on the unknown frequency:

$$\pi(\nu \mid I) \;=\; \kappa\frac{\mathrm{d}\nu}{\nu}, \quad \text{with} \;\; \kappa \;=\; 1/\log[\frac{\nu_{MAX}}{\nu_{MIN}}].$$

The limits can come either from non-simultaneous radio ephemerides, from the minimum and maximum imposed by the data sampling; or from the grand minimum and maximum seen from all observations of similar pulsars.

Marginalized posterior odds.

But from the preceding section, we can already write down the marginalized posterior likelihood $\mathcal{O}$, given a rotation frequency. Hence the posterior for the frequency looks like:

$$\Lambda(\nu \mid \mathcal{H}_2, \{y_i\}, I)\mathrm{d}\nu \;=\; \kappa\frac{\mathrm{d}\nu}{\nu} \times \mathcal{O}(\{y_i\}) \;\mid\; \mathcal{H}_2, \mathcal{H}_0, I).$$

Notice for Bayesian methods it is the *range* of the trial frequencies and derivatives that are important; not the total *number of trials*. Also, this is usually multi–peaked. Hence, even once one has found the mode (largest peak), using χ^2 tables to tell one what drop in log–likelihood corresponds to 67.43%, 95.45%, or 99.73% (i.e. 1, 2, and 3σ) will not work. But with Bayesian inference, we can simply step through a grid of ν values to directly calculate the drop in log-likelihood that encloses each volume of probability (see [4, 18, 56]). We show a figure of the log–likelihood vs ν. and delineate the Credible Regions containing 95.45%, and 99.73% of the total volume posterior probability.

2.4 Conclusion: Challenges Subtle and Grand

There are many simple, low–dimension problems, of great importance yet basic enough to give a student, left to be done. (Typically The worst problem for the student will be data I/O.) For more ambitious, higher dimensional problems one will need more sophisticated computer techniques (such as the kind of modular EM/DA/MCMC presented here [57]). Familiarity with Bayesian fundamentals and computation details can also give one a fresh perspective on frequentist methods (even bootstrapping! [33]). And, finally, your work, as you collaborate further with experts in statistics, may open up ways of looking at the data that most of us cannot begin to visualize yet.

Acknowledgments: Most of the worked example on high energy pulsars was freely taken from work with A. Carramiñana. N. Kaiser kindly provided a

sketch of the CBR history. A. C. acknowledges the hospitality of Wellesley College, UNH, and MIT; her collaborators in AstroStatistics (esp. J. Scargle, V. Kashyap, T. Loredo, E. Kolaczyk, A. Siemiginowska, and D. van Dyk); NASA contract NAG5-7984; and the AISRP "AS-DATA" grant.

2.5 References

[1] T. Loredo. In *Statistical Challenges in Modern Astronomy*, Springer-Verlag, 1992.
[2] J. Berger, in *Statistical Challenges in Modern Astronomy II*, ed. E. Feigelson and G. Babu, p 15.
[3] Connors, A., in *Statistical Challenges in Modern Astronomy II*, ed. E. Feigelson and G. Babu, p 39.
[4] Lampton, M., Margon, B. and Bowyer, S., 1976, *ApJL*, **208**, L177.
[5] Cash, W. 1979, *ApJ* 228, 939.
[6] Wachter, K., Leach, R., Kellog, E., 1979, it ApJ, 230, 274.
[7] Avni, Y., 1978, A&A, **66**, 307.
[8] Feigelson, E.D., Babu, G.J.,, and Guth, J. 1992, *ApJ*, 397, 55.
[9] Protassov et al., 2001, to be published in ApJ.
[10] E. T. Jaynes. *Probability Theory: The Logic of Science*. Kluwer, in process. http://bayes.wustl.edu/
[11] Bijaoui, A.B., *A & A*, **13**, 226, 1971.
[12] Bijaoui, A.B., 1971, ,*A & A*, **13**, 232.
[13] Richardson, W. H., 1972, *J. Opt. Soc. Am.*, 62, 55.
[14] Lucy, L.B., 1974, *AJ*, 79, 745L.
[15] van Dyk, D. A. 2001, submitted to *ApJ*.
[16] van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. 2001, *ApJ*, 548. 224.
[17] Cornwell, T.J.,1979, in *Image Formation from Coherence Functions in Astronomy, Proceedings of IAU Colloq. 49*, Ed., C. van Schooneveld, (C. D. Reidel ,(Astrophysics and Space Science Library) Volume 76, 227
[18] J. Skilling. in *Maximum Entropy and Bayesian Methods*, ed. P. Fougere, Kluwer, Dordrecht, 341, 1990.
[19] Uson, J.M. and Wilkinson, D.T., 1984, *Nature*, 312, 427.
[20] Boynton, P.E. and Partridge, R.B., 1973, *ApJ*, 181, 243.
[21] Kaiser, N., and Lasenby, A.N., 1988 "unpublished paper; available from authors on request".
[22] Davies, R.D., et al., 1987, *Nature*, 326, 462.
[23] Lasenby, R.N., and Davies, R.D., 1988, in *Large Scale Motions in the Universe*, ed. Rubin, V.C., Coyne, G.V. (Vatican Press, Princeton U. Press), p 277.
[24] Church, S.E., Lasenby, A.N., & Hills, R.E., 1993, *MNRAS*, 261, 705.
[25] Loredo, T.J., http://www.astro.cornell.edu/staff/loredo/bayes/.

[26] Freeman, P. E., et al., 1999, *ApJL* 524:753-771
[27] Kashyap, V., Drake, J. J., 1998, *ApJ*, 503, 450.
[28] Graziani et al., 1992, In: *Gamma-ray bursts; AIP Conf Proc. 265* (New York: AIP), p211.
[29] J. Berger, invited talk, AAS 194.
http://www.stat.duke.edu/∼berger/papers.html
[29] Feigelson, E., 1996, Comment at the public Bayes session at SCMA II, Penn State.
[30] (1971) Eadie, W.T., Drijard, D., James, F.E., Roos, M., and Sadoulet, B., Statistical Methods in Experimental Physics (New York: North-Holland)
[31] Bevington, P. R., 1969, *Data Reduction and Error Analysis for the Physical Sciences*. (New York: McGraw-Hill)
[32] Tanner, M.A., 1985, *Tools for Statistical Inference*. Kluwer.
[33] Young, C.A., et al, this conference
[34] Nowak and Kolaczyk, 2001 (preprint).
[35] Kolaczyk, E., this conference
[36] Esch, D., et al, this conference
[37] Connors et al., 2001, *Proc. of the 5th Huntsville Gamma–Ray Burst Symposium, AIP Conference Proc.*, ed. Kippen, R.M., Mallozzi, R.S., Connaughton, V. (New York: AIP).
[38] Seward, F.D., 1989, *Space Science Reviews*, 49, 385.
[39] Thompson, D.J., 2001 in High Energy Gamma-Ray Astronomy, 103.
[40] de Jager et al. 2001, in High Energy Gamma-Ray Astronomy, 613
[41] Halpern, J. & Holt, S., 1992, *Nature*, 357, 222.
[42] Bertsch, D. et al., 1992, *Nature*, 357, 306.
[43] Gehrels, N., et al., 2000, *Nature*, 404, 363.
[44] Halpern et al., 2001, *ApJL*, 552, L125.
[45] Loredo, T.J., these proceedings.
[46] R. Buccheri and B. Sacco, in: *Data Analysis in Astronomy*, eds: L. Scarsi, V. Di Gesu, P. Crane, and S. Levialdi, Plenum Press, p 15, 1985.
[47] Leahy, D.A., Elsner, R.F., Weisskopf, M.C., 1983, *ApJL*, 272, 256.
[48] O.C. De Jager, J.W.H. Swanepoel, and B. C. Rubenheimer, *A & A*, **221**, 80, 1989.
[49] Connors, A. 1997, in "Proceedings of the Fifth Workshop in Data Analysis in Astronomy (DAA V)", ed. Di Gesu, V., Duff, M.J.B., Heck, A, Maccarone, M.C., Scarsi, L. and Zimmeran, H.U., (World Scientific Publishing Co., London), 251.
[50] Gregory, P. C. & Loredo, T. J. 1992, *ApJ*, 398, 146
[51] Jefferys, W. and Berger, J., 1992, *American Scientist*, 80, 64.
[52] Scargle, J, 1998., ApJ, 504, 405 .
[53] M. West, in: ed. E. Feigelson and G. Babu, Springer-Verlag, New York, 299, 1992.

[54] Jaynes, E.T., *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. Kluwer, 1978.
[55] Connors and Carramiñana 2001, these proceedings
[56] Bretthorst, G. L., 1988, *Bayesian Spectrum Analysis and Parameter Estimation*, in Lecture Notes in Statistics, 48, (Springer-Verlag, NY, New York). http://bayes.wustl.edu/
[57] van Dyk, D. A. 2001, these proceedings.

Commentary by Eric D. Kolaczyk[2]

The problem selected here by Dr. Connors, that of detecting the presence of changepoints in Poisson time series data, is somehow both simple and rich ... at the same time ... and thus a wonderful "tutorial" problem. And the fact that solutions to this problem have the potential for real scientific impact makes it even more interesting. Reference was made in the paper to a number of ways in which the basic principles and techniques outlined may be extended to deal with structures more complex than those addressed therein. I would like to describe one such extension – multiscale changepoint detections – and in particular show how it in fact derives from the same principles and techniques. In doing so, I also expand upon the summary presented in [1].

Basic Modeling Framework

Implicit in the formulation of Connors is the presence of binned photon counts. A total of N_{TOT} bins on the unit interval $[0, 1]$ are assumed. With this condition, one cannot hope to have the data indicate the location of possible changepoints ϕ beyond the resolution of these bins. So we introduce the notion of a model with two parts: (1) a hypothesis $\mathcal{H}_i$ that there are i changepoints, for $i = 0, 1, \ldots, N_{TOT} - 1$, and (2) a set of changepoint locations $\phi^{(i)} = (\phi_1, \ldots, \phi_i)$, restricted to some subset of the bin endpoints. Our models are thus of the form $\mathcal{M}_i = (\mathcal{H}_i, \phi^{(i)})$.

Now suppose that one wishes to find simultaneously the most likely number of changepoints *and* their location. This then, in the Bayesian paradigm, becomes a question of maximizing the quantity $p(\mathcal{M}_i|\mathbf{y})$ over all models $\mathcal{M}_i$, where $\mathbf{y} = (y_1, \ldots, y_{N_{TOT}})$ are the observed (binned) counts. The Bayes factor (BF) is often used as a device for comparison of models, and is defined as the ratio of the posterior to the prior odds. For example, the Bayes factor for comparing $\mathcal{M}_1 = (\mathcal{H}_1, \phi_1)$ and $\mathcal{M}_0 = \mathcal{H}_0$ is

$$\mathrm{BF} = \frac{p(\mathcal{M}_1|\boldsymbol{y}) \,/\, p(\mathcal{M}_0|\boldsymbol{y})}{p(\mathcal{M}_1) \,/\, p(\mathcal{M}_0)} = \frac{p(\boldsymbol{y}|\mathcal{M}_1)}{p(\boldsymbol{y}|\mathcal{M}_0)} = \frac{p(\boldsymbol{y}|\phi_1, \mathcal{H}_1)}{p(\boldsymbol{y}|\mathcal{H}_0)}\,.$$

[2]Department of Statistics, Boston University

Note that the last expression above shows that the Bayes factor here is essentially the statistic $\Lambda_2(\cdot|\cdot)$ in section 2.2 of Connors i.e., without having marginalized over the location of the changepoint. [Note too that, for simplicity, I have adopted a non-periodic model here.]

Re-Parameterizing the Model

Let's consider our comparison of $\mathcal{M}_1$ and $\mathcal{M}_0$ in more detail. Suppose that $\mathcal{M}_1$ is true i.e., there is a single changepoint and it is located at ϕ_1. This leaves the data parameterized simply by two means, say, μ_L and μ_R. And the statistics containing all relevant information in the data for these two parameters (technically, the *sufficient* statistics) are simply y_L and y_R, say, corresponding to the total counts in the bins to the left and right of ϕ_1, respectively. Similarly, in the case that $\mathcal{M}_0$ instead is true, the data are parameterized by a single mean $\mu_{TOT} = \mu_L + \mu_R$ and the sufficient statistic is just $y_{TOT} = y_L + y_R$. Hence our Bayes factor actually has the form

$$\mathrm{BF} = \frac{p(y_L, y_R|\mathcal{M}_1)}{p(y_{TOT}|\mathcal{M}_0)} \,.$$

But now consider that the pair (y_{TOT}, y_L) clearly contains the same information as (y_L, y_R) in model $\mathcal{M}_1$ and write the Bayes factor as

$$\mathrm{BF} = \frac{p(y_{TOT}|\mathcal{M}_1)}{p(y_{TOT}|\mathcal{M}_0)}\, p(y_L|y_{TOT}, \mathcal{M}_1) \,. \tag{2.1}$$

Writing with respect to our original parameterization (μ_L, μ_R), the term in the numerator of this expression is

$$p(y_{TOT}|\mathcal{M}_1) = \int p(y_{TOT}|\mu_L, \mu_R, \mathcal{M}_1)\, p(\mu_L, \mu_R|\mathcal{M}_1) d\mu_L d\mu_R \,.$$

But $p(y_{TOT}|\mu_L, \mu_R, \mathcal{M}_1)$ is just the probability mass function of a Poisson random variable with mean μ_{TOT} i.e., it depends on μ_L and μ_R only through their sum. And certainly the term $p(y_{TOT}|\mathcal{M}_0)$ in the denominator of equation (2.1) similarly only depends on μ_{TOT}. So if we choose prior distributions $p(\mu_{TOT}|\mathcal{M}_1) = p(\mu_{TOT}|\mathcal{M}_0)$ i.e., reflecting a belief that the total expected counts in the data is unaffected by whether there is a changepoint or not, then we obtain that $p(y_{TOT}|\mathcal{M}_1) = p(y_{TOT}|\mathcal{M}_0)$ and the Bayes factor becomes

$$\mathrm{BF} = p(y_L|y_{TOT}, \mathcal{M}_1) \,.$$

Finally, writing

$$p(y_L|y_{TOT}, \mathcal{M}_1) = \int p(y_L|\mu_L, \mu_R, y_{TOT}, \mathcal{M}_1) p(\mu_L, \mu_L|Y_{TOT}, \mathcal{M}_1) d\mu_L d\mu_R \,,$$

and noting that the first probability in the integral above is the probability mass function of a binomial random variable with parameters y_{TOT} and $f = \mu_L/\mu_{TOT}$, we see that

$$\text{BF} = \int p(y_L|f, Y_{TOT}, \mathcal{M}_1)p(f|Y_{TOT}, \mathcal{M}_1)df\,.$$

Which is just to say that our goal of testing for a changepoint at a given location ϕ_1 reduces to a comparison of two Poisson means, which in turn reduces to a statistic based on the binomial distribution (which also is the case, say, in the frequentist-based Neyman-Pearson theory for this reduced problem).

Comparing the above then to section 2.3 of Connors, we see (1) that the re-parameterization (μ_{TOT}, f) is a very natural one to make, and (2) why the parameter β in the exponential prior drops out of the statistic $\Lambda_2(\cdot|\cdot)$. In fact, the prior on μ_{TOT} as a whole, whatever it is, will drop out of this Bayes factor, as long as it chosen to be the same under both $\mathcal{M}_1$ and $\mathcal{M}_0$.

Extending the Basic Model

The principles above may be applied in a recursive manner to deal with multiple changepoint models as well i.e., models $\mathcal{M}_i$ with $2 \leq i \leq N_{TOT}-1$. Take the case of two changepoints, with $\mathcal{M}_2 = (\mathcal{H}_2, (\phi_1, \phi_2))$. There will be three intervals of interest, and hence three mean parameters (μ_L, μ_C, μ_R) and three summary statistics (y_L, y_C, y_R) of relevance. Now, without loss of generality, we can (exploiting the independence inherent in Poisson sampling) consider just the first two sub-intervals as a sub-problem of our full problem, and note that it is simply the single changepoint problem from above. In following our strategy of re-parameterization, we are led to the alternate statistics $(y_{L,C}, y_L)$ in place of (y_L, y_C), where $y_{L,C} = y_L + y_C$. Then treat the sub-intervals underlying $y_{L,C}$ and y_R as a pair and again apply our results from the single changepoint problem.

Although the above argument is heuristic, one can show formally that, for example, in comparing $\mathcal{M}_2$ to $\mathcal{M}_0$ we are led a Bayes factor of

$$\begin{aligned} \text{BF}_{2,0} &= \frac{p(y_{TOT}|\mathcal{M}_2)\,p(y_{L,C}|y_{TOT}, \mathcal{M}_2)\,p(y_L|y_{L,C}, \mathcal{M}_2)}{p(y_{TOT}|\mathcal{M}_0)} \\ &= p(y_{L,C}|y_{TOT}, \mathcal{M}_2)\,p(y_L|y_{L,C}, \mathcal{M}_2)\,, \end{aligned}$$

or in comparing $\mathcal{M}_2$ to $\mathcal{M}_1 = (\mathcal{H}_1, \phi_2)$ we get

$$\begin{aligned} \text{BF}_{2,1} &= \frac{p(y_{TOT}|\mathcal{M}_2)\,p(y_{L,C}|y_{TOT}, \mathcal{M}_2)\,p(y_L|y_{L,C}, \mathcal{M}_2)}{p(y_{TOT}|\mathcal{M}_1)\,p(y_{L,C}|y_{TOT}, \mathcal{M}_1)} \\ &= p(y_L|y_{L,C}, \mathcal{M}_2)\,. \end{aligned}$$

In both cases we exploit the assumption that priors are chosen for μ_{TOT} and the parameters f (of which there are now two) so as to be independent

of the underlying $\mathcal{M}$, as in the single changepoint example above. For example, a natural family to use here is the family of beta distributions with density function

$$p(f|\gamma_1, \gamma_2) = \frac{\Gamma(\gamma_1 + \gamma_2)}{\Gamma(\gamma_1)\Gamma(\gamma_2)} f^{\gamma_1 - 1} (1 - f)^{\gamma_2 - 1}$$

and positive parameters (γ_1, γ_2). Included in this family is the special case $\gamma_1 = \gamma_2 = 1$, which corresponds to the uniform distribution on $[0, 1]$, which is the prior used in section 2.3 of Connors.

Multiscale Changepoint Detection

The above arguments generalize to an arbitrary number of changepoints and it is not hard to see that a comparison of any two nested models involves a Bayes factor that is a product of conditional probabilities across various *scales* or resolutions of aggregations. For non-nested models one obtains a ratio of such products.

Kolaczyk and Nowak have studied probability models with this sort of multiscale structure in some detail. Formal links between them and wavelet-based methods can be made, including an analogue of multiresolution analysis (MRA) and various efficient computational algorithms for estimation and testing. See [1] and the references therein.

Here in the present context, in searching for an optimal number of changepoints and their locations, the product structure of the Bayes factors and a strong degree of redundancy among candidate intervals over all possible changepoint locations allows (somewhat surprisingly!) for the search over all possible models to be solved exactly using a dynamic programming algorithm of complexity $O(N_{TOT}^3)$. An example of results from applying this technique to gamma-ray burst data can be found in [1]. Interestingly, the same approach also can be derived from the perspective of minimum description length (MDL).

References

[1] Kolaczyk, E.D. *This volume.*

This page intentionally left blank

3

Hierarchical Models, Data Augmentation, and Markov Chain Monte Carlo

David A. van Dyk[1]

ABSTRACT The ever increasing power and sophistication of today's high energy astronomical instruments is opening a new realm of high quality data that is quickly pushing beyond the capabilities of the "classical" data-analysis methods in common use. In this chapter we discuss the use of highly structured models that not only incorporate the scientific model (e.g., for a source spectrum) but also account for stochastic components of data collection and the instrument (e.g., background contamination and pile up). Such hierarchical models when used in conjunction with Bayesian or likelihood statistical methods offer a systematic solution to many challenging data analytic problems (e.g., low count rates and pile up). Hierarchical models are becoming increasingly popular in physical and other scientific disciplines largely because of the recent development of sophisticated methods for statistical computation. Thus, we discuss such methods as the EM algorithm, data augmentation, and Markov chain Monte Carlo in the context of high energy high resolution low count data.
This paper is followed by a commentary by astronomer Michael Strauss.

3.1 Introduction

Today's highly sophisticated astronomical instruments offer a new window into the complexities of the visible and invisible universe. As the state of instrumentation evolves to produce ever finer resolution in spectral, spatial, and temporal data ever more sophisticated statistical techniques are required to properly handle this data. For example, standard off-the-shelf methods such as χ^2 fitting and background subtraction are ill-equipped to handle the high resolution low count per bin data available from such instruments as the Chandra X-ray Observatory. See Siemiginowska et al. (1997) and van Dyk et al. (2001) for a general discussion of such issues. The Gaussian assumptions implicit in such methods are not justified with low counts and the resulting fits and error bars are therefore unreliable. Testing

[1]Department of Statistics, Harvard University

for model features such as spectral lines or a source above background is always a challenging task and standard methods such as the F-test, likelihood ratio test, and Cash statistic though commonly used in practice are inappropriate (Protassov et al. 2002). An even greater challenge is properly accounting for pile-up in X-ray detectors, a task that confounds standard techniques and thus demands more sophisticated statistical methods.

In this chapter, we outline a paradigm for data analysis that we believe is robust enough to systematically handle these and many other statistical challenges presented by modern astronomical instruments. It is important conceptually to break any data analysis scheme into (at least) three components, all of which are critical and must be done thoughtfully to ensure sound inference. These components are *model building*, *statistical inference*, and *statistical computation*.

The importance of careful model building is evident in the complexity and subtlety of the physical mechanisms giving rise to the observed data of modern instrumentation. The instrument response blurs the energy and sky coordinates of photons, counts are contaminated with background, the effective area of the instrument and the propensity of photons to be absorbed vary with energy, pile-up masks the energy and count of incoming photons, source spectral models are complex and may include emission and absorption features as well as a continuum. A statistical model should aim to describe all such components of data generation. Thus, by a *model* we mean much more than a parametric description of how the mean source flux varies with energy and/or sky coordinates. Models that include statistical descriptions of the processes that degrade the data can guide us in accounting for these degradations and eliminate the need for ad-hoc corrections, e.g., for pile-up and background. Because of the complexity of these models, we organize them into a hierarchical structure, which is formulated in terms of various unobserved quantities (e.g., counts without background contamination). Such unobserved quantities are often called *augmented data* and play an important role in the computational methods we suggest.

Once a model is formulated, statistical inference involves drawing inferences (e.g., point estimates and error bars) regarding unobserved quantities such as the model parameters describing the flux of the source. Important model-based modes of statistical inference include maximum likelihood and Bayesian inference. With large samples the asymptotic Gaussian behavior of the maximum likelihood estimate can be the basis for sound frequentist inference. Nevertheless, we generally take a Bayesian perspective for a number of practical reasons such as a ready mechanism for combining information from multiple sources, mathematical justification in small samples, and an obvious framework for handling nuisance parameters. Despite the placement of this chapter in a Bayesian section, we say very little about the relative merit of Bayesian and frequentist methods; our emphasis is on model building and statistical computation. Because of the aforementioned

practical advantages of Bayesian methods, they are often the only tractable methods available for fitting complex models—which is motivation enough for many practical minded statisticians to "be Bayesian." Here we give only enough details of Bayesian and likelihood methods to motivate the computational tools, giving somewhat more emphasis to Bayesian methods. For further reading on Bayesian methods, we recommend one of the several high-quality recent texts on the subject such as Gelman et al. (1995), Carlin & Louis (1996) and Gilks et al. (1996), as well as other chapters in this volume including those by Connors, Loredo & Chernoff, and Berger *et al.*.

Because of the highly-structured nature of the statistical models that we propose, sophisticated computational methods (e.g., the EM algorithm, the data augmentation algorithm, and Markov chain Monte Carlo) are often required. The methods we suggest are designed to be computationally stable and generally easy to implement. The details of the algorithm often follow directly from the hierarchical model specification via simple statistical calculations.

The remainder of this chapter is organized into five sections. In Section 3.2 we introduce a simple example, accounting for background contamination of counts. We use this example to motivate hierarchical modeling and the method of data augmentation, which are in turn generalized and more fully developed in Section 3.3. The computational methods are introduced and illustrated using the motivating example of background contamination in Section 3.4. In Section 3.5 we outline how these methods can be used to tackle the difficult problem of photon pile-up. Concluding remarks regarding the direction of modern statistical analysis appear in Section 3.6.

3.2 A Motivating Example

In this section we introduce a simple example that is used throughout the chapter to motivate ideas and methods. The example is simple so as not to distract attention from the statistical methods. As illustrated in Section 3.5, however, hierarchical models, data augmentation, and MCMC can tackle much more complicated problems.

Suppose we have observed counts, Y, contaminated with background in a (source) exposure and have observed a second exposure of pure background resulting in counts, Z. Throughout we assume the source exposure is τ_S seconds and the pure background exposure is τ_B seconds with both exposures using the same area of the detector. To model the source exposure, we assume Y follows a Poisson distribution[2] with intensity $\lambda_B + \lambda_S$, where

[2]Recall $Y \overset{\mathrm{d}}{\sim}$ Poisson (λ) (read as Y is distributed as Poisson with intensity λ) indicates that Y follows the Poisson distribution with intensity and expectation λ, i.e.,

λ_B and λ_S represent the expected counts during the source exposure due to background and source respectively. Thus, the distribution function for Y given λ_B and λ_S is

$$p(Y|\lambda_B, \lambda_S) = e^{-(\lambda_B+\lambda_S)}(\lambda_B + \lambda_S)^Y/Y! \text{ for } Y = 0, 1, 2, \ldots. \tag{3.1}$$

We wish to estimate λ_S and treat λ_B as a *nuisance* parameter, a parameter that is of little interest, but must be included in the model. The expected counts during the background exposure are assumed to be the same as in the source exposure, but corrected for the exposure time, $\lambda_B\tau_B/\tau_S$. I.e.,

$$p(Z|\lambda_B, \lambda_S) = e^{-(\lambda_B\tau_B/\tau_S)}(\lambda_B\tau_B/\tau_S)^Z/Z! \text{ for } Z = 0, 1, 2, \ldots. \tag{3.2}$$

Maximum likelihood estimation involves estimating λ_B and λ_S by the values the maximize the likelihood function, i.e., the product of Equations 3.1 and 3.2. Under certain regularity conditions (e.g., $\lambda_B, \lambda_S > 0$), maximum likelihood estimates asymptotically follow a Gaussian distribution. This result leads to confidence intervals and error bars with (asymptotic) frequentist properties.

Bayesian inference is based on the *posterior* distribution,

$$p(\lambda_S, \lambda_S B|Y, Z) \propto p(Y|\lambda_B, \lambda_S)p(Z|\lambda_B, \lambda_S)p(\lambda_B, \lambda_S), \tag{3.3}$$

where $p(\lambda_B, \lambda_S)$ is the *prior* distribution which quantifies information regarding the values of the λ_S and λ_B available prior to observing the data. The posterior distribution combines such prior information with the information in the observed counts. The posterior distribution is a complete summary of our information, but if it is similar to Gaussian in shape, it is often summarized by its mean vector and variance matrix that can be used as point estimates and to compute error bars. The posterior distribution can also be used to compute a ζ-level credible region, R, such that $\int_R p(\lambda_S, \lambda_B|Y, Z)d\lambda_S d\lambda_B = \zeta$. Such probability statements should be regarded as summaries of the available information for the model parameters, in contrast to the frequentist interpretation of a confidence interval.

Implicitly, the counts from the source exposure, Y, are made up of two components, $Y = Y_S + Y_B$, where Y_S are counts from the source exposure due to the source and Y_B are the counts due to background. Since neither Y_S nor Y_B are observed, we call these counts *missing data*. We note that if Y_S and Y_B had been observed, our statistical analysis would be greatly simplified since we could confine attention to $Y_S \overset{\text{d}}{\sim} \text{Poisson}(\lambda_S)$. Of course, it is impossible to observe Y_S and Y_B. Nonetheless, this "thought experiment" offers insight into computational methods that are useful both for Bayesian and likelihood-based inference. In particular, the method of data augmentation is an elegant computational construct allowing us to take

$p(Y = y) = e^{-\lambda}y^\lambda/y!$.

advantage of the fact that if it were possible to collect additional data, statistical analysis could be greatly simplified. This is true regardless of why the so-called "missing-data" are not observed. There is a large class of powerful statistical methods designed for "missing-data" problems. These methods have broad application in astrophysics (and in the physical sciences generally) once we note that quantities observed with measurement error can be regarded as "missing-data".

To illustrate the method of data augmentation, we begin by reformulating our model in terms of Y_S and Y_B. In particular, consider the multi-level or *hierarchical model*

Level 1: $Y|Y_B, \lambda_S \overset{\text{d}}{\sim} \text{Poisson}(\lambda_S) + Y_B$,

Level 2: $Y_B|\lambda_B \overset{\text{d}}{\sim} \text{Poisson}(\lambda_B)$ and $Z|\lambda_B \overset{\text{d}}{\sim} \text{Poisson}(\lambda_B \tau_B/\tau_S)$,

Level 3 (optional): specify a prior distribution for λ_B and λ_S.

Notice that in each level of the model, we specify the distribution of random quantities conditioning on unobserved quantities whose distribution is specified in lower levels of the model. For example, in Level 1, we condition on Y_B, the distribution of which is specified in Level 2. *The power of such a hierarchical model is that it separates a complex model into a number of easy to handle smaller parts.*

If Y_S and Y_B were observed, Level 1 specifies the form of the likelihood for λ_S, i.e.,

$$L(\lambda_S|Y_S) = e^{-\lambda_S}\lambda_S^{Y_S}, \tag{3.4}$$

and Level 2 specified the form of the likelihood for λ_B, i.e.,

$$L(\lambda_B|Y_B, Z) = e^{-\lambda_B k}\lambda_B^{Y_B+Z}, \tag{3.5}$$

where $k = (\tau_S + \tau_B)/\tau_S$. Notice that Equations 3.1 and 3.2 are relatively complex functions of λ_S and λ_B and are harder to, for example, maximize than are Equations 3.4 and 3.5.

It is also easy to estimate the "missing data" in this hierarchical model. In particular, if λ_B and λ_S were known, the conditional distribution of Y_B given Y can be computed using Bayes Theorem,

$$p(Y_B|Y, \lambda_S, \lambda_B) = \frac{p(Y|Y_B, \lambda_S, \lambda_B)\; p(Y_B|\lambda_S, \lambda_B)}{p(Y|\lambda_S, \lambda_B)} \tag{3.6}$$

$$= \binom{Y}{Y_B}\left(\frac{\lambda_B}{\lambda_S + \lambda_B}\right)^{Y_B}\left(\frac{\lambda_S}{\lambda_S + \lambda_B}\right)^{Y-Y_B}. \tag{3.7}$$

That is,

$$Y_B|Y, \lambda_S, \lambda_B \overset{\text{d}}{\sim} \text{Binomial}^{3}\left[Y, \lambda_B/(\lambda_S + \lambda_B)\right]. \tag{3.8}$$

[3]Recall $Y \overset{\text{d}}{\sim} \text{Binomial}(n, P)$ indicates that Y follows a binomial distribution with n

Thus, given the model parameters, we can predict the "missing data" (e.g., by its conditional expectation with error bars based on its conditional standard deviation). Likewise, given the "missing data" we can estimate the model parameters (e.g., using maximum likelihood or a Bayesian estimate). This leads to an iterative strategy that updates the "missing data" given the model parameters and then the model parameters given the "missing data." Such computational methods include the EM algorithm and the Data Augmentation (DA) algorithm and are referred to generally as the method of data augmentation.

In the next two sections we abstract and generalize the important features of this example to construct robust tools for analysis of the high resolution high quality data available with today's sophisticated instruments. In Section 3.4 we show how data augmentation can be used to compute maximum likelihood estimates, Bayesian posterior modes and means, as well as error bars. Generally these methods involve maximizing, simulating, and computing expectations of standard distribution functions. Such simple distributions often arise naturally from a hierarchical model expressed in terms of the "missing data," e.g., Equations 3.4, 3.5, and 3.8. Details of the computation stability as well as examples which illustrate the computational simplicity appear in the following sections.

3.3 Data Augmentation and Hierarchical Models

The term "data augmentation" originated with computational methods designed to handle missing data, but as illustrated in Section 3.2, the method is really quite general and often useful when there is no missing data per se. In particular, for Monte Carlo integration in Bayesian data analysis we aim to obtain a sample from the posterior distribution, $p(\boldsymbol{\theta}|\mathrm{Y})$. In some cases, we can *augment* the model to $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$, where X may be missing data or any other unobserved quantity (e.g., counts due to background). With judicious choice of X, it may be much easier to obtain a sample from $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$ than directly from $p(\boldsymbol{\theta}|\mathrm{Y})$. Once we have a sample from $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$, we simply discard the sample of X to obtain a sample from $p(\boldsymbol{\theta}|\mathrm{Y})$. The notation here is more general, but the idea is exactly that of Section 3.2; we use statistical insight to construct $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$ so that both $p(\boldsymbol{\theta}|\mathrm{X}, \mathrm{Y})$ and $p(\mathrm{X}|\boldsymbol{\theta}, \mathrm{Y})$ are simple or at least standard distributions.

Absorption Lines. Absorption can be accounted for by supposing the expected counts in energy bin i are $\mathcal{F}_i\pi_i$, where $\mathcal{F}_i$ would be the expected counts if there were no absorption and π_i is the expected proportion of

independent trials each with probability p, i.e., $\Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$. As an example, Y may be the random number of heads in n independent flips of a (possibly unfair) coin that has probability p of coming up heads on each flip.

counts in energy bin i that are not absorbed. (We might, for example, parameterize $\mathcal{F}_i$ as a power law.) In particular, we might model the counts in energy bin i as $Y_i \overset{\mathrm{d}}{\sim}$ Poisson $(\mathcal{F}_i \pi_i)$. To formulate this model using data augmentation, we let Y_i^+ be the unobserved counts that the detector would have detected if no photons were absorbed. We can then formulate the hierarchical model,

Level 1: $Y_i | Y_i^+, \mathcal{F}_i, \pi_i \overset{\mathrm{d}}{\sim}$ Binomial (Y_i^+, π_i),

Level 2: $Y_i^+ | \mathcal{F}_i \overset{\mathrm{d}}{\sim}$ Poisson $(\mathcal{F}_i)$,

Level 3 (optional): specify prior distributions for $\mathcal{F}_i$ and π_i.

Again, the power of the data augmentation is the ability to partition the model complexity into simpler pieces, in this case a binomial absorption model and a Poisson spectral model with no absorption.

Many standard absorption models (including absorption lines) and continuum spectral models (e.g., power laws and bremsstrahlung emission) can be formulated using simple transformations of π_i and $\mathcal{F}_i$ that are linear functions of energy. In this case, given the "missing" absorbed photon counts both Level 1 and Level 2 specify Generalized Linear Models that are well studied and generally easy to fit. Likewise, given the model parameters and the observed data, the absorbed photons follow a simple model, $Y_i^+ \overset{\mathrm{d}}{\sim}$ Poisson $[(1 - \pi_i)\mathcal{F}_i] + Y_i$.

Emission Lines. Spectral models often include emission lines,

$$\mathcal{F}_i = c(E_i) + \sum_{k=1}^{K} \delta_{ik}$$

where $c(E_i)$ is the expected continuum counts in energy bin i and δ_{ik} is the expected counts from emission like k in energy bin i. For each photon, we postulate a variable that specifies whether the photon is due to the continuum or a particular emission line. This unobserved specification variable is treated as "missing data." Given this variable we can fit the continuum using the counts due to the continuum without the complication of emission lines. Likewise we can fit each of the emission lines (e.g., parameters specifying a Gaussian or Lorenzian distribution) using the counts attributed to that line. Conversely, given the parameter of continuum and the emission lines, the specification variable for each photon follows a simple multinomial distribution.

Multiple Model Components. So far, we have divided the unobserved quantities into two groups, the model parameters and the "missing data." More generally, we may partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$, where some component of θ are model parameters of scientific interest, others may be

nuisance parameters, and still others may be "missing data" or other unobserved quantities. The key is that we select the unobserved quantities and the partition of $\boldsymbol{\theta}$ so that $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_{k+1}, \ldots, \boldsymbol{\theta}_p, \mathrm{Y})$ is a standard distribution for each k. In this way we partition a complex problem into a sequence of simpler standard problems which we handle iteratively and one at a time. Thus, we can easily account for absorption, emission lines, instrument response, and background, all in the context of a Poisson model without sacrificing numerical stability, computational simplicity, or sound statistical inference. Details of such a model appear in van Dyk et al. (2001); see also van Dyk's discussion of Strauss (this volume).

3.4 Model Fitting

In Sections 3.2 and 3.3 we emphasize repeatedly that judicious choice of the "missing data," X, can lead to simple conditional models, $p(\boldsymbol{\theta}|\mathrm{X}, \mathrm{Y})$ and $p(\mathrm{X}|\boldsymbol{\theta}, \mathrm{Y})$, even when $p(\boldsymbol{\theta}|\mathrm{Y})$ is much more complex. In this section we show how these simple conditional models can be used to construct computation tools for likelihood-based and Bayesian model fitting. In recent years, these tools have become popular throughout the social, physical biological and engineering sciences primarily because of their computational stability and simplicity.

3.4.1 The EM Algorithm

Dempster et al. (1977) formulated the expectation maximization (EM) algorithm to compute a maximum likelihood estimate, that is

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\mathrm{Y}), \tag{3.9}$$

where Y is the observed data, $\boldsymbol{\theta}$ is a model parameter, $L(\boldsymbol{\theta}|\mathrm{Y})$ is the likelihood function, and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate. (More generally, we can replace $L(\boldsymbol{\theta}|\mathrm{Y})$ with a posterior distribution in Equation 3.9 and use EM to compute the posterior mode, $\hat{\theta}$.) In particular, Dempster et al. (1977) considered maximum likelihood estimation in the presence of incomplete data or problems that can be formulated as such (e.g., spectral imaging with background or degraded counts). In this context, the EM algorithm builds on the intuitive idea that (i) if there were no "missing data," maximum likelihood estimation would be easy, and (ii) if the model parameters were known, the "missing data" could easily be imputed (i.e., predicted) by its (conditional) expectation.

These two steps take on a simple form in the context of the background example described in Section 3.2. In particular, if Y_S had been observed, we could estimate λ_S with Y_S. Likewise, if λ_S and λ_B were known, Y_S could be estimated as the proportion of the observed counts, Y, implied by

λ_S and λ_B, i.e., the conditional expectation of Y_S, $Y\lambda_S/(\lambda_B + \lambda_S)$. This leads naturally to a two-step iteration which converges to the maximum likelihood estimate. It should be noted that this procedure necessarily leads to a non-negative estimate of λ_S, whereas the common estimate resulting from "subtracting background," $Y - Z\tau_S/\tau_B$, may be negative.

The two steps in this simple iteration correspond to the M-step (i.e., maximization step) and the E-step (i.e., expectation step) of EM respectively, with the proviso that not the missing data, but rather the so-called augmented-data log likelihood should be imputed by its conditional expectation. In general, we begin by defining the *missing data*, X, and the corresponding loglikelihood, $L(\boldsymbol{\theta}|\mathrm{Y}, \mathrm{X})$. EM starts with an initial value[4] $\boldsymbol{\theta}^{(0)} \in \Theta$ and iterates the following two steps for $t = 0, 1, \ldots$

E-step: Compute the conditional expectation of the loglikelihood corresponding to the augmented data (Y, X), given the observed data and the current parameter value,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}\left[\log L(\boldsymbol{\theta}|\mathrm{Y}, \mathrm{X})\middle|\mathrm{Y}, \boldsymbol{\theta}^{(t)}\right] \ ; \tag{3.10}$$

M-step: Determine $\boldsymbol{\theta}^{(t+1)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, that is, find $\boldsymbol{\theta}^{(t+1)}$ so that $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ for all $\boldsymbol{\theta} \in \Theta$;

until convergence. The usefulness of the EM algorithm is apparent when both of these steps can be accomplished with minimal analytic and computation effort but the direct maximization in Equation 3.9 is difficult. In many common models (e.g., multivariate Gaussian, Poisson, binomial, exponential, etc.) $\log L(\boldsymbol{\theta}|\mathrm{Y}, \mathrm{X})$ is linear in a set of simple "augmented-data sufficient statistics." Thus, as will be illustrated below, computing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ involves routine calculations. The M-step then only requires computing the maximum likelihood estimates as if there were no "missing data," by using the predicted augmented-data sufficient statistics from the E-step as data.

To illustrate these ideas, we return to the example of Section 3.2. We set $\mathrm{X} = \{Y_S, Y_B\}$, $\mathrm{Y} = \{Y, Y_B\}$, and $\boldsymbol{\theta} = (\lambda_B, \lambda_S)$. In this case, $\log L(\boldsymbol{\theta}|\mathrm{Y}, \mathrm{X}) = \log L(\lambda_S|Y_S) + \log(\lambda_B|Y_B, Z)$; see Equations 3.4 and 3.5. Thus, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =$

$$-\lambda_S + \mathrm{E}\left(Y_S|Y, \boldsymbol{\theta}^{(t)}\right)\log\lambda_S - k\lambda_B + \left[Z + \mathrm{E}\left(Y_B|Y, \boldsymbol{\theta}^{(t)}\right)\right]\log\lambda_B. \tag{3.11}$$

Elementary calculations show the expectations in Equation 3.11 are given by $Y\lambda_S/(\lambda_B + \lambda_S)$ and $Y\lambda_B/(\lambda_B + \lambda_S)$, which is the E-step, and $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is maximized by $\lambda_S^{(t+1)} = \mathrm{E}(Y_S|Y, \lambda_S^{(t)})$ and $\lambda_B^{(t+1)} = [Z + \mathrm{E}(Y_B|Y, \lambda_S^{(t)})]/k$, which is the M-step.

[4]Parenthetic superscripts indicate iteration number.

3.4.2 The Data Augmentation Algorithm

In the context of Bayesian data analysis, numerical summaries of the posterior distributions are often computed via numerical integration. Because of the high dimension of the parameter space in most practical problems, Monte Carlo integration is really the only useful method. If we can obtain a sample from the posterior distribution, $\{\boldsymbol{\theta}^{(t)}, t = 1, \ldots, T\}$, Monte Carlo integration approximates the posterior mean of any function, g, of the parameter with

$$\mathrm{E}[g(\boldsymbol{\theta})|\mathrm{Y}] = \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathrm{Y}) d\boldsymbol{\theta} \approx \frac{1}{T} \sum_{t=1}^{T} g(\boldsymbol{\theta}^{(t)}), \tag{3.12}$$

where we assume $\mathrm{E}[g(\boldsymbol{\theta})|\mathrm{Y}]$ exists. For example, $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $g(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathrm{E}(\boldsymbol{\theta}|\mathrm{Y}))(\boldsymbol{\theta} - \mathrm{E}(\boldsymbol{\theta}|\mathrm{Y}))'$ lead to the posterior mean and variance respectively. Probabilities, such as $\zeta = \Pr(\boldsymbol{\theta} \in R)$ can be computed using $g(\boldsymbol{\theta}) = I\{\boldsymbol{\theta} \in R\}$, where the function I takes on value 1 if the condition in curly brackets holds and zero otherwise. Likewise, quantiles of the distribution can be approximated by the corresponding quantiles of the posterior sample. In short, a robust data analysis requires only a sample from the posterior distribution.

In the highly structured models we described in Section 3.3 we must use sophisticated algorithms to obtain a posterior sample. Here we introduce the powerful *Data Augmentation (DA) algorithm* (Tanner & Wong 1987). A description of the more general *Gibbs sampler* (Metropolis et al. 1953) and *Metropolis-Hastings algorithms* (Hastings 1970) with applications in astronomy can be found in (van Dyk et al. 2001). All of these algorithms construct a Markov chain with stationary distribution equal to the posterior distribution (e.g., Gelfand & Smith 1990); i.e., once the chain has reached stationarity, it generates samples which are identically (but not independently) distributed according to the posterior distribution. These samples can then be used for Monte Carlo integration; hence these algorithms are known as Markov chain Monte Carlo or MCMC methods. (See Tierney [1996] for regularity conditions for using Equation 3.12 with MCMC draws [11].) From the onset then, it is clear that three important concerns when using MCMC in practice are (1) selecting starting values for the Markov chain, (2) detecting convergence of the Markov chain to stationarity, and (3) the effect of the lack of independence in the posterior draws. Space does not allow us to address all of these practical issues. Instead we direct interested readers to van Dyk et al. (2001) and the references therein.

In order to obtain a sample from $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$, the DA algorithm uses an iterative sampling scheme that samples first X conditional on $\boldsymbol{\theta}$ and Y and second samples $\boldsymbol{\theta}$ given (X, Y). Clearly, the DA algorithm is most useful when both of these conditional distributions are easily sampled from. The iterative character of the resulting chain naturally leads to a Markov chain,

which we initialize at some starting value, $\boldsymbol{\theta}^{(0)}$. For $t = 1, \ldots, T$, where T is dynamically chosen, we repeat the following two steps:

Step 1: Draw $\mathrm{X}^{(t)}$ from $p(\mathrm{X}|\mathrm{Y}, \boldsymbol{\theta}^{(t-1)})$,

Step 2: Draw $\boldsymbol{\theta}^{(t)}$ from $p(\boldsymbol{\theta}|\mathrm{Y}, \mathrm{X}^{(t)})$.

Since the stationary distribution of the resulting Markov chain is the desired posterior distribution, for large t, $\boldsymbol{\theta}^{(t)}$ approximately follows $p(\boldsymbol{\theta}|\mathrm{Y})$.

To illustrate the utility of the algorithm, we return to the background contamination model introduced in Section 3.2. Given some starting value, $\boldsymbol{\theta}^{(0)} = (\lambda_B^{(0)}, \lambda_S^{(0)})$ the two steps of the algorithm at iteration t become

Step 1: Draw $Y_B^{(t)}$ using the binomial distribution given in Equation 3.8 and set $Y_S^{(t)} = Y - Y_B^{(t)}$.

Step 2: Draw $\lambda_B^{(t)}$ and $\lambda_S^{(t)}$ from independent γ distributions[5]

$$\lambda_B^{(t)} \Big| Y_B^{(t)} \overset{\mathrm{d}}{\sim} \gamma\left(\alpha_B + Y_B + Z, \beta_B + k\right) \text{ and } \lambda_S^{(t)} \Big| Y_S^{(t)} \overset{\mathrm{d}}{\sim} \gamma\left(\alpha_S + Y_S, \beta_S + 1\right). \tag{3.13}$$

Here $\alpha_B, \beta_B, \alpha_S$, and β_S are hyperparameters which quantify prior information via a prior γ distribution on λ_S and λ_B; see van Dyk et al. (2001) for guidance in selecting these parameters. In the first step, we stochastically divide the source count into source counts and background counts based on the current values of λ_B and λ_S. In the second step we use this division to update λ_B and λ_S. Markov chain theory tells us the iteration converges to the desired draws from the posterior distribution.

3.5 Accounting for Pile Up

Pile-up occurs in X-ray detectors when two or more photons arrive in a single spatial cell during the same time frame (i.e., the discrete time units). Such coincident events are counted as a single event with energy equal to the sum of the energies of each of the individual photons. Thus, for bright sources pile-up can seriously distort both the count rate and the energy spectrum. Accounting for pile-up is perhaps the most important outstanding data-analytic challenge for Chandra. Conceptually, however, there is no difficulty in addressing pile-up in a hierarchical Bayesian framework using MCMC; we must stochastically separate a subset of the observed

[5]The γ (α, β) distribution is a continuous distribution on the positive real line with probability density function $p(Y) = \beta^{\alpha} Y^{\alpha-1} e^{-\beta Y}/\Gamma(\alpha)$, expected value α/β, and variance α/β^2 for positive α and β.

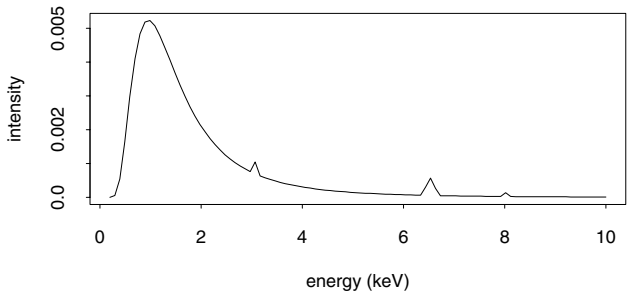


FIGURE 3.1. A Typical Energy Spectrum. We plot the expected photon count per bin per time frame as a function of energy and illustrate the smooth continuum with three small emission lines. This spectrum is plotted at low resolution (100 energy bins) to reduce the computational burden required for handling pile-up; see Figures 3.2.

counts into multiple counts of lower energy while conditioning on the current iteration of the model being fit. The attraction of hierarchical models in this setting is that they allow us to handle pile up ignoring all other model components. That is, when we separate counts into multiple counts of lower energy, the spectral model is completely specified and all the other degradations of the data (e.g., instrument response and background contamination) are accounted for by conditioning on the appropriate "missing data." Thus, we can attack pile up as an isolated problem.

Unfortunately, even in isolation handling pile up is challenging. The difficulty lies in computation. Simply enumerating the set of photons that could result in a particular observed event, let alone their relative probabilities, is an enormous task. Nonetheless, we believe there is great promise in Monte Carlo techniques which if carefully designed, can automatically exclude numerous possibilities that have minute probability. As an illustration, Figure 3.2 plots the conditional distribution of the energy of one of two photons with energy summing to 10 keV, assuming the energy spectra is as in Figure 3.1 and the point spread function is uniform across some area of the detector. The symmetry of the distribution in Figure 3.2 reflects the exchangeability of the component photon energies and the modes arises from the spectral emission lines in Figure 3.1. In practice, an observed energy can be the sum of more than two actual photon energies; in this case there is an 8% chance that there are three photons (and a 61% chance of only one photon, 29% chance of two photons, and 1% chance of four photons).

Care must be taken to efficiently sample from such complex distributions.

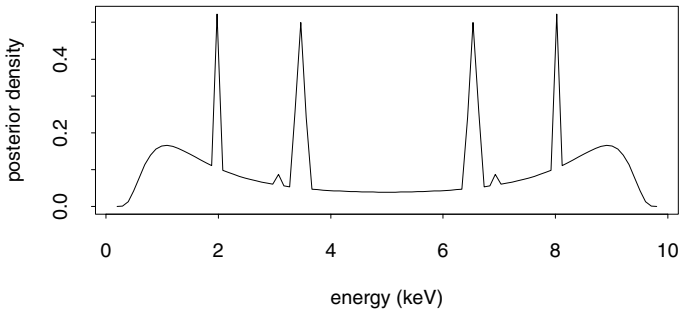


FIGURE 3.2. Un-piling Two Photons. The plot illustrates the conditional distribution of the energy of one of two photons with energy summing to 10 keV, assuming the energy spectra is as in Figure 3.1 and a uniform point spread function. Sophisticated Monte Carlo methods are required to simulate such a highly multi-modal distribution.

Development of Monte Carlo samplers for this task is an area of current research. Nonetheless, even with substantial simplifying assumptions (e.g., at most two photons can pile) preliminary results from our hierarchical model fit via MCMC show great promise. An example is given in the contributed paper by Kang et al. (this volume).

3.6 The Future of Data Analysis

The highly structured models described in this chapter reflect a new trend in applied statistics—it is becoming ever more feasible to build application specific models which are designed to account for the hierarchical and latent structures inherent in any particular data generation mechanism. Such multi-level models have long been advocated on theoretical grounds, but recently the development of new computational tools such as those described here has begun to bring such model fitting into routine practice. Although these methods offer great promise, they are by no means statistical black boxes that will automatically solve any problem. The flexibility of such models and computational methods require users to be statistically savvy. We, however, believe the benefits of superior scientific modeling far outweigh these costs. Indeed the future of data analysis lies in sophisticated application-specific modeling and methods.

Acknowledgments: The author gratefully acknowledges funding for this project partially provided by NSF grant DMS-01-04129 and by NASA contract NAS8-39073 (CXC). This chapter is a result of a joint effort of the members of the Astro-Statistics group at Harvard University whose members include A. Connors, D. Esch, P. Freedman, C. Hans, H. Kang, V. L. Kashyap, R. Protassov, D. Rubin, A. Siemiginowska, N. Sourla, and Y. Yu.

3.7 References

[1] Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

[2] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc., Ser. B,* **39**, 1–37.

[3] Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

[4] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

[5] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in Practice*. Chapman & Hall, London.

[6] Hastings, W. K. (1970). Monte Carlo sampling methods usings Markov chains and their applications. *Biometrika* **57**, 97–109.

[7] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phy.* **21**, 1087–1092.

[8] Protassov, R., van Dyk, D., Connors, A., Kashyap, V., and Siemiginowska, A. (2002). Statistics: Handle with care – detecting multiple model components with the likelihood ratio test. *Astrophysical J.* to appear.

[9] Siemiginowska, A., Elivis, M., Alanna, C., Freeman, P., Kashyap, V., and Feigelson, E. (1997). in *Statistical Challenges in Modern Astronomy II (eds. E. Feigelson and G. Babu)*, 241–253. Springer-Verlag, New York.

[10] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82**, 528–550.

[11] Tierney, L. (1996). in *Markov Chain Monte Carlo in Practice (eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter)*. Chapman & Hall, London.

[12] van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *Astrophysical J.* **548**, 224–243.

Commentary by Michael A. Strauss[6]

Astronomers often find themselves tackling complicated likelihood problems. With some basic knowledge of the underlying statistics of a given astronomical problem, and some familiarity with likelihood functions and Bayesian statistics, we often are able to write down a likelihood function in closed form. However, if the problem is complicated enough (read "interesting", as it usually is), we are stymied when it comes time to maximize this likelihood, especially if there is an interesting and complicated parameter space to fit for. This paper describes useful techniques for solving exactly this sort of problem, which are common in astronomy, by a "divide and conquer" approach, doing the problem iteratively. The very nasty problem of deconvolving the effects of "pile-up" in X-ray spectra is a particularly good example of this.

Another problem which may be amenable to this approach is illustrated in Figure 3.3, which shows the spectrum of a quasar from the Sloan Digital Sky Survey (see my contribution to these proceedings). The spectrum shows a blue continuum with strong superposed emission lines. Blueward (to the left) of the Lyα emission line of hydrogen are superposed a large number of absorption lines of Lyα, due to filaments and wisps of hydrogen gas at redshifts between that of the quasar and zero. Astronomers very much want to measure the statistics of the Lyα forest absorption, but are stymied in part because of the lack of complete understanding of the unabsorbed continuum of the quasar itself. That is, the observations represent the convolution of two unknowns: the quasar spectrum, and the Lyα forest absorption spectrum, and it is not clear how optimally to separate the two. It would be interesting to know if the methods described in this paper could allow an optimal solution to this problem.

[6]Princeton University Observatory

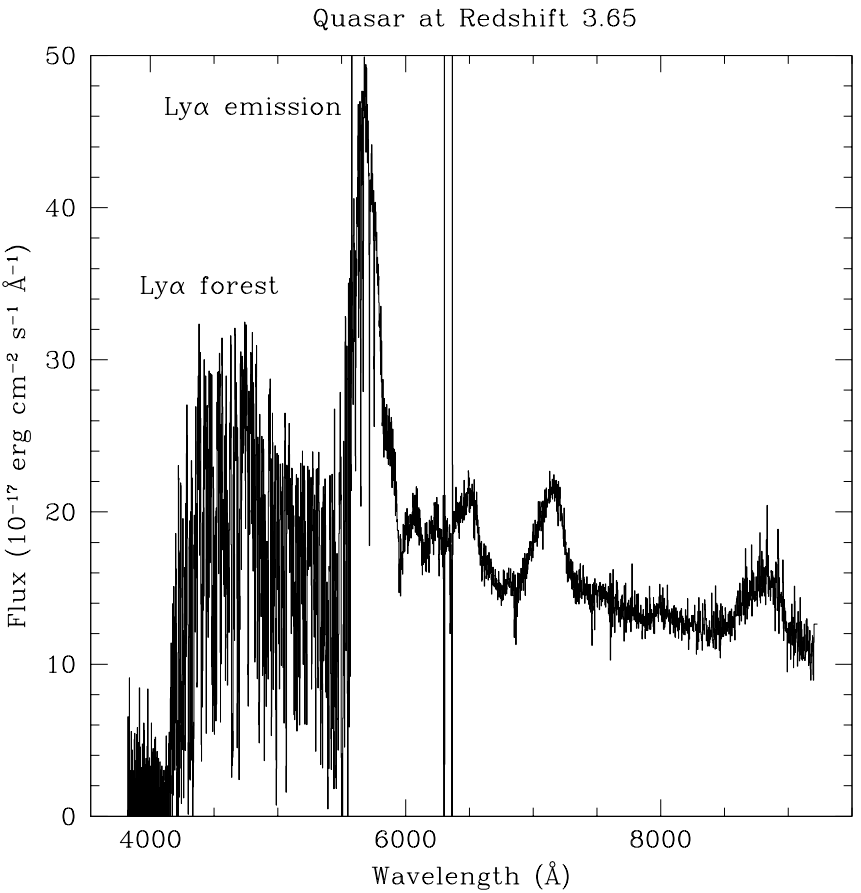


FIGURE 3.3. The spectrum of a high-redshift quasar from the Sloan Digital Sky Survey.

4

Bayesian Adaptive Exploration

Thomas J. Loredo[1] **and David F. Chernoff**

ABSTRACT We describe a framework for adaptive astronomical exploration based on iterating an *Observation–Inference–Design* cycle that allows adjustment of hypotheses and observing protocols in response to the results of observation on-the-fly, as data are gathered. The framework uses a unified Bayesian methodology for the inference and design stages: Bayesian inference to quantify what we have learned from the available data; and Bayesian decision theory to identify which new observations would teach us the most. In the design stage, the utility of possible future observations is determined by how much information they are expected to add to current inferences as measured by the (negative) entropies of the probability distributions involved. Such a Bayesian approach to experimental design dates back to the 1970s, but most existing work focuses on linear models. We use a simple *nonlinear* problem—planning observations to best determine the orbit of an extrasolar planet—to illustrate the approach and demonstrate that it can significantly improve observing efficiency (i.e., reduce uncertainties at a rate faster than the familiar "root-N" rule) in some situations. We highlight open issues requiring further research, including dependence on model specification, generalizing the utility of an observation (e.g., to include observing "costs"), and computational issues.
This paper is followed by a commentary by David A. van Dyk.

4.1 Introduction

Incremental learning from experience, where one proceeds step by step to a desired goal, making decisions and asking questions on the basis of available information, is a basic aspect of human behavior. The classical paradigm for the scientific method, with its rigid sequence of hypothesis formation, followed by experiment and then analysis, bears little resemblance to this adaptive, self-adjusting learning behavior. The classical paradigm has served science well but its limitations are apparent in settings where data collection and analysis may proceed in concert, where learning proceeds on-the-fly and what has been learned from past data may be profitably used to alter the collection of future data.

[1]Department of Astronomy, Cornell University

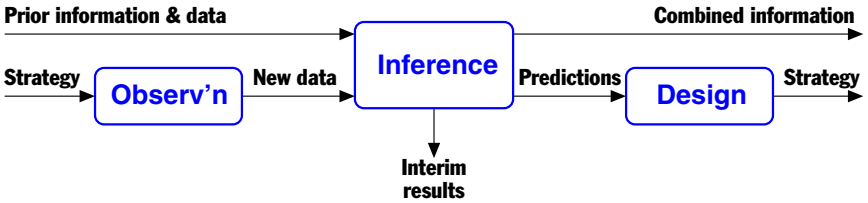


FIGURE 4.1. Information flow through one cycle of the adaptive exploration process. Information (e.g., data) and an observing strategy are input from a previous cycle on the left; combined information and a new observing strategy are output to the next cycle on the right.

We describe here an adaptive extension of the scientific method built on a model for scientific exploration where, after an initial setup phase, exploration proceeds by iterating a three-stage cycle: *Observation–Inference–Design*. Figure 1 depicts the flow of information through one such cycle. In the observation stage, new data are obtained based on an observing strategy produced by the previous cycle of exploration. The inference stage synthesizes the information provided by previous and new observations to assess hypotheses of interest. This synthesis produces interim results such as signal detections, parameter estimates, or object classifications. Finally, in the design stage the results of inference are used to predict future data for a variety of possible observing strategies; the strategy that offers the greatest predicted improvement in inferences (subject to any resource constraints) is passed on to the next Observation–Inference–Design cycle.

The Bayesian approach to statistics provides ideal tools for developing a unified framework for adaptive exploration: Bayesian inference for the inference stage, and Bayesian experimental design for the design stage. Bayesian inference—using probability theory to combine prior information and data to produce posterior probabilities for hypotheses of interest—is a formal description of learning perfectly suited for the tasks of the inference stage of the exploration cycle. It is now widely used in several astronomical disciplines and its basic features will be familiar to many astronomers. In contrast, formal methods for experimental design (Bayesian or otherwise) will likely be new to most astronomers. Bayesian design—an application of Bayesian decision theory—identifies an optimal experimental or observational design by first specifying the purpose for a study, and then comparing how well candidate designs achieve that purpose by using the techniques of Bayesian inference to predict and analyze future data. A main goal of this brief paper is to introduce astronomers to Bayesian design, in the context of adaptive exploration.

In 1956, Lindley described how one could use tools from information theory and Bayesian statistics to compare experimental designs when one's purpose is simply to gain knowledge about a phenomenon [Lin56]. He later incorporated these ideas into a more general theory of Bayesian experimen-

tal design, described in his influential 1972 review of Bayesian statistics [Lin72]. Although non-Bayesian methods for optimal design predate Lindley's work (standard references are [Fed72, Che72, AF97]), the Bayesian approach provides a more fundamental rationale for many earlier methods, and unifies and generalizes them (see [CV95] for discussion of the relationships between Bayesian and non-Bayesian design). In the three decades since Lindley's review, the theory of design has matured significantly. But as noted in Toman's recent review, "unfortunately much of the work in this area remains purely theoretical" [Tom99]. This is largely due to the computational complexity of Bayesian design, an obstacle noted already in Lindley's foundational work. In experimental design, one must account for both uncertainty regarding the hypotheses under consideration, and uncertainty about the values of future data. For the former, one must perform the difficult parameter space integrals that are characteristic of Bayesian inference [Lor99]; for the latter, one must additionally integrate in the sample space as is typically done in frequentist calculations. In a sense, experimental design is the arena in which the Bayesian and frequentist outlooks meet, producing problems with the combined complexity of both approaches.

As a result of this complexity, the vast majority of research in optimal design (Bayesian or non-Bayesian) has focused on simple models for which the required integrals can be evaluated analytically, such as linear models with additive Gaussian errors. Existing work on nonlinear design typically linearizes about a best-fit model [Mac92, SS98]. But the last decade has seen enormous strides in Bayesian computation due largely to the development of sampling-based methods for evaluating parameter space integrals, particularly Markov Chain Monte Carlo (MCMC) methods. Such methods not only facilitate rigorous calculations with complicated models; they also provide results in a form that can be readily interpreted and processed by end-users, even when the hypothesis space is of large dimension. We describe them further below.

Only recently have sampling-based algorithms that combine parameter and data sampling been brought to bear on Bayesian design [MP95, CMP95, MP96, Mul99]. Here we use simple sampling algorithms to implement the adaptive exploration strategy outlined above in the context of a simple but realistic *non*linear astronomical design problem. The sampling approach not only allows us to evaluate integrals without approximating the integrands, but also allows straightforward graphical display of all elements of the calculation. We hope this example provides an accessible introduction to Bayesian experimental design for astronomers, as well as a demonstration of the potential of adaptive exploration.

The following section describes the motivation for our interest in adaptive exploration—optimal allocation of observing resources for the Space Interferometry Mission—and then introduces adaptive exploration by example. We follow the strategy through one full cycle and through the observation and inference stages of a second cycle, using as an example radial velocity

observations of a star with the goal of determining the orbital parameters of an unseen planetary companion. The final section discusses several directions for future research.

4.2 Example: Measuring an Exoplanet Orbit

Our work on adaptive exploration is motivated by the Space Interferometry Mission, the first main mission of NASA's *Origins* program.[2] SIM is designed to measure the directions to astronomical sources with unprecedented accuracy. In its highest precision mode it is expected to achieve 1 microarcsecond astrometric accuracy. This will allow detection of the reflex motion "wobble" of a star with an Earth-like planet at a distance of several parsecs, or with a Jupiter-like planet at kiloparsecs. But SIM's high-accuracy measurements are time consuming, seriously restricting the number of stars that can be examined in a search for extrasolar planets. SIM observations are thus a precious resource that must be optimally allocated (not only for planet searches, but also for other diverse science SIM will undertake). During the mission, targets with no planets must be quickly weeded out, and observations of targets with companions must be scheduled to optimally determine the number of planets and their orbital parameters so that SIM can characterize as many systems as possible. In addition, before the launch of the SIM spacecraft in 2009, the SIM project will undertake extensive preparatory observations in order to carefully select both science target stars and reference stars against which the motions of the science targets will be measured. Reference stars must be free of planetary companions that would complicate their motion. The SIM Extrasolar Planet Interferometric Survey (EPIcS) key project is considering using binary stars with eccentric orbits as reference stars, since planets will have been swept from such systems. The preparatory observing campaign must identify hundreds of such stars and measure their orbits with high precision. This will require a huge expenditure of observational resources that must be optimized.

As a simple example of the kind of problem that must be addressed for optimizing SIM mission and preparatory observing, we consider here the problem of making radial velocity (RV) measurements of a star in order to best determine the parameters of the orbit of an unseen Jupiter-mass companion. Observations of this type will comprise much of SIM preparatory observing; similar ideas will apply to analysis of astrometric data. We consider observations of a 1 $M_\odot$ star known to have a single planetary companion; our goal is to choose future observations to best improve our

[2]For detailed information about SIM, see the SIM web site:
`http://sim.jpl.nasa.gov/`

estimates of the planet's orbital parameters. The function giving the radial velocity vs. time for a star exhibiting Keplerian reflex motion has six parameters. To simplify the calculations, we focus here on the three most important parameters—the orbital period, τ, the eccentricity, e, and the velocity amplitude, K—and we presume the remaining geometric parameters are known a priori (these include the time of periastron crossing, the longitude of periastron, and the orbital inclination). We model the value of each datum d_i as having additive noise, so that

$$d_i = v(t_i; \tau, e, K) + e_i, \tag{4.1}$$

where $v(t; \tau, e, K)$ gives the velocity at time t as a function of the parameters, and e_i represents the unknown noise contribution for datum i. We take the noise to have independent Gaussian distributions with standard deviation $\sigma = 8$ m s^{-1} (typical of current RV surveys).

The first cycle of exploration requires a "setup" strategy specifying the initial observations. Ideally, such a strategy would be developed using design theory and predictions based solely on prior information about the possible orbits (e.g., an assumed period distribution for orbits). For simplicity, the setup strategy here specifies 10 equally-spaced velocity measurements.

4.2.1 Cycle 1: Observation

Figure 2a shows the results of the observation stage of the first Observation-Inference-Design cycle. The points with error bars show the results of 10 simulated observations. For reference, the dashed curve shows the true velocity curve, with $\tau = 800$ d, $e = 0.5$, and $K = 50$ m s^{-1} (typical parameters for current observations of Jupiter-like extrasolar planets). The observations span somewhat less than two periods.

4.2.2 Cycle 1: Inference

For the inference stage, we calculate the posterior probability density for the parameters given the available data. Bayes's theorem gives this as

$$p(\tau, e, K|D, I) \propto p(\tau, e, K|I)\, \mathcal{L}(\tau, e, K), \tag{4.2}$$

where $p(\tau, e, K|I)$ is the prior probability density for the orbital parameters, $\mathcal{L}(\tau, e, K)$ is the likelihood function (the probability for the data presuming τ, e, and K are known), and I denotes the modeling assumptions (Keplerian orbit, noise properties, etc.). We assume we have no significant prior knowledge of the parameters, and take the prior to be a constant. Our assumption of Gaussian noise probabilities leads to a likelihood proportional to $\exp[-\chi^2(\tau, e, K)/2]$, where $\chi^2(\tau, e, K)$ is the familiar goodness-of-fit statistic given by a weighted sum of squared residuals. Thus,

$$p(\tau, e, K|D, I) \propto \exp[-\chi^2(\tau, e, K)/2]. \tag{4.3}$$

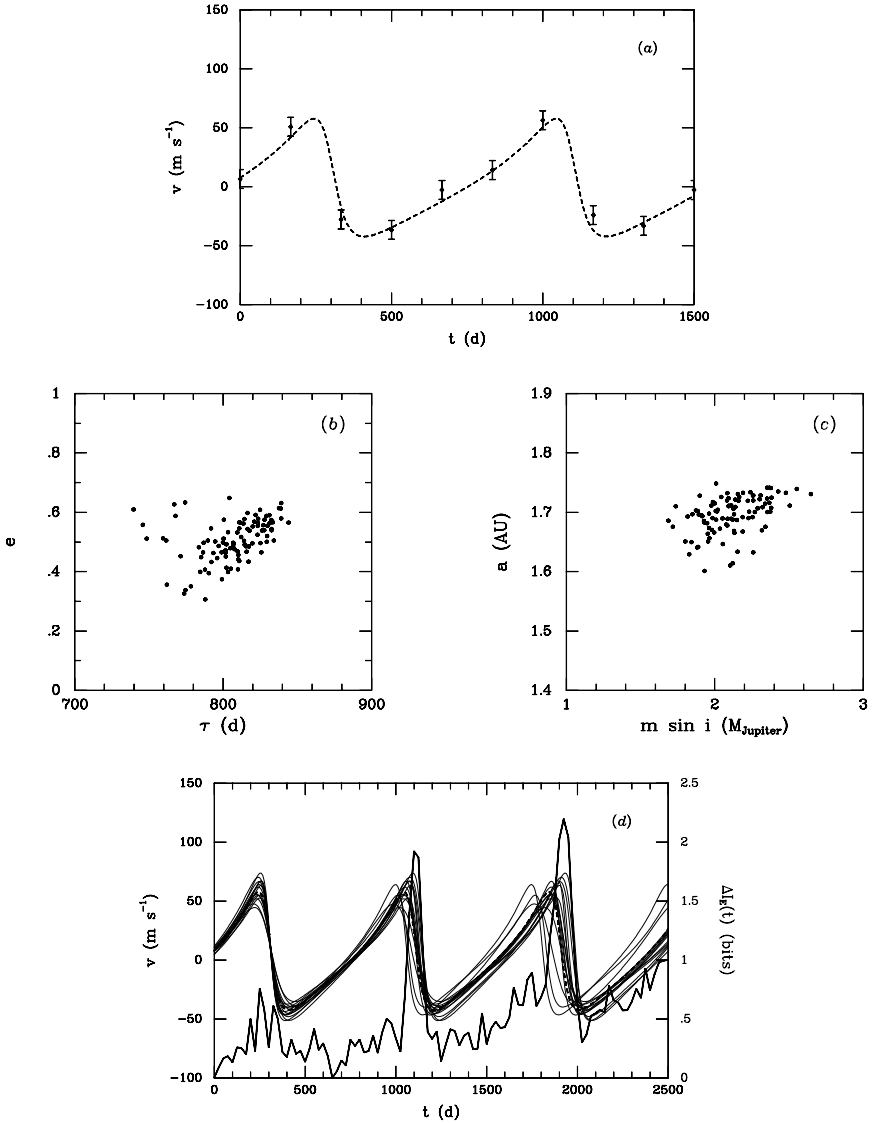


FIGURE 4.2. One cycle of the exploration process for simulated planet search data. (a) Observation stage, showing 10 simulated observations and true velocity curve (dashed). (b,c) Inference stage, showing samples from the posterior distribution for two velocity curve parameters (b) and two derived orbital parameters (c). (d) Design stage, showing predicted velocity curves (thin solid curves), true velocity curve (dashed curve), and the expected information gain for a sample at each time (thick solid curve, right axis).

To find best-fit parameters, we could maximize the posterior density (corresponding to minimizing χ^2). To constrain the parameters, we could locate the constant-χ^2 surface that encloses, say, 90% of the posterior probability for all three parameters; such a region is called a 90% (joint) credible region. If we were primarily interested in just the period, we could separately focus on it by calculating the marginal distribution for τ, given by integrating out the other parameters;

$$p(\tau|D,I) \propto \int de \int dK \ \exp[-\chi^2(\tau,e,K)/2]. \tag{4.4}$$

A 90% credible region for τ alone would be a region of the τ axis containing 90% of this marginal density.

All of these summaries of the posterior distribution could be calculated with common numerical methods (optimization and quadrature). But for problems with more dimensions, such calculations can be challenging. A more flexible approach is to use *posterior sampling* (see [Lor99] for a brief introduction and references). In this approach one constructs a random number generator that samples from the parameter space according to the posterior distribution (in contrast to more common Monte Carlo methods that sample from the data space). In this case, each sample would be a triplet (τ, e, K) drawn from $p(\tau, e, K|D, I)$; repeated sampling will produce a set of values, $\{\tau_j, e_j, K_j\}$. Once a set of such samples is available, many quantities of interest can be found by simple manipulations of the samples. In addition, posterior samples can be used directly to report results in a way that is easy to interpret and easy to use in future calculations.

Figures 2b and 2c are examples of interim results from the inference stage of the exploration cycle based on the observations shown in Figure 2a. We used a simple rejection method [PTVF92] to sample the posterior distribution; Figure 2b shows the τ and e coordinates of 100 such samples, displaying the marginal distribution $p(\tau, e|D, I)$. In a more careful calculation, we would use more samples and smoothing to find contours of credible regions; here it suffices to note that the displayed cloud of points should conservatively bound a 90% credible region. We see that the period and eccentricity are usefully constrained by the 10 data points, although significant uncertainty remains. Also, the posterior distribution is obviously not well-approximated by a Gaussian. Figure 2c shows how easily a complicated marginal distribution can be found using the samples; it displays the marginal distribution for the planet's semimajor axis, a, and $m \sin i$, the product of its mass and the sine of its orbital inclination. These are each nonlinear functions of the three model parameters. To produce Figure 2c we simply evaluated these functions for each of the 100 samples of (τ, e, K) already produced; this is much simpler than numerically evaluating the multiple integral defining the marginal distribution over a $(m \sin i, a)$ grid. By reporting the actual sample values, other investigators could use the results of these observations in their own calculations and fully account for

the uncertainties simply by evaluating any quantities of interest over the set of samples.

4.2.3 Cycle 1: Design

For the design stage, we locate the time at which to make the next observation so that we have the best chance of significantly reducing our uncertainty in the parameters. We accomplish this in three steps: predict future data at various times, find the effect of the predicted data on inferences, and then identify the time for which the expected improvement in precision is greatest. We discuss each step in turn.

To predict the value, d, of a future datum at time t, we calculate the *predictive distribution*. To find it, we first predict d assuming we know the true parameter values, and then account for parameter parameter uncertainty by averaging over the parameter space. For given values of (τ, e, K), the predictive probability density for d is just the likelihood for d (a Gaussian centered at $v(t;\tau,e,K)$). The averaging weight we must use to account for parameter uncertainty is the posterior distribution from the inference stage. The predictive distribution is thus the convolution of the Gaussian likelihood for d and the posterior from the inference stage;

$$\begin{aligned} p(d|t,D,I) &= \int d\tau \int de \int dK\; p(\tau,e,K|D,I) \\ &\quad \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{[d-v(t;\tau,e,K)]^2}{2\sigma^2}\right) \\ &\approx \frac{1}{N} \sum_{\{\tau_j,e_j,K_j\}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{[d-v(t;\tau_j,e_j,K_j)]^2}{2\sigma^2}\right) \end{aligned} \quad (4.5)$$

where the last line gives a Monte Carlo integration estimate of the predictive distribution using N posterior samples from the inference stage. To give some sense of what the predictive distribution looks like for various values of time, Figure 2c shows the $v(t)$ curves for the first 15 sampled parameter points as thin solid lines; the true curve is again displayed as a thick dashed curve. The ensemble of thin curves depicts our uncertainty in $v(t)$. The predicted data values at each time are additionally uncertain due to the noise which "blurs" the curves by 8 m s^{-1}. The ensemble of blurred curves represents the predictive distribution as a function of time. The uncertainty is greatest near times of periastron crossing when the velocity is changing most quickly (it is minimal at 300 d, the initial time of periastron crossing we assumed was known). Also, the uncertainty in the period makes the velocity uncertainty at periastron crossing grow with time as predictions with different periods fall increasingly out of synchronization.

Next we must measure how future data would affect our inferences. If datum d at time t were available, we could update our inferences simply

by multiplying the posterior distribution from the previous stage by the likelihood function based on the single new datum (the Gaussian factor in equation (4.5)), and renormalizing. (This is equivalent to doing a new χ^2 calculation considering all 11 data points at once.) The new posterior, $p(\tau, e, K|d, t, D, I)$, will hopefully be more informative about the parameters than the current one. The information in the posterior is given by the negative Shannon entropy of the posterior distribution,[3]

$$\mathcal{I}(d,t) = \int d\tau \int de \int dK\ p(\tau, e, K|d, D, I) \log[p(\tau, e, K|d, t, D, I)]. \quad (4.6)$$

This is the information gain for a particular datum at time t; to account for prediction uncertainty, we must calculate the *expected* information gain, averaging over d using the predictive distribution of equation (4.5):

$$\mathcal{EI}(t) = \int dd\ \mathcal{I}(d,t) p(d|t, D, I). \quad (4.7)$$

The best sampling time is the one that maximizes the information gain, so we must evaluate $\mathcal{EI}(t)$ as a function of time. For problems such as this where the width of the noise distribution does not depend on the value of the underlying signal, one can show that the expected information gain is equal to the entropy of the predictive distribution [SW97, SW00],

$$\mathcal{EI}(t) = -\int dd\ p(d|t, D, I) \log[p(d|t, D, I)]. \quad (4.8)$$

Thus the best sampling time is the time at which the entropy (uncertainty) of the predictive distribution is maximized. This is an eminently reasonable criterion: Bayesian design is telling us that we will learn the most by sampling where we know the least.

We use nested Monte Carlo methods to calculate $\mathcal{EI}(t)$ as a function of time. At each time, we sample a datum from the predictive distribution by first drawing a set of parameter values from the posterior, and then drawing a data value from the sampling distribution with those parameters. We then estimate $p(d|t, D, I)$ for that datum using equation (4.5). Repeating this process and averaging the logarithm of the estimates provides a Monte Carlo estimate of equation (4.8). The thick solid curve in Figure 2d shows this estimate of $\mathcal{EI}(t)$, using base-2 logarithms so that the relative information gain is measured in bits (with an offset so the smallest $\mathcal{EI}(t)$ is at

[3]For a Gaussian distribution, $\mathcal{I}$ is proportional to $-\log(\sigma)$ and thus increases with decreasing σ as one would expect; but it is a more general measure of spread than the standard deviation. To be formally correct, the argument of the logarithm in equation (4.6) should be divided by a measure on the parameter space so the argument is dimensionless; this has no significant effect on our results. An alternative definition of information is the cross-entropy or Kullback-Leibler divergence between the posterior and prior; it gives the same results as the Shannon entropy for this calculation [Mac92].

0 bits; the raggedness in the curve reflects the Monte Carlo uncertainties). $\mathcal{EI}(t)$ quantifies the uncertainty that is apparent in the set of thin sampled $v(t)$ curves. It is maximized near the periastron crossing subsequent to the available data, at $t = 1925$ d. Thus the observing strategy produced by this observation–inference–design cycle is: observe at $t = 1925$ d.

4.2.4 Cycle 2: Observation and Inference

Figure 3 shows the consequences of following this strategy. Figure 3a shows the previous data and a new datum obtained by simulating an observation at $t = 1925$ d. Incorporating this new datum into the posterior yields posterior samples shown in Figure 3b. We also used these samples to produce 15 predicted $v(t)$ curves in Figure 3a to display the velocity curve uncertainty after incorporating the new datum. Finally, Figure 3c shows the updated marginal distribution for the planet's mass and semimajor axis. Comparing to the corresponding panels in Figure 2, we see very significant reduction in uncertainty. In particular, the period uncertainty has decreased by more than a factor of two and the semi-major axis uncertainty is also drastically decreased; this was accomplished by incorporating the information *from a single well-chosen datum.* This is a dramatically larger increase in precision than one might have expected using rule-of-thumb "root-n" arguments based on random sampling. This is typical behavior for this problem; we have not chosen the simulated data set in any special way to obtain this behavior. It continues for subsequent cycles.

4.3 Challenges

This simple example illustrates the adaptive exploration methodology and demonstrates its potential. Several issues need to be addressed to make adaptive exploration useful in more complicated settings. Befitting a conference on statistical challenges, we close with a list of topics for future research. The field of experimental design has a wide and diverse literature spread across several disciplines, and some of these topics are being addressed in current research under such titles as sequential design, active data selection, and active, adaptive, or incremental learning.

In our example the goal was inference of the parameters of a system known to contain a single planet. In reality, the goals of inference may not be so clear-cut. Observers may not be sure a system has a planetary companion at the start of an exploration, so the goal is initially detection of a planet. Or if a system is chosen because it is known to have a companion, the goals may include detection of possible additional planets. At some point, the goal may shift from detection to estimation. How do design criteria for detection compare to those for estimation? When and how should

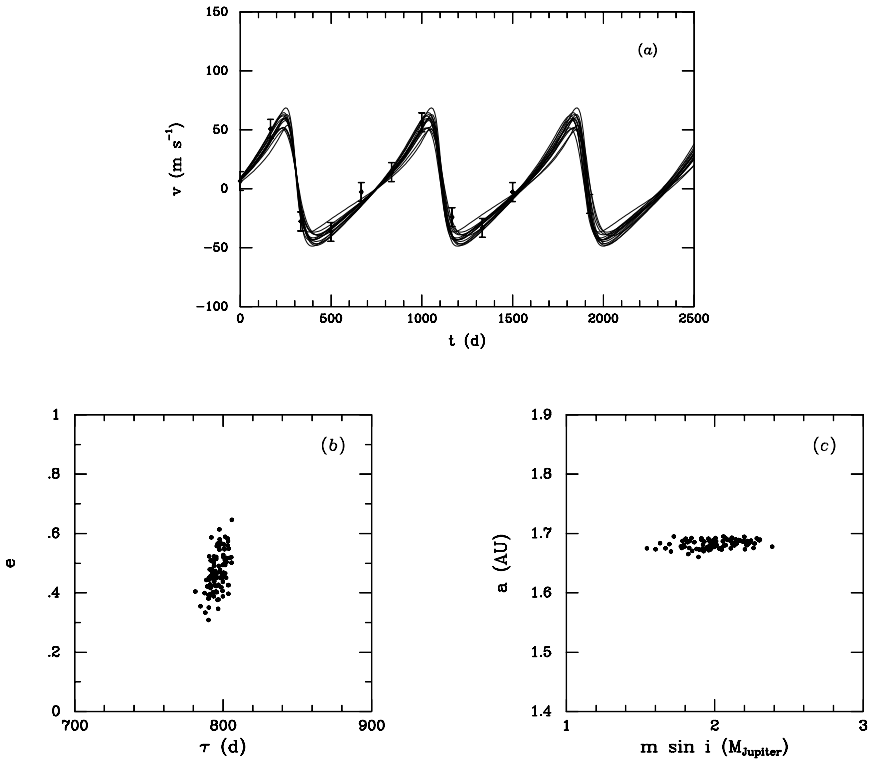


FIGURE 4.3. The beginning of the next cycle of the exploration process for simulated planet search data. (a) Observation stage, showing original 10 simulated observations, a new datum at 1925 d. Also shown are predicted velocity curves from the inference stage. (b,c) Inference stage, showing samples from the posterior distribution for two velocity curve parameters (b) and two derived orbital parameters (c). The single new datum has greatly increased the precision of inferences due to optimal selection of the observing epoch.

the adaptive methodology shift its goal from detection to estimation? The work of Toman [Tom96] on Bayesian design for multiple hypothesis testing provides a starting point for addressing these questions.

Our utility function was simply the information provided by new data. In some settings, one may wish to incorporate other elements in the utility function, such as the cost of observing as a function of time or sample size. How can an observer map such costs to an information scale so that information and other costs or benefits can be combined into a single utility function?

We used a simple rejection method for generating posterior samples in our example. While attractively simple, in our experience such an approach will not be useful for problems with more than five or six parameters (even fairly sophisticated envelope functions will waste too many sam-

ples). The obvious tool for addressing this is MCMC, but the Markov chain must ultimately sample over both the parameter space and the sample space (of future observations). Are there MCMC algorithms uniquely suited to adaptive exploration? Müller and Parmigiani and their colleagues [MP95, CMP95, MP96, Mul99] have developed a variety of Monte Carlo approaches to Bayesian design in various settings that should be helpful in this regard. Also, since adaptive exploration offers the hope of quickly reducing uncertainties, at some point it may make sense to linearize about the best-fit model and use analytic methods. Criteria need to be developed to identify when this is useful.

Finally, in our example, the observing strategy for the first cycle was chosen somewhat arbitrarily. Ideally, it would be chosen using design principles and prior information. This raises many practical and theoretical questions. What should the size of a "setup" sample be? Should adaptive exploration start after a single sample, or are there benefits (perhaps associated with computational complexity) for starting with larger samples? Can the algorithms used for analysis when several samples are available also be used for designing the setup strategy, or are different algorithms required if prior information is very vague? Clearly, there is overlap between these issues and those already raised. This kind of design issue has been addressed informally for planning observations for the Hubble Space Telescope Cepheid key project [FHM+94]. Can a more formal approach improve on such a priori designs?

We hope this brief introduction will encourage astronomers and statisticians to explore these issues together in a variety of astronomical contexts.

4.4 References

[AF97] A. C. Atkinson and V. V. Fedorov. Optimum design of experiments. In Samuel Kotz, Campbell B. Read, and David L. Banks, editors, *Encyclopedia of statistical sciences*, pages 107–114. Wiley, New York, 1997.

[Che72] H. Chernoff. *Sequential Analysis and Optimal Design.* SIAM, Philadelphia, 1972.

[CMP95] M. Clyde, P. Muller, and G. Parmigiani. Exploring expected utility surfaces by markov chains, 1995.

[CV95] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Stat. Sci.*, 10:273–304, 1995.

[Fed72] V. V. Fedorov. *Theory of Optimal Experiments.* Academic, New York, 1972.

[FHM+94] W. L. Freedman, S. M. Hughes, B. F. Madore, J. R. Mould, M. G. Lee, P. Stetson, R. C. Kennicutt, A. Turner, L. Ferrarese, H. Ford, J. A. Graham, R. Hill, J. G. Hoessel, J. Huchra, and

G. D. Illingworth. The hubble space telescope extragalactic distance scale key project. 1: The discovery of cepheids and a new distance to m81. *Ap. J.*, 427:628–655, June 1994.

[Lin56] D. V. Lindley. On the measure of information provided by an experiment. *Ann. Stat.*, 27:986–1005, 1956.

[Lin72] D. V. Lindley. *Bayesian statistics–a review.* SIAM, Philadelphia, 1972.

[Lor99] T. J. Loredo. Computational technology for bayesian inference. In *ASP Conf. Ser. 172: Astronomical Data Analysis Software and Systems VIII*, volume 8, pages 297+, 1999.

[Mac92] D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[MP95] P. Muller and G. Parmigiani. Numerical evaluation of information theoretic measures. In D. A. Berry, K. M Chaloner, and J. F. Geweke, editors, *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner*, pages 397–406. Wiley, New York, 1995.

[MP96] Peter Muller and Giovanni Parmigiani. Optimal design via curve fitting of monte carlo experiments. *J. Am. Stat. Assoc.*, 90:1322–1330, 1996.

[Mul99] P. Muller. Simulation based optimal design. In J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, 1999.

[PTVF92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, 1992.

[SS98] P. Sebastiani and R. Settimi. First-order optimal designs for non-linear models. *J. Stat. Plan. Inf.*, 74:177–192, 1998.

[SW97] P. Sebastiani and H. P. Wynn. Bayesian experimental design and shannon information. In *1997 Proceedings of the Section on Bayesian Statistical Science*, pages 176–181. American Statistical Association, 1997.

[SW00] P. Sebastiani and H. P. Wynn. Maximum entropy sampling and optimal bayesian experimental design. *J. Roy. Stat. Soc. B*, 62:145–157, 2000.

[Tom96] B. Toman. Bayesian experimental design for multiple hypothesis testing. *J. Am. Stat. Assoc.*, 91:185–190, 1996.

[Tom99] B. Toman. Bayesian experimental design. In Samuel Kotz and N. L. Johnson, editors, *Encyclopedia of statistical sciences, Update Vol. 3*, pages 35–39. Wiley, New York, 1999.

Commentary by David A. van Dyk[4]

Loredo and Chernoff should be congratulated for their thoughtful Monte Carlo implementation of Bayesian decision analysis. Their proposal promises to significantly improve the scientific information obtained by Origins and other programs. Here I offer only some fine tuning of their proposed method.

Loredo and Chernoff suggest choosing an observation time, t, by maximizing the expected negative Shannon entropy, $\mathrm{E}\,[\mathcal{I}(d,t)|t]$, with d the observed datum. Here I suggest two potentially useful and easy-to-use generalizations, namely, to treat the negative entropy as a *value* function rather than a *utility* function and to consider other functions with more direct scientific interpretation. To clarify these issues, I use Loredo and Chernoff's example involving the measurement of an exoplanet orbit.

For any selected t, there is a distribution for the observed d, denoted $p(d|t, D, I)$. The *value* of d can be measured by a *value function* such as the negative entropy—the larger $\mathcal{I}(d,t)$, the more information that is gained by d. Since d has a distribution so does $\mathcal{I}(d,t)$—there is variability in the information gained from the selected t depending on the observed d. Loredo and Chernoff suggest selecting t by maximizing the expected information gained. That is, they treat $\mathcal{I}(d,t)$ as a *utility function*—a function whose expected value determines the preferred choice. A more general strategy is to consider the full posterior distribution of $\mathcal{I}(d,t)$, namely $p(\mathcal{I}(d,t)|t, D, I)$. One observation time may maximize the expected information gained but with a relatively high variance and thus seem more risky; see Figure 4.4.

Shannon entropy is a generic measure of value and somewhat removed from quantities of scientific interest. When using MCMC, however, it easy to simulate the distribution of other value functions such as the maximum or mean error bars on velocity or the error bars on some specific function of the model parameters. Multivariate value functions can also be considered which can include the statistical value of the data (e.g., entropy or error bars), costs of the the data in dollars or satellite time, and waiting time for the data. Such quantities may be easier to interpret and should be easy to compute—though computation may be slower because the analytic simplifications of Loredo and Chernoff are not applicable.

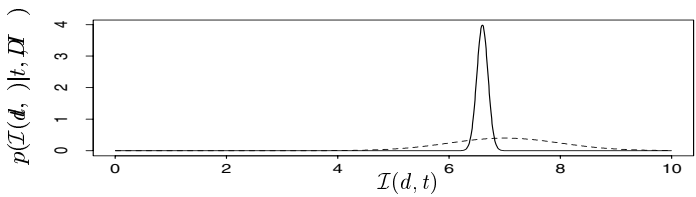


FIGURE 4.4. The dashed density corresponds to an observation time that may seem more risky than that of the solid density.

[4]Department of Statistics, Harvard University

5

Bayesian Model Selection and Analysis for Cepheid Star Oscillations

James O. Berger[1], William H. Jefferys, Peter Muller and Thomas G. Barnes

ABSTRACT Cepheid variables are a class of pulsating variable stars with the useful property that their periods of variability are strongly correlated with their absolute luminosity. Once this relationship has been calibrated, knowledge of the period gives knowledge of the luminosity. This makes these stars useful as "standard candles" for estimating distances in the universe. Available data consists of photometric and velocity information for a number of Cepheid variables, at unequally spaced points in their periods. Note that photometry and velocity are connected by nonlinear relations involving the physical parameters of interest. Bayesian analysis is used to provide inferences for useful physical features, such as the absolute luminosity of the star, its distance, its radius, and other parameters.

In the absence of reliable physical models of the pulsation of Cepheid variables, we model the photometry and velocity curves as (i) a trigonometric polynomial with an unknown number of terms; or (ii) via a wavelet basis with an unknown number of terms. Bayesian analysis allows computation of the posterior probabilities of these varying dimensional models, and results in inferences on the physical parameters that are based on 'averaging' over the possible models. Computations are done using reversible-jump Markov chain Monte Carlo methodology.

This paper is followed by a commentary by Thomas J. Loredo.

5.1 Introduction

5.1.1 Bayesian Model Selection and Model Averaging

The Bayesian approach to hypothesis testing and model selection is conceptually straightforward. One assigns *prior* probabilities to all unknown hypotheses or models, as well as to unknown parameters or quantities within models, and uses probability theory to compute the *posterior* probabilities

[1]Institute of Statistics & Decision Sciences, Duke University

of the hypotheses or models, given observed data. One attractive feature of this approach is simplicity of interpretation: stating, at the end of the analysis, that the only tenable models are Models 5, 6, and 7, and that they have probabilities 0.34, 0.56 and 0.10, respectively, has appealing clarity.

A second attractive feature of this approach is that one need not choose a fixed model. One could select Model 6 above (it is the model most likely to be true), but the data also gives considerable support to Model 5, and even Model 7 should not be ignored. One deals with this uncertainty by 'Bayesian model averaging,' in which predictions or desired estimates from models are averaged according to the model posterior probabilities. Thus if Models 5, 6, and 7 provided distance estimates (posterior means) to a star of 750, 790, and 800 parsecs, respectively, the 'model-averaged' distance estimate would be $0.34 \times 750 + 0.56 \times 790 + 0.10 \times 800 = 777.4$ parsecs.

The accuracy associated with a model-averaged estimate will also incorporate the model uncertainty. For instance, suppose Model j yields the distance estimate $\hat{d}_j$, with associated posterior variance V_j, and that p_j is the posterior probability of Model j. Then the overall variance of the model-averaged distance estimate $\hat{d}^* = \sum p_j \hat{d}_j$ is given by

$$V^* = \sum p_j [V_j + (\hat{d}_j - \hat{d}^*)^2].$$

For the case in the previous paragraph, if the individual model posterior variances were $V_5 = V_6 = V_7 = 400$ (corresponding to standard errors of 20 parsecs), then the overall variance of $\hat{d}^* = 777.4$ would be 795.24, almost twice the variance that would be associated with any specific model. (Indeed, it is a general advantage of the Bayesian approach that inaccuracies in all unknown parameters are incorporated automatically.)

Note, also, that the Bayesian approach to model selection acts as a natural "Ockham's razor," in the sense of favoring a simpler model over a more complex model if the data provides roughly comparable fits for the models. And this is without having to introduce any explicit penalty for the more complex models. (For an interesting historical example of Ockham's razor, and general discussion and references, see Jefferys and Berger, 1992.)

5.1.2 Cepheid Star Oscillations

Cepheid variable stars pulsate, varying their luminosity (light output) and size with a very regular periodic behavior. It is possible to measure both the velocity of the surface of the star as it pulsates and the variable luminosity and color of the star. For instance, Figure 1 presents the data concerning the radial velocity of the surface of the star T Moncerotis, at various phases of the star's period. (The actual data are indicated by the x's.)

There is a mathematical relationship between surface velocity, luminosity and color that enables one to determine the distance to the star. The considerable uncertainty in these measurements and the limited data that

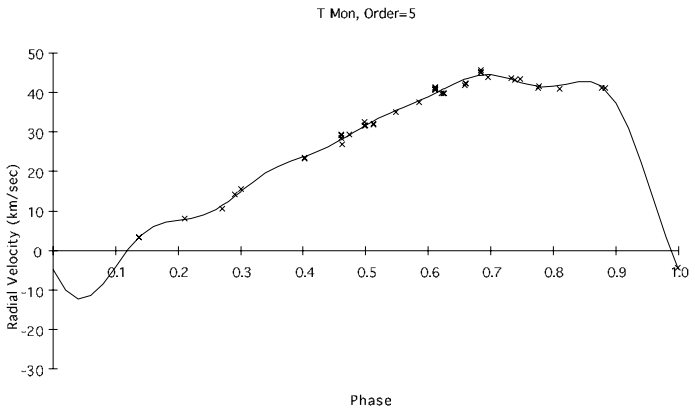


FIGURE 5.1. The radial velocity data (the x's) for T Mon, and their fit to a fifth-order trigonometric polynomial.

is available for each star suggest that analysis which fully incorporates these uncertainties is desirable.

5.1.3 Challenges in the Bayesian Approach

There are three significant challenges in implementing the Bayesian approach for complex problems. The first challenge is common to all statistical analyses, namely the need to find appropriate statistical models for the data. For a Cepheid star, the most challenging features to model are the radial velocity and the photometric information. For instance, Figure 1 clearly indicates that the radial velocity of the star is a quite complex function of its phase, but existing physical theories for Cepheid stars do not provide guidance as to the form of this function. Hence one must resort to generic statistical modelling, such as Fourier analysis. Figure 1 shows that a fifth-order trigonometric polynomial fits this particular data quite well, but the needed order of the polynomial changes from star to star and, indeed, there are typically several different orders that fit a particular star well (without overfitting). The different models that will be under consideration in our analysis are simply the different possible orders of the trigonometric polynomials (which will also be used to model the photometric data). Later we will also consider wavelet models of the radial velocity.

The second challenge in Bayesian analysis is to choose prior distributions for unknown quantities in the analysis (for instance, for the unknown Fourier coefficients of the trigonometric polynomial). The most common choices are noninformative or objective priors; these will be primarily utilized in the analysis here and are discussed in the next subsection.

The third challenge is computational. Bayesian analysis can require the computation of high-dimensional integrals, and is especially costly when

model selection is involved. (For instance, when using trigonometric polynomials in the Cepheid modeling, it is necessary to compute up to 50-dimensional integrals for up to 40 models; in the wavelet version of the analysis discussed in section 5.4, these numbers increase by orders of magnitudes.) The modern approach to such computation is Markov chain Monte Carlo (MCMC) analysis. This is a computational paradigm that can be easily described in simple cases, but which requires study and experience for successful application in complex cases (such as that considered here). Thus we will content ourselves, in this paper, with only a higher-level description of the particular steps needed in the Cepheid problem. Recent general books on the subject are Robert and Casella (1999) and Chen, Shao, and Ibrahim (2000).

5.2 Objective Bayesian Model Selection

5.2.1 Statistical Notation

The data, $\boldsymbol{Y}$, is assumed to have arisen from one of several possible models $M_1, \ldots, M_k$. Under M_i, the density of $\boldsymbol{Y}$ is $f_i(\boldsymbol{y} \,|\, \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i$ is an unknown vector of parameters.

The Bayesian approach to model selection begins by assigning prior probabilities, $P(M_i)$, to each model; often, equal prior probabilities are used, i.e. $P(M_i) = 1/k$, and this will be done here. It is also necessary to choose prior distributions $p_i(\boldsymbol{\theta}_i)$ for the unknown parameters of each model; sometimes these can also be chosen in a "default" manner, as will be illustrated later. The analysis then proceeds by computing the posterior probabilities of each model, which elementary probability theory (Bayes theorem) shows to be equal to

$$P(M_i|\boldsymbol{y}) = \frac{P(M_i)m_i(\boldsymbol{y})}{\sum_{j=1}^{k} P(M_j)m_j(\boldsymbol{y})}, \tag{5.1}$$

where $m_j(\boldsymbol{y}) = \int f_j(\boldsymbol{y} \,|\, \boldsymbol{\theta}_j)p_j(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j$ is the *marginal density* of $\boldsymbol{y}$. See Kass and Raftery (1995) for a general discussion of Bayesian model selection.

5.2.2 Choice of Prior Distributions

It may well be the case that subjective knowledge about the $\boldsymbol{\theta}_i$ is available, and can be incorporated into subjective proper priors for the $\boldsymbol{\theta}_i$. This is clearly desirable if it can be done. Indeed, for Cepheid stars we will see that subjective prior information concerning their distance can be utilized.

For most of the unknown parameters in models it will typically be infeasible to utilize subjective prior distributions. Frequently this is because

subjective information is simply unavailable. (Thus, for Cepheid stars, turning the physical principles that underlie oscillatory behavior into models for the velocity and photometric curves is so difficult to accomplish that, in actuality, there is little subjective information about the Fourier coefficients of the curves.) Even if subjective prior information is available, it can be very difficult to utilize in high-dimensional problems.

For these and other reasons, the most popular Bayesian methods are default or 'objective Bayesian' methods. For estimation and prediction problems, objective Bayesian theories are well developed. The most famous of these are the *Jeffreys prior* (cf. Jeffreys, 1961), *maximum entropy* priors (cf. Jaynes, 1999), and *reference priors* (which prove remarkably successful in higher dimensional problems; cf., Berger and Bernardo, 1992).

Testing and model selection have proved to be much more resistant to the development of default Bayesian methods. This is because the objective priors discussed above are typically improper distributions (i.e., their integrals are infinite). This does not typically pose a problem in estimation and prediction, but it does for testing and model selection. See Berger and Pericchi (2001) for discussion of these difficulties and possible solutions. Here are some guidelines for choosing default priors in model selection.

1. *Common Parameters:* If all models have certain common parameters (see Berger and Pericchi, 2001, for discussion of what it means to be 'common') these parameters can typically be assigned the same improper objective prior. For instance, all the models for Cepheid radial velocity will have a common unknown mean radial velocity u_0, and this can be assigned the usual objective (improper) prior $p_i^*(u_0) = 1$.

2. *Conventional proper priors* are sometimes available in the literature. For instance, in the Cepheid problem, we will model the observed radial velocities as arising from a trigonometric polynomial subject to error; in statistical language, the ensuing model can formally be written as a general linear model of the form

$$\boldsymbol{Y} = \theta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ is the vector of observations (radial velocities in the Cepheid problem), $\boldsymbol{X}$ is the corresponding design matrix of covariates (sines and cosines evaluated at multiples of the phases of the observations, corresponding to the trigonometric polynomial in the Cepheid problem), $\boldsymbol{\theta}$ is an unknown vector of parameters (the Fourier coefficients of the trigonometric polynomial in the Cepheid problem), $\mathbf{1}$ is the column vector of ones, θ_0 is the unknown mean level of the observations, and $\boldsymbol{\varepsilon}$ is a multivariate normal vector of errors with mean zero and covariance matrix $\sigma^2\boldsymbol{G}$, $\boldsymbol{G}$ a known matrix (i.e., $\boldsymbol{\varepsilon}$ is $\boldsymbol{N}(\mathbf{0}, \sigma^2\boldsymbol{G})$).

The recommended prior (from Zellner and Siow, 1980) for the unknown θ_0 is $p(\theta_0) = 1$, while that for $\boldsymbol{\theta}$, given σ^2, can be written in two stages (for

later convenience) as:

$$p(\boldsymbol{\theta} \,|\, \sigma^2, \tau) \text{ is } \boldsymbol{N}(\mathbf{0}, \tau n \sigma^2 (\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1}); \;\; p(\tau) = \frac{1}{\sqrt{2\pi}\,\tau^{3/2}} \exp(-\frac{1}{2\tau}). \tag{5.2}$$

3. A *general default prior* for model selection is the "empirical expected posterior prior" (Perez and Berger, 2000). For a given model M_i with unknown parameters $\boldsymbol{\theta}_i$, the most convenient form for this prior, when computation is to be done via the Markov chain Monte Carlo method, arises from introducing 'latent' random variables $\mathbf{y}^*$, which can be thought of as random subsamples of the data with sample size (typically) equal to the dimension of $\boldsymbol{\theta}_i$. Then the desired prior distribution is

$$p_i(\boldsymbol{\theta}_i, \mathbf{y}^*) = f_i(\mathbf{y}^* \,|\, \boldsymbol{\theta}_i) p_i^*(\boldsymbol{\theta}_i) m^E(\mathbf{y}^*) / m^*(\mathbf{y}^*), \tag{5.3}$$

where $p_i^*(\boldsymbol{\theta}_i)$ is a standard (improper) objective prior, $m^E(\mathbf{y}^*)$ refers to the empirical distribution of subsamples (i.e., choose each subsample of the given size with equal probability), and $m^*(\mathbf{y}^*)$ is the marginal density of $\mathbf{y}^*$ under the prior p_i^*. (The actual prior for $\boldsymbol{\theta}_i$ is the marginal density found by summing over $\mathbf{y}^*$ in (5.3), but it is actually more convenient computationally to work with the 'latent' joint distribution.)

5.3 Cepheid Stars

5.3.1 The Model and Likelihood

For a given star, the data consists of m observed radial velocities $\mathbf{U}_i, i = 1, \ldots, m$, at unequally spaced phases of the star's period (cf. Figure 1), together with n vectors of photometry data consisting of magnitude $\mathbf{V}_i, i = 1, \ldots, n$, and color index $C_i, i = 1, \ldots, n$. (It is to be understood that, attached to each observation, is the phase at which it was observed; note that the radial velocity and photometry data were typically observed at different phases of the star's period.) Each observation has a standard deviation specified by the observer; denote these by σ_{U_i}, σ_{V_i}, and σ_{C_i}, respectively. It is generally wise to be somewhat skeptical of such specified standard errors, and so we take the variances of the data to, instead, be given by $\sigma^2_{U_i}/\tau_u$, $\sigma^2_{V_i}/\tau_v$, $\sigma^2_{C_i}/\tau_c$, where the *parameters* τ_u, τ_v, and τ_c are unknown.

To complete the modelling of the data, let u_i, v_i, and c_i denote the true unknown mean velocity, magnitude, and color index, respectively, corresponding to each data point. We assume normality and independence of the measurement errors, so that

$$\mathbf{U}_i \sim N(u_i, \sigma^2_{U_i}/\tau_u), \;\; \mathbf{V}_i \sim N(v_i, \sigma^2_{V_i}/\tau_v), \;\; \text{and} \;\; C_i \sim N(c_i, \sigma^2_{C_i}/\tau_c). \tag{5.4}$$

Since the velocities u and photometry (v, c) are periodic functions of time, an obvious strategy is to model them as trigonometric polynomials.

For the velocity u at phase ϕ, this would lead to the representation

$$u = u_0 + \sum_{j=1}^{M} [\theta_{1j} \cos(j\phi) + \theta_{2j} \sin(j\phi)], \tag{5.5}$$

where u_0 is the mean radial velocity of the star and M is the (unknown) order of the trigonometric polynomial. A similar equation holds for the luminosity data v. (We need to do this only for u and v, since the colors c are mathematically related to u and v through (5.7) below.) Let N denote the (unknown) order of the trigonometric polynomial for v. These polynomials contain $2M+1$ and $2N+1$ terms, respectively, including the leading constant terms.

Let $\boldsymbol{u}$ and $\boldsymbol{v}$ denote column vectors of the velocity and luminosity data, respectively; define $\boldsymbol{X}_u$ and $\boldsymbol{X}_v$ to be the $(m \times 2M)$ and $(n \times 2N)$ design matrices consisting of the sines and cosines of multiple angles, evaluated at the phases of the corresponding data; and let $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$ be the corresponding vectors of unknown Fourier coefficients. With the normality assumption above, we can summarize the model (corresponding to M and N) for the velocity and luminosity data by the statistical linear models

$$\begin{aligned} \boldsymbol{U} &= u_0 \mathbf{1} + \boldsymbol{X}_u \boldsymbol{\theta}_u + \boldsymbol{\varepsilon}_u \\ \boldsymbol{V} &= v_0 \mathbf{1} + \boldsymbol{X}_v \boldsymbol{\theta}_v + \boldsymbol{\varepsilon}_v, \end{aligned} \tag{5.6}$$

where u_0 and v_0 are the (unknown) mean radial velocity and luminosity, respectively, and $\boldsymbol{\varepsilon}_u$ and $\boldsymbol{\varepsilon}_v$ are independently $\boldsymbol{N}(\mathbf{0}, \boldsymbol{G}_u/\tau_u)$ and $\boldsymbol{N}(\mathbf{0}, \boldsymbol{G}_v/\tau_v)$ multivariate errors, with $\boldsymbol{G}_u$ and $\boldsymbol{G}_v$ being the known diagonal matrices of the variances $\sigma_{U_i}^2$ and $\sigma_{V_i}^2$, respectively. (Note that τ_u and τ_v would have had to be introduced at this stage of the modeling, in any case, to account for the fact that u and v cannot be expected to exactly follow a trigonometric polynomial of finite order.)

The phases in the above likelihoods (entering through the design matrices) were assumed to be known exactly. In practice, however, the velocity data and photometry data are taken independently, and 'translated' to the same phase scale. The period of the star is not known perfectly, however, so that there is an unknown phase error $\Delta\phi$ (the difference between the zero-point of the phase for the velocity data and that for the photometric data) that is introduced. Thus we include that additional unknown in (say) the phase for the photometric data.

The (nonlinear) relationship between the radius of the star and the photometry is given by

$$c_i = a[-0.1 v_i - b - 0.5 \log(\phi_0 + \Delta r_i/s)], \tag{5.7}$$

where a and b are known constants, ϕ_0 and s are the angular size and distance of the star (the latter being of primary interest to us), and Δr,

the change in radius corresponding to phase ϕ, is given by

$$\Delta r = -g \sum_{j=1}^{M} \frac{1}{j} [\theta_{1j} \sin(j(\phi - \Delta\phi)) - \theta_{2j} \cos(j(\phi - \Delta\phi))], \tag{5.8}$$

found by integrating the nonconstant part of (5.5) term by term with respect to the phase; here g is another known constant. These expressions are to be inserted in the likelihood terms arising from the C_i in (5.4).

5.3.2 Choice of Prior Distributions

The unknown parameters in the above model are:

(1) The orders of the trigonometric polynomials, M and N.

(2) The parameters τ_u, τ_v, and τ_c, adjusting the measurement standard errors.

(3) The angular diameter ϕ_0 and the unknown phase error $\Delta\phi$.

(4) The distance s.

(5) The mean velocity and luminosity, u_0 and v_0, and the Fourier coefficients, $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$.

Some additional 'hyperparameters' will be introduced through the prior distributions for these unknowns, and the hyperparameters will also require prior distributions.

The orders of the models are expected to be modest (given the limited amount of data and the strong Ockham's razor effect of Bayesian analysis); we thus chose a uniform prior on the model orders (M, N) up to some cutoff (e.g., $(10, 10)$), with zero probability assigned to higher orders.

The parameters τ_u, τ_v, τ_c are given the standard objective priors for 'scale parameters,' namely the Jeffreys-rule prior $p(\tau) = 1/\tau$. Similarly, the priors on the 'location parameters' u_0 and v_0 are chosen to be the standard objective priors $p(u_0) = 1$ and $p(v_0) = 1$. Note that we are employing Rule 1 of subsection 5.2.2; since these parameters are common scale and location parameters for all models, they have an essentially fixed interpretation across models and can be assigned standard objective priors (even though improper).

For the parameters $\Delta\phi$ and ϕ_0, we also chose the objective priors $p(\Delta\phi) = 1$ and $p(\phi_0) = 1$. While it is unclear if these are 'optimal' objective priors for these parameters, preliminary investigations showed that the choice of priors for these parameters is almost irrelevant for the Cepheid data sets, so that additional effort was not expended in their development.

Failure to take the spatial distribution of the stars into account would result in the so-called *Lutz-Kelker bias*, which is a bias in the estimated

distance. Bayesian analysis takes care of such biases through the straightforward process of incorporating the cause of the bias in the prior distribution. If Cepheid stars were distributed uniformly over a region, the prior distribution of distances s from the observer would be proportional to s^2. However, the spatial distribution of Cepheid variables is known to be flattened with respect to the galactic plane. We thus modify the s^2 prior by using a spatial distribution of stars that is exponentially stratified as one moves away from the galactic plane. In particular, the prior distribution on the distance s, given a hyperparameter z_0, is

$$p(s) \propto s^2 \exp\left(-|z|/z_0\right),$$

where $z = s \sin\beta$, with β being the galactic latitude of the star (its angle above the galactic plane, another known covariate), and z_0 being the imperfectly known 'scale height.' This 'hyperparameter' z_0 is known to be $z_0 = 97 \pm 7$ parsecs, so we simply assigned it a gamma prior distribution with mean 97 and standard deviation 7.

The priors on the Fourier coefficients $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$ must be chosen carefully, to avoid overfitting or underfitting. Luckily, the models in (5.6) are exactly of the form (5.2), so that the conventional priors described there can be utilized directly. (We are slightly cheating here, in that $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$ also occur in the likelihood terms arising from the C_i, when (5.7) is used in 5.4), and one could argue that the appropriate default priors should reflect this. We ignore this complication, in part because we think it would make little difference and, in part, because it is unclear how to take this into account in defining a default prior. Also, in the computations reported here, we utilized the simpler hyperprior $p(\tau) = 1/\tau^{3/2}$.)

A possible alternative prior for the Fourier coefficients would be the empirical expected posterior prior, also defined in section 5.2.2. Note that, for the normal linear model, $m^*(\mathbf{y}^*)$ can be found in closed form. Space precludes our presenting these results here.

5.3.3 Computation

Space limitations preclude a full description of the MCMC computation that is used to analyze the Bayesian model. We thus limit the discussion to presentation of the major steps in the analysis, especially those that are non-standard. Familiarity with MCMC computation is assumed.

A reversible-jump MCMC algorithm of the type reviewed in Dellaportas *et. al.* (2000) is used to generate posterior distributions and estimates. The full conditional distributions for the variance and precision parameters and hyperparameters are standard χ^2 distributions and so the sampling of these parameters can be accomplished with straightforward Gibbs sampling.

For $\Delta\phi$, ϕ_0 and s, we employ a random-walk Metropolis algorithm using, as the proposal distribution, a multivariate normal distribution centered on

the currently imputed parameter values and with a covariance matrix that is proportional to the covariance matrix for the linearized least-squares problem for just these three parameters. (This means linearizing the logarithm in the expression for c_i in (5.7)). This proposal distribution leads to a fast mixing Markov chain, which implies fast convergence of the computational algorithm.

The Fourier coefficients $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$, as well as u_0 and v_0, are sampled via an independence-chain Metropolis step. The natural proposal distributions are found by combining the normal likelihoods in (5.6) with the normal priors (given τ) in (5.2), leading to conjugate normal posterior distributions. Note that these are not the actual full conditionals from the posterior, because of the nonlinear way in which $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$ appear in the full likelihood. However, the acceptance probabilities for these proposals are well over 90%, and the sampling of the Fourier parameter spaces is very effective.

The Metropolis steps for $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$ are included within a step that proposes a jump between models. Thus, if the current model has a certain number of parameters, we propose a jump to a model with a (possibly different) number of parameters, and simultaneously propose new values for all the Fourier coefficients. To make the sampling efficient, during the burn-in phase we also estimate the posterior probabilities of the individual models, and use them as the basis for the proposal probabilities of new models during the computation phase of the calculation. Thus models of higher posterior probability are proposed with greater frequency. A total of 10,000 iterations of the MCMC computation were performed.

5.3.4 Results

Figures 2 and 3 give the posterior probabilities of the orders of the trigonometric models for the radial velocity and the photometry, respectively. The fifth-order model is clearly overwhelmingly preferred for velocity. For the photometry model, on the other hand, the third and fourth-order models are nearly equally supported. The MCMC computational strategy discussed above will automatically perform 'model-averaging' over these models, when computing posterior quantities of interest.

Estimates, standard errors, etc., for any of the unknowns or parameters in the model are also available from the MCMC computation. Here we simply show, in Figure 4, the posterior distribution of the distribution of the parallax (the inverse of the distance) for T Mon. Figure 5 shows the simulation history of the parallax, i.e., the values of the parallax that were generated at each trial of the MCMC computation. The very random appearance of this history strongly indicates that the MCMC computation was accurate.

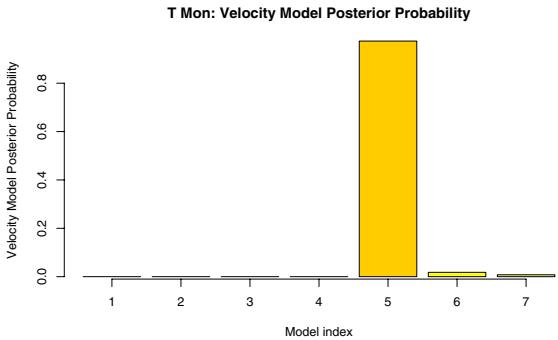


FIGURE 5.2. Posterior marginal distribution of velocity models for T Mon.

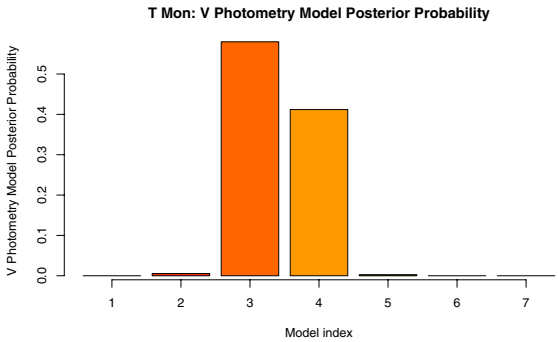


FIGURE 5.3. Posterior marginal distribution of photometry models for T Mon.

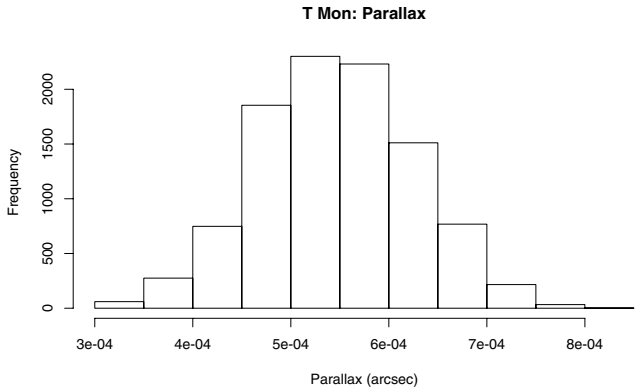


FIGURE 5.4. Posterior marginal distribution of the parallax of T Mon.

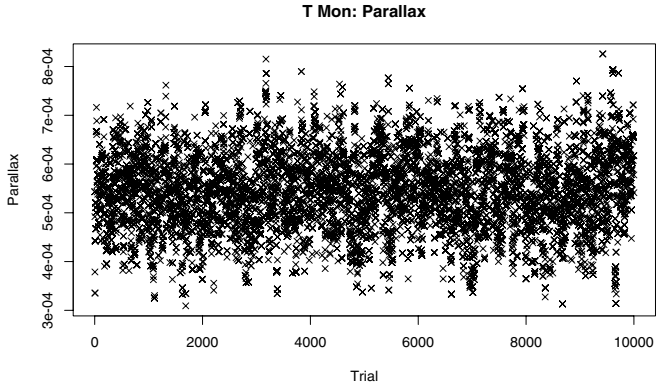


FIGURE 5.5. Simulation history of the parallax of T Mon.

5.4 A Wavelet Approach

Examination of Figure 1 suggests a potential concern. It turns out that the analysis is quite sensitive to the extent of the 'dip' in the velocity curve that occurs between phases 0 and 0.1. Notice also that there is no data between 0.9 and 0.1 (phases 1.0 and 0.0 being, of course, considered to be equal). Because Fourier analysis is non-local (each term in the trigonometric polynomial influencing the curve over the entire domain), there is concern that Fourier analysis may over-accentuate or under-accentuate the dip, in order to find a slightly better fit at points distant from the dip.

An approach that avoids this difficulty is the wavelet approach, since wavelet bases are local. To date, we have only applied this approach to the problem of fitting the velocity curve. Space precludes a detailed description here (see Müller, Berger, and Jefferys, 2001, for details and results), but we can, at least, outline the needed steps.

Step 1. A *function space* prior is needed, i.e., a prior on the space of possible velocity curves. The idea is to develop the prior in terms of intuitively accessible features of the function, and then transform this prior into the wavelet domain (a domain in which it is not as natural to construct priors). Adapting a suggestion of Vannucci and Corradi (1999), we chose the function space prior to be a Gaussian process (since this allows easy transformation into the wavelet domain). We actually construct the prior on differences of the function, since this makes it easier to (i) build periodicity into the Gaussian process and (ii) build smoothness into the function.

Step 2. One *transforms* this Gaussian process prior on the function space into the wavelet domain, using a bivariate wavelet decomposition, as suggested in Vannucci and Corradi (1999). The resulting prior on the wavelet coefficients is multivariate normal with a non-diagonal covariance matrix

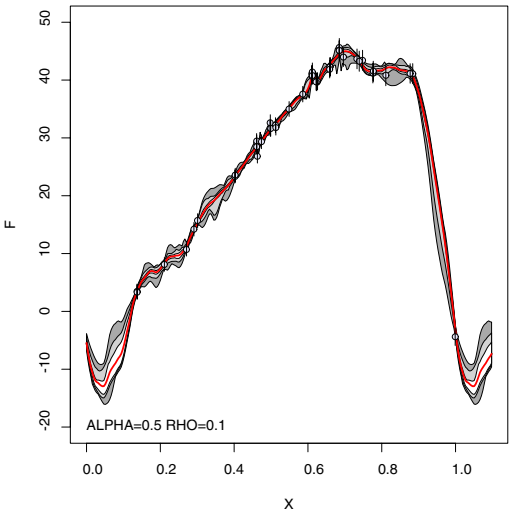


FIGURE 5.6. Wavelet fit of velocity data for T Mon. Shown are contour lines for the posterior distribution of the velocity as a function of phase; the thick smooth line in the center is the posterior mean curve. The grey shaded margins show central 50% (light grey) and central 90% (dark grey) intervals. The points are the observed data points, with little error bars showing 2 standard deviations for the measurement error.

(i.e., the wavelet coefficients are apriori dependent).

Step 3. A *model* in the wavelet domain is defined by some subset of all the wavelets in the basis. We specify the prior probability of a model through the device of allowing each wavelet coefficient to be zero, with specified probability $p(k)$, where k is the 'level' of the wavelet coefficient. (In practice, we used $p(k) = 1 - \alpha^{k+1}$, and tried various values of α.) Then, with probability $1 - p(k)$, the coefficient would be in the model. The prior distribution of the coefficients in the model is obtained from the multivariate normal prior found in Step 2, by conditioning on the other coefficients being zero.

Step 4. A *Metropolis-Hastings MCMC* analysis is implemented, in which moves are made to adjacent models (i.e., either a nonzero wavelet coefficient is set equal to zero, or a zero coefficient is made nonzero). The key is that, utilizing properties of the multivariate normal distribution, the computations involved in these 'small' steps are of relatively low cost to implement. (Wavelet models are large enough that it would be prohibitively expensive to compute, from scratch, the posterior model probabilities.)

5.5 Acknowledgments

This research was supported by the National Science Foundation, under Grants DMS-9802261 and DMS-0103265.

References

Berger, J. and Bernardo, J. (1992). On the development of the reference prior method. In J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith (editors), *Bayesian Statistics 4*, Oxford University Press, London.

Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison (with Discussion). To appear in the IMS Lecture Notes volume on *Model Selection*, P. Lahiri, editor.

Chen, M.H., Shao, Q.M. and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag.

Dellaportas, P., Forster, J. and Ntzoufras, I. (2000). On Bayesian model and variable selection using MCMC. To appear in *Statistics and Computing*.

Jaynes, E.T. (1999). *Probability Theory: The Logic of Science*. Accessible at the website http://bayes.wustl.edu/etj/prob.html.

Jefferys, W. H. and Barnes, T. G. (1999). Bayesian analysis of Cepheid variable data. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 777–783.

Jefferys, W. and Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist* **80**, 64–72.

Jeffreys, H. (1961). *Theory of Probability* (3rd edition), Oxford University Press, London.

Kass, R. and Raftery, A. (1995). Bayes factors and model uncertainty. *J. Amer. Statist. Assoc.*, 90, 773–795.

Müller, P., Berger, J., and Jefferys, W. (2001). Nonparametric regression with wavelet based priors: efficient posterior simulation for unequally spaced data and dependent priors. ISDS Discussion Paper, Duke University.

Pérez, J.M. and Berger, J. (2000). Expected posterior prior distributions for model selection. ISDS Discussion Paper 00-08, Duke University

Robert, C.P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Vannucci, M. and Corradi, F. (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J. Roy. Statist. Soc. B* , 971–986.

Zellner, A. and Siow (1980). Posterior odds for selected regression hypotheses. In *Bayesian Statistics 1*, eds. J.M. Bernardo, et. al., Valencia: Valencia University Press, pp. 585–603.

Commentary by Thomas J. Loredo[2]

The work of Berger et al. reported here is an exciting achievement. In these comments I highlight a few aspects of their approach that may help readers new to this kind of Bayesian modeling better appreciate the significance of this work and the applicability of elements of the approach to problems other than the Cepheid calibration problem.

5.6 Uncertain noise levels

A minor detail in the analysis that is nevertheless worth highlighting is the treatment of uncertainty in measurement errors. It is common in many astronomical disciplines for measurements to have standard errors associated with them that the analyst may consider to be only rough estimates (typically underestimates) of the actual standard errors. This is particularly often the case with cutting-edge observations. An example is high accuracy stellar radial velocity measurements, as used for detection of extrasolar planets. These measurements sometimes require spectroscopy capable of measuring velocities with few m s^{-1} accuracy. The formal standard errors (estimated, e.g., from photon counting statistics and instrument performance in test situations) often underestimate the actual errors because of unpredictable influences (e.g., stellar activity or atmospheric effects). This reveals itself by producing unacceptably large minimum χ^2 values in fits of velocity time series to models believed to be highly reliable (Keplerian reflex motion models).

The usual approach in such situations is to rescale the errors to make χ^2 have its expected value, and then proceed with the errors fixed at this rescaled value. This procedure is flawed; in general it leads one to underestimate the uncertainties in other parameters because it ignores uncertainty in the standard errors.

Berger et al. handle this by explicitly introducing scale parameters for the standard errors (their τ parameters), and treating them on an equal footing with other parameters. For astrophysical inferences, the τ parameters are uninteresting; the authors' MCMC calculation marginalizes (integrates) over them, fully accounting for their uncertainty in inferences of other parameters of interest.

This is the proper way to handle such uncertainty, and in the limit when the data allow precise inference of the scaling parameters, it reduces to the standard practice of simply rescaling the errors. This proper treatment does not always require the complexity of an MCMC calculation. A simple example illustrates these points. Suppose data x_i measure an unknown

[2]Department of Astronomy, Cornell University

constant, μ, and suppose that the reported standard errors are all the same, σ. If the true errors are σ/τ, the likelihood for μ and τ is the product of N Gaussians with width σ/τ,

$$\mathcal{L}(\mu,\tau) \propto \left(\frac{\tau}{\sigma}\right)^N \exp\left[-\frac{\tau^2}{2\sigma^2}\sum_i (x_i-\mu)^2\right]. \tag{5.9}$$

We are ultimately interested only in μ, so we multiply by a prior for τ (use the standard scale-invariant $1/\tau$ prior) and integrate over τ. The result is

$$\mathcal{L}(\mu) \propto \left[1+\frac{(\mu-\bar{x})^2}{s^2}\right]^{-N/2}, \tag{5.10}$$

where $\bar{x}$ is the sample mean and s^2 is the root-mean-square deviation from the mean. This has the form of Student's t-distribution. This likelihood has power-law tails, and is broader than the Gaussian likelihood that would result if we just fixed τ at some best-fit value. But if N is very large, equation (5.10) is well-approximated by

$$\mathcal{L}(\mu) \propto \exp\left[-\frac{(\mu-\bar{x})^2}{2s^2}\right]. \tag{5.11}$$

This is just what one would get from the standard fixed-τ approach. Thus marginalization accounts for τ uncertainty by broadening the likelihood; but when τ is well-determined, it effectively just plugs in its estimate.

5.7 Systematic error

The most important innovation in the analysis by Berger et al. is their extensive and rigorous accounting for *model uncertainty*. It is the uniquely Bayesian concept of the probability for a model, combined with the ability to marginalize over unknowns (i.e., model choice), that makes such an accounting possible. Although they do not use the term in the paper, to properly understand the significance of their calculation I think it is important to use it here: they have shown how to account for an important source of *systematic error* (see Drell et al. 2000 for a simpler Bayesian treatment of systematic error in cosmology).

Systematic error is an embarassment to frequentist statistics. It is not "random," and therefore cannot be described with (frequentist) probabilities. It is thus difficult to carry out calculations that account for it. Taylor (1997) summarizes the situation thus: "No simple theory tells us what to do about systematic errors. In fact, the only theory of systematic errors is that they must be identified and reduced until they are much less than the required precision." In regard to quantitative accounting for it, he continues,

" . . . there are various possible ways to proceed. None can be rigorously justified. . . . "

In Bayesian inference, probabilities describe uncertainty, not (necessarily) "randomness" or experiment-to-experiment fluctuations. Systematic error is thus amenable to probabilistic treatment. This was noted half a century ago by Jeffreys (1961). In an example concerning estimation of a location parameter, he wrote:

> *Systematic error has a meaning only if we understand by the true value something different from the location parameter. It is therefore an additional parameter, and requires a significance test for its assertion.* There is no epistemological difference between the Smith effect and Smith's systematic error; the difference is that Smith is pleased to find the former, while he may be annoyed at the discovery of the latter. Now with a proper understanding of induction there is no need for annoyance.

Translating to more modern terminology, systematic error can in principle be accounted for by modifying the model for the data. Uncertainty in such error can thus be quantified by using Bayesian methods to account for model uncertainty.

Jeffreys stumbled in cases where Bayesian model comparison (his "significance test") could not conclusively determine whether a particular systematic effect was present or not: "The problem that remains is, how should we deal with possible systematic errors that are *not* yet established and whose values are unknown?" Today this problem is routinely dealt with via Bayesian model averaging (rather than choosing a single best model), the key ingredient of the Cepheid analysis reported here.

Systematic error has been the bane of cosmological research for decades, leading investigators analyzing similar data to reach discrepant conclusions due to the influences of modeling assumptions. There may be important sources of systematic uncertainty in Cepheid calibration beyond the light curve model uncertainties accounted for in this study. But this approach should go a long way toward further resolving discrepancies in this field, and will hopefully motivate further use of Bayesian methods to quantify systematic uncertainties in astronomy.

5.8 Computational complexity

The authors state that their computational methods are too complicated to describe in detail in the limited space available here, and indeed the few details provided indicate that significant effort and not a little artistry were required to perform the calculations. Many presentations of Bayesian methods at the conference shared this level of computational complexity,

leading to the oft-repeated remark in discussion sessions that Bayesian calculations are much more challenging than frequentist calculations.

This statement is misleading. In problems amenable to both Bayesian and frequentist analyses with similar models, Bayesian and frequentist calculations typically have similar complexity *when carried out at the same level of approximation* (in fact, the Bayesian calculation is sometimes much simpler in such cases). The key observation here is that in most problems of realistic complexity, rigorous frequentist calculations are not hard—they are impossible. Typically, no rigorous frequentist result exists for a finite sample size, and the analyst must rely on asymptotic approximations. When such an approximation is adequate, it should be compared in complexity, not with a full Bayesian calculation, but with an asymptotic Bayesian calculation. Such calculations are straightforward and involve quantities and manipulations familiar from standard frequentist analyses (e.g., locating maxima and finding Hessian matrices). The primary tool is called the *Laplace approximation* (see Loredo 1999 for a brief overview and references). Interestingly, because Bayesian calculations typically require ratios of probabilities, an asymptotic Bayesian calculation is sometimes accurate to higher order than its frequentist counterpart, because the lowest order error cancels in the ratio.

The difficult Bayesian calculations described in the work of Berger et al. and in other presentations are difficult because they implement calculations that would be difficult or impossible even to frame in frequentist terms—calculations that are exact for finite sample size, or that rigorously account for model uncertainty. Instead of bemoaning the complexity of such calculations, we should be grateful that we are at last able to perform them at all. Even so, their complexity is a legitimate stumbling block to potential users. The success of calculations like those reported here will hopefully motivate the community to support development of software that hides implementation details from users, and research in Bayesian computation to develop general-purpose algorithms that reduce the computational complexity of Bayesian analyses.

P. S. Drell, T. J. Loredo, and I. M. Wasserman. Type Ia supernovae, evolution, and the cosmological constant. *Ap. J.*, 530:593–617, 2000.

H. Jeffreys. *Theory of Probability.* Clarendon Press, Oxford, 1961.

T. J. Loredo. Computational technology for bayesian inference. In D. M. Mehringer, R. L. Plante, and D. A. Roberts, editors, *ASP Conf. Ser. 172: Astronomical Data Analysis Software and Systems VIII*, pages 297–306, San Francisco, 1999. Astronomical Society of the Pacific.

J. R. Taylor. *An introduction to error analysis.* University Science Books, Sausalito, CA, 1997.

6

Bayesian Multiscale Methods for Poisson Count Data

Eric D. Kolaczyk[1]

ABSTRACT We present an overview of recent work on a flexible framework for multiscale modeling of Poisson count data, such as is encountered regularly in the field of high-energy astrophysics, that allows for intuitive, easily interpretable, computationally efficient implementations of Bayesian inference for standard tasks like smoothing, deconvolution, and segmentation. At the foundation of this approach is a multiscale factorization of the Poisson likelihood, which can be viewed formally as deriving from a blending of concepts from the literatures on wavelets, recursive partitioning, and graphical models.

6.1 Introduction

Astronomers, especially those studying phenomena in the higher energy levels (e.g., x-ray and γ-ray), are faced with the challenge of analyzing increasingly vast amounts of photon counting data (typically with temporal and/or spatial labels), whose statistical properties generally are characterized as Poisson in nature. Methods of analysis must necessarily be computationally efficient and scalable, particularly those intended to serve as instrument-based or preliminary ground-based tools. These requirements can present a significant challenge to the development and adoption of Bayesian methods in such settings.

Consider, for example, the task of conducting multiscale analyses, standard methods for which derive typically from some manner of wavelet-based representation of the data. A multiscale analysis of Poisson data with wavelets leads to technical statistical challenges not necessarily encountered with, say, data following a standard Gaussian (i.e., "signal plus noise") model, due to fundamental differences in how the underlying statistical distributions "interact" with wavelet structures. These challenges in turn have a direct impact on issues of analytical and computational tractability of resulting methods.

We present here an overview of recent work on ways to meet these chal-

[1]Department of Mathematics and Statistics, Boston University

lenges, based on the use of likelihood factorizations. The resulting statistical framework allows for the creation of methods for standard tasks such as smoothing, deconvolution, and segmentation that are intuitive and interpretable, as well as analytically and computationally tractable, even for posterior-based Bayesian inference. The basic modeling structure is introduced in section 6.2, illustration of how that structure may be used for standard inferential tasks is given in section 6.3, and some additional discussion regarding extensions and generalizations can be found in section 6.4.

6.2 The basic multiscale modeling structure

The goal in this paper is to communicate the fundamental usefulness of certain structural characteristics in multiscale modeling, with less emphasis being placed on more detailed alterations that would necessarily have to be made in the context of various specific applications. Hence, we will work with the following generic modeling structure throughout. Let $X(t), t \in [0,1)$ be a Poisson process with intensity function $\lambda(t) \geq 0$. Additionally, assume that through convention and/or design the interval $[0,1)$ is discretized into N equispaced bins $I_n = [n/N, (n+1)/N)$, $n = 0, \ldots, N-1$. There then results from this discretization an $N \times 1$ vector $\boldsymbol{X}$ of independent Poisson random variables $X_n \sim \text{Poisson}(\Lambda_n)$, where $\Lambda_n = \int_{I_n} \lambda(t)dt$ and '$\sim$' is to be read 'distributed as'. Our focus in this paper will be on "low level" data processing tasks involving statistical inference on the vector $\boldsymbol{\Lambda}$ (i.e., on $\lambda(\cdot)$ up to the resolution of the binning).

6.2.1 Factorizing the data likelihood

It is more or less commonplace now to have tools in the astronomer's data analysis toolbox for doing scale-sensitive inference – that is, inference on the characteristics of an object(s) (e.g., time series of photon arrivals, image mapping of point sources, etc.) for which there are potentially structural components at multiple scales. Wavelet-based methods are by far the most common such tools, and there have been numerous contributions in this direction. See, for example, the book by Starck, Murtagh, and Bijaoui (1998) or the chapter by Starck in this volume, for an overview.

Our own approach is intimately related to, yet distinct from, wavelets and such wavelet-based methods. To better motivate both this connection and the inherent differences, consider the simple case in which an orthonormal wavelet transform is computed for data observed from a signal-plus-noise model i.e., $\boldsymbol{W} = \mathcal{W}\boldsymbol{Y}$, where $Y_n = \Lambda_n + Z_n$ and the Z_n are independent and identically distributed Gaussian random variables of mean zero and unit variance. Because the matrix $\mathcal{W}$ is an orthogonal matrix and the data $\boldsymbol{Y}$ are independent Gaussian, the vector $\boldsymbol{W}$ too is independent Gaussian.

Hence, from a likelihood based perspective we might write something like

$$\prod_{n=0}^{N-1} \Pr\left(Y_n|\Lambda_n\right) = \prod_{j,k} \Pr\left(W_{j,k}|\omega_{j,k}\right), \tag{6.1}$$

where (j,k) refers to the standard scale-position indexing resulting from the definition of orthonormal wavelet functions $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$ with respect to a single function ψ, and $\boldsymbol{\omega} = \mathcal{W}\boldsymbol{\Lambda}$.

The key point to note here is that the joint (i.e., N dimensional) likelihood is *factorized* in both the time (left hand side) and multiscale (right hand side) domains into a product of N component likelihoods. And furthermore, each component relies on a single pairing of observation and parameter — Y_n with Λ_n in the time domain and $W_{j,k}$ with $\omega_{j,k}$ in the multiscale domain. This deceptively simple fact has both analytical and computational implications. For example, it can be seen to motivate the standard idea of thresholding individual empirical wavelet coefficients $W_{j,k}$ in order to denoise the signal $\boldsymbol{Y}$ as a whole, which is essentially an $O(N)$ algorithm (e.g., Donoho and Johnstone 1994, but see also Johnstone and Silverman 1997 for extensions to certain types of correlated data). And much of the corresponding analysis of the statistical risk of such estimators boils down to understanding the aggregate behavior of the individual risks associated with such thresholding. Additionally, most Bayesian methods in this context consist, for similar reasons, of making a posterior inference on $\boldsymbol{\Lambda}$ implicitly through component-wise posterior inferences on the $\omega_{j,k}$ (e.g., Chipman, Kolaczyk, and McCulloch 1997; Clyde, Parmigiani, and Vidakovic 1998; Abramovich, Sapatinas, and Silverman 1998).

Now consider the case in which the same wavelet transform is applied to our Poisson observations i.e., $\boldsymbol{W} = \mathcal{W}\boldsymbol{X}$. With the change from Gaussian to Poisson observations, the orthogonality of $\mathcal{W}$ is no longer sufficient to ensure the statistical independence of the components of $\boldsymbol{W}$. Hence, a factorization of the form given in (6.1) does not hold. While the effect of this point on thresholding methods might be simply to adjust the level of the thresholds used, its impact on the development of Bayesian methodologies is more substantial, as the full likelihood must be used in an explicit manner.

Generally speaking, the existence of factorizations of complex probability functions involves a delicate combination of issues concerning both the underlying distribution and its parameterization. The study of such factorizations is of central interest to the area of *graphical modeling* in the statistics literature (e.g., Lauritzen 1996), wherein models are formulated by specifying conditional independence relationships among the relevant variables through the absence of edges connecting vertices (representing the variables) on a mathematical graph. From this perspective, one can view the factorization on the right hand side of (6.1) as a factorization of the joint distribution of $\boldsymbol{Y}$ with respect to the binary tree graph generated

by the index pairs (j,k), for $j = 0, 1, \ldots, J-1$, $k = 0, 1, \ldots, 2^j - 1$, and $J = \log_2(N)$.

Although such a factorization does not exist for the Poisson data $\boldsymbol{X}$ when the $W_{j,k}$ are the empirical wavelet coefficients, it *does* exist when instead the $W_{j,k}$ are replaced by a certain conditional distribution. Specifically, let $I_{j,k} \equiv [k/2^j, (k+1)/2^j)$, for all (j,k), and define $X_{j,k}$ to be the sum of the X_n for which $I_n \subseteq I_{j,k}$. Let $\Lambda_{j,k}$ be defined similarly in terms of the Λ_n. It can be shown then that the factorization

$$\prod_{n=0}^{N-1} \Pr(X_n|\Lambda_n) = \Pr(X_{0,0}|\Lambda_{0,0}) \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} \Pr(X_{j+1,2k}|X_{j,k}, \rho_{j,k}) \qquad (6.2)$$

holds, where $\rho_{j,k} \equiv \Lambda_{j+1,2k}/\Lambda_{j,k}$. The marginal distribution of $X_{0,0}$ is just Poisson$(\Lambda_{0,0})$, while the conditional distributions $X_{j+1,2k}|X_{j,k}$ are binomial$(X_{j,k}; \rho_{j,k})$. This result may be derived directly using well-known relations between the binomial and Poisson distributions (Kolaczyk 1999; Timmerman and Nowak 1999), or more formally using a probabilistic analogue of the type of multiresolution analysis (MRA) that underlies orthonormal wavelet bases (Kolaczyk and Nowak 2000).

To better understand how (6.2) compares to (6.1), consider the case in which $\mathcal{W}$ corresponds to the Haar wavelet transform. There the $W_{j,k}$ are simply (proportional to) the difference of $X_{j+1,2k}$ and $X_{j+1,2k+1}$ i.e., the sums of counts in the left and right "children" intervals $I_{j+1,2k}$ and $I_{j+1,2k+1}$ of the "parent" interval $I_{j,k}$. This difference provides some notion of the information in the data localized to the scale/position combination (j,k). However, consideration of the conditional distribution of one of the children, $X_{j+1,2k}$, given the value of the parent $X_{j,k}$, provides a similar notion of such local information. And it is with respect to this latter notion that a multiscale factorization exists for Poisson data, in which case the accompanying re-parameterization of $\boldsymbol{\Lambda}$ is not with respect to its Haar coefficients $\omega_{j,k}$ but rather the ratios $\rho_{j,k}$.

6.2.2 Prior distributions on the multiscale parameters

The factorization of the likelihood in (6.2) may be thought of in analogy to a wavelet decomposition of a function. In other words, it provides an alternative, position/scale representation of an object of interest. When the underlying structure of that object is well-captured in this representation, it may prove beneficial to conduct inference on the $\Lambda_{j,k}$ indirectly through direct inference on the $\rho_{j,k}$. In order to conduct such inference on the $\rho_{j,k}$ using Bayesian methodologies, an appropriate prior distribution structure must be specified for these parameters (hence making them random variables).

In the literature on wavelets and the Gaussian signal-plus-noise model there is already a sizeable literature on Bayesian approaches. Most begin

with the observation that, across many contexts, distributions of wavelet coefficients $\omega_{j,k}$ have been observed to be "heavy-tailed" and centered at zero (e.g., Mallat 1998). Various authors therefore have suggested the use of zero-mean Laplacian distributions, mixtures of zero-mean Gaussians, and generalized Gaussian distributions to capture this behavior. Most methods assume independence among the coefficient distributions, citing the ability of wavelets to roughly "decorrelate" the structure in an object, but more sophisticated models attempt to capture weak dependencies through the use of multivariate distributions or model the persistence of edges across scales of coefficients through the use of tree-based hidden Markov models. See Chipman and Wolfson (1999) for a recent survey.

Now consider the nature of the $\rho_{j,k}$. If in a certain region most of the Λ_i are roughly equal, then many of the $\Lambda_{j,k}$ will be roughly equal across locations k for some range of scales j. We will then have $\rho_{j,k} \approx 1/2$ for many of the (j,k), which is the analogue of having $\omega_{j,k} \approx 0$ in the case of wavelets. Also note that, by definition, $\rho_{j,k} \in [0,1]$, for all (j,k). These observations, combined with the fact that each $\rho_{j,k}$ arises as the parameter of a binomial distribution, suggest the use of the beta distribution

$$f(\rho) = \frac{1}{B(\alpha,\beta)}\rho^{\alpha-1}(1-\rho)^{\beta-1}, \tag{6.3}$$

as a prior family (being conjugate to the binomial family), where $\alpha, \beta > 0$ and $B(\alpha,\beta)$ is just the standard beta function. For example, through choice of $\alpha = \beta$ a distribution arises with symmetry about the point $1/2$, where α less than, equal, or greater than one yields U-shaped, uniform, and unimodal distributions, respectively. More flexibility in shape results from the use of mixtures of such betas, for example, a point mass at $1/2$ and a uniform distribution, where the weight on each of the two components may be adjusted to reflect a balancing of prior beliefs in relative homogeneity (i.e., $\rho_{j,k} \approx 1/2$) and ignorance (i.e., $\rho_{j,k} \sim \text{Unif}(0,1)$). See Kolaczyk (1999) and Timmerman and Nowak (1999), for more discussion along these lines. Additionally, cross-scale dependencies may be incorporated through use of hidden Markov model tree structures, as in the Gaussian case, as described in Nowak (1999).

On a final note we mention that, because one often has relevant information on the character of $\lambda(\cdot)$ i.e., of the intensity in the original time domain, it is important to understand the nature of the prior distribution induced on the $\Lambda_{j,k}$ through our specification of priors on the $\rho_{j,k}$. Study of this issue may be found within Kolaczyk (1999), Timmerman and Nowak (1999), and Nowak and Kolaczyk (2000), and in Louie and Kolaczyk (2002) in more generality, in which conditions for such characteristics as (non)stationarity, long-range dependence, and asymptotic convergence are explored.

6.3 Illustration of methods

In this section we will consider three common inferential tasks — smoothing, deconvolution, and segmentation – and show how the basic modeling structure of the previous section may be adapted in each context to obtain efficient algorithms for posterior-based inference.

6.3.1 Smoothing

Figure 6.1(a) shows a plot of the photon arrival times for a gamma-ray burst observed by the BATSE instruments, as part of the recently completed Compton Gamma Ray Observatory (CGRO) mission. Counts for the first $N = 256$ 64ms time bins are displayed. Norris *et al.* (1996) fit this and similar bursts with linear combinations of asymmetric pulse functions, from which aggregate information on number, location, amplitude, and width of peaks is used to discern commonality across what has been found to be a highly variable class of signals. Methods such as these are inherently parametric, of course, and in situations such as this, where reliable physical models are lacking, a nonparametric m ethod often can be employed usefully in a complementary fashion to gain insight into features perhaps missed by the parametric method.

It is standard to model such observations as Poisson, in the manner outlined at the start of section 6.2. Recalling the factorization of the Poisson distribution in (6.2), consider a model for the multiscale parameters $\rho_{j,k}$ that specifies

$$\rho_{j,k} \,|\, \gamma_{j,k} \;\sim\; \gamma_{j,k}\, \delta_{1/2} + (1 - \gamma_{j,k})\, B_{j,k} \tag{6.4}$$

$$\gamma_{j,k} \,|\, p_j \;\sim\; \text{Bernoulli}(p_j) \tag{6.5}$$

$$B_{j,k} \,|\, \alpha_j \;\sim\; \text{Beta}(\alpha_j, \alpha_j) \tag{6.6}$$

In other words, each $\rho_{j,k}$ is modeled independently of the others as a mixture of a point mass at $1/2$ and a beta random variable, where the mixing parameter p_j and the beta parameter α_j are indexed by scale j (but not position).

Under these conditions it is not difficult to show (Kolaczyk 1999, Lemma 2) that the posterior has the factorization

$$\Pr\left(\mathbf{\Lambda}|\boldsymbol{X}, \Lambda_{0,0}\right) = \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} \Pr(\rho_{j,k}|X_{j+1,2k}, X_{j+1,2k+1}) \;. \tag{6.7}$$

That is, the time domain posterior (left hand side) is actually a product of local posteriors in the multiscale domain (right hand side), each of which actually are mixtures of beta distributions like the prior. Therefore, posterior-based inference on $\mathbf{\Lambda}$ may be accomplished through performing

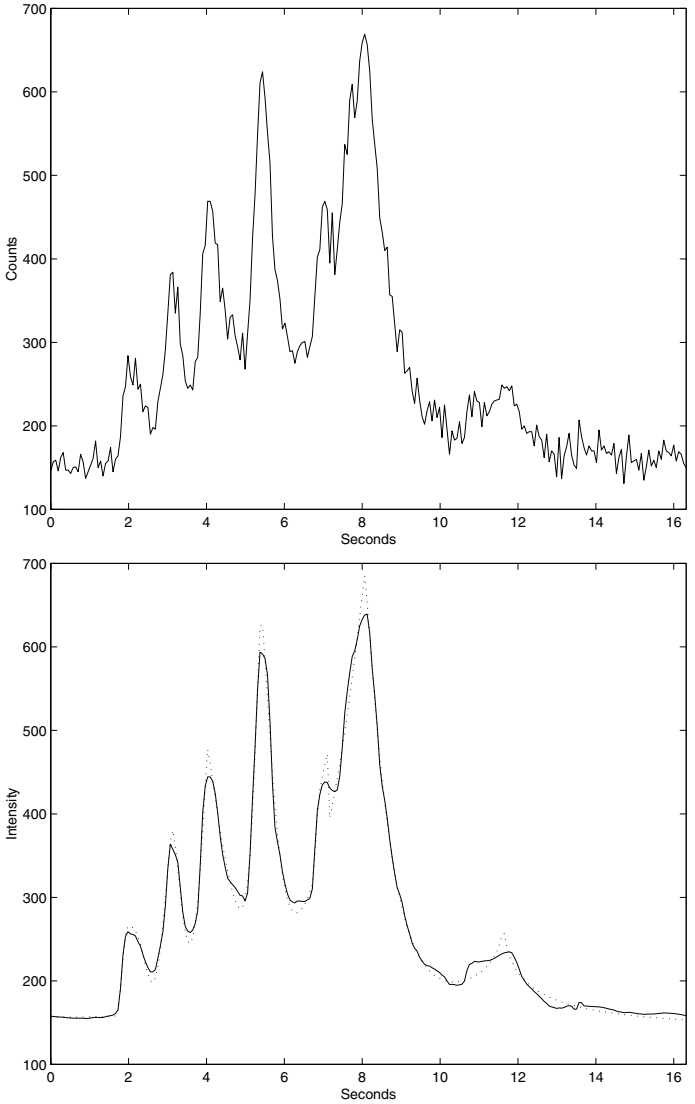


FIGURE 6.1. On the top (a) is the GRB BATSE trigger # 1425, with photon arrival times binned into 256 64ms bins. On the bottom (b) is our translation-invariant multiscale posterior mean estimate (solid) and an estimate based on the parametric fitting of asymmetric pulse shapes (dotted). Priors for the multiscale method were independent mixtures of a point mass at 1/2 and a uniform distribution (i.e., $\alpha_j \equiv 1$), with the mixing parameters p_j fit using an empirical Bayes method.

posterior-based inference on each multiscale component $\rho_{j,k}$ and then inverting the underlying multiscale transformation. For example, if the posterior mean is to be used as an estimate of $\mathbf{\Lambda}$, this relates to the posterior means of the $\rho_{j,k}$ via the formula

$$E\left[\Lambda_n|\boldsymbol{X},\Lambda_{0,0}\right] = \Lambda_{0,0}\prod_{j=0}^{J-1} E\left[\tilde{\rho}_{j,j(n)}\,|\,X_{j+1,2j(n)},X_{j+1,2j(n)+1}\right] \quad , \qquad (6.8)$$

where $j(n)$ represents the position index at scale j of the ancestor of the n-th component of $\mathbf{\Lambda}$ and $\tilde{\rho}_{j,j(n)}$ is equal to either $\rho_{j,j(n)}$ or $1-\rho_{j,j(n)}$, depending on whether these ancestors are left or right children of their parents, respectively.

In Figure 6.1(b) is shown a translation-invariant version of this posterior mean estimate (see Kolaczyk 1999) for the intensity underlying BATSE trigger #1425. Super-imposed upon that is the estimate obtained by the method of Norris *et al.*, in which seven distinct pulse shapes were fit. Note that while our nonparametric method confirms the presence and general form of the first six of those seven, it suggests evidence of there being in fact two pulses in the region of their seventh. Such sections of the data with notable degrees of pulse overlap are particularly difficult to fit parametrically (J. Norris, personal communication).

6.3.2 Deconvolution

Due to effects associated with the measurement process and instrumentation, often it is not possible to observe the data $\boldsymbol{X} \sim \text{Poisson}(\mathbf{\Lambda})$ directly. Instead it may be more appropriate to model the data as "indirect" observations $Y_m \sim \text{Poisson}(\mu_m)$, for $m = 0,\ldots,M-1$, where $\boldsymbol{\mu} = P\mathbf{\Lambda}$ and P is some $M \times N$ transition matrix. That is, we specify a Poisson linear inverse problem, where the underlying mean vector $\boldsymbol{\mu}$ is a "blurred" version of the object $\mathbf{\Lambda}$ in which our interest truly lies.

Although there are a variety of methods for dealing with inverse problems, a now-commonplace one for those involving Poisson data is through use of the Expectation-Maximization (EM) algorithm framework. To review briefly, in the present context one can introduce an auxiliary set of random variables $\{Z_{m,n}\}$, where $Z_{m,n}$ is the number of (e.g., photon) counts associated with X_n that contribute to the total in Y_m. Clearly $\boldsymbol{Y}$ and $\boldsymbol{X}$ can be obtained as the marginal row/column totals of the matrix $\boldsymbol{Z}$ and, by properties of the Poisson distribution, we have that $Z_{m,n} \sim \text{Poisson}(P_{m,n}\Lambda_n)$. Given a prior distribution $\Pr(\mathbf{\Lambda})$ for $\mathbf{\Lambda}$, the EM algorithm may be used to iterate to a Bayesian maximum *a posteriori* (MAP) estimate by computing the conditional expectation

$$Q\left(\mathbf{\Lambda},\mathbf{\Lambda}^{(i)}\right) \equiv E_{\mathbf{\Lambda}^{(i)}}\left[\log\Pr\left(\boldsymbol{Z}|\mathbf{\Lambda}\right)\,|\,\boldsymbol{Y}\right] + \log\Pr\left(\mathbf{\Lambda}\right), \qquad (6.9)$$

as the E-step, and then maximizing $Q\left(\mathbf{\Lambda}, \mathbf{\Lambda}^{(i)}\right)$ as a function of $\mathbf{\Lambda}$ to produce $\mathbf{\Lambda}^{(i+1)}$, as the M-step. The computational feasibility of such approaches typically is limited by that of the M-step, which in turn is linked to the nature of the prior and how it interacts with the likelihood. The computational complexity of this step may range from calculation of simple closed-form solutions to running a full Monte Carlo simulation at each iteration i.

Now consider doing multiscale Bayesian inference on $\mathbf{\Lambda}$ in a manner similar to that of section 6.3.1, but based on the observations $\boldsymbol{Y}$ and using the EM framework. A simple conditioning argument shows that the distribution of our auxiliary data $\boldsymbol{Z}$ may be expressed as

$$\begin{aligned}\Pr(\boldsymbol{Z} \mid \mathbf{\Lambda}) \;=\; & \Pr(X_{0,0}|\Lambda_{0,0}) \;\times\; \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} \Pr(X_{j+1,2k} \mid X_{j,k}, \rho_{j,k}) \;\times \\ & \prod_{n=0}^{N-1} \Pr(Z_{0,n}, \ldots, Z_{M-1,n}|X_n, p_{0,n}, \ldots, p_{M-1,n}). \qquad (6.10)\end{aligned}$$

Two characteristics of the expression in (6.10) are important to note: (i) the third term (second line) on the right hand side does not involve $\mathbf{\Lambda}$, and (ii) the first two terms (first line) on the right hand side are identical to those in equation (6.2). Hence, with respect to optimizations involving $\mathbf{\Lambda}$, only the first two terms are relevant and these instruct us effectively to act as if we had observed the data $\boldsymbol{X}$ in the first place.

The end result is an EM algorithm in the Bayesian multiscale framework that is no more computationally intensive than the standard EM algorithm for maximum likelihood estimation, with closed form expressions for both E- and M-steps. For example, suppose we choose to induce a prior distribution on $\mathbf{\Lambda}$ by placing independent beta priors on the multiscale parameters i.e., $\rho_{j,k} \sim \text{beta}(\alpha_j, \alpha_j)$. Then the E-step in (6.9) boils down to calculating

$$Z^{(i)}(m,n) \;=\; \frac{Y_m \Lambda_n^{(i)} p_{m,n}}{\sum_{l=0}^{N-1} \Lambda_l^{(i)} p_{m,l}}, \qquad (6.11)$$

due to the fact that $\boldsymbol{Z}|\boldsymbol{Y}$ is multinomial in distribution and the linearity of the logarithm of this distribution in the $Z_{m,n}$. Furthermore, the M-step results in the $(i+1)$-th iteration estimates

$$\rho_{j,k}^{(i+1)} = \frac{X_{j+1,2k} + \alpha_j - 1}{X_{j,k} + 2(\alpha_j - 1)}, \qquad (6.12)$$

from which the estimate $\mathbf{\Lambda}^{(i+1)}$ may be constructed in a manner similar to that in equation (6.8). A more detailed derivation of these results, as well as results establishing convergence of the EM algorithm under various choice of the α_j, may be found in Nowak and Kolaczyk (2000).

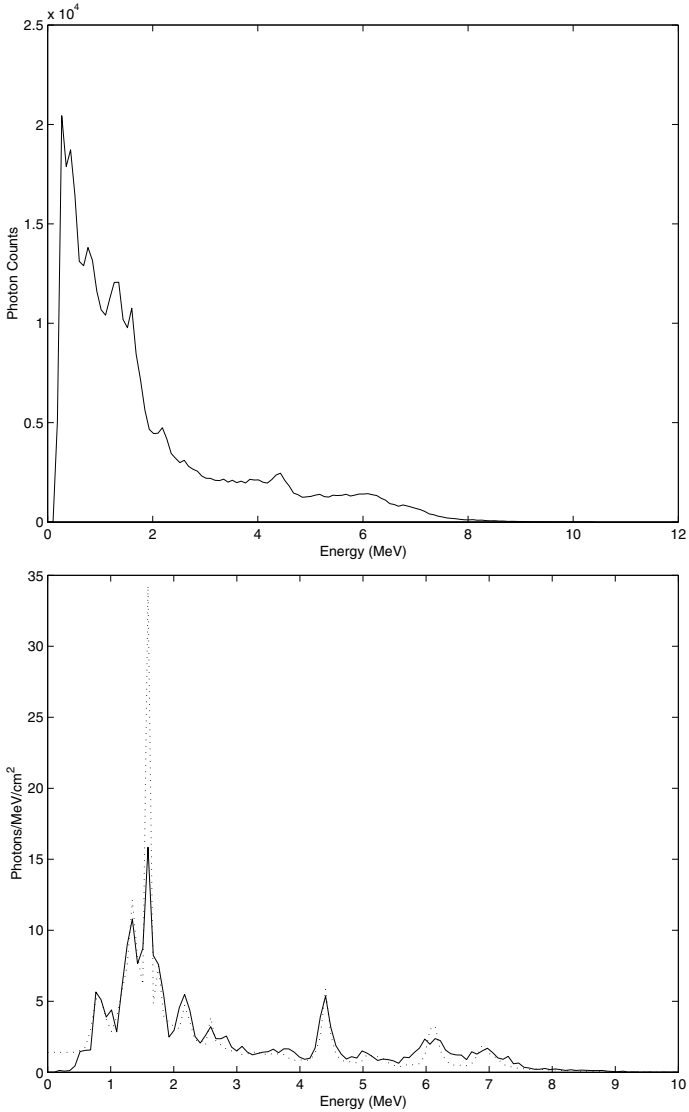


FIGURE 6.2. On the top (a) is simulated solar flare data, as might be measured by the COMPTEL instruments. On the bottom (b) are the Bayesian multiscale estimate of the underlying energy spectrum (solid) and the spectrum itself (dotted).

By way of illustration, Figure 6.2(a) shows a collection of counts corresponding to a certain theoretical energy spectrum for the production of gamma rays by energetic particles interacting with the ambient solar atmosphere (Murphy *et al.* 1991). The counts were simulated from this model as if having been observed by the COMPTEL instruments (also part of the the CGRO mission). Due to the underlying physics of the measurement devices, an arriving photon in fact has a good chance of being recorded at some lower energy level than that at which it obtains. Figure 6.2(b) shows an estimate of the underlying energy spectrum $\mathbf{\Lambda}$ resulting from our multiscale deconvolution algorithm. The true spectrum is super-imposed upon this plot – although there is the expected attenuation in the heights of the various spectral peaks, note that the relative number, location, and width of each are well-recovered.

6.3.3 Segmentation

Our final illustration involves the task of segmentation. For the sake of simplicity, consider now again the case in which we directly observe the measurements $\boldsymbol{X} \sim \text{Poisson}(\mathbf{\Lambda})$. Often it is of interest to divide the domain of observation, which we have generically taken to be the interval $[0, 1]$, into disjoint regions within which the vector $\mathbf{\Lambda}$ has some sort of locally similar behavior. The simplest example is that in which we wish to identify regions in which $\mathbf{\Lambda}$ is piecewise constant – that is, in which the underlying Poisson process is locally homogeneous. This problem can also be referred to as one of finding multiple changepoints in $\mathbf{\Lambda}$.

One can envision for this problem, in principle at least, the generation of data $\boldsymbol{X}$ as a three stage process in which (i) a collection of segmentation points are laid down in $[0, 1]$ at some subset of the locations n/N, for $n = 1, \ldots, N-1$, (ii) values for $\mathbf{\Lambda}$ are chosen for the resulting subintervals of constant intensity, and (iii) $\boldsymbol{X}$ is sampled as $\text{Poisson}(\mathbf{\Lambda})$. This three-step procedure lends itself naturally to hierarchical modeling in a Bayesian setting. Moreover, if one pictures the segmentation points being laid down in a recursive fashion, then structural and conceptual connections between recursive partitioning and our multiscale modeling framework may be exploited to obtain a Bayesian multiscale method for simultaneously selecting the most likely number of segmentation points and their locations.

This search for an optimal segmentation of a given dataset $\boldsymbol{X}$ can be viewed as a Bayesian model selection problem. Specifically, we seek the most likely member $\mathcal{M}$ of, say, the collection $\mathcal{L}$ of all possible recursive partitions, i.e.,

$$\mathcal{M}^{opt} \equiv \arg\max_{\mathcal{M} \in \mathcal{L}} \Pr\left(\mathcal{M} | \boldsymbol{X}\right). \tag{6.13}$$

Due to similar reasons associated with our factorizations of the likelihood and choice of priors upon which rested our results in sections 6.3.1 and 6.3.2, it turns out that the optimization in (6.13) can be solved in an efficient man-

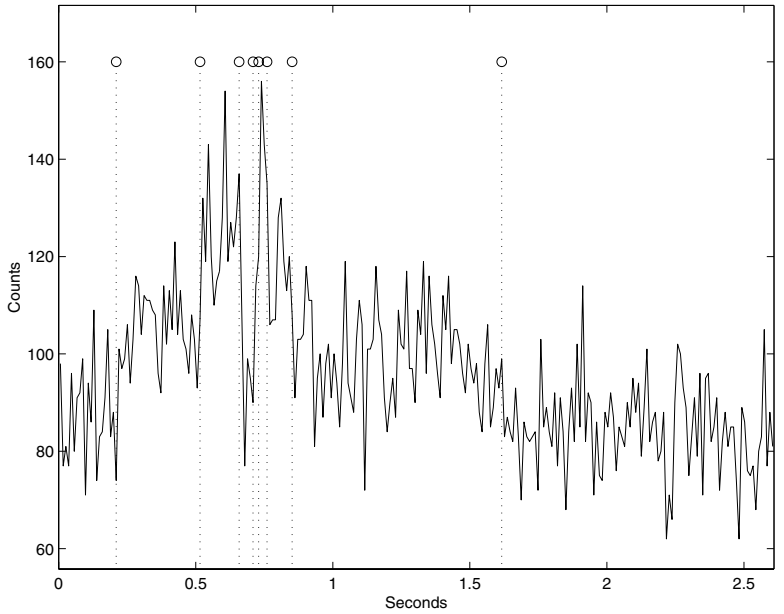


FIGURE 6.3. Bayesian multiscale segmentation of GRB BATSE trigger #845.

ner. A redundancy among many of the recursive partitions and the prevalence of binary tree structures allows a search for $\mathcal{M}^{opt}$ in roughly $O(N^3)$ operations using a type of probability propagation algorithm. Details may be found in Kolaczyk and Nowak (2001). Figure 6.3 shows an illustration of this algorithm when applied to a gamma-ray burst of a rather different character than that encountered in Figure 6.1.

6.4 Discussion

The goal here has been to provide an overview of a general framework for statistical analysis of Poisson count data, in a manner sensitive to structure at multiple scales, with an emphasis on Bayesian methods. Of course, in a specific application there is likely to be additional information beyond that used in the applications described herein, including data from other instruments, different wavelengths, and physical models. Bayesian methods often are particularly convenient for incorporating this type of information into the inferential process. On the other hand, it is common for such methods to become computationally burdensome, which can be a serious disadvantage for some of the high data-throughput applications that now characterize many aspects of modern satellite-based astronomy. Hence the emphasis in our methods on the use of likelihood factorizations, whose decoupling of

the underlying probability structure facilitates the development of efficient computational algorithms.

On a final note, we mention two other related pieces of work. David Esch and David van Dyk have adapted the deconvolution methodology to the processing of Chandra x-ray image data, and are exploring the use of MCMC for adaptively setting the prior parameters (i.e., the α_j's, in the notation of this paper). Alex Young is studying the performance of the same methodology in the context of solar flare data at the γ-ray level, and is using parametric bootstrapping methods to obtain confidence statements on the reconstructions. Readers are referred to the papers by these authors in this volume.

Software for performing analyses like those described in this paper is available at http://math.bu.edu/people/kolaczyk/astro.html .

References

Abramovich, F., Sapatinas, T., and Silverman, B.W. (1998). Wavelet thresholding via a Bayesian approach, *Journal of the Royal Statistical Society, Series B.,* 60, 725 - 750.

Chipman, H.A., Kolaczyk, E.D., and McCulloch, R.E. (1997). Adaptive Bayesian wavelet shrinkage, *Journal of the American Statistical Society,* 92, 1413-1421.

Chipman, H.A., and Wolfson, L.J. (1999). Prior elicitation in the wavelet domain. In *Bayesian Inference in Wavelet-Based Models,* Müller, P. and Vidakovic, B. (eds.). New York: Springer-Verlag.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets, *Biometrika,* 85, 391 - 402.

Donoho, D.L. and Johnstone, I.M. (1994), Ideal spatial adaptation via wavelet shrinkage, *Biometrika* **81** 425-455.

Johnstone, I.M. and Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B.,* **59** 319 -351.

Kolaczyk, E.D. (1999b). Bayesian multiscale models for Poisson processes. *Journal of the American Statistical Association,* 94, 920-933.

Kolaczyk, E.D. and Nowak, R.D. (2000), "A multiresolution analysis for likelihoods: theory and methods." Submitted to *Annals of Statistics.*

Lauritzen, S.L. (1996), *Graphical Models,* New York: Oxford University Press.

Louie, M.M. and Kolaczyk, E.D. (2002). Multiscale spatial process models. Submitted to the *Journal of the American Statistical Association.*

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. PAMI* **11** 674 - 693.

Murphy, R.J., Ramaty, R., Reames, D.V., and Kozlovsky, B. (1991). Solar abundances from gamma-ray spectroscopy – Comparisons with energetic particle, photospheric, and coronal abundances. *The Astrophysical Journal,* **371** 793-803.

Norris, J.P. *et al.* (1996), "Attributes of pulses in long bright gamma-ray bursts," *The Astrophysical Journal,* 459, 393 - 412.

Nowak, R.D. (1999). Multiscale hidden Markov models for Bayesian image analysis. In *Bayesian Inference in Wavelet-Based Models,* Müller, P. and Vidakovic, B. (eds.). New York: Springer-Verlag.

Nowak, R.D. and Kolaczyk, E.D. (2000). A statistical multiscale framework for Poisson inverse problems. *IEEE Transactions on Information Theory* **46:5** 1811-1825.

Starck, J.L., Murtagh, F. and Bijaoui, A. (1998). *Image Processing and Data Analysis: The Multiscale Approach.* Cambridge: Cambridge University Press.

Timmermann, K. and Nowak, R. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory,* 45, 846-862.

7

NASA's Astrophysics Data Environment

Joseph H. Bredekamp[1] **and Daniel A. Golombek**

ABSTRACT NASA has a comprehensive space science data management program to assure that science data assets acquired from space missions are expediently available and utilized by scientists, educators, and the general public. This paper will discuss the guiding principles and approach for space science data archives, and describe the current landscape of astronomical data centers and services. It will conclude with prospects and opportunities to mine and exploit the emerging collective "digital sky" in all wavelengths for new scientific discoveries.

7.1 Space science data management

NASA's Office of Space Science (OSS) is committed to the preservation and utilization of data assets acquired from its space flight missions. Space science data are "open" resources as they ultimately belong to the research community and public at large, and not to individual investigators or instrument builders. OSS strives to provide a coherent and coordinated space science data environment to maximize the quality, accessibility, and usability of the vast space science data holdings for scientists, educators, and the general public.

Data archive and dissemination systems have been established for the major space science disciplines, guided by the principle of putting the data holdings under the jurisdiction of active science users in order to provide science "wrap- around" expertise and guidance.

The interface and flow of science data products from space flight projects and experiments to the appropriate discipline data archive system is included in a Project Data Management Plan developed by each project at its onset to address all aspects of data handling through the mission life cycle. Data management continues to be addressed as a key topic throughout the implementation and operation phases of the project. Indeed, science

[1]Office of Space Science, NASA Headquarters

productivity, along with timely delivery of science data to archives for open access by the community are two key evaluation criteria for determining priority for continuing the operation of on-going missions.

The coherent data environment that OSS strives toward is much more than access to science data assets only, but rather access to the data along with all ancillary information, software, tools, and capabilities to locate, retrieve, and analyze the data and convert it to meaningful information leading to scientific insight. That environment could thus be more accurately described as "data, computing, tools" organic infrastructure to support the scientific research endeavor. Evolving that robust infrastructure requires a significant investment in a wide range of computer science and technology, ranging from standards, interoperability, and commonality issues, to database and storage technologies, computational methods and algorithms, grid technologies, collaborative tools, etc.

7.2 Current astrophysics data landscape

The astrophysics data environment represents perhaps the fullest realization of the OSS science data management philosophy and approach. There is a long history of open archives and sharing in that community. Much of the current structure can be traced to the 1987 Astrophysics Data System Workshops. The concept of Science Archive Research Centers (SARC's) for astrophysics sub- disciplines organized by wavelength regimes was one of the recommendations coming out of these workshops and incrementally implemented by OSS. New software tools, research aids and services, and other advanced technology capabilities have been developed and infused over the years, many of them sponsored by the Applied Information Systems Research Program and/or Astrophysics Data Program open solicitations. There have been significant advances over the past several years in terms of federation, coordination, interoperability, and sharing across the various elements of the system, and the system is poised for the next level toward the concept of a seamless "digital sky".

The principal elements of the astrophysics data architecture are mission science centers, data archive centers, integrating information services, and the permanent archive. The relationships between these, and the user communities are depicted in Figure 1.

7.2.1 Mission science centers

The mission science centers are generally responsible for all phases of a missions science operations, from overseeing the peer-reviewed proposal selection process, to the execution of the observations, to the calibration of the data, and ultimately, to the dissemination of the data to the professional

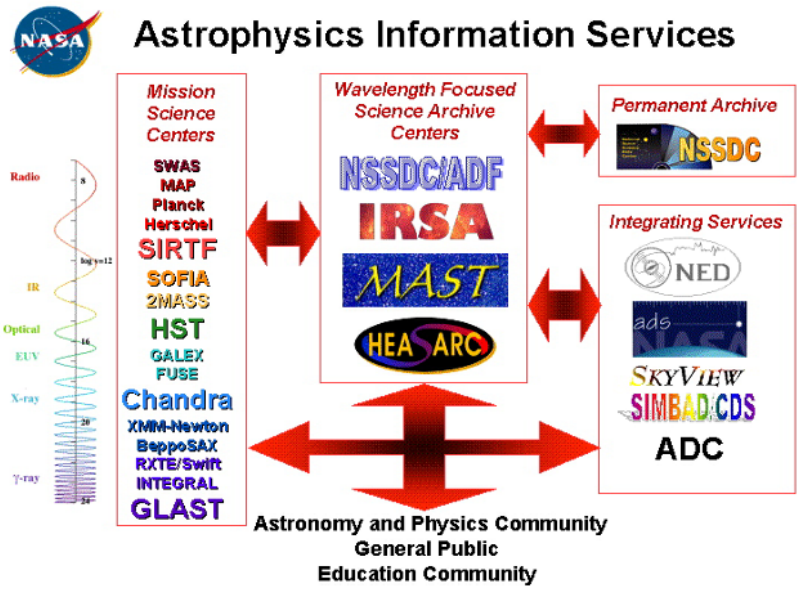


FIGURE 7.1. NASA's Astrophysics Data and Information Services

community (via high-capacity science archive facilities) and the general public. The staff of a NASA mission science center will typically consist of astronomers, technicians, software engineers, administrators, and educators. Mission centers may also manage Guest Observer grant programs, sponsor postdoctoral fellowship programs, develop science data analysis software, and host visiting astronomers from around the world. All science mission centers have a public outreach office that assures that the publics interest in astronomy is regularly rewarded with the latest images and results from the observatories. As part of these outreach efforts, a very successful, by its use and the number of students it reaches, education program is also conducted at these centers.

Space Telescope Science Institute (STScI): STScI is the science center for the Hubble Space Telescope mission (`http://www.stsci.edu`). The HST was the first of NASA's Great Observatories and was launched in April 1990. The Institute was established in 1981 and is located in Baltimore, MD. In addition to the services expected of a NASA mission science center, STScI hosts an annual symposium dedicated to HST-based research, manages the prestigious Hubble Fellow Program, and supports a vigorous research staff. The HST archive presently contains over 7 TB of data and is growing with 100 new science exposures every day. The next HST servicing mission will see the installation of the Advanced Camera for Surveys and the re-activation of the Near Infrared Camera and Multi-Object Spectrograph. STScI has been se-

lected as the science center for the Next Generation Space Telescope - a 6 to 8 meter class observatory, slated for launch within the decade, with instrumentation in the 0.65 micron wavelength range.

Chandra X-ray Center
(CXC): The Chandra X-Ray Observatory is the latest of NASA's Great Observatories: a high-resolution imaging and spectrographic telescope operating in the X-ray part of the electro-magnetic spectrum. Chandra was launched on July 23, 1999. The Chandra Data Archive (CDA) is part of the Chandra X-Ray Observatory Science Center (CXC; `http://cxc.harvard.edu`) which is operated for NASA by the Smithsonian Astrophysical Observatory in Cambridge, MA. The current holdings of the CDA amount to approximately 3 million data products with a total volume of 1 TB, in addition to an extensive collection of databases that hold mission information and metadata on the data products. The Chandra archive volume is expected to expand by almost 1 TB per year of active mission.

SIRTF Science Center
The Space InfraRed Telescope Facility (SIRTF) is the fourth and final element in NASA's family of Great Observatories. SIRTF consists of a 0.85-meter telescope and three cryogenically cooled science instruments capable of performing imaging and spectroscopy in the 3-180 micron wavelength range. Incorporating the latest in large-format infrared detector arrays, SIRTF offers orders-of-magnitude improvements in capability over existing programs. While SIRTF's mission lifetime requirement remains 2.5 years, recent programmatic and engineering developments have brought a 5-year cryogenic mission within reach. A fast-track development schedule will lead to a launch in July 2002. The SIRTF Science Center (`http://sirtf.caltech.edu`) is co-located with the Infrared Processing and Analysis Center (IPAC) on the campus of the California Institute of Technology.

7.2.2 Science archive centers

Besides the mission-specific centers listed above NASA also hosts wavelength-specific data archive centers. All these centers not only provide the data, but also software tools for its reduction and analysis, documentation and expert assistance to the user both to the professional astronomer as well as to educators and students or the public at large

Infrared Science Archive
(IRSA - `http://irsa.ipac.caltech.edu`) is located at the IPAC at Caltech and houses all the infrared and submillimeter data obtained by NASA-supported missions. The extracted source catalogs, images and spectra are available from the Infrared Space Observatory (ISO), the Two Micron All-Sky Survey (2MASS), the Midcourse Space Experiment (MSX), and the Infrared Astronomical Satellite (IRAS) missions. IRSA will also host the science data archives for the SIRTF and Stratospheric Observatory for Infrared Astronomy (SOFIA) missions when they become operational.

Multi-mission Archive at STScI
(MAST - `http://archive.stsci.edu/index.html`) hosts the collection of optical and UV datasets and catalogs from past and present NASA missions. In addition to HST data, it includes data from the International Ultraviolet Explorer (IUE), Far Ultraviolet Explorer (FUSE), Copernicus, three ASTRO and ORFEUS missions, , the Digitized Sky Survey, and the VLA FIRST survey. HSTs Guide Star Catalog (GSC) can be queried from this site as well. Once released, the Sloan Digital Sky Survey (SDSS) images, spectra and catalogs will be available from the MAST.

High Energy Astronomy Science Archive
(HEASARC - `http://heasarc.gsfc.nasa.gov/`) is located at NASA's Goddard Space Flight Center (GSFC) and includes all gamma-ray, X-ray, and extreme ultraviolet observations of cosmic (non-solar) sources obtained by currenty operating and past NASA-supported missions. The data available include those obtained from the Compton Gamma Ray Observatory (NASA's second Great Observatory which was decommissioned in 2000)), the Rossi X-Ray Timing Experiment (RXTE), Roentgen Satellite (ROSAT), Extreme Ultraviolet Explorer (EUVE), Advanced Satellite for Cosmology and Astrophysics (ASCA), BeppoSAX, and the X-Ray Multi-Mirror (XMM) missions. HEASARC provides a very large volume of multi-mission software tools such as the HEAsoft package as well as SkyView and AstroBrowse tools to search for and obtain multi-wavelength images of the sky.

7.2.3 Integrating services

To complement the data archives, and to facilitate an even easier dissemination of the science results, NASA supports several catalog, bibliographic, and thematic information services.

Astronomical Data Center
(ADC - `http://adc.gsfc.nasa.gov`) is located within the National Space Science Data Center at NASA/GSFC and is the custodian of the many hundreds of standard catalogs that astronomers use in support of their research. The ADC has developed significant search, access, and cross-correlation software tools (e.g., IMPReSS, CatsEye, Viewer). ADC has played a lead role in the application of XML (eXtensible Markup Language) technology to NASA's needs in astrophysics data management and has, in particular, developed XML-based tools for the automated ingestion of catalogs and tables and for facilitating the retrievability of their contents.

NASA Extragalactic Database
(NED - `http://nedwww.ipac.caltech.edu`) is hosted at the IPAC/Caltech and provides combined bibliographic and database services. It provides a thematic view of extragalactic astronomy and contains positions, name resolution, basic data, and bibliographic references for more than four million extragalactic objects. NED also includes almost 4 million photometric

measurements, more than three million position measurements, more than 200,000 redshift and radial velocity measurements. Finally, to complement this impressive catalog, more than 700,000 images from 2MASS and DSS (generated on-the-fly) are available.

Astrophysics Data System
(ADS- `http://adsabs.harvard.edu/abstract_service.html`)
is an abstract service that includes almost 700,000 abstracts from all the major astrophysics journals and conference series with links to the whole paper. It also includes instrumentation, physics and geophysics abstracts as well as links to the Astrophysics Preprint server. ADS can be searched by author, title, words in the abstract or object name.

Centre Données de astronomiques de Strasbourg
(CDS - `http://cdsweb.u-strasbg.fr`) is located in Strasbourg, France and is a notable international partner for astrophysics information services. The Set of Identifications, Measurements, and Bibliography for Astronomical Data Basic (SIMBAD) is mirrored in the US at `http://simbad.harvard.edu/Simbad`. SIMBAD includes almost 3 million galactic objects whose characteristics and bibliography can be searched from almost 8 million identifiers or by their positions.

7.2.4 Permanent archive

The *National Space Science Data Center*
(NSSDC) at the Goddard Space Flight Center serves as the permanent data archive for all space science disciplines, including astrophysics, space physics, solar physics, and planetary science. The NSSDC also provides other multidisciplinary services such as master data catalogs and information services, standards support, and photographic resources.

7.2.5 Productivity and interoperability

The astrophysics data centers and services are heavily utilized and productive in contributing to new research results. The concept of "archival science" has grown in popularity in recent years to where existing archived data is used for investigations different from the originally proposed investigations, or combined for interdisciplinary investigations, assimilated into theoretical models, etc. The archives and information services have also proved to be very valuable tools and resources for observation planning and analysis for operating missions as well.

As an example of archive utilization, the daily ingest and retrieval rates for HST data are shown in Figure 3. Note that retrieval rates now far exceed the ingest rate. The average retrieval rate of 15 GB/day for HST data is typically a factor of 4 larger than the ingest rate. Similarly across all the data centers and services volumes of data are growing rapidly, and expected

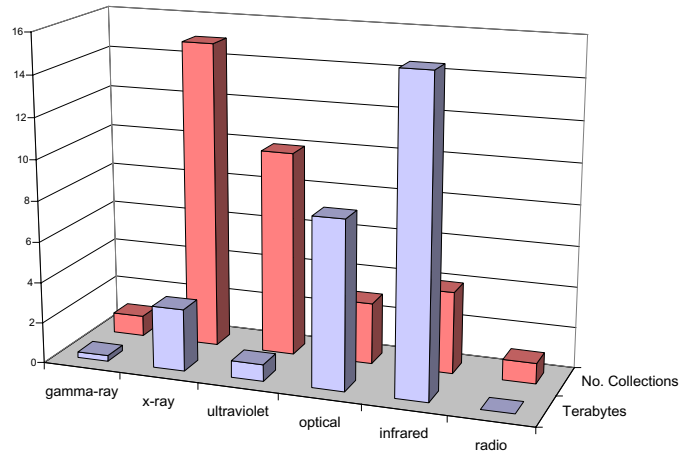


FIGURE 7.2. Content of the NASA Archives and Data Centers

to continue to grow dramatically into the future. And the utilization of those volumes of data are growing at even a faster rate.

Another productivity measure is the number of scientific publications that have resulted from the use of the data archives. Figure 4 provides the number of publications based on data within the MAST and HEASARC archives for the period 1999-2000. It is estimated that approximately 4000 scientific papers per year are based, at least in part on data and information services within the astrophysics system.

The astrophysics data and information services operate as a federation to improve overall productivity and efficiency, and enhance interoperability and interdisciplinary access to data assets and services. There is strong coordination and collaboration across the various elements to plan and evolve an integrated system with the goal to afford users a "world view" of consistent interfaces and paths to allow data discovery and exploration as a whole.

A common front-end user interface which would provide a searchable web- based browser among the various data centers is a critical element for such interoperability. "AstroBrowse" is such an interface layer and was conceived by R. Hanisch (STScI) and S. Murray (SAO) and implemented as a prototype at HEASARC, MAST, and CDS. Other efforts are also proceeding across the broad consortium of astrophysics data providers to build upon and extend these discovery capabilities to include fuller functionality for users to locate, retrieve, and correlate data resources.

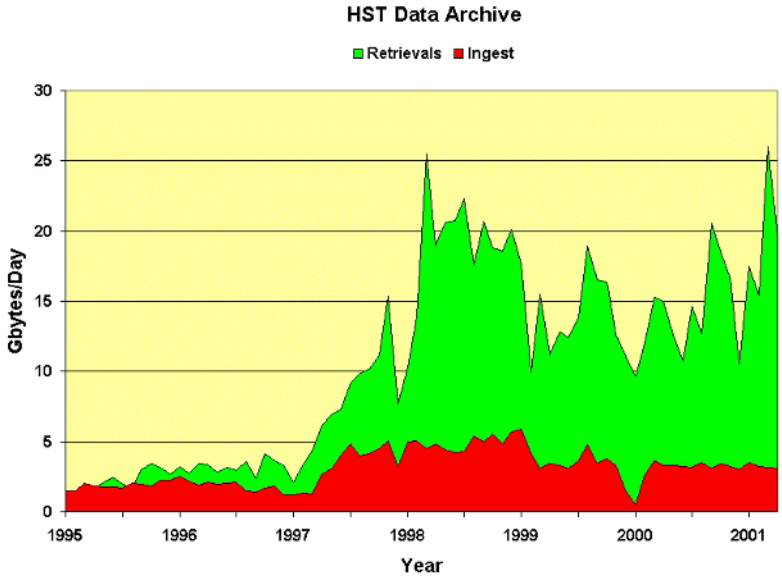


FIGURE 7.3. Usage of the HST Data Archive

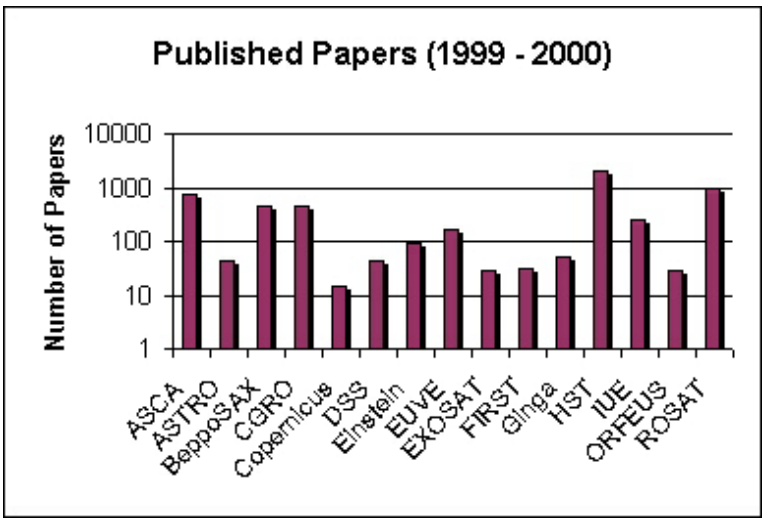


FIGURE 7.4. Number of publications based on data within HEASARC and MAST

7.3 Emerging prospects and opportunities

The trend in data volumes and complexity will only increase into the future as we look deeper and with finer resolution into the sky. Refinements in measuring fluctuations and anisotropies in the cosmic microwave background with data from missions such as the recently launched Microwave Anisotropy Probe and the future European Space Agency Planck mission have exciting prospects for unlocking the structure of the early universe and more precise estimation of cosmological parameters.

The collective set of digital sky data, both space- and ground-based spanning the entire electromagnetic spectrum has enormous potential for data mining and exploration. The resulting "digital sky" is now within the venue of observational astronomy, albeit not altogether easy, straightforward, and transparent. The basic technologies are in hand to exploit the data archive as a whole, but to realize the full and enormous potential for scientific discovery calls for significant advances in our current frameworks, both technological and scientific. These challenges provide an opportunity to drive productive interdisciplinary partnerships and collaborations involving space scientists, computer scientists and technologists , mathematicians and statisticians.

7.4 Summary

NASA's Office of Space Science supports a vigorous and robust system of data and information services which are heavily used by the world-wide community. This infrastructure enhances the productivity of the research endeavor, as well as extending utilization to benefit educators and the public. The data archive and information services are poised for the next challenge to exploit the collective and seamless "digital sky", and to engage the broad range of requisite partnerships involving astronomers and astrophysicists, computer scientists and technologists, mathematicians, and statisticians, as well as with international collaborators around the world to meet the challenge. References

7.5 Appendix A: URL listings

Office of Space Science
NASA Office of Space Science - `http://spacescience.nasa.gov/`
Space Science Missions - `http://spacescience.nasa.gov/missions/`

Mission-specific Archives
Chandra X-ray Center - `http://cxc.harvard.edu`

SIRTF Science Center - http://sirtf.caltech.edu
Space Telescope Science - Institute - http://www.stsci.edu

Wavelength-specific Archives
High Energy - http://heasarc.gsfc.nasa.gov
Infrared/Sub-mm - http://irsa.ipac.caltech.edu
Optical /UV - http://archive.stsci.edu/index.html

Integrated Services
ADS - http://adsabs.harvard.edu/abstract_service.html
NED - http://nedwww.ipac.caltech.edu
NSSDC - http://nssdc.gsfc.nasa.gov
SIMBAD - http://simbad.harvard.edu/Simbad

8

Statistical and Astronomical Challenges in the Sloan Digital Sky Survey

Michael A. Strauss[1]

ABSTRACT The Sloan Digital Sky Survey is an ambitious effort to map the entire Northern sky at high Galactic latitudes, using modern CCD cameras to take images in five photometric bands, and a pair of multi-object spectrographs to measure redshifts for 10^6 galaxies and 10^5 quasars. I describe some of the recent scientific results from the survey, focusing on quasars and galaxies, with an emphasis on the statistical challenges that they raise. The data are very rich, with potential impact on a large variety of astronomical problems, but most analyses to date have been carried out using rather unsophisticated statistical tools. These data are thus ideal to foster collaboration between astronomers and statisticians.
This paper is followed by a commentary by statistician David A. van Dyk.

8.1 Introduction

Astronomy is traditionally done by individuals or small groups of astronomers, who use their handful of telescope nights a year to carry out focussed projects. However, we need massive datasets gathered uniformly, on a scale much larger than any small group of workers could collect in the traditional mode, to answer the big questions which currently face astronomy: How did the first objects form after the Big Bang? What is the distribution of galaxies on the largest scales? What is the full range of properties of galaxies and stars, and what are the relationships between them? The Sloan Digital Sky Survey (SDSS) addresses this need. It uses a dedicated 2.5-meter telescope at Apache Point, New Mexico, with a wide-field CCD imaging camera which operates in drift-scan mode, taking images of 20 square degrees of sky per hour in five broad photometric bands (u, g, r, i and z) covering the wavelength range accessible to CCDs from the ground. These data are reduced by a series of interconnected software pipelines; from the resulting lists of detected objects, the brightest galaxies and quasars are chosen to

[1]Princeton University Observatory

be observed by a pair of fiber-fed multi-object double spectrographs, which obtain spectra of 640 objects at a time. The hardware of the project is summarized by York et al. (2000), while Stoughton et al. (2001) discuss the outputs of the pipelines. To date, the survey has imaged roughly 2000 of the planned 10,000 square degrees of sky, and has obtained spectra of 200,000 objects. The first of these data are now public, and can be accessed on the web from links off the project web site, `http://www.sdss.org`.

The scientific goals of the project are focussed on the large-scale distribution of galaxies: the survey was designed to obtain as uniform as possible a sample of galaxies for which to measure spectra and therefore redshifts, thus obtaining a three-dimensional map of the distribution of galaxies. However, the data are richer than this scientific goal alone would imply: in the data obtained thus far, there are over 5×10^7 detected objects with five-color photometry, allowing investigations of the nature of galaxies, quasars, stars, the structure of our Milky Way, asteroids, and many other exciting areas of astronomy.

In this paper, I will outline some of the recent scientific developments in SDSS (with some emphasis on work in which I personally have been involved), describing some of the interesting statistical questions that the data raise. The statistically astute reader will notice that for the most part, we have not been using the most modern and powerful statistical methods for our analyses (although see Bob Nichol's contribution to these proceedings for a welcome exception to this!); the message is that these data are rich enough to allow far more sophisticated analyses on a variety of scientific problems. Thus this is fertile ground for close collaboration between astronomers and statisticians.

8.2 Stellar photometry: Statistical challenges in data reduction

There exist a number of software packages in astronomy for analyzing CCD images, such as IRAF, FOCAS (Jarvis & Tyson 1981), Sextractor (Bertin & Arnouts 1996), VISTA, and others. Our collaboration has worked more or less from scratch in developing our image reduction software (Lupton et al. 2001; Stoughton et al. 2001). The goal is to reliably find and measure the properties of all statistically significant objects in the imaging data, self-consistently in the five photometric bands. As we observe through the Earth's atmosphere, the light from point sources is smeared to a disk of diameter typically $1'' - 1.5''$. Thus our first challenge is to characterize this smearing, or Point Spread Function (PSF) accurately. This PSF has a non-trivial shape, roughly described by the sum of two Gaussians, plus a power-law tail. Moreover, it can vary appreciably on the scale of arcminutes, due to the optics of the telescope and camera, and changing conditions in the

atmosphere. Our approach (Lupton et al. 2001), which seems to work fairly well, is to expand the measured PSF as a function of position in Karhunen-Loève eigenmodes, and then fit the coefficients to low-order polynomials in position. With an accurate model of the PSF, one can then measure the properties of detected objects quite well. In particular, for point sources, this allows an optimal measurement of the total flux of the object (once aperture corrections are applied; see Stoughton et al. 2001).

Galaxies are not point sources, but have radial profiles that can often be characterized by an exponential ($\exp[-r/r_0]$), or the mathematically awkward $r^{1/4}$ law, $\exp[-(r/r_0)^{1/4}]$. The software fits every object to these two models, convolved with the PSF and allowing for arbitrary ellipticity and orientation. The difference between this so-called *model magnitude* and the PSF magnitude turns out to be a powerful measure of extendedness in the images (this method can be extended using a Bayesian approach, knowing the relative numbers of stars and galaxies as a function of magnitude; see Scranton et al. 2001).

One measure of the accuracy of the resulting photometry can be found in the distribution of colors of stars. Stellar colors are determined to first order solely by their surface temperature, so (ordinary) stars lie on a one-dimensional locus in the four-dimensional color space spanned by our filters ($u-g, g-r, r-i, i-z$; Newberg & Yanny 1997; Fan 1999; Finlator et al. 2000). Figure 8.1 illustrates this; notice the thinness of the stellar locus, and the relatively small number of outliers. The errors are close to those expected from photon statistics; in particular, we have been successful in recognizing, flagging, and in some cases, correcting, a wide variety of systematic errors would would cause stars to scatter from the locus: cosmic rays, bad columns on the CCDs, bleed trails and diffraction spikes from saturated stars, overlapping images, and so on.

Incidentally, a full statistical characterization of the stellar locus has much to teach us about different stellar populations and the structure of our Galaxy (one is looking at various populations of stars, each with its own spatial distribution, as one looks in different directions, due to the different contributions of thin disk, thick disk, and halo). A first stab at modelling Galactic structure with SDSS data was carried out by Chen et al. (2001). The Galactic halo turns out not to be smooth, but to show substructure (Ivezić et al. 2000; Yanny et al. 2000; Newberg et al. 2001), which is believed to be due to the cannibalization of dwarf galaxies by the Milky Way. Thus far, this substructure has been seen by simply plotting the distribution of stars in color-magnitude-position space, and looking for overdensities by eye; this is an area ripe for a more sophisticated statistical treatment.

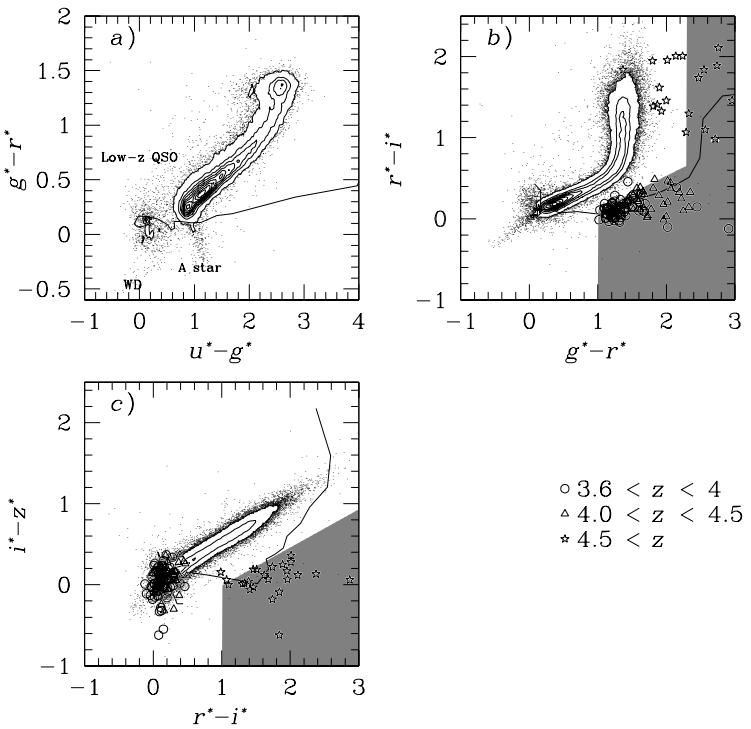


FIGURE 8.1. A series of projections of the stellar locus in SDSS color space. Stars brighter than $i^* = 21$ are shown. Also shown is the distribution of high-redshift quasars, the predicted colors of quasars as a function of redshift (line), and the region of color space in which high-redshift quasars are selected (shaded region).

8.3 Finding high-redshift quasars

If ordinary stars fall along a one-dimensional sequence in color space, objects which do *not* lie on this sequence are inevitably interesting. The most numerous class of such objects are the quasars. Quasars are the very luminous nuclei of galaxies which include supermassive black holes into which material is streaming; this material is heated up so much by viscosity that it can outshine its parent galaxy by orders of magnitude. The SDSS is finding these objects in great numbers; they have intrinsically bluer spectra than do stars, and thus are easy to pick out from their distinctly blue colors. Indeed, quasars are selected for spectroscopic follow-up with the SDSS by a conceptually simple algorithm that characterizes the stellar locus as a 1-dimensional sausage in color space (Newberg & Yanny 1997); a quasar candidate is anything that falls far from this locus. See Bob Nichol's contribution to these proceedings for a rather different approach to this problem, based on the mixture model.

At high redshift, the observed colors of quasars change systematically. Neutral hydrogen along the line of sight to the quasar systematically absorbs blue light, causing the quasar to appear red; the higher the redshift, the greater the reddening of the quasar. This effect is shown schematically in Figure 8.1; the thin line is a model for the median color of quasars at ever-increasing redshift. High-redshift quasars are intrinsically interesting, because they are observed at an epoch when the universe was quite a bit younger than it is today. Thus we have been carrying out a survey of the very reddest objects in the SDSS imaging database, looking for the very highest redshift objects (which turn out to be very rare; the quasar number density drops off dramatically at redshifts above 3 or so; Fan et al. 2001a). Figure 8.1 shows the colors of some of the quasars we've found at redshift $z > 3.6$; we now have discovered over 200 such objects (Anderson et al. 2001 and references therein), by far the largest sample of high-redshift quasars that exists. Again, it is worth emphasizing that this is successful because we've managed to keep systematic errors down to a manageable level, such that outliers in color-color space are there for astrophysical reasons, not due to glitches in the data. Given that the parent sample from which these quasars were selected contains many million stars, this is a non-trivial statement.

This technique works well to redshifts somewhat larger than 5. To go still further requires more work. For $z \geq 5.8$, a quasar is so red that it is likely to be detected only in the z band (not to be confused with the symbol for redshift!), our longest-wavelength band at 9000 A. These objects are very rare, and we are dominated by systematics in trying to find them. In particular, most cosmic rays (high-energy particles which CCD detectors are quite effective at detecting) are recognized by the fact that they are confined to a single pixel, thus looking quite different from the PSF. However, occasionally, cosmic rays splatter over several pixels, and can mimic

stars. If they hit the z band chips, they will thus appear as a z-band only object, and thus are candidates for high-redshift quasars.

Even in the absence of glitches like this, selecting the very reddest objects (as measured by the ratio of the fluxes in the z and i bands) will preferentially pick up the many sigma positive tail of errors in z. That is, if an object's intrinsic magnitude in z is 20.2 with an estimated error of 0.1, a 4-sigma event (which happens a non-negligible amount even with Gaussian statistics, when one has a parent sample of 10^7 objects!) will make this appear to be at 19.8 magnitude, and thus much redder than it really is. Finally, there is an interesting astrophysical contaminant to our red objects, namely extremely cool stars (brown dwarfs; Leggett et al. 2001), with surface temperatures of order 1000 K.

With all of these effects acting, we have had to do a tremendous amount of sifting to find real quasars (Fan et al. 2001b). We made our rejection of cosmic ray events much more stringent than under normal processing, throwing out any object that showed any hint of being a cosmic ray. Eyeballing the remaining images rejected many more candidates. We then obtained follow-up images in z to determine if the object was really there (many cosmic rays had still survived all this winnowing), and to check the photometric measurements. Finally, observations at yet longer wavelengths, at J ($1.3\,\mu$m) allowed us to distinguish quasars from brown dwarfs. When all was said and done, we were left with four objects (from a parent sample of $> 10^7$, selected over 1500 square degrees of sky), every one of which was a high-redshift quasar. Indeed, these four objects are the most distant quasars known, with redshifts of 5.74, 5.82, 5.99, and 6.28, respectively. For standard cosmological models, the highest-redshift object is observed only 900 million years after the Big Bang; thus we're looking back 94% of the age of the universe. I mention in passing that the $z = 6.28$ quasar shows evidence in its spectrum of the Gunn-Peterson (1965) effect, due to neutral hydrogen in the intergalactic medium between the quasar and us (Becker et al. 2001; see also Djorgovski et al. 2001). This is evidence that we're probing to an epoch before substantial numbers of stars and quasars formed: stars and quasars emit ultraviolet photons which ionize the intergalactic medium.

The mere presence of very luminous quasars so soon after the Big Bang is a challenge to cosmological models, and to statistics as well. One can estimate a lower limit to the mass of the black hole powering a quasar from its luminosity (based on an argument originally due to Arthur Eddington, that says that gravity has to be stronger than radiation pressure to allow material to fall in); the quasars we're observing all have inferred black-hole masses of order a few billion times the mass of the Sun.

However, the universe was very close to homogeneous (to a part in 10^5) soon after the Big Bang, as we observe directly from the smoothness of the Cosmic Microwave Background. The present-day structure of the universe, from individual galaxies to the largest walls and voids in the galaxy distri-

bution, is believed to have grown from these 10^{-5} fluctuations via the process of gravitational instability. In modern inflationary models for the Big Bang, these fluctuations arose from quantum fluctuations, with a Gaussian distribution by the Central Limit Theorem (Peacock 1999). Astronomers use this fact to estimate the number of virialized structures of a given scale at a given redshift, essentially by asking for the fraction of space that is overdense by a certain amount, given the Gaussian distribution (Press & Schechter 1974). In these calculations, the high-redshift quasars inevitably are interpreted as many-sigma fluctuations, which requires that we believe that the extreme tail of the distribution is accurately Gaussian (cf., the discussion in Chiu et al. 1998; Willick 2000). It is an interesting statistical question to ask about the validity of the Central Limit Theorem in predicting the extreme tail of the distribution under these circumstances (see the comment by Licia Verde at the end of this paper).

8.4 Describing the manifold of galaxy properties

High-redshift quasars are among the rarest objects in the SDSS database. But there is a variety of interesting scientific and statistical questions that arise from the more common objects, such as galaxies. We wish to describe the properties of galaxies, with the goal of understanding the physical processes that give rise to the observational properties that they have, and what the correlations between these properties are. Among the salient properties of galaxies, one might list their luminosity, color (as measured in several bands), extent, ellipticity, asymmetry (i.e., deviations from elliptical isophotes), their internal velocity dispersion, their surface brightness profile, their morphology (relative strength of bulge and disk, strength and number of spiral arms, etc.), the strength of emission and absorption lines in the spectra, and the large-scale environment in which these galaxies find themselves (i.e., are they in clusters? Walls? Filaments? Voids?). This is a rather complicated multi-parameter space, and we wish to understand the physical relationships between all these quantities. There is a fair amount of empirical knowledge in the literature, much of it looking at these various quantities two at a time: for example, bluer galaxies tend to be of lower surface brightness, and the internal velocity dispersion of an elliptical galaxy is correlated with its luminosity and size (Djorgovski & Davis 1987; Dressler et al. 1987; see Bernardi et al. 2001 for a spectacular demonstration of this with 10,000 SDSS galaxies).

We are still struggling with ways to explore this manifold in its full glory, however, and most analyses in the literature get no more sophisticated than using Principle Component Analysis, which will not address questions such as possible curvature in any relations found between parameters, and whether galaxies naturally divide up into distinct classes (cf., Kormendy

& Djorgovski 1989; Strateva et al. 2001). The SDSS data again are very rich; we already have high-quality spectroscopy for over 10^5 galaxies with five-band images, and images alone for literally millions more, and thus demand more powerful statistical tools to analyze them. This is a field in which astronomers not only do not have sufficient statistical tools to tackle these data, but do not yet know what the proper astronomical questions to ask are; we simply haven't explored the data in enough detail to formulate the questions properly.

One of the galaxy attributes mentioned above was their large-scale environment. We have known for two decades that the galaxy distribution shows a rich array of structures, the cosmic web (see Rien van de Weygaert's contribution to these proceedings). We do see correlations between the nature of galaxies and where they find themselves with respect to this web; the best-known of these correlations states that elliptical galaxies are preferentially found in clusters of galaxies (the morphology-density relation; Dressler 1980). There are also correlations found between the clustering strength of galaxies and their color, surface brightness, and luminosity (in the sense that low-luminosity, blue, low surface brightness galaxies are somewhat more weakly clustered). However, all these attributes are correlated with one another. For example, elliptical galaxies tend to be red and of relatively high surface brightness, thus it isn't known whether the fact that blue galaxies are more weakly clustered is just a manifestation of the morphology-density relation, or whether it has a component independent of that. Astronomers have not yet had sufficiently voluminous data (until now) to address this question, and even now, we struggle with the somewhat crude statistical techniques we have at our disposal to try to characterize clustering strength (see the contribution by Vicent Martínez in these proceedings).

Figure 8.2 shows the distribution of galaxies in the public release of the SDSS main galaxy sample, showing the now-familiar cosmic web. As mentioned earlier, astronomers believe that this structure arose from an initially Gaussian set of fluctuations (that is, the density field on any given smoothing scale has a Gaussian distribution, and the individual Fourier modes have random phases). In this picture, one gets a complete statistical distribution of the density field using two-point statistics, in particular, the power spectrum (see Zehavi et al. 2001 and Scranton et al. 2001 for first analyses of the SDSS data along these lines). However, as gravitational instability continues to work, structures become non-linear, and two-point statistics are no longer adequate for a full description of the galaxy density field (Martínez & Saar 2001). We are able to quantify this into the mildly non-linear regime using perturbation theory. We also have a heuristic model, based on a hierarchical clustering model, to describe the set of high-order correlation functions, in the very non-linear regime. And we have a variety of statistical tools, including Minkowski functionals, void statistics, Voronoi Tessellations, measurements of fractal dimensions, and

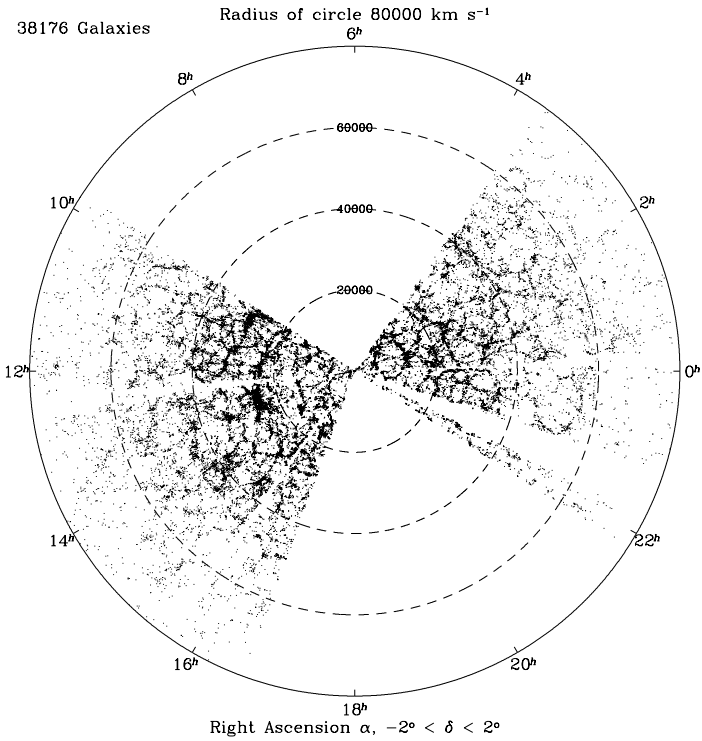


FIGURE 8.2. The distribution of galaxies from the Early Data Release (Stoughton et al. 2001) of the SDSS galaxy redshift survey. Most of these data were taken on the Celestial Equator ($\delta = 0$) in a narrow slice, so right ascension is plotted as the angular coordinate, and redshift as the radial coordinate.

so on, which attempt to give a handle on various aspects of the non-linear structures that we see. Unfortunately, the problem of how fully to describe statistically the beautiful structures that we see still evades us, and we are still only able to make a qualitative comparison of the observed galaxy distribution to that predicted in specific cosmological models. This problem is made more complicated yet by remembering that each of the points in Figure 8.2 is a galaxy, with its own morphology, luminosity, spectral properties, etc., and we wish to describe how its physical properties are related to the large-scale structure in which it is embedded. This is a great challenge for statisticians and astronomers alike, especially in the face of datasets like that of the SDSS.

Acknowledgments: The Sloan Digital Sky Survey (SDSS) is a joint project of The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Princeton University, the United States Naval Observatory, and the University of Washington. Apache Point Observatory, site of the SDSS telescopes, is operated by the Astrophysical Research Consortium (ARC). Funding for the project has been provided by the Alfred P. Sloan Foundation, the SDSS member institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is http://www.sdss.org. I wish to thank my colleagues on the SDSS for making these wonderful data possible. I acknowledge support from NSF grant AST-0071091.

8.5 References

[1] Anderson, S. et al. 2001, AJ, 122, 503

[2] Becker, R. et al. 2001, AJ, 122, 2850 (astro-ph/0108097)

[3] Bernardi, M. et al. 2001, AJ, submitted

[4] Bertin, E., & Arnouts, S. et al. 1996, A&AS, 117, 393

[5] Chen, B., et al. 2001, ApJ, 553, 184

[6] Chiu, Ostriker, J.P., & Strauss, M.A. 1998, ApJ, 494, 479

[7] Djorgovski, S., & Davis, M. 1987, ApJ, 313, 59

[8] Djorgovski, S. et al. 2001, ApJ, 560, L5

[9] Dressler, A. 1980, ApJ, 236, 351

[10] Dressler, A., Lynden-Bell, D., Burstein, D., Davies, R. L., Faber, S. M., Terlevich, R. J., & Wegner, G. 1987, ApJ, 313, 42

[11] Fan, X. 1999, AJ, 117, 2528

[12] Fan, X., Strauss, M. A., Richards, G. T., et al. 2001a, AJ, 121, 54

[13] Fan, X., Narayanan, V., Lupton, R.H., et al. 2001b, AJ, AJ, 122, 2933

[14] Finlator, K., Ivezić, Ž., Fan, X., et al. 2000, AJ, 120, 2615

[15] Gunn, J.E., & Peterson, B. 1965, ApJ, 142, 1633

[16] Ivezić, Ž., Goldston, J., Finlator, K., et al. 2000, AJ, 120, 963

[17] Jarvis, J.F., & Tyson, J.A. 1981, AJ, 86, 476

[18] Kormendy, J. & Djorgovski, S. 1989, ARA&A, 27, 235

[19] Leggett, S. et al. 2001, ApJ, 564, 452

[20] Lupton, R., Gunn, J. E., Ivezić, Z., Knapp, G. R., Kent, S., & Yasuda, N. 2001, in ASP Conf. Ser. 238, Astronomical Data Analysis Software and Systems X, ed. F. R. Harnden, Jr., F. A. Primini, and H. E. Payne (San Francisco: Astr. Soc. Pac.), 269 (astro-ph/0101420)

[21] Martínez, V.J. & Saar, E. 2001, Statistics of the Galaxy Distribution (Boca Raton: Chapman & Hall/CRC)

[22] Newberg, H.J., & Yanny, B. 1997, ApJS, 113, 89

[23] Newberg, H.J. et al. 2001, AJ, submitted

[24] Peacock, J.A. 1999, Cosmological Physics (Cambridge: Cambridge University Press)

[25] Press, W. H., & Schechter, P. 1974, ApJ, 187, 425

[26] Scranton, R., et al. 2001, ApJ, submitted (astro-ph/0107416)

[27] Stoughton, C. et al. 2001, AJ, submitted

[28] Strateva, I. et al. 2001, AJ, 122, 1861

[29] Willick, J.A. 2000, ApJ, 530, 80

[30] Yanny, B., Newberg, H. J., Kent, S., et al. 2000, ApJ, 540, 825

[31] York, D. G., Adelman, J., Anderson, J. E., et al. 2000, AJ, 120, 1579

[32] Zehavi, I., Blanton, M.R., Frieman, J.A. et al. 2001, ApJ, submitted (astro-ph/0106476)

Commentary by David A. van Dyk[2]

8.6 Data Mining in Space

Data Mining refers to methods for automatically or semi-automatically scouring a very large dataset for useful information; see Hand (1998) and Hand et al. (2000) for good reviews of the statistical perspective. Generally speaking data mining has a negative connotation to statisticians. The term conjures up images of automated methods trawling through large data sets looking for features or patterns with little regard to implicit multiple testing. Thus, the methods employed have little ability to distinguish chance fluctuations from real patterns or to uncover underlying structure in the data. Unfortunately, for my prepared comment, Michael Strauss, Bob Nichol, and others working on the Sloan Digital Sky Survey (SDSS) are clearly taking the statistical challenges of this mammoth data analysis project seriously. They are to be commended for their model-based approach which is clearly bearing fruit in the form of their impressive astronomical discoveries.

Hand (1998) identifies two basic components that characterize data mining in a wide range of applications, *modeling* and *pattern recognition*. Modeling involves looking for large scale structure in the data. In the context of the SDSS, this may include comparing the distribution of stars with galactic models, empirically characterizing the large scale structure of the universe, and classification of features (i.e., objects). Although, there are many standard statistical methods that are designed for such modeling tasks, many of the astronomical models which are relevant to SDSS are highly complex and do not fall into any standard statistical modeling framework. Nonetheless, statistical model building techniques and highly structured hierarchical models are potentially useful even in such complex settings. An example which outlines a model for the large scale structure of the universe is described below. The approach the SDSS scientists take to classification is to use model-based classification algorithms (e.g., fitting finite mixture models using the EM algorithm; Uribe et al. this volume). Such methods are useful not only in their ability to classify objects but also in their model-based approach which is designed to shed light on the mechanisms and structure underlying the classification. Again the SDSS group is to be commended for there emphasis on fast computational methods (e.g., kd-trees; see Nichol, this volume) that do not sacrifice the model-based methods.

Searching for local features or patterns in the data, i.e., feature detection, is the second standard task in data mining. This may include searching for

[2]Department of Statistics, Harvard University

faint objects, anomalous objects, or clusters (e.g., of galaxies). This can be an especially challenging task owing to errors in the data (e.g., contamination by background, cosmic events, and asteroids as well as measurement and data recording errors) and the inherent multiple testing problem. As described by Strauss and Nichol (both in this volume) the SDSS team is both taking great care in cleaning the data and developing new statistical methodology (the False Discovery Rate Procedure) for handling the multiple testing problem.

8.7 Modeling the Large-Scale Structure of the Universe

In this section I outline a model for the large-scale distribution of galaxies in the universe. This model is by no means meant to be a finished product—it is based on a most rudimentary understanding of the cosmic structure. Rather, I hope to illustrate how highly structured hierarchical models (van Dyk, this volume) can be used to model complex structure and the incompleteness in the observed data. Such a model can be fit directly to the observed data which leads to direct estimates of parameter uncertainties and standard methods for model adjustment and improvement (e.g., Protassov et al. 2002).

In the first levels of the model, we describe the large-scale structure itself. This could be done in a variety of ways (e.g, using Voroni tessellations as described by van de Weygaert and Icke, this volume). As a first step, three dimensional data visualization techniques (e.g., Cook, this volume) should improve our understanding of the structure and perhaps answer questions such as whether nodes are connected by filaments or walls (Strauss, this volume). We use standard statistical models that aim to describe two dimensional slices and projections of the galactic distribution.

Level 1: Nodes and Filaments. We might model the nodes as a three dimensional spatially inhomogeneous Poisson process, the nodes becoming more sparse with distance. Given the node locations, filaments connect pairs of nodes with the probability of a connecting filament decreasing as the distance between nodes increases.

Level 2: Galactic Locations Along Filaments. Galaxies are placed along the filaments according to a second inhomogeneous Poisson process with intensity increasing with proximity to the nodes.

Level 3: Distance and Direction from Filaments. Given the location along the filament the center of the galaxies are distributed according to a bivariate Gaussian or Lorentzian distribution.

Additional modeling of the distribution of galaxy type or other galactic specifications can easily be added to such a model.

The final two levels of the model account for the data collection process.

Level 4: Stochastic Censoring of Data. The likelihood that a particular galaxy is observed depends on its distance, direction (e.g., relative to our own galaxy), and magnitude as well as observation patterns. Such censoring can be modeled to account for the missing data.

Level 5: Errors in Variables. The distance to galaxies is generally measured with error bars which can easily be taken into account by such a model. If the distance is not observed (i.e., the spectrum is not observed/analyzed) the observed direction can still be accounted for by such a model.

Such a hierarchical model can be fit in a Bayesian paradigm via Markov chain Monte Carlo. Although this would be a demanding computational task the rewards could be great. Typically such complex systems are modeled using computer simulations which try to mimic patterns in the observed data. Unfortunately, error bars and model improvement techniques are not generally forthcoming. Fitting a model to the data in Bayesian setting yields not only (model-dependent) error bars on fitted parameters but also ready methods to check the model which offer advice as to how to improve the model which can then be refitted and rechecked.

8.8 References

[1] Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician* **52**, 112–118.

[2] Hand, D. J., Gordon, B., Kelly, M. G., and Adams, N. M. (2000). Data mining for fun and profit (with discussion). *Statistical Science* **15**, 111–131.

[3] Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2002). Statistics: Handle with care – detecting multiple model components with the likelihood ratio test. *The Astrophysical Journal* to appear.

9

Challenges for Cluster Analysis in a Virtual Observatory

S. G. Djorgovski[1], R. Brunner, A. Mahabal, R. Williams, R. Granat and P. Stolorz

ABSTRACT There has been an unprecedented and continuing growth in the volume, quality, and complexity of astronomical data sets over the past few years, mainly through large digital sky surveys. Virtual Observatory (VO) concept represents a scientific and technological framework needed to cope with this data flood. We review some of the applied statistics and computing challenges posed by the analysis of large and complex data sets expected in the VO-based research. The challenges are driven both by the size and the complexity of the data sets (billions of data vectors in parameter spaces of tens or hundreds of dimensions), by the heterogeneity of the data and measurement errors, the selection effects and censored data, and by the intrinsic clustering properties (functional form, topology) of the data distribution in the parameter space of observed attributes.
Examples of scientific questions one may wish to address include: objective determination of the numbers of object classes present in the data, and the membership probabilities for each source; searches for unusual, rare, or even new types of objects and phenomena; discovery of physically interesting multivariate correlations which may be present in some of the clusters; etc. This paper is followed by a commentary by statistician Dianne Cook.

9.1 Towards a Virtual Observatory

Observational astronomy is undergoing a paradigm shift. This revolutionary change is driven by the enormous technological advances in telescopes and detectors (e.g., large digital arrays), the exponential increase in computing capabilities, and the fundamental changes in the observing strategies used to gather the data. In the past, the usual mode of observational astronomy was that of a single astronomer or small group performing observations of a small number of objects (from single objects and up to some hundreds of objects). This is now changing: large digital sky surveys over a range of wavelengths, from radio to x-rays, from space and ground are becoming the

[1]Palomar Observatory, Caltech

dominant source of observational data. Data-mining of the resulting digital sky archives is becoming a major venue of the observational astronomy. The optimal use of the large ground-based telescopes and space observatories is now as a follow-up of sources selected from large sky surveys. This trend is bound to continue, as the data volumes and data complexity increase. The very nature of the observational astronomy is thus changing rapidly. See, e.g., Szalay & Gray (2001) for a review.

The existing surveys already contain many Terabytes of data, from which catalogs of many millions, or even billions of objects are extracted. For each object, some tens or even hundred parameters are measured, most (but not all) with quantifiable errors. Forthcoming projects and sky surveys are expected to deliver data volumes measured in Petabytes. For example, a major new area for exploration will be in the time domain, with a number of ongoing or forthcoming surveys aiming to map large portions of the sky in a repeated fashion, down to very faint flux levels. These synoptic surveys will be generating Petabytes of data, and they will open a whole new field of searches for variable astronomical objects.

This richness of information is hard to translate into a derived knowledge and physical understanding. Questions abound: How do we explore datasets comprising hundreds of millions or billions of objects each with dozens of attributes? How do we objectively classify the detected sources to isolate subpopulations of astrophysical interest? How do we identify correlations and anomalies within the data sets? How do we use the data to constrain astrophysical interpretation, which often involve highly nonlinear parametric functions derived from fields such as physical cosmology, stellar structure, or atomic physics? How do we match these complex data sets with equally complex numerical simulations, and how do we evaluate the performance of such models?

The key task is now to enable an efficient and complete scientific exploitation of these enormous data sets. The problems we face are inherently statistical in nature. Similar situations exist in many other fields of science and applied technology today. This poses many technical and conceptual challenges, but it may lead to a whole new methodology of doing science in the information-rich era.

In order to cope with this data flood, the astronomical community started a grassroots initiative, the National (and ultimately Global) Virtual Observatory (NVO). The NVO would federate numerous large digital sky archives, provide the information infrastructure and standards for ingestion of new data and surveys, and develop the computational and analysis tools with which to explore these vast data volumes. Recognising the urgent need, the National Academy of Science Astronomy and Astrophysics Survey Committee, in its new decadal survey *Astronomy and Astrophysics in the New Millennium* (McKee, Taylor, *et al.* 2001) recommends, as a first priority, the establishment of a National Virtual Observatory (NVO).

The NVO would provide new opportunities for scientific discovery that

were unimaginable just a few years ago. Entirely new and unexpected scientific results of major significance will emerge from the combined use of the resulting datasets, science that would not be possible from such sets used singly. In the words of a "white paper" [2] prepared by an interim steering group the NVO will serve as *an engine of discovery for astronomy.*

Implementation of the NVO involves significant technical challenges on many fronts, and in particular the *data analysis.* Whereas some of the NVO science would be done in the image (pixels) domain, and some in the interaction between the image and catalog domains, it is anticipated that much of the science (at least initially) will be done purely in the catalog domain of individual or federated sky surveys. A typical data set may be a catalog of $\sim 10^8 - 10^9$ sources with $\sim 10^2$ measured attributes each, i.e., a set of $\sim 10^9$ data vectors in a ~ 100-dimensional parameter space.

Dealing with the analysis of such data sets is obviously an inherently multivariate statistical problem. Complications abound: parameter correlations will exist; observational limits (selection effects) will generally have a complex geometry; for some of the sources some of the measured parameters may be only upper or lower limits; the measurement errors may vary widely; some of the parameters will be continuous, and some discrete, or even without a well-defined metric; etc. In other words, analysis of the NVO data sets will present many challenging problems for multivariate statistics, and the resulting astronomical conclusions will be strongly affected by the correct application of statistical tools.

We review some important statistical challenges raised by the NVO. These include the classification and extraction of desired subpopulations, understanding the relationships between observed properties within these subpopulations, and linking the astronomical data to astrophysical models. This may require a generation of new methods in data mining, multivariate clustering and analysis, nonparametric and semiparametric estimation and model and hypothesis testing.

9.2 Clustering analysis challenges in a VO

The exploration of observable parameter spaces, created by combining of large sky surveys over a range of wavelengths, will be one of the chief scientific purposes of a VO. This includes an exciting possibility of discovering some previously unknown types of astronomical objects or phenomena (see Djorgovski *et al.* 2001a, 2001b, 2001c for reviews).

A complete observable parameter space axes include quantities such as the object coordinates, velocities or redshifts, sometimes proper motions,

[2] Available at http://www.arXiv.org/abs/astro-ph/0108115, and also published in Brunner, Djorgovski, & Szalay (2001), p. 353.

fluxes at a range of wavelength (i.e., spectra; imaging in a set of bandpasses can be considered a form of a very low resolution spectroscopy), surface brightness and image morphological parameters for resolved sources, variability (or, more broadly, power spectra) over a range of time scales, etc. Any given sky survey samples only a small portion of this grand observable parameter space, and is subject to its own selection and measurement limits, *e.g.,* limiting fluxes, surface brightness, angular resolution, spectroscopic resolution, sampling and baseline for variability if multiple epoch observations are obtained, etc.

A major exploration technique envisioned for the NVO will be unsupervised clustering of data vectors in some parameter space of observed properties of detected sources. Aside from the computational challenges with large numbers of data vectors and a large dimensionality, this poses some highly non-trivial statistical problems. The problems are driven not just by the *size* of the data sets, but mainly (in the statistical context) by the *heterogeneity and intrinsic complexity of the data*.

A typical VO data set may consist of $\sim 10^9$ data vectors in $\sim 10^2$ dimensions. These are measured source attributes, including positions, fluxes in different bandpasses, morphology quantified through different moments of light distribution and other suitably constructed parameters, etc. Some of the parameters would be primary measurements, and others may be derived attributes, such as the star-galaxy classification, some may be "flags" rather than numbers, some would have error-bars associated with them, and some would not, and the error-bars may be functions of some of the parameters, e.g., fluxes. Some measurements would be present only as upper or lower limits. Some would be affected by "glitches" due to instrumental problems, and if a data set consists of a merger of two or more surveys, e.g., cross-matched optical, infrared, and radio (and this would be a common scenario within a VO), then some sources would be misidentified, and thus represent erroneous combinations of subsets of data dimensions. Surveys would be also affected by selection effects operating explicitly on some parameters (e.g., coordinate ranges, flux limits, etc.), but also mapping onto some other data dimensions through correlations of these properties; some selection effects may be unknown.

Physically, the data set may consist of a number of distinct classes of objects, such as stars (including a range of spectral types), galaxies (including a range of Hubble types or morphologies), quasars, etc. Within each object class or subclass, some of the physical properties may be correlated, and some of these correlations may be already known and some as yet unknown, and their discovery would be an important scientific result by itself. Some of the correlations may be spurious (*e.g.,* driven by sample selection effects), or simply uninteresting (e.g., objects brighter in one optical bandpass will tend to be brighter in another optical bandpass). Correlations of independently measured physical parameters represent a reduction of the statistical dimensionality in a multidimensional data parameter space, and

their discovery may be an integral part of the clustering analysis.

Typical scientific questions posed may be:

- How many statistically distinct classes of objects are in this data set, and which objects are to be assigned to which class, along with association probabilities?
- Are there any previously unknown classes of objects, i.e., statistically significant "clouds" in the parameter space distinct from the "common" types of objects (e.g., normal stars or galaxies)? An application may be separating quasars from otherwise morphologically indistinguishable normal stars.
- Are there rare outliers, i.e., individual objects with a low probability of belonging to any one of the dominant classes? Examples may include known, bur relatively rare types of objects such as high-redshift quasars, brown dwarfs, etc., but also previously unknown types of objects; finding any such would be a significant discovery.
- Are there interesting (in general, multivariate) correlations among the properties of objects in any given class, and what are the optimal analytical expressions of such correlations? An example may be the "Fundamental Plane" of elliptical galaxies, a set of bivariate correlations obeyed by this Hubble type, but no other types of galaxies (see, e.g., Djorgovski 1992, 1993, and Djorgovski *et al.* 1995, for reviews).

The complications include the following:

1. Construction of these complex data sets, especially if multiple sky surveys, catalogs, or archives are being federated (an essential VO activity) will inevitably be imperfect, posing quality control problems which must be discovered and solved first, before the scientific exploration starts. Sources may be mismatched, there will be some gross errors or instrumental glitches within the data, subtle systematic calibration errors may affect pieces of the large data sets, etc.
2. The object classes form multivariate "clouds" in the parameter space, but these clouds in general need not be Gaussian: some may have a power-law or exponential tails in some or all of the dimensions, and some may have sharp cutoffs, etc.
3. The clouds may be well separated in some of the dimensions, but not in others. How can we objectively decide which dimensions are irrelevant, and which ones are useful?
4. The *topology* of clustering may not be simple: there may be clusters within clusters, holes in the data distribution (negative clusters?), multiply-connected clusters, etc.

5. All of this has to take into the account the heterogeneity of measurements, censored data, incompleteness, etc.

The majority of the technical and methodological challenges in this quest derive from the expected heterogeneity and intrinsic complexity of the data, including treatment of upper an lower limits, missing data, selection effects and data censoring, etc. These issues affect the proper statistical description of the data, which then must be reflected in the clustering algorithms.

Related to this are the problems arising from the data modeling. The commonly used mixture-modeling assumption of clusters represented as multivariate Gaussian clouds is rarely a good descriptor of the reality. Clusters may have non-Gaussian shapes, *e.g.,* exponential or power-law tails, asymmetries, sharp cutoffs, etc. This becomes a critical issue in evaluating the membership probabilities in partly overlapping clusters, or in a search for outliers (anomalous events) in the tails of the distributions. In general, the proper functional forms for the modeling of clusters are not known *a priori*, and must be discovered from the data. Applications of non-parametric techniques may be essential here. A related, very interesting problem is posed by the *topology* of clustering, with a possibility of multiply-connected clusters or gaps in the data (*i.e.,* negative clusters embedded within the positive ones), hierarchical or multi-scale clustering (*i.e.,* clusters embedded within the clusters) etc.

The clusters may be well separated in some of the dimensions, but not in others. How can we objectively decide which dimensions are irrelevant, and which ones are useful? An automated and objective rejection of the "useless" dimensions, perhaps through some statistically defined entropy criterion, could greatly simplify and speed up the clustering analysis.

Once the data are partitioned into distinct clusters, their analysis and interpretation starts. One question is, are there interesting (in general, multivariate) correlations among the properties of objects in any given cluster? Such correlations may reflect interesting new astrophysics (*e.g.,* the stellar main sequence, the Tully-Fisher and Fundamental Plane correlations for galaxies, etc.), but at the same time complicate the statistical interpretation of the clustering. They would be in general restricted to a subset of the dimensions, and not present in the others. How do we identify all of the interesting correlations, and discriminate against the "uninteresting" observables?

Here we describe some of our experiments to date, and outline some possible avenues for future exploration.

9.3 Examples and some possible approaches

Separation of the data into different types of objects, be it known or unknown in nature, can be approached as a problem in automated classifi-

cation or clustering analysis. This is a part of a more general and rapidly growing field of Data Mining (DM) and Knowledge Discovery in Databases (KDD). We see here great opportunities for collaborations between astronomers and computer scientists and statisticians. For an overview of some of the issues and methods, see, e.g., Fayyad *et al.* (1996b) .

If applied in the catalog domain, the data can be viewed as a set of n points or vectors in an m-dimensional parameter space, where n can be in the range of many millions or even billions, and m in the range of a few tens to hundreds. The data may be clustered in k statistically distinct classes, which could be modeled, *e.g.,* as multivariate Gaussian clouds, and which hopefully correspond to physically distinct classes of objects (*e.g.,* stars, galaxies, quasars, etc.). This is schematically illustrated in Figure 1.

If the number of object classes k is known (or declared) *a priori*, and training data set of representative objects is available, the problem reduces to supervised classification, where tools such as Artificial Neural Nets or Decision Trees can be used. This is now commonly done for star-galaxy separation in sky surveys (*e.g.,* Odewahn *et al.* 1992, or Weir *et al.* 1995). Searches for known types of objects with predictable signatures in the parameter space (*e.g.,* high-z quasars) can be also cast in this way.

However, a more interesting and less biased approach is where the number of classes k is not known, and it has to be derived from the data themselves. The problem of unsupervised classification is to determine this number in some objective and statistically sound manner, and then to associate class membership probabilities for all objects. Majority of objects may fall into a small number of classes, *e.g.,* normal stars or galaxies. What is of special interest are objects which belong to much less populated clusters, or even individual outliers with low membership probabilities for any major class. Some initial experiments with unsupervised clustering algorithms in the astronomical context include, *e.g.,* Goebel *et al.* (1989), Weir *et al.* (1995), de Carvalho *et al.* (1995), and Yoo *et al.* (1996), but a full-scale application to major digital sky surveys yet remains to be done. Intriguing applications which addressed the issue of how many statistically distinct classes of GRBs are there (Mukherjee *et al.* 1998, Rogier *et al.* 2000, Hakkila *et al.* 2000).

In many situations, scientifically informed input is needed in designing the clustering experiments. Some observed parameters may have a highly significant, large dynamical range, dominate the sample variance, and naturally invite division into clusters along the corresponding parameter axes; yet they may be completely irrelevant or uninteresting scientifically. For example, if one wishes to classify sources of the basic of their broad-band spectral energy distributions (or to search for objects with unusual spectra), the mean flux itself is not important, as it mainly reflects the distance; coordinates on the sky may be unimportant (unless one specifically looks for a spatial clustering); etc. Thus, a clustering algorithm may divide the data set along one or more of such axes, and completely miss the really

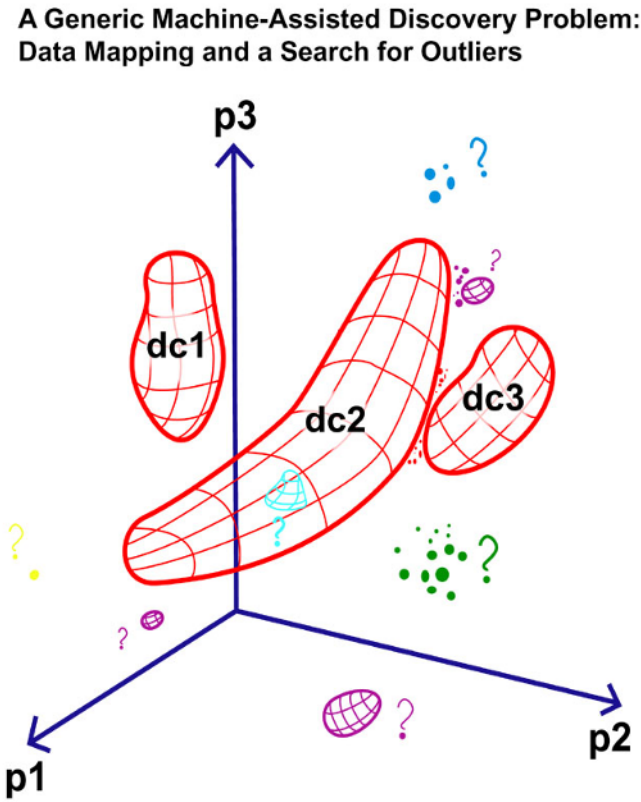


FIGURE 9.1. A schematic illustration of the problem of clustering analysis in some parameter space. In this example, there are 3 dimensions, $p1$, $p2$, and $p3$ (*e.g.,* some flux ratios or morphological parameters), and most of the data points belong to 3 major clusters, denoted $dc1$, $dc2$, and $dc3$ (*e.g.,* stars, galaxies, and ordinary quasars). One approach is to isolate these major classes of objects for some statistical studies, *e.g.,* stars as probes of the Galactic structure, or galaxies as probes of the large scale structure of the universe, and filter out the "anomalous" objects. A complementary view is to look for other, less populated, but statistically significant, distinct clusters of data points, or even individual outliers, as possible examples of rare or unknown types of objects. Another possibility is to look for holes (negative clusters) within the major clusters, as they may point to some interesting physical phenomenon – or to a problem with the data.

scientifically interesting partitions, e.g., according to the colors of objects.

One method we have been experimenting with (applied on the various data sets derived from DPOSS) is the Expectation Maximisation (EM) technique, with the Monte Carlo Cross Validation (MCCV) as the way of determining the maximum likelihood number of the clusters.

This may be a computationally very expensive problem. For the simple K-means algorithm, the computing cost scales as $K \times N \times I \times D$, where K is the number of clusters chosen *a priori*, N is the number of data vectors (detected objects), I is the number of iterations, and D is the number of data dimensions (measured parameters per object). For the more powerful Expectation Maximisation technique, the cost scales as $K \times N \times I \times D^2$, and again one must decide *a priori* on the value of K. If this number has to be determined intrinsically from the data, *e.g.,* with the Monte Carlo Cross Validation method, the cost scales as $M \times K_{max}^2 \times N \times I \times D^2$ where M is the number of Monte Carlo trials/partitions, and K_{max} is the maximum number of clusters tried. Even with the typical numbers for the existing large digital sky surveys ($N \sim 10^8 - 10^9$, $D \sim 10 - 100$) this is already reaching in the realm of Terascale computing, especially in the context of an interactive and iterative application of these analysis tools. Development of faster and smarter algorithms is clearly a priority.

One technique which can simplify the problem is the multi-resolution clustering. In this regime, expensive parameters to estimate, such as the number of classes and the initial broad clustering are quickly estimated using traditional techniques, and then one could proceed to refine the model locally and globally by iterating until some objective statistical (*e.g.,* Bayesian) criterion is satisfied.

One can also use intelligent sampling methods where one forms "prototypes"of the case vectors and thus reduces the number of cases to process. Prototypes can be determined from simple algorithms to get a rough estimate, and then refined using more sophisticated techniques. A clustering algorithm can operate in prototype space. The clusters found can later refined by locally replacing each prototype by its constituent population and reanalyzing the cluster.

Techniques for dimensionality reduction, including principal component analysis and others can be used as preprocessing techniques to automatically derive the dimensions that contain most of the relevant information.

9.4 Concluding comments

Given this computational and statistical complexity, blind applications of the commonly used (commercial or home-brewed) clustering algorithms could produce some seriously misleading or simply wrong results. The clustering methodology must be robust enough to cope with these problems,

and the outcome of the analysis must have a solid statistical foundation.

In our experience, design and application of clustering algorithms must involve close, working collaboration between astronomers and computer scientists and statisticians. There are too many unspoken assumptions, historical background knowledge specific to the given discipline, and opaque jargon; constant communication and interchange of ideas are essential.

The entire issue of discovery and interpretation of multivariate correlations in these massive data sets has not really been addressed so far. Such correlations may contain essential clues about the physics and the origins of various types of astronomical objects.

Effective and powerful data visualization, applied in the parameter space itself, is another essential part of the interactive clustering analysis. Good visualisation tools are also critical for the interpretation of results, especially in an iterative environment. While clustering algorithms can assist in the partitioning of the data space, and can draw the attention to anomalous objects, ultimately a scientist guides the experiment and draws the conclusions. It is very hard for a human mind to really visualise clustering or correlations in more than a few dimensions, and yet both interesting clusters and multivariate correlations with statistical dimensionality > 10 or even higher are likely to exits, and possibly lead to some crucial new astrophysical insights. Perhaps the right approach would be to have a good visualisation embedded as a part of an interactive and iterative clustering analysis.

Another key issue is interoperability and reusability of algorithms and models in a wide variety of problems posed by a rich data environment such as federated digital sky surveys in a VO. Implementation of clustering analysis algorithms must be done with this in mind.

Finally, scientific verification and evaluation, testing, and follow-up on any of the newly discovered classes of objects, physical clusters discovered by these methods, and other astrophysical analysis of the results is essential in order to demonstrate the actual usefulness of these techniques for a VO or other applications. Clustering analysis can be seen as a prelude to the more traditional type of astronomical studies, as a way of selecting of interesting objects of samples, and hopefully it can lead to advances in statistics and applied computer science as well.

9.5 Acknowledgments

We wish to thank numerous collaborators, including R. Gal, S. Odewahn, R. de Carvalho, T. Prince, J. Jacob, D. Curkendall, and many others. This work was supported in part by the NASA grant NAG5-9482, and by private foundations. Finally, we thank the organizers for a pleasant and productive meeting.

9.6 References

Boller, T., Meurs, E., & Adorf, H.-M. 1992, A&A, 259, 101

Brunner, R.J., Djorgovski, S.G., & Szalay, A.S. (editors) 2001a, *Virtual Observatories of the Future*, ASPCS vol. 221.

Brunner, R., Djorgovski, S.G., Gal, R.R., Mahabal, A., & Odewahn, S.C. 2001b, in: *Virtual Observatories of the Future*, eds. R. Brunner, S.G. Djorgovski & A. Szalay, ASPCS, 225, 64

Burl, M., Asker, L., Smyth, P., Fayyad, U., Perona, P., Crumpler, L., & Aubelle, J. 1998, Mach. Learning, 30, 165

de Carvalho, R., Djorgovski, S., Weir, N., Fayyad, U., Cherkauer, K., Roden, J., & Gray, A. 1995, in *Astronomical Data Analysis Software and Systems IV*, eds. R. Shaw *et al.* , ASPCS, 77, 272

Djorgovski, S.G. 1992, in: *Cosmology and Large-Scale Structure in the Universe*, ed. R. de Carvalho, ASPCS, 24, 19

Djorgovski, S.G. 1993, in: *The Globular Cluster – Galaxy Connection*, eds. G. Smith & J. Brodie, ASPCS, 48, 496

Djorgovski, S.G., Pahre, M.A., & de Carvalho, R.R. 1995, in: *Fresh Views of Elliptical Galaxies*, eds. A. Buzzoni *et al.* , ASPCS, 86, 129

Djorgovski, S.G., Mahabal, A., Brunner, R., Gal, R.R., Castro, S., de Carvalho, R.R., & Odewahn, S.C. 2001a, in: *Virtual Observatories of the Future*, eds. R. Brunner, S.G. Djorgovski & A. Szalay, ASPCS, 225, 52 [astro-ph/0012453]

Djorgovski, S.G., Brunner, R., Mahabal, A., Odewahn, S.C., de Carvalho, R.R., Gal, R.R., Stolorz, P., Granat, R., Curkendall, D., Jacob, J., & Castro, S. 2001b, in: *Mining the Sky*, eds. A.J. Banday *et al.* , ESO Astrophysics Symposia, Berlin: Springer Verlag, p. 305 [astro-ph/0012489]

Djorgovski, S.G., Mahabal, A., Brunner, R., Williams, R., Granat, R., Curkendall, D., Jacob, J., & Stolorz, P. 2001c, in: *Astronomical Data Analysis*, eds. J.-L. Starck & F. Murtagh, *Proc. SPIE* **4477**, p. 43 [astro-ph/0108346]

Fayyad, U., Djorgovski, S.G., & Weir, W.N. 1996a, in *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad *et al.* , Boston: AAAI/ MIT Press, p. 471

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (eds.) 1996b, *Advances in Knowledge Discovery and Data Mining*, Boston: AAAI/MIT Press

Goebel, J., Volk, K., Walker, H., Gerbault, F., Cheeseman, P., Self, M., Stutz, J., & Taylor, W. 1989, A&A, 222, L5

Hakkila, J., Haglin, D., Pendleton, G., Mallozzi, R., Meegan, C., & Rogier, R. 2000, ApJ, 538, 165

Mukherjee, S., Feigelson, E., Babu, J., Murtagh, F., Fraley, C., & Raftery, A. 1998, ApJ, 508, 314

Odewahn, S.C., Stockwell, E., Pennington, R., Humphreys, R., & Zumach, W. 1992, AJ, 103, 318

Paczyński, B. 2000, PASP, 112, 1281
Rogier, R., Hakkila, J., Haglin, D., Pendleton, G., & Mallozzi, R. 2000, in: *Gamma-Ray Bursts, 5th Hunsville Symp.*, eds. R. Kippen *et al.* , AIP Conf. Proc. 526, 38
Szalay, A., & Gray, J. 2001, Science, 293, 2037
Weir, N., Fayyad, U., & Djorgovski, S. 1995, AJ, 109, 2401
Yoo, J., Gray, A., Roden, J., Fayyad, U., de Carvalho, R., & Djorgovski, S. 1996, in: *Astronomical Data Analysis Software and Systems V*, eds. G. Jacoby & J. Barnes, ASPCS, 101, 41

Commentary by Dianne Cook[3]

This paper provides a detailed description of the development of a virtual observatory. The objective is to build an archive that coordinates large quantities of digital sky survey data from a variety of sources, and ultimately make new discoveries that improve our society's understanding about the universe.

The paper raises several questions from the perspective of a non-astronomer: Is there any data currently available? Where should one look to monitor the activity of the National Virtual Observatory?

A main focus of the paper is outlining the tasks for cluster analysis in extracting information from the virtual observatory data. My commentary focused on this aspect of the paper.

As it is commonly practiced, cluster analysis is a fuzzy science, that is often thought to magically extract structure. Cluster analysis is a collection of algorithms that group observations into similarity groups. All depend on an interpoint (intercluster) distance metric that defines the proximity of two observations (clusters). The way observations are then grouped together varies from algorithm to algorithm: hierarchical methods work sequentially through from closest neighbors to most distant; k-means requires an initial choice of k and then iteratively assigns observations to the nearest mean, and then recalculates the means; model-based hierarchical clustering overlays a probability distribution on the data and then estimates parameters to the distribution. Many types of cross-validation methods are available to ascertain the "best" results. But ultimately they may all produce inadequate results. The issue underlying the fuzziness is that the term "cluster" is itself a fuzzy concept. Ideally the analyst has a precise definition of "cluster". In practice, this information needs to be extracted from the data too, and the analyst begins a cluster analysis with little idea of what is being sought.

[3]Department of Statistics, Iowa State University

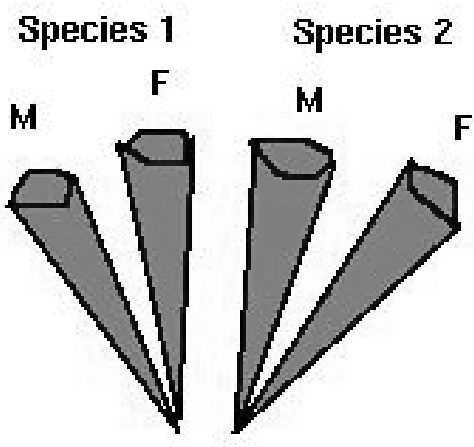


FIGURE 9.2. Schematic diagram of the crabs data.

Here is a simple example of the complications with clustering using Australian crabs data. There are 5 variables and 200 observations, and 4 real clusters in the data corresponding to males and females in two species. The 5 variables are strongly linearly dependent, and the cluster structure lies parallel to the linear dependency. And the clusters corresponding to the sexes are joined at the smallest values (Figure 9.2). The cluster structure can be intuitively modeled using 4 pencils, where pairs of the pencils are joined at one end, then diverge from each other at the other end. In Figure 53.1 the top left plot shows a pairwise plot of two variables (CL vs RW) where the sex separation can be seen. The right plot shows a tour projection where the 4 cluster can be seen reasonably distinctly, rather like looking down the "barrels of the pencil clusters". Assuming that the variables are standardized to have zero mean and unity variance, virtually all cluster algorithms will carve data up into clusters along the line of correlation (bottom left plot). Hence if we were to use cross-validation or comparison of results between several algorithms we might mistakenly believe that we have produced a consistent, and useful result. But it cannot be further from the truth. Now, an astute analyst might expect that model-based clustering using equal elliptical variance-covariance structure might extract the 4 real clusters, but alas it also fails (bottom right plot). The BIC criterion for model-based clustering does indeed suggest equal elliptical variance-covariance but the number of cluster is predicted to be 3, not 4.

This data is a strong candidate for clustering in the principal components space. And indeed the results are somewhat better. Figure 9.4 shows the true groups (left) and the results from hierarchical clustering in the principal components (right). The sexes of one species of crabs ("x" and

"+") gets seriously confused but generally clusters corresponding to the two species are extracted and the sexes of one species are reasonably well-extracted. The cluster algorithm was run in the space of the first 3 principal components. The first 3 principal components capture the variation and the cluster structure due to the species and sexes quite adequately: the 4 "joined-pencils" shape is visible in the first 3 principal components rather than the more awkward 5D space of the original data. This somewhat unusual.

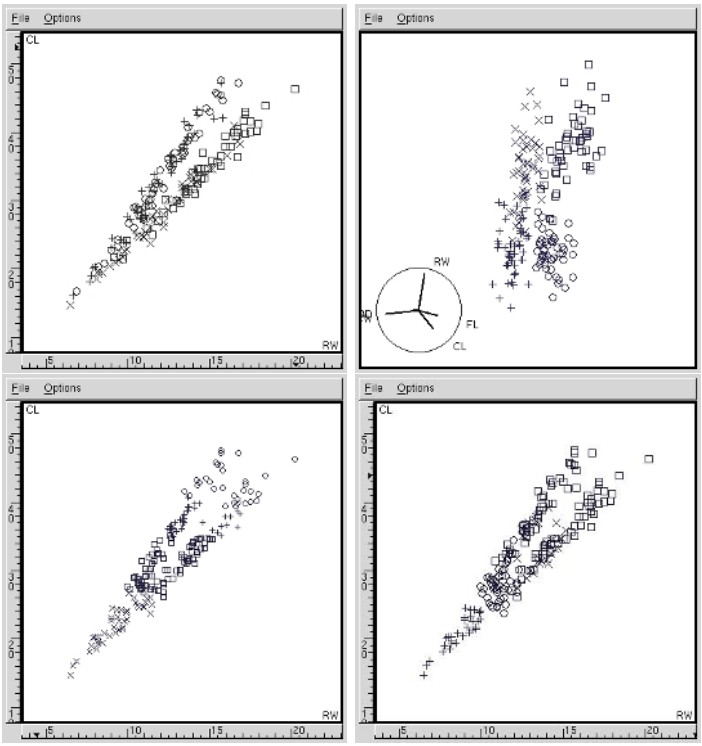


FIGURE 9.3. Clustering difficulties in even a simple data set.

In general, reducing data to a small number of principal components can throw the cluster structure to the wind. Often the cluster structure can be found in the lower principal components. The reason is that principal components is a linear structure extractor, but cluster structure is often non-linear. This is an observation made by Donnell et al (1994). So beware of using principal components analysis as a dimension reduction technique.

Some additional background to clustering with the k-means can be found in Tarpey et al (1995). In this paper is a careful study of the way the algorithm works under several data distribution assumptions. A interesting clustering method that is not well-known can be found in Osbourne

et al (1995). Ultimately good cluster analysis benefits from a heavy use of graphics and a good subject matter knowledge. We used the software ggobi (`www.ggobi.org`) to generate the plots in this paper. GGobi includes tour methods which help delineate the shape of clusters in high-dimensional Euclidean space. Cook et al (1995) contains another cluster analysis example on 7D particle physics data. This data lies in a neat geometric shape that can be extracted using tour methods.

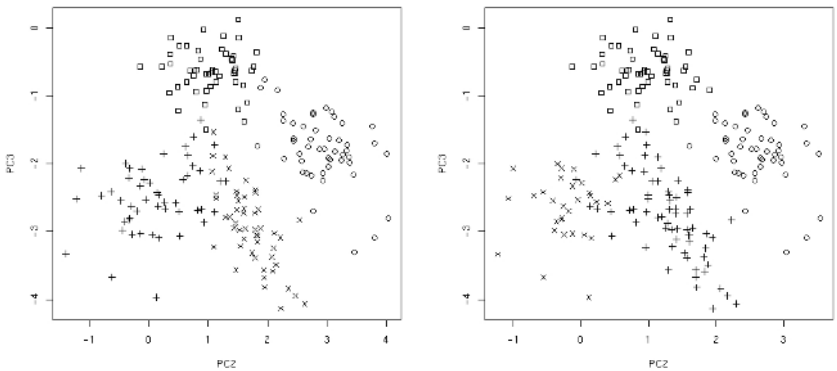


FIGURE 9.4. Clustering in principal components.

In summary, my challenge to astronomers is this: How do you quantitatively define what is interesting in astronomical data? When you say "outlier" what do you mean mathematically? When you say "cluster" what do you mean mathematically?

9.7 References

Cook, D. et al, 1995, Grand Tour and Projection Pursuit, Journal of Computational and Graphical Statistics, 4(3):155-172.

Donnell, D. et al, 1994, Analysis of Additive Dependencies using Smallest Additive Principle Components (with discussion), The Annals of Statistics, 22:1636-1673.

Osbourn, G. C. et al, 1995, Empirically defined regions of influence for clustering analysis, Pattern Recognition, 28(11):1793-1806.

Tarpey, T. et al, 1995, Principal Points and Self–Consistent Points of Elliptical Distributions, Annals of Statistics, 23:103–112.

This page intentionally left blank

10

Statistics of Galaxy Clustering

Vicent J. Martínez[1] and Enn Saar

ABSTRACT In this introductory talk we will establish connections between the statistical analysis of galaxy clustering in cosmology and recent work in mainstream spatial statistics. The lecture will review the methods of spatial statistics used by both sets of scholars, having in mind the cross-fertilizing purpose of the meeting series. Special topics will be: description of the galaxy samples, selection effects and biases, correlation functions, nearest neighbor distances, void probability functions, Fourier analysis, and structure statistics.
This paper is followed by a commentary by Rien van de Weygaert.

10.1 Introduction

One of the most important motivations of these series of conferences is to promote vigorous interaction between statisticians and astronomers. The organizers merit our admiration for bringing together such a stellar cast of colleagues from both fields. In this third edition, one of the central subjects is cosmology, and in particular, statistical analysis of the large-scale structure in the universe. There is a reason for that — the rapid increase of the amount and quality of the available observational data on the galaxy distribution (also on clusters of galaxies and quasars) and on the temperature fluctuations of the microwave background radiation.

These are the two fossils of the early universe on which cosmology, a science driven by observations, relies. Here we will focus on one of them — the galaxy distribution. First we briefly review the redshift surveys, how they are built and how to extract statistically analyzable samples from them, considering selection effects and biases. Most of the statistical analysis of the galaxy distribution are based on second order methods (correlation functions and power spectra). We comment them, providing the connection between statistics and estimators used in cosmology and in spatial statistics. Special attention is devoted to the analysis of clustering in Fourier space, with new techniques for estimating the power spectrum, which are becoming increasingly popular in cosmology. We show also the results of

[1]Observatori Astronòmic, Universitat de València

applying these second-order methods to recent galaxy redshift surveys.

Fractal analysis has become very popular as a consequence of the scale-invariance of the galaxy distribution at small scales, reflected in the power-law shape of the two-point correlation function. We discuss here some of these methods and the results of their application to the observations, supporting a gradual transition from a small-scale fractal regime to large-scale homogeneity. The concept of lacunarity is illustrated with some detail.

We end by briefly reviewing some of the alternative measures of point statistics and structure functions applied thus far to the galaxy distribution: void probability functions, counts-in-cells, nearest neighbor distances, genus, and Minkowski functionals.

10.2 Cosmological datasets

Cosmological datasets differ in several respects from those usually studied in spatial statistics. The point sets in cosmology (galaxy and cluster surveys) bear the imprint of the observational methods used to obtain them.

The main difference is the systematically variable intensity (mean density) of cosmological surveys. These surveys are usually magnitude-limited, meaning that all objects, which are brighter than a pre-determined limit, are observed in a selected region of the sky. This limit is mainly determined by the telescope and other instruments used for the program. Apparent magnitude, used to describe the limit, is a logarithmic measure of the observed radiation flux.

It is usually assumed that galaxies at all distances have the same (universal) luminosity distribution function. This assumption has been tested and found to be in satisfying accordance with observations. As the observed flux from a galaxy is inversely proportional to the square of its distance, we can see at larger distances only a bright fraction of all galaxies. This leads directly to the mean density of galaxies that depends on their distance from us r.

This behaviour is quantified by a selection function $\phi(r)$, which is usually found by estimating first the luminosity distribution of galaxies (the luminosity function).

One can also select a distance limit, find the minimum luminosity of a galaxy, which can yet be seen at that distance, and ignore all galaxies that are less luminous. Such samples are called volume-limited. They are used for some special studies (typically for counts-in-cells), but the loss of hard-earned information is enormous. The number of galaxies in volume-limited samples is several times smaller than in the parent magnitude-limited samples. This will also increase the shot (discreteness) noise.

In addition to the radial selection function $\phi(r)$, galaxy samples also are frequently subject to angular selection. This is due to our position in the

Galaxy — we are located in a dusty plane of the Galaxy, and the window in which we see the Universe, also is dusty. This dust absorbs part of galaxies' light, and makes the real brightness limit of a survey dependent on the amount of dust in a particular line-of-sight. This effect has been described by a $\phi(b) \sim (\sin b)^{-1}$ law (b is the galactic latitude); in reality the dust absorption in the Galaxy is rather inhomogeneous. There are good maps of the amount of Galactic dust in the sky, the latest maps have been obtained using the COBE and IRAS satellite data [Schlegel et al. 1998].

Edge problems, which usually affect estimators in spatial statistics, also are different for cosmological samples. The decrease of the mean density towards the sample borders alleviates these problems. Of course, if we select a volume-limited sample, we select also all these troubles (and larger shot noise). From the other side, edge effects are made more prominent by the usual observing strategies, when surveys are conducted in well-defined regions in the sky. Thus, edge problems are only partly alleviated; maybe it will pay to taper our samples at the side borders, too?

Some of the cosmological surveys have naturally soft borders. These are the all-sky surveys; the best known is the IRAS infrared survey, dust is almost transparent in infrared light. The corresponding redshift survey is the PSCz survey, which covers about 85% of the sky [Saunders et al. 2000]. A special follow-up survey is in progress to fill in the remaining Galactic Zone-of-Avoidance region, and meanwhile numerical methods have been developed to interpolate the structures seen in the survey into the gap [Schmoldt et al. 1999, Saunders & Ballinger 2000].

Another peculiarity of galaxy surveys is that we can measure exactly only the direction to the galaxy (its position in the sky), but not its distance. We measure the radial velocity v_r (or redshift $z = v_r/c$, c is the velocity of light) of a galaxy, which is a sum of the Hubble expansion, proportional to the distance d, and the dynamical velocity v_p of the galaxy, $v_r = H_0 d + v_p$. Thus we are differentiating between redshift space, if the distances simply are determined as $d = v_r/H_0$, and real space. The real space positions of galaxies could be calculated if we exactly knew the peculiar velocities of galaxies; we do not. The velocity distortions can be severe; well-known features of redshift space are fingers-of-God, elongated structures that are caused by a large radial velocity dispersion in massive clusters of galaxies. The velocity distortions expand a cluster in redshift space in the radial direction five-ten times.

For large-scale structures the situation is different, redshift distortions compress them. This is due to the continuing gravitational growth of structures. These differences can best be seen by comparing the results of numerical simulations, where we know also the real-space situation, in redshift space and in real space.

The last specific feature of the cosmology datasets is their size. Up to recent years most of the datasets have been rather small, of the order of 10^3 objects; exceptions exist, but these are recent. Such a small number of

points gives a very sparse coverage of three-dimensional survey volumes, and shot noise has been a severe problem.

This situation is about to change, swinging to the other extreme; the membership of new redshift surveys already is measured in terms of 10^5 (160,000 for the 2dF survey, quarter of a million planned) and million-galaxy surveys are on their way (the Sloan Survey). More information about these surveys can be found in their Web pages: *http://www.mso.anu.edu.au/2dFGRS/* for the 2dF survey and *http://www.sdss.org/* for the Sloan survey. This huge amount of data will force us to change the statistical methods we use. Nevertheless, the deepest surveys (e.g., distant galaxy cluster surveys) will always be sparse, so discovering small signals from shot-noise dominated data will remain a necessary art.

10.3 Correlation analysis

There are several related quantities that are second-order characteristics used to quantify clustering of the galaxy distribution in real or redshift space. The most popular one in cosmology is the two-point correlation function, $\xi(\mathbf{r})$. The infinitesimal interpretation of this quantity reads as follows:

$$dP_{12} = \bar{n}^2[1 + \xi(\mathbf{r})]dV_1 dV_2 \tag{10.1}$$

is the joint probability that in each one of the two infinitesimal volumes dV_1 and dV_2, with separation vector $\mathbf{r}$, lies a galaxy. Here $\bar{n}$ is the mean number density (intensity). Assuming that the galaxy distribution is a homogeneous (invariant under translations) and isotropic (invariant under rotations) point process, this probability depends only on $r = |\mathbf{r}|$. In spatial statistics, other functions related with $\xi(r)$ are commonly used:

$$\lambda_2(r) = \bar{n}^2\xi(r) + 1, \qquad g(r) = 1 + \xi(r), \qquad \Gamma(r) = \bar{n}(\xi(r) + 1), \tag{10.2}$$

where $\lambda_2(r)$ is the second-order intensity function, $g(r)$ is the pair correlation function, also called the radial distribution function or structure function, and $\Gamma(r)$ is the conditional density proposed by Pietronero (1987).

Different estimators of $\xi(r)$ have been proposed so far in the literature, both in cosmology and in spatial statistics. The main differences are in correction for edge effects. Comparison of their performance can be found in several papers [Pons-Bordería et al. 1999, Kerscher et al. 2000, Stoyan & Stoyan 2000]. There is clear evidence that $\xi(r)$ is well described by a power-law at scales $0.1 \leq r \leq 10\,h^{-1}$ Mpc where h is the Hubble constant in units of 100 km $\mathrm{s}^{-1}\,\mathrm{Mpc}^{-1}$:

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma},$$

with $\gamma \simeq 1.8$ and $r_0 \simeq 5.4\,h^{-1}$ Mpc. This scaling behavior is one of the reasons that have lead some astronomers to describe the galaxy distribution as fractal. A power-law fit for $g(r) \propto r^{3-D_2}$ permits to define the correlation dimension D_2. The extent of the fractal regime is still a matter of debate in cosmology, but it seems clear that the available data on redshift surveys indicate a gradual transition to homogeneity for scales larger than 15–20 h^{-1} Mpc [Martínez 1999]. Moreover, in a fractal point distribution, the correlation length r_0 increases with the radius of the sample because the mean density decreases [Pietronero 1987]. This simple prediction of the fractal interpretation is not supported by the data, instead r_0 remains constant for volume-limited samples with increasing depth [Martínez et al. 2001].

Several versions of the volume integral of the correlation function are also frequently used in the analysis of galaxy clustering. The most extended one in spatial statistics is the so-called Ripley K-function

$$K(r) = \int_0^r 4\pi s^2 (1 + \xi(s)) ds \tag{10.3}$$

although in cosmology it is more frequent to use an expression which provides directly the average number of neighbors an arbitrarily chosen galaxy has within a distance r, $N(< r) = \bar{n} K(r)$ or the average conditional density

$$\Gamma^*(r) = \frac{3}{r^3} \int_0^r \Gamma(s) s^2 ds$$

Again a whole collection of estimators are used to properly evaluate these quantities. Pietronero and coworkers recommend to use only minus–estimators to avoid any assumption regarding the homogeneity of the process. In these estimators, averages of the number of neighbors within a given distance are taken only considering as centers these galaxies whose distances to the border are larger than r. However, caution has to be exercised with this procedure, because at large scales only a small number of centers remain, and thus the variance of the estimator increases.

Integral quantities are less noisy than the corresponding differential expressions, but obviously they do contain less information on the clustering process due the fact that values of $K(r_1)$ and $K(r_2)$ for two different scales r_1 and r_2 are more strongly correlated than values of $\xi(r_1)$ and $\xi(r_2)$. Scaling of $N(< r) \propto r^{D_2}$ provides a smoother estimation of the correlation dimension. If scaling is detected for partition sums defined by the moments of order q of the number of neighbors

$$Z(q, r) = \frac{1}{N} \sum_{i=1}^{N} n_i(r)^{q-1} \propto r^{D_q/(q-1)},$$

the exponents D_q are the so-called generalized or multifractal dimensions [Martínez et al. 1990]. Note that for $q = 2$, $Z(2, r)$ is an estimator of

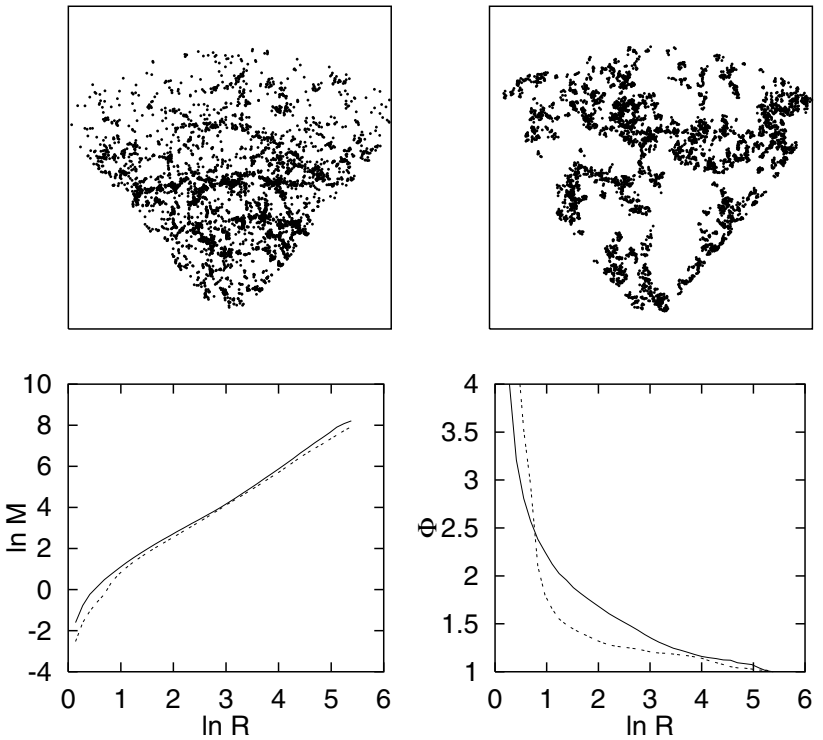


FIGURE 10.1. Comparison of a Las Campanas survey slice (upper left panel) with the Rayleigh-Lévy flight model (upper right panel). The fractal dimensions of both distributions coincide, as shown by the $\ln M$–$\ln R$ curves in the lower left panel, but the lacunarity curves (in the lower right panel) differ considerably. The solid lines describe the galaxy distribution, dotted lines – the model results. From (Martínez & Saar 2002).

$N(< r)$ and therefore D_q for $q = 2$ is simply the correlation dimension. If different kinds of cosmic objects are identified as peaks of the continuous matter density field at different thresholds, we can study the correlation dimension associated to each kind of object. The multiscaling approach [Jensen et al. 1991] associated to the multifractal formalism provides a unified framework to analyze this variation. It has been shown [Martínez et al. 1995] that the value of D_2 corresponding to rich galaxy clusters (high peaks of the density field) is smaller than the value corresponding to galaxies (within the same scale range) as prescribed in the multiscaling approach.

Finally we want to consider the role of lacunarity in the description of the galaxy clustering [Martínez & Saar 2002]. In Fig. 10.1, we show the space distribution of galaxies within one slice of the Las Campanas redshift survey, together with a fractal pattern generated by means of a Rayleigh-

Lévy flight [Mandelbrot 1982]. Both have the same mass-radius dimension, defined as the exponent of the power-law that fits the variation of mass within concentric spheres centered at the observer position.

$$M(R) = FR^{D_M}. \tag{10.4}$$

The best fitted value for both point distributions is $D_M \simeq 1.6$ as shown in the left bottom panel of Fig. 10.1. The different appearance of both point distributions is a consequence of the different degree of lacunarity. Blumenfeld & Mandelbrot (1997) have proposed to quantify this effect by measuring the variability of the prefactor F in Eq. 10.4,

$$\Phi = \frac{E\{(F - \bar{F})^2\}}{\bar{F}^2}$$

The result of applying this lacunarity measure is shown in the right bottom panel of Fig. 10.1. The visual differences between the point distributions are now well reflected in this curve.

10.4 Power spectra

The current statistical model for the main cosmological fields (density, velocity, gravitational potential) is the Gaussian random field. This field is determined either by its correlation function or by its spectral density, and one of the main goals of spatial statistics in cosmology is to estimate those two functions.

In recent years the power spectrum has attracted more attention than the correlation function. There are at least two reasons for that — the power spectrum is more intuitive physically, separating processes on different scales, and the model predictions are made in terms of power spectra. Statistically, the advantage is that the power spectrum amplitudes for different wavenumbers are statistically orthogonal:

$$E\left\{\widetilde{\delta}(\mathbf{k})\widetilde{\delta}^{\star}(\mathbf{k}')\right\} = (2\pi)^3\delta_D(\mathbf{k} - \mathbf{k}')P(\mathbf{k}).$$

Here $\widetilde{\delta}(\mathbf{k})$ is the Fourier amplitude of the overdensity field $\delta = (\rho - \bar{\rho})/\bar{\rho}$ at a wavenumber $\mathbf{k}$, ρ is the matter density, a star denotes complex conjugation, $E\{\}$ denotes expectation values over realizations of the random field, and $\delta_D(\mathbf{x})$ is the three-dimensional Dirac delta function. The power spectrum $P(\mathbf{k})$ is the Fourier transform of the correlation function $\xi(\mathbf{r})$ of the field.

Estimation of power spectra from observations is a rather difficult task. Up to now the problem has been in the scarcity of data; in the near future there will be the opposite problem of managing huge data sets. The development of statistical techniques here has been motivated largely by the

analysis of CMB power spectra, where better data were obtained first, and has been parallel to that recently.

The first methods developed to estimate the power spectra were direct methods — a suitable statistic was chosen and determined from observations. A good reference is Feldman et al. (1994).

The observed samples can be modeled by an inhomogeneous point process (a Gaussian Cox process) of number density $n(\mathbf{x})$:

$$n(\mathbf{x}) = \sum_i \delta_D(\mathbf{x} - \mathbf{x}_i),$$

where $\delta_D(\mathbf{x})$ is the Dirac delta-function. As galaxy samples frequently have systematic density trends caused by selection effects, we have to write the estimator of the density contrast in a sample as

$$D(\mathbf{x}) = \sum_i \frac{\delta_D(\mathbf{x} - \mathbf{x}_i)}{\bar{n}(\mathbf{x}_i)} - 1,$$

where $\bar{n}(\mathbf{x}) \sim \bar{\rho}(\mathbf{x})$ is the selection function expressed in the number density of objects.

The estimator for a Fourier amplitude (for a finite set of frequencies $\mathbf{k}_i$) is

$$F(\mathbf{k}_i) = \sum_j \frac{\psi(\mathbf{x}_j)}{\bar{n}(\mathbf{x}_j)} e^{i\mathbf{k}_i \cdot \mathbf{x}} - \widetilde{\psi}(\mathbf{k}_i),$$

where $\psi(\mathbf{x})$ is a weight function that can be selected at will. The raw estimator for the spectrum is

$$P_R(\mathbf{k}_i) = F(\mathbf{k}_i) F^\star(\mathbf{k}_i),$$

and its expectation value

$$E\left\{\langle |F(\mathbf{k}_i)|^2 \rangle\right\} = \int G(\mathbf{k}_i - \mathbf{k}') P(\mathbf{k}') \frac{d^3k'}{(2\pi)^3} + \int_V \frac{\psi^2(\mathbf{x})}{\bar{n}(\mathbf{x})} d^3x,$$

where $G(\mathbf{k}) = |\tilde{\psi}(\mathbf{k})|^2$ is the window function that also depends on the geometry of the sample volume. Symbolically, we can get the estimate of the power spectra $\widehat{P}$ by inverting the integral equation

$$G \otimes \widehat{P} = P_R - N,$$

where $\otimes$ denotes convolution, P_R is the raw estimate of power, and N is the (constant) shot noise term.

In general, we have to deconvolve the noise-corrected raw power to get the estimate of the power spectrum. This introduces correlations in the estimated amplitudes, so these are not statistically orthogonal any more. A sample of a characteristic spatial size L creates a window function of

width of $\Delta k \approx 1/L$, correlating estimates of spectra at that wavenumber interval.

As the cosmological spectra are usually assumed to be isotropic, the standard method to estimate the spectrum involves an additional step of averaging the estimates $\widehat{P}(\mathbf{k})$ over a spherical shell $k \in [k_i, k_{i+1}]$ of thickness $k_{i+1} - k_i > \Delta k = 1/L$ in wavenumber space. The minimum-variance requirement gives the FKP [Feldman et al. 1994] weight function:

$$\psi(\mathbf{x}) \sim \frac{\bar{n}(\mathbf{x})}{1 + \bar{n}(\mathbf{x})P(k)},$$

and the variance is

$$\frac{\sigma_P^2(k)}{P_R^2(k)} \approx \frac{2}{\mathcal{N}},$$

where $\mathcal{N}$ is the number of coherence volumes in the shell. The number of independent volumes is twice as small (the density field is real). The coherence volume is $V_c(k) \approx (\Delta k)^3 \approx 1/L^3 \approx 1/V$.

As the data sets get large, straight application of direct methods (especially the error analysis) becomes difficult. There are different recipes that have been developed with the future data sets in mind. A good review of these methods is given in Tegmark et al. (1998).

The deeper the galaxy sample, the smaller the coherence volume, the larger the spectral resolution and the larger the wavenumber interval where the power spectrum can be estimated. The deepest redshift surveys presently available are the PSCz galaxy redshift survey (15411 redshifts up to about $400h^{-1}$ Mpc, see Saunders et al. (2000), the Abell/ACO rich galaxy cluster survey, 637 redshifts up to about $300\,h^{-1}$ Mpc [Miller & Batuski 2001]), and the ongoing 2dF galaxy redshift survey (141400 redshifts up to $750h^{-1}$ Mpc [Peacock et al. 2001]). The estimates of power spectra for the two latter samples have been obtained by the direct method [Miller et al. 2001, Percival et al. 2001]. Fig. 10.2 shows the power spectrum for the 2dF survey.

The covariance matrix of the power spectrum estimates in Fig. 10.2 was found from simulations of a matching Gaussian Cox process in the sample volume. The main new feature in the spectra, obtained for the new deep samples, is the emergence of details (wiggles) in the power spectrum. While sometime ago the main problem was to estimate the mean behaviour of the spectrum and to find its maximum, now the data enables us to see and study the details of the spectrum. These details have been interpreted as traces of acoustic oscillations in the post-recombination power spectrum. Similar oscillations are predicted for the cosmic microwave background radiation fluctuation spectrum. The CMB wiggles match the theory rather well, but the galaxy wiggles do not, yet.

Thus, the measurement of the power spectrum of the galaxy distribution is passing from the determination of its overall behaviour to the discovery and interpretation of spectral details.

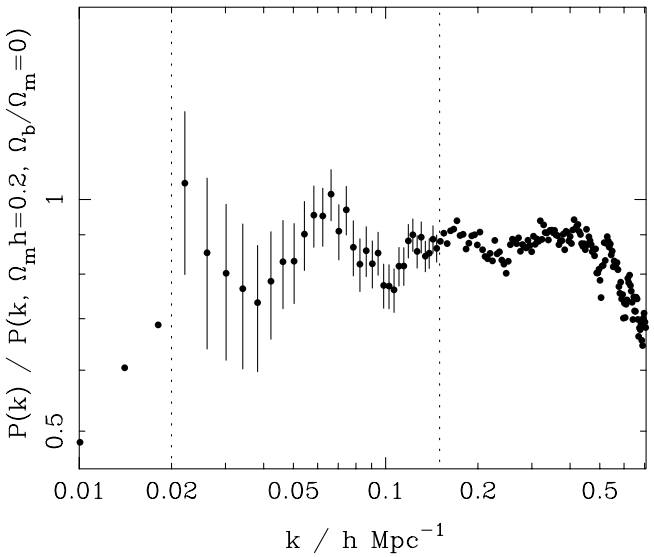


FIGURE 10.2. Power spectrum of the 2dF redshift survey, divided by a smooth model power spectrum. The spectrum is not deconvolved. Error bars are determined from Gaussian realizations; the dotted lines show the wavenumber region that is free of the influence of the window function and of the radial velocity distortions and nonlinear effects. (Courtesy of W. J. Percival and the 2dF galaxy redshift survey team.)

10.5 Other clustering measures

To end this review we briefly mention other measures used to describe the galaxy distribution.

10.5.1 Counts-in-cells and void probability function

The probability that a randomly placed sphere of radius r contains exactly N galaxies is denoted by $P(N, r)$. In particular, for $N = 0$, $P(0, r)$ is the so-called void probability function, related with the empty space function or contact distribution function $F(r)$, more frequently used in the field of spatial statistics, by $F(r) = 1 - P(0, r)$. The moments of the counts-in-cells probabilities can be related both with the multifractal analysis [Borgani 1993] and with the higher order n-point correlation functions [White 1979, Stoyan et al. 1995, Szapudi et al. 1999].

10.5.2 Nearest-neighbor distributions

In spatial statistics, different quantities based on distances to nearest neighbors have been introduced to describe the statistical properties of point

processes. $G(r)$ is the distribution function of the distance r of a given point to its nearest neighbor. It is interesting to note that $F(r)$ is just the distribution function of the distance r from an arbitrarily chosen point in $\mathbb{R}^3$ — not being an event of the point process — to a point of the point process (a galaxy in the sample in our case). The quotient

$$J(r) = \frac{1 - G(r)}{1 - F(r)}$$

introduced by van Lieshout & Baddeley (1996) is a powerful tool to analyze point patterns and has discriminative power to compare the results of N-body models for structure formation with the real distribution of galaxies [Kerscher et al. 1999].

10.5.3 Topology

One very popular tool for analysis of the galaxy distribution is the genus of the isodensity surfaces. To define this quantity, the point process is smoothed to obtain a continuous density field, the intensity function, by means of a kernel estimator for a given bandwidth. Then we consider the fraction of the volume f which encompasses those regions having density exceeding a given threshold ρ_t. The boundary of these regions specifies an isodensity surface. The genus $G(S)$ of a surface S is basically the number of holes minus the number of isolated regions plus 1. The genus curve shows the variation of $G(S)$ with f or ρ_t for a given window radius of the kernel function. An analytical expression for this curve is known for Gaussian density fields. It seems that the empirical curve calculated from the galaxy catalogs can be reasonably well fitted to a Gaussian genus curve [Canavezes et al. 1998] for window radii varying within a large range of scales.

10.5.4 Minkowski functionals

A very elegant generalization of the previous analysis to a larger family of morphological characteristics of the point processes is provided by the Minkowski functionals. These scalar quantities are useful to study the shape and connectivity of a union of convex bodies. They are well known in spatial statistics and have been introduced in cosmology by Mecke et al. (1994). On a clustered point process, Minkowski functionals are calculated by generalizing the Boolean grain model into the so-called germ-grain model. This coverage process consists in considering the sets $A_r = \cup_{i=1}^{N} B_r(\mathbf{x}_i)$ for the diagnostic parameter r, where $\{\mathbf{x}_i\}_{i=1}^{N}$ represents the galaxy positions and $B_r(\mathbf{x}_i)$ is a ball of radius r centered at point $\mathbf{x}_i$. Minkowski functionals are applied to sets A_r when r varies. In $\mathbb{R}^3$ there are four functionals: the volume V, the surface area A, the integral mean curvature H, and the

Euler-Poincaré characteristic χ, related with the genus of the boundary of A_r by $\chi = 1 - G$. Application of Minkowski functionals to the galaxy cluster distribution can be found in Kerscher et al. (1997). These quantities have been used also as efficient shape finders by Sahni et al. (1998).

Acknowledgments: This work was supported by the Spanish MCyT project AYA2000-2045 and by the Estonian Science Foundation under grant 2882. Enn Saar is grateful for the invited professor position funded by the Vicerrectorado de Investigación de la Universitat de València.

10.6 References

[Blumenfeld & Mandelbrot 1997] Blumenfeld R & Mandelbrot B 1997 *Phys. Rev. E* **56**, 112–118.

[Borgani 1993] Borgani S 1993 *Mon. Not. R. Astr. Soc.* **260**, 537–549.

[Canavezes et al. 1998] Canavezes A, Springel V, Oliver S J, Rowan-Robinson M, Keeble O, White S D M, Saunders W, Efstathiou G, Frenk C S, McMahon R G, Maddox S, Sutherland W & Tadros H 1998 *Mon. Not. R. Astr. Soc.* **297**, 777–793.

[Feldman et al. 1994] Feldman H A, Kaiser N & Peacock J A 1994 *Astrophys. J.* **426**, 23–37.

[Jensen et al. 1991] Jensen M H, Paladin G & Vulpiani A 1991 *Phys. Rev. Lett.* **67**, 208–211.

[Kerscher et al. 1999] Kerscher M, Pons-Bordería M, Schmalzing J, Trasarti-Battistoni R, Buchert T, Martínez V J & Valdarnini R 1999 *Astrophys. J.* **513**, 543–548.

[Kerscher et al. 1997] Kerscher M, Schmalzing J, Retzlaff J, Borgani S, Buchert T, Gottlober S, Muller V, Plionis M & Wagner H 1997 *Mon. Not. R. Astr. Soc.* **284**, 73–84.

[Kerscher et al. 2000] Kerscher M, Szapudi I & Szalay A S 2000 *Astrophys. J.* **535**, L13–L16.

[Mandelbrot 1982] Mandelbrot B B 1982 *The fractal geometry of nature* W.H. Freeman San Francisco.

[Martínez 1999] Martínez V J 1999 *Science* **284**, 445–446.

[Martínez et al. 1990] Martínez V J, Jones B J T, Domínguez-Tenreiro R & van de Weygaert R 1990 *Astrophys. J.* **357**, 50–61.

[Martínez et al. 2001] Martínez V J, López-Martí B & Pons-Bordería M J 2001 *Astrophys. J.* **554**, L5–L8.

[Martínez et al. 1995] Martínez V J, Paredes S, Borgani S & Coles P 1995 *Science* **269**, 1245–1247.

[Martínez & Saar 2002] Martínez V J & Saar E 2002 *Statistics of the Galaxy Distribution* Chapman and Hall/CRC Press Boca Raton.

[Mecke et al. 1994] Mecke K R, Buchert T & Wagner H 1994 *Astron. Astrophys.* **288**, 697–704.

[Miller & Batuski 2001] Miller C J & Batuski D J 2001 *Astrophys. J.* **551**, 635–642.

[Miller et al. 2001] Miller C J, Nichol R C & Batuski D J 2001. *Astrophys. J.* **555**, 68

[Peacock et al. 2001] Peacock J A, Cole S, Norberg P, Baugh C M, Bland-Hawthorn J, Bridges T, Cannon R D, Colless M, Collins C, Couch W, Dalton G, Deeley K, Propris R D, Driver S P, Efstathiou G, Ellis R S, Frenk C S, Glazebrook K, Jackson C, Lahav O, Lewis I, Lumsden S, Maddox S, Percival W J, Peterson B A, Price I, Sutherland W & Taylor K 2001 *Nature* **410**, 169–173.

[Percival et al. 2001] Percival W J, Baugh C M, Bland-Hawthorn J, Bridges T, Cannon R, Cole S, Colless M, Collins C, Couch W, Dalton G, Propris R D, Driver S P, Efstathiou G, Ellis R S, Frenk C S, Glazebrook K, Jackson C, Lahav O, Lewis I, Lumsden S, Maddox S, Moody S, Norberg P, Peacock J A, Peterson B A, Sutherland W & Taylor K 2001. *Mon. Not. R. Astr. Soc.*, **327**, 1297

[Pietronero 1987] Pietronero L 1987 *Physica A* **144**, 257.

[Pons-Bordería et al. 1999] Pons-Bordería M J, Martínez V J, Stoyan D, Stoyan H & Saar E 1999 *Astrophys. J.* **523**, 480–491.

[Sahni et al. 1998] Sahni V, Sathyaprakash B S & Shandarin S F 1998 *Astrophys. J.* **495**, L5–L8.

[Saunders & Ballinger 2000] Saunders W & Ballinger B E 2000 *in* R. C Kraan-Korteweg, P. A Henning & H Andernach, eds, 'The Hidden Universe, ASP Conference Series' Astronomical Society of the Pacific, San Francisco, 181. astro-ph/0005606, .

[Saunders et al. 2000] Saunders W, Sutherland W J, Maddox S J, Keeble O, Oliver S J, Rowan-Robinson M, McMahon R G, Efstathiou G P, Tadros H, White S D M, Frenk C S, Carramiñana A & Hawkins M R S 2000 *Mon. Not. R. Astr. Soc.* **317**, 55–64.

[Schlegel et al. 1998] Schlegel D J, Finkbeiner D P & Davis M 1998 *Astrophys. J.* **500**, 525–553.

[Schmoldt et al. 1999] Schmoldt I M, Saar V, Saha P, Branchini E, Efstathiou G P, Frenk C S, Keeble O, Maddox S, McMahon R, Oliver S, Rowan-Robinson M, Saunders W, Sutherland W J, Tadros H & White S D M 1999 *Astron. J.* **118**, 1146–1160.

[Stoyan et al. 1995] Stoyan D, Kendall W & Mecke J 1995 *Stochastic Geometry and its Applications* John Wiley & Sons Chichester.

[Stoyan & Stoyan 2000] Stoyan D & Stoyan H 2000 *Scand. J. Satist.* **27**. 641–656.

[Szapudi et al. 1999] Szapudi I, Colombi S & Bernardeau F 1999 *Mon. Not. R. Astr. Soc.* **310**, 428–444.

[Tegmark et al. 1998] Tegmark M, Hamilton A J S, Strauss M A, Vogeley M S & Szalay A S 1998 *Astrophys. J.* **499**, 555–576.

[van Lieshout & Baddeley 1996] van Lieshout M N M & Baddeley A 1996 *Stat. Neerlandica* **50**, 344.

[White 1979] White S D M 1979 *Mon. Not. R. Astr. Soc.* **186**, 145–154.

Commentary by Rien van de Weygaert

10.7 Spatial Statistics and the Galaxy Distribution

Following the contribution by V. Martínez providing a nice and extensive overview of the large variety of statistical methods, along the lines of the excellent textbook he and E. Saar have just published[2] on methods that have been developed over the years to describe and characterize the evidently nontrivial patterns in the spatial distribution of galaxies, it may be worthwhile to add some additional characteristic issues on spatial statistics within a cosmological context. I want to point out two (and a half) issues – or, rather, details – concerning the study of cosmological point processes.

The first issue concerns the very motivation behind the cosmologists' diligence in studying the aspects of the spatial clustering of galaxies and other cosmologically relevant objects. What answer do we expect to extract from the spatial point distribution mapped out by galaxies ? How can it

[2]Statistics of the Galaxy Distribution, V. Marínez & E. Saar, 2002, Chapman & Hall

be applied towards discrimation between cosmological theories ? The basic reason behind this brings us to the *ergodic theorem*.

The second issue concerns the issue that physical theories in general make predictions on continuous physical fields. In order to mould the data into a readily interpretable form the usual practice involves the use of filtering the discrete distribution of measurements. The choice and technique of the filtering, however, is critical for this process to produce valid answers.

10.7.1 The Ergodic Theorem

The overriding reason for cosmologists to spend a large degree of attention on the spatial statistics of the galaxy distribution is that the theory of structure and galaxy formation provides us with statistical predictions, ensemble expectations, instead of predictions on the formation of particular objects. No structure formation theory will ever be able to predict an object like the Virgo or Coma cluster; they are mere realizations arisen from a primordial density field which itself is a stochastic sample from a given stochastic distribution. The latter is what a viable cosmological theory will be able to predict on the basis of appropriate physical laws and primordial cosmological processes.

A principal stumbling block for any cosmological theory therefore might be the fact that we only have one sole realization of the relevant physical system at hand. Unlike the experimental physicist testing his/her probe under the conditional circumstances of the laboratory, the cosmologist must settle for this one realization.

To solve the dilemma of comparing theoretical predictions in terms of stochastic distributions with the one realization we have at our disposal, the Universe in which we live, the *ERGODIC THEOREM* is the necessary *condition*. Stating that we may equate *Spatial Averages* with the *Ensemble Averages* predicted by the physical theories as long as we can probe a sufficient amount of representative spatial volumes in the observable Universe provides us with the means of testing cosmological theories.

In this sense it may be good to realize that it is only with the advent of major systematic redshift surveys like the Las Campanas redshift survey, the 2dF redshift survey and the SDSS survey that we can hope to compare the spatial patterns in the Universe with those of theoretical predictions, or those of numerical simulations. Uniform sky surveys like the APM survey (Maddox et al. 1990) did provide us already with sufficient information to assure ourselves of the condition of having probed a representative volume of the Universe – on the basis of the depth scaling of the two-point correlation function – for inferring statistically meaningful measures of the underlying power spectrum.

On the other hand, the interpretation of the Cosmic Microwave Background fluctuations on the largest available scales still does pose us with issues concerning “cosmic variance”.

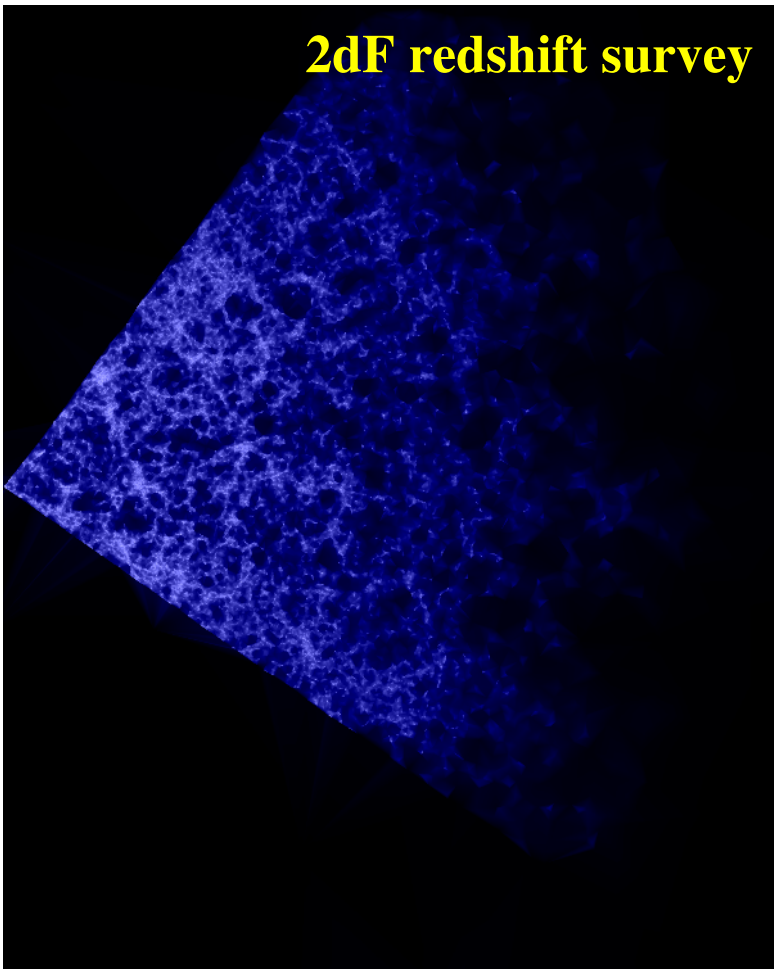


FIGURE 10.3. The Delaunay Tesselation Field Estimator reconstruction of the 2dF survey field south. The DTFE reconstruction shows, more clearly than the galaxy distribution, the coherence of the cosmic foam discretely "sampled" by the galaxy distribution. Notice the detailed and refined structure which appears to be specifically strengthened by this fully adaptive method (from Schaap & van de Weygaert 2002b). Data courtesy: the 2dF consortium.

10.7.2 Phaedo

Once we have assured ourselves of the sensibility of testing cosmological theories through statements on the probability of the realizations we find in nature, we also have to be aware of the limitations of such considerations. As cosmological theory is often concerned with "Statements of Truth" on the workings of the Universe, yet we are confined to assessments in terms of statistics, we should always be aware of possible pitfalls. It will be very hard to extract information on influences which are not taken into account in the theories being tested. That is, missing out in possibly significant parameters discarded from the statistical tests.

In this sense, the statement by Plato ($\approx$ 380 BC) might be a sobering one for claims too audacious:
"any statement about Truth based on likelihood considerations cannot be held as decisive" (freely transcribed).

10.7.3 Continuous Cosmological Fields versus Discrete Data Samples

An important aspect of spatial statistical analyses in cosmology is the fact that cosmological data quite often concern *discretely* sampled datasets while theoretical predictions concern statements on the basis of *continuous* physical fields.

The most frequent example is the galaxy distribution itself. It is supposed to reflect an underlying continuous density field. Another example concerns the measurement of cosmic flow fields, almost exclusively on the basis determined on the basis of galaxy peculiar velocities. The latter is then supposed to be a measurement of the continuous matter flow field at a few (galaxy site) discrete cosmic locations.

Discarding major overriding questions concerning the fact whether the galaxy distribution may indeed be regarded as a genuine reflection of the underlying matter field – given the fact that we still lack a convincing theory of galaxy formation and are therefore condemned to taking into account a possible "biasing" on the basis of a mere ad-hoc and heuristic description – we are still posed with the question how to infer objectively information on a continuous underlying field.

Many approaches base themselves on filtering the measured data onto some previously defined grid, which then can be processed by often sophisticated procedures yielding well-defined answers. One problem with these filtering procedures, a well-known issue, is that one usually incurs considerable loss of information through artificially defined filters which do not adapt to the inherent properties of the discrete point distribution. A telling example is how isotropically defined filters manage to dilute the signals of anisotropic features like filaments or walls. Another one is that by lack on information on inherent spatial scales in the distribution, the filter tends

to erase signatures of substructures at spatial scales lower than the filter characteristic scale. This is in particular worrisome once it gets towards distributions involving a hierarchy of scales. Precisely the latter is supposedly the case for most popular theories of structure formation.

In this comment I therefore would like to point out the virtues of a new technique that has been developed by Schaap & Van de Weygaert (2000), the *Delaunay Tessellation Field Estimator (DTFE)* of the corresponding spatial point process. Based on the earlier work by Bernardeau & Van de Weygaert (1996) for reconstruct complete volume-covering and volume-weighted velocity fields from a set of point-sampled velocities – which proved to yield a significant improvement in reproducing the statistics of the underlying continuous velocity field – the DTFE reconstructs the full and cohesive density field of which the discrete galaxy distribution is supposed to be a sparse sample. Without invoking any artificial and often structure diluting filter it is able to render both the *ANISOTROPIC* nature of the various foam elements as well as the *HIERARCHICAL* character of the distribution in full contrast (see Schaap & van de Weygaert 2002 and in this volume).

The potential promise of the DTFE may be amply appreciated from its succesful reconstruction of a density field from the galaxy distribution in the southern part of the 2dF survey (see Figure, data courtesy: 2dF consortium). Evidently, it manages to bring out any fine structural detail of the intricate and often tenuous filamentary structures. Notice the frequently razor-sharp rendition of thin edges surrounding void-like regions. Hence, it defines a volume-covering density field reconstruction that retains every structural detail, which will enable us to study in a much improved fashion the statistical and geometric properties of the foam. Indeed, it even appears to "clean" the original discrete galaxy distribution map by suppressing its shot noise contribution.

Bernardeau F., van de Weygaert R., 1996, MNRAS, 279, 693

Martínez V., Saar E., 2002, Statistics of the Galaxy Distribution, Chapman & Hall

Maddox S.J., Sutherland W.J., Efstathiou G., and Loveday J., 1990, MNRAS, 243, 692

Plato, $\approx 380BC$, Phaedo, (Penguin version)

Schaap W., van de Weygaert R., 2000, A&A, 363, L29

Schaap W., van de Weygaert R., 2002b, MNRAS, in prep.

11

Analyzing Large Data Sets in Cosmology

Alexander S. Szalay[1] **and Takahiko Matsubara**

ABSTRACT We describe the issues related to the analysis of the large scale distribution of galaxies. The emerging huge data sets from wide field sky surveys pose interesting issues, both statistical and computational. One needs to reconsider the notion of optimal statistics. We discuss the power spectrum analysis of wide area galaxy surveys using the Karhunen-Loeve transform as a case study.

11.1 Introduction

There is a very distinct trend in astronomy today, driven by the development in instrumentation, in particular detector size. The result is that the size of astronomy data is growing exponentially, doubling every year. This even exceeds the rate of Moore's law describing the speedup of computer's CPUs. This trend is resulting in the emergence of large scale surveys, like 2MASS (Two Micron Sky Survey), SDSS (Sloan Digital Sky Survey) or 2dFGRS (Two Degree Field Galaxy Redshift Survey). Soon there will be almost all-sky data in more than ten wavebands. These large scale surveys have another important characteristics: they are done by a single group, with sound statistical plans and well-controlled systematics.

As a result, the data are becoming increasingly more homogeneous, and approach a fair sample of the Universe. This trend has brought a lot of advances in the analysis of the large scale galaxy distribution. Our goal today is to reach an unheard-of level of accuracy in measuring both the global cosmological parameters and the shape of the power spectrum of primordial fluctuations.

These large, homegenous datasets are also changing the way we are approaching their analysis. Traditionally, statistics in cosmology has been primarily dealing with how to extract the most information from the small samples of galaxies we had. This is no longer the case: redshift surveys are

[1]Department of Physics and Astronomy, Johns Hopkins University

approaching the 300,000 mark today and will soon exceed a million galaxies, while angular catalogs today have samples in excess of 50 million galaxies and are soon approaching 10 billion (the proposed Large-aperture Synoptic Survey Telescope, http://www.lssto.org). Whereas the cosmic background radiation (CMB) observations of the COBE satellite had a few thousand pixels on the sky, the recently launched Microwave Anisotropy Probe (MAP, http://map.gsfc.nasa.gov) will have a million and the forthcoming Planck satellite (http://astro.estec.esa.nl/Planck) will have more than 10 million. Thus, shot noise and sample size is not an issue any more. The limiting factor in these data sets are the systematic uncertainties like photometric zero points, effects of seeing, uniformity of filter, and so forth (Eisenstein et al. 2001).

The statistical issues related to this are also changing accordingly: it is increasingly important to find techniques that can be de-sensitized to certain systematic uncertainties. Many of the traditional statistical techniques in astronomy have been focusing on 'optimal' techniques. It was generally understood, that these minimized the statistical noise in the result, but they mey have been quite sensitive to various systematics.

Statistical considerations also often assume infinite computational resources. This was not an issue in the past, when sample sizes were in the thousands. But, many of these techniques involve matrix diagonalizations or inversions, with computations scaling as the 3rd power of matrix size, so that computing costs are a billion times higher for as data samples increase thousand times. Even if the speedup of our computers keeps up with the growth of our data volumes, it cannot keep of with traditional matrix calculations. We need to find algorithms which scale more gently. In the near future, we hypothesize that only algorithms with $N \log N$ scaling will remain feasible.

As the statistical noise is going down, due to the larger samples, another effect is emerging: 'cosmic variance'. This error term reflects the fact that our observing position is fixed at the Earth, and at any time we can only study a fixed – albeit ever increasing – region of the Universe. This provides an ultimate bound on the accuracy of any astronomical measurement. We should carefully keep this effect in mind where designing new experiments.

In this paper we will discuss our goals, and the current state-of-the-art techniques in extracting cosmological information from our large data sets. In particular, we use the Karhunen-Loeve (KL) transform as a case study, showing step by step improvements needed to turn an optimal method into a useful one.

11.2 Precision Cosmology

Today we are entering the era of precision cosmology. The large new surveys with their well-defined systematics are key to this transition. There are many different measurements we can make which each constrain various combinations of the cosmological parameters. For example, the fluctuations in the CMB around multipole l values of a few hundred are very sensitive to the overall curvature of the Universe, determined by both dark matter and dark energy (de Bernardis et al. 2001, Netterfield et al. 2001).

Due to the expansion of the Universe, we can use redshifts to measure distances of galaxies. Since galaxies are not at rest in the frame of the expanding Universe, their motions cause an additional distortion in the line-of-sight coordinate. This property can be used to study the dynamics of galaxies, inferring the underlying mass density. Local redshift surveys can measure the amount of gravitating dark matter, but they are insensitive to the dark energy. Combining these different measurements (CMB + redshift surveys), each with their own degeneracy can yield considerably tighter constraints than either of them.

We know most cosmological parameters to an accuracy of about 10% or somewhat better today. Soo we will be able to reach the regime of 2-5% relative errors, through both better data but also better statistical techniques.

11.2.1 The Global Parameters

The relevant parameters include the age of the Universe, t_0, the expansion rate of the Universe, also called as Hubble's constant H_0, the deceleration parameter q_0, the density parameter Ω, and its components, the dark energy, or cosmological constant Ω_Λ, the dark matter Ω_m, the baryon fraction f_B, and the curvature Ω_k. These are not independent from one another, of course. Together, they determine the dynamic evolution of the Universe, assumed to be homogeneous and isotropic, described by a single scale factor $a(t)$:

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \left[\frac{\Omega_m}{a^3} + \Omega_\Lambda + \frac{\Omega_k}{a^2}\right] \tag{11.1}$$

Today, at $t = t_0$ the three components of the density add up to 1,

$$\Omega_m + \Omega_\Lambda + \Omega_k = 1, \tag{11.2}$$

thus for a Eucledian (flat) Universe $\Omega_m + \Omega_\Lambda = 1$.

One can use the both dynamics, luminosities and angular sizes of objects observable at high redshift to constrain the cosmological parameters. Distant supernovae have been used as standard candles to get the first hints about a large cosmological constants. The angular size of the Doppler-peaks

in the CMB fluctuations gave the first conclusive evidence for a flat universe, using the angular diameter-distance relation. The gravitational infall manifested in redshift-space distortions of galaxy surveys has been used to constrain the amount of dark matter.

These all seem to add up to a remarkably consistent picture today: a flat Universe, with

$$\Omega_\Lambda = 0.65 \pm 0.05, \qquad \Omega_m = 0.35 \pm 0.05. \tag{11.3}$$

It would be nice to have several independent measurements for the above quantities.

Recently, new interpretations have emerged about the nature of the cosmological constant – it appears that there are many possibilities, like quintessence, that can be the dark energy. Now we are facing the challenge of coming up with measuremnts and statistical techniques to distinguish among them.

11.2.2 The Fluctuation Spectrum

There are several parameters used to specify the shape of the fluctuation spectrum. These include: the amplitude σ_8, the rms value of the density fluctuations in a sphere of 8 Mpc radius; the shape parameter Γ; the redshift-distrotion parameter β; the bias parameter b; and the baryon fraction $f_B = \Omega_B/\Omega_m$. Other quantites like the neutrino mass also affect the shape of the fluctuation spectrum, although in more subtle ways than the ones above (Seljak and Zaldarriega 1996).

The shape of the fluctuation spectrum is another sensitive measure of the Big Bang at early times. Galaxy surveys have traditionally measured the fluctuations over much smaller scales (below 100 Mpc) where the fluctuations are nonlinear, and even the shape of the spectrum has been altered by gravitational infall and the dynamics of the Universe. The expected spectrum on very large spatial scales (over 200 Mpc) is revealed by precision CMB measurements. COBE showed that the spectrum is scale-invariant, reflecting the primordial initial conditions, remarkably close to the predicted Zeldovich-Harrison shape. There are several interesting physical effects that will leave an imprint on the fluctuations: the scale of the horizon at recombination, the horizon at matter-radiation equality, and the sound-horizon — all between 100-200 Mpc (Eisenstein and Hu 1998).

These scales have been rather difficult to measure: they used to be too small for CMB and too large for redshift surveys. This is rapidly changing as new higher-resolution CMB experiments are now covering sub-degree scales, corresponding to less than 100 Mpc comoving, and redshift surveys like 2dF and SDSS are reaching scales well above 300 Mpc.

We have yet to measure the overall contribution of baryons to the mass content of the Universe. We expect to find the counterparts of the CMB

Doppler bumps in galaxy surveys as well, since these are the remnants of horizon scale fluctuations in the baryons at the time of recombination. The Universe behaved like a resonant cavity at the time. Due to the dominance of the dark matter over baryons the amplitude of these fluctuations is suppressed, but with high precision measurements they should be detectable.

A small neutrino mass of a few electron volts is well within the realm of possibilities. Due to the very large cosmic abundance of relic neutrinos, even such a small mass would have an observable effect on the shape of the power spectrum of fluctuations. It is likely that the sensitivity of current redshift surveys will enable us to make a meaningful test of such a hypothesis.

One can also use large angular catalogs, projections of a 3-dimensional random field to the sphere of the sky, to measure the projected power spectrum. This technique has the advantage that dynamical distortions due to the peculiar motions of the galaxies do not affect the projected distribution. The first such analyses show a lot of promise.

11.3 Large Redshift Surveys

As mentioned in the introduction, some of the issues related to the statistical analysis of large redshift surveys, like 2dF (Percival et al. 2001) or SDSS (York et al. 2000) are quite different from their predecessors with only a few thousand galaxies. The foremost difference is that shot-noise, the usual hurdle of the past is irrelevant.

Astronomy is different from laboratory science in that we cannot change the position of the observer at will. Our experiments in studying the Universe will never approach an ensemble average, there will always be an unavoidable *cosmic variance* in our analysis. By studying a larger region of the Universe (going deeper and/or wider) can decrease this term, but it will always be present in our statistics.

The dominant source of uncertainties in large redshift surveys today is in the systematics, like photometric calibrations, or various instrumental and natural foregrounds and backgrounds. There are also effects, like nonlinearities on smaller scales or redshift space distortions, which turn an otherwise homogeneous and isotropic random process into a non-isotropic one. As a result, it is increasingly important to find statistical techniques which can reject or incorporate some of these effects into the analysis.

11.3.1 Statistical Techniques Used

The most frequent techniques used in analyzing data about spatial clustering are the two-point correlation functions and various power spectrum estimators. There is an extensive literature about the relative merits of each of the techniques. For an infinitely large data set, in principle both

techniques are equivalent. In practice, however, there are subtle differences: finite sample size affects the two estimators somewhat differently, edge effects show up in a slightly different fashion, and practical issues about computability and hypothesis testing differ for the two techniques.

The most often used estimator for the two point correlations is the LS estimator (Landy & Szalay 1992),

$$\xi(r) = \frac{DD - 2DR + RR}{RR} \tag{11.4}$$

which has a minimal variance for a Poisson process. Here DD, DR and RR describe the respective normalized pair count in a given distance range. For this estimator, and for correlation functions in general, hypothesis testing is somewhat cumbersome. If the correlation function is evaluated over a set of differential distance bins, these values are not independent, and their correlation matrix depends also on the three and four-point correlation functions which are less known than the two-point function itself. The brute-force technique involves the computation of all pairs and binning them up, so it would scale as $\mathcal{O}(N^2)$. In terms of modelling systematic effects, it is very easy to comoute the two-point correlation function between two points.

Another popular second order statistic is the power spectrum $P(k)$, usually measured by using the FKP estimator (Feldman et al. 1994). This is the Fourier-space equivalent of the LS estimator for correlation functions. It has both advantages and disadvantages over correlation functions. Hypothesis testing is much easier, since in Fourier space the power spectrum at two different wavenumbers are correlated, but the correlation is compact. It is determined by the 'window-function', the Fourier transform of the sample volume, which is usually very well-understood. For most realistic surveys the window function is rather anisotropic, making angular averaging of the three-dimensional power spectrum estimator somewhat complicated. During hypothesis testing one is using the estimated values of $P(k)$, either directly in 3D Fourier space, or compressed into quadratic sums binned by bands. Again, the 3rd and 4th order terms are appearing in the correlation matrix. The effects of systematic errors are much harder to estimate.

Hypothesis testing is usually performed in a parametric fashion, with the assumption that the underlying random process is Gaussian. We evaluate the log likelihood as

$$\ln L(\pi) = -\frac{1}{2}\mathbf{x^T C^{-1} x} - \frac{1}{2}\ln|\mathbf{C}| \tag{11.5}$$

where $\mathbf{x}$ is the data vector, and $\mathbf{C}$ is its correlation matrix, dependent on the parameter vector π. There is a fundamental lower bound on the statistical

error, given by the Fisher matrix **F**, defined by

$$F_{\alpha\beta} = \int dP(\mathbf{x}) \left(\frac{\partial L}{\partial \pi_\alpha}\right) \left(\frac{\partial L}{\partial \pi_\beta}\right). \tag{11.6}$$

The famous Cramer-Rao bound states that Var $\pi_\alpha \geq 1/\sqrt{F}_{\alpha\alpha}$. The Fisher matrix can be easily computed. This is a common tool used these days to evaluate the sensitivity of a given experiment to measure various cosmological parameters. For more detailed comparisons of these techniques, see Tegmark et al. (1998).

What would an ideal method be? It would be useful to retain much of the advantages of the 2-point correlations where the systematics are easy to model, and those of the power spectra where the modes are only weakly correlated. Third, we would like to have a hypothesis testing method where the correlation matrix does not involve 3rd and 4th order quantitites. Interestingly, there is such a method given by the Karhunen-Loeve transform. In the following subsection we describe the method, and show why does it provide such a useful framework for the analysis of the galaxy distribution, and then we discuss some of the detailed issues we had to deal with over the years to turn this into a practical tool.

One can also argue about parametric and non-parametric techniques, like using bandpowers to characterize the shape of the fluctuation spectrum. We would like to postulate, that for the specific case of redshift surveys it is not possible to have a purely non-parametric analysis. While the shape of the power spectrum itself can be described in a non-parametric way, the distortions along the redshift direction are dependent on a physical model (gravitational infall), thus without an explicit parametrization or ignoring this effect no analysis is possible.

11.4 Karhunen-Loeve Analysis of Redshift Surveys

The Karhunen-Loève (KL) eigenfunctions (Karhunen 1947, Loève 1948) provide a basis set in which the distribution of galaxies can be expanded. These eigenfunctions are computed for a given survey geometry and fiducial model of the power spectrum. For a Gaussian galaxy distribution, the KL eigenfunctions provide optimal estimates of model parameters, i.e. the resulting error bars are given by the inverse of the Fisher matrix for the parameters (Vogeley & Szalay 1996). This is achieved by finding the orthonormal set of eigenfunctions that optimally balance the ideal of Fourier modes with the finite and peculiar geometry and selection function of a real survey. In this section, we present the formalism for the KL analysis following the notation of Vogeley & Szalay (1996) who introduced this approach to galaxy clustering. The KL method has been applied to the Las

Campanas redshift survey by Matsubara, Szalay & Landy (2000) and to the PSCz survey by Hamilton, Tegmark & Padmanabhan (2001).

11.4.1 Details of the Method

The distribution of galaxies is pixelized by dividing the survey volume into a set of N cells. The data vector can then be defined as

$$d_i = n_i^{-1/2}(m_i - n_i) \tag{11.7}$$

where m_i is the number of galaxies in the i-th cell, $n_i = \langle m_i \rangle$ is the expected number of galaxies and the factor $n_i^{-1/2}$ is included to whiten the shot noise as explained below. The data vector $\boldsymbol{d}$ is expanded into the set of KL eigenfunctions $\boldsymbol{\Psi}_n$ as

$$\boldsymbol{d} = \sum_n B_n \boldsymbol{\Psi}_n. \tag{11.8}$$

The eigenfunctions $\boldsymbol{\Psi}_n$ are obtained by solving the eigenvalue problem (Vogeley & Szalay 1996):

$$\boldsymbol{R}\boldsymbol{\Psi}_n = \lambda_n \boldsymbol{\Psi}_n, \tag{11.9}$$

where $\lambda_n = \langle B_n^2 \rangle$ and

$$R_{ij} = \langle d_i d_j \rangle = n_i^{1/2} n_j^{1/2} \omega_{ij} + \delta_{ij} \ . \tag{11.10}$$

The second term is the whitened shot noise correlation matrix. The correlation matrix $\boldsymbol{R}$ is computed for a fiducial model using the cell-averaged angular correlation function

$$\omega_{ij} \equiv \frac{1}{V_i V_j} \int \int d^2\theta_i \ d^2\theta_j \ \omega(|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j|) \ , \tag{11.11}$$

where the integral extends over the i-th and j-th cells, and V_i and V_j are the corresponding cell volumes. Forming the eigenmodes $\boldsymbol{\Psi}_n$ requires assuming an a priori model for $\omega(\theta)$ but, as discussed by Vogeley & Szalay (1996), this choice does not bias the estimated parameters below.

The KL eigenmodes defined above satisfy the conditions of orthonormality $\boldsymbol{\Psi}_n \cdot \boldsymbol{\Psi}_m = \delta_{nm}$, and statistical orthogonality, $\langle B_n B_m \rangle = \langle B_n^2 \rangle \delta_{nm}$. Further, they sort the data in decreasing signal-to-noise ratio if they are ordered by the corresponding eigenvalues (Vogeley & Szalay 1996). What this means in the measurement of model parameters will be clarified below.

The KL expansion is used to estimate model parameters by computing the covariance matrix $\boldsymbol{C}$ of the KL coefficients. We use the first N_{mode} of the KL eigenmodes and choose to parameterize the model. The theoretical covariance matrix is then given by

$$C_{mn} = \langle B_m B_n \rangle_{\rm model} = \boldsymbol{\Psi}_m^T \boldsymbol{R}_{\rm model} \boldsymbol{\Psi}_m \ . \tag{11.12}$$

11.4.2 Advantages of the KL Transform

The KL transform is often called *optimal subspace filtering* (Therrien 1992) describing the fact that during the analysis some of the modes are discarded. This does offer distinct advantages. If the measurement is composed of a signal of interest (gravitational clustering) superposed on various backgrounds (shot-noise, selection effects, photometric errors, etc.) which have slightly different statistical properties, the diagonalization of the correlation matrix can potentially segregate these different types of processes into their own subspaces. If we select our subspace carefully, we can actually improve on the signal to noise of our analysis.

The biggest advantage is that hypothesis testing is very easy and elegant. First of all, all KL modes are orthogonal to one another, even if the survey geometry is extremely anisotropic. Of course, none of the KL modes can be narrower than the survey window, and they shape is clearly affected by the survey geometry. The orthogonality of the modes represents a repulsion between the modes; they cannot get too close, otherwise they could not be orthogonal. As a result, the KL modes are dense-packed into Fourier-space, thus optimally representing the information enabled by the survey geometry.

Secondly, the KL transform is a linear transformation. If we do our likelihood testing over the KL-transform of the data, the correlation matrix involved in the likelihood computation contains only second order quantities. Thus the problems with 3 and 4-point correlation functions do not apply at all. All these advantages became very apperent when we applied the KL method to real data.

11.4.3 Redshift Space Distortions

Since galaxies are observed in redshift space, it is essential that we account for the redshift space distortions. This is straight-forward for surveys of small angular extent (plane-parallel case, see Kaiser 1987). It is much harder to derive a similar expression for wide-angle surveys although finally several alternative formulations, leading to identical results, have been proposed. These expressions involve the redshift-space distortion parameter β which describes the relation between the large scale gravitational infall and the overdensity.

The calculation has been extended by Matsubara and Suto (1996) to the case of higher redshifts. The SDSS survey will have about 100,000 galaxy redshift for luminous ellipticals, with a typical redshift of about 0.4. At this distance the effects of cosmological curvature are becoming important (light propagates along geodesics) and this may result in a distortion of the transverse coordinates since we can only observe angles. Interstingly, it was possible to derive a closed expression for the two-point correlation function in curved spacetime, when the two lines-of-sight are separated at an arbi-

trary angle. This expression can only be integrated numerically, but luckily a very accurate numerical fitting formula has been found (Matsubara & Suto 1996, Matsubara 2001).

The forward computation of the KL transform is very advantageous in this respect; if we have an analytic expression for the two-point correlation function in redshift space, the problem is solvable numerically. If we tried to evaluate the same expression in Fourier space, we get considerably more complicated expressions. We experimented with different pixel shapes, from the tophat window to higher order Epanichnikov kernels.

11.4.4 Pixelization

The detailed calculation of the likelihood can be quite demanding. A typical likelihood fit will involve the computation of the correlation matrix at a few hundred thousand values of the parameter vector. Our first computations could easily take several days on relatively fast computers. In the beginning, we used a contiguous layout of rectangular pixels on the sky and slightly elongated splits along the radial direction. The calculation of the correlation matrix was rather complex, since it involved Monte-Carlo integration for the expectation of $\xi(r)$ over the finite sized cells. For more distant cells, it was enough to use the correlation function evaluated at the center of the pair of cells. In order to speed up the calculation for these 'hard' pixels, we have built a lookup table indexed by the relative geometry of the two cells. This resulted in a 100-fold speedup in our computations.

The next breakthrough came with the introduction of spherical pixels. If we use pixels with spherical symmetry, the computation of the average of the correlation function when its two endpoints are drawn from the pixels can be written as a convolution with the kernel corresponding to the pixel shape. This means that, by including a multiplicative factor in the power spectrum, we can directly evaluate the expectation value of the correlation function. We have created a lookup table for the correlation function with this kernel, and then used a cubic spline interpolator to get the precise values. This has yielded another order of magnitude speedup. Now we have a toolbox where we can easily run a full analysis of a given data set in a matter of a few hours.

11.4.5 Survey Design and Accuracy

It is interesting to consider how different choices, like the intrinsic clustering strength and abundance of objects in a cosmological sample, affect the accuracy of how the cosmological parameters can be determined. In the maximum-likelihood method, one can easily evaluate the expected parameter estimation errors in any sample from the Fisher information matrix. We have used our KL technique to consider the seven-dimensional Fisher

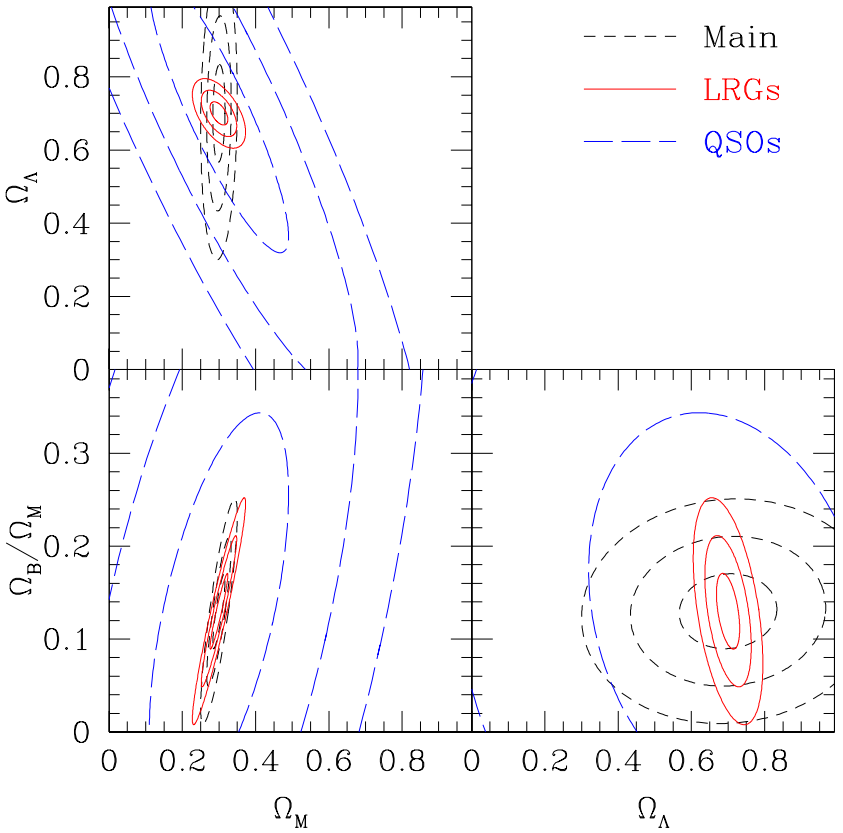


FIGURE 11.1. The marginalized concentration ellipses for three cosmological parameters (Matsubara & Szalay 2002)

matrix for three types of objects in the SDSS survey: main galaxies, luminous red galaxies (LRGs) and quasars. To illustrate the behavior of the multi-dimensional Fisher matrix, we have used concentration ellipses in marginalized two-dimensional parameter space.

In a recent paper (Matsubara a& Szalay 2002), we divided the survey volume into generic boxes in order to simplify the Fisher matrix estimation. We ignored the correlations between these sub-regions, so the constraints will improve somewhat if those correlations are properly included. However, the inversion of the resulting huge matrices can become extremely time-consuming. The use of the KL transform is a practical strategy in this case. Such methods can also be used in a targeted data-compression role to find linear combinations of counts which retain as much information about the parameters as possible.

The choice of the cell radius is somewhat arbitrary in this work. We

choose the spherical cell with radius of 10 Mpc for galaxies and LRGs, which is the border of the linear regime. With a larger cell radius, the validity of the linear theory increases and the shot noise is reduced. The cosmic variance, however, increases with cell radius. The parameter estimation is dominated by the highest signal-to-noise modes which are at large wavelengths, in particular for the case of the LRG sample. The high frequency modes close to pixel scales mostly contain shot noise after the KL transformation. As a result, we believe our conclusions are not sensitive to the choice of the cell radius. A fully accurate determination of the optimal choice of the cell radius depends on the behavior of the nonlinear effects, so that a comparison with numerical simulation is needed, beyond the scope of the current work.

We have considered three subsets of the SDSS redshift data, spanning a wide range of depth, sampling density and intrinsic clustering strength. We found, that for measuring cosmological parameters in the linear regime there is a clear optimum, represented by the intermediate-redshift LRG's. The low spatial density of quasars is not overcome by their much larger depth, and the relatively small depth of the main SDSS galaxies is not compensated by their high sampling density – the redshift is not high enough to test curvature, and their cosmic variance is too large. The LRG sample, much smaller in numbers than the main sample and much shallower than the quasars, is an excellent compromise between sampling density and cosmological depth. The constraints derived from the LRGs are much tighter than for the other two samples.

The advantage of these intermediate-redshift objects, and the logic behind this optimum goes beyond the SDSS. In designing future redshift surveys, it is important to find the right balance between the density of objects and the survey depth. Their interplay can be quite complex, as we have shown here. The relation between accuracy and sky coverage is simple and can be estimated analytically.

11.5 Trends and Computational Issues

The problems we are facing with the exponentially growing astronomy data are serious. Most statistical techniques labeled as 'optimal' are based on several assumptions which have been correct in the past but not necessarily in the near future. The assumptions include that the dominant contribution to the variance is statistical, and that the computational resources are infinite compared to the cost of computation, and they ignore the cosmic variance.

Many of these optimal algorithms are based upon maximum likelihood estimators, and thus they involve inversions of large matrices, an approximately N^3 operation. The increase in CPU power will not be able to keep

up with such a scaling.

What are the possibilities? We can use clever data structures, borrowed from computer science to pre-organize our data into a tree-hierarchy, and having the computational cost dominated by the cost of sorting, an $N \log N$ process. This is the approach taken by A. Moore and collaborators in their tree-code (see paper by R. Nichol et al. in this volume).

Another approach might be to use approximate statistics, as advocated by I. Szapudi (2001). In the presence of a cosmic variance term, an alorithm that spends an enormous amount of CPU time to minimize the statistical variance to a level substantially below the cosmic variance can be very wasteful. One can define a cost function that includes all terms in the variance and a computational cost $Q(\epsilon)$, as a function of the accuracy ϵ of the estimator. Minimizing this cost-function $C(\epsilon)$ will give the best possible results, given the nature of the data and our finite computational resources.

$$C(\epsilon) = \sigma^2_{cosmic} + \sigma^2_{stat}(\epsilon) + Q(\epsilon) \tag{11.13}$$

We expect to see more and more of these algorithms emerging over the next few years. One nice example of these these ideas is the fast CMB analysis developed by Szapudi et al. (2002) which will reduce the computations for a survey of the size expected from the Planck satellite from 10 million years to approximately 1 day!

11.6 Summary

Several important new trends are becoming apparent in modern cosmology and astrophysics: the amount of data available is doubling every year, the data are well understood, and much of the low level processing is already done by the time the data is published. This makes it much easier to perform additional statistical analyses.

At the same time many of the current outstanding problems in cosmology are inherently statistical, either studying the distributions of typical objects (in parametric or non-parametric fashion) or finding the atypical objects: extremes and/or outliers. Many of the necessary algorithms are scaling with powers of N, the size of the data. Today, we find that more and statistical tools use advanced data structures and/or approximate techniques to achive fast computability.

In the not too distant future, when our data sets are going through another order of magnitude growth, only $N \log N$ algorithms will remain feasible — the cost of computation will become a very important ingredient of an optimal algorithm. Such an evolution in our approach to astrostatistics can only be accomplished with an active and intense collaboration of astronomers, statisticians and computer scientists.

Acknowledgements We would like to acknowledge useful discussions

with Dan VandenBerk, Daniel Eisenstein and Adrian Pope. TM acknowledges support from the Ministry of Education, Culture, Sports, Science, and Technology, Grant-in-Aid for Encouragement of Young Scientists, 13740150, 2001. AS acknowledges support from grants NSF AST-9802 980 and NASA LTSA NAG-53503.

Alcock, C. & Paczyński, B. 1979, Nature, 281, 358

Ballinger, W. E. & Peacock, J. A. & Heavens, A. F. 1996, MNRAS, 282, 877

de Bernardis, P. et al. 2000, Nature, 404, 955

Eisenstein, D. J. & Hu, W. 1998, ApJ, 496, 605

Feldman, H. A., Kaiser, N. & Peacock, J. A. 1994, ApJ, 426, 23

Hamilton, A. J. S. 1992, ApJ, 385, L5

Hamilton, A. J. S., Tegmark, M., Padmanabhan, N., 2000, MNRAS, 317, L23

Kaiser, N. 1987, MNRAS, 227, 1

Karhunen, H. 1947, Ann. Acad. Science Finn. Ser. A.I. 37

Landy, S.D. and Szalay, A.S. 1992, ApJ, 394, 25

Loève, M. 1948, Processes Stochastiques et Mouvement Brownien, (Hermann, Paris France)

Matsubara, T., Szalay, A. S., Landy, S. D., 2000, ApJ, 535, L1

Matsubara, T. & Suto, Y. 1996, ApJ, 470, L1

Matsubara, T. 2000, ApJ, 535, 1

Matsubara, T. & Szalay, A. S. 2001, ApJ, 556, L67

Matsubara, T. & Szalay, A. S. 2002, ApJ, 556, 67, also astro-ph/0203358

Netterfield, C.B. et al. 2001, submitted to ApJ, astro-ph/0104460

Peebles, P. J. E. 1980, The Large-Scale Structure of the Universe (Princeton: Princeton University Press)

Percival, W.J., et al. 2001, MNRAS, 327, 1297, astro-ph/0105252

Seljak, U. & Zaldarriaga, M. 1996, ApJ, 469, 437

Szapudi, I., Prunet, S. & Colombi, S. 2001, ApJ, 561, L11, also astro-ph/0107383

Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997, ApJ, 480, 22

Tegmark,M. et al 1998, ApJ, 499, 555

Therrien, C. W. 1992, Discrete Random Signals and Statistical Signal Processing, (New Jersey: Prentice-Hall).

Vogeley, M.S., Szalay, A.S., 1996, ApJ, 465, 34

York, D. G. et al. 2000, AJ, 120, 1579

12

The Cosmic Foam: Stochastic Geometry and Spatial Clustering across the Universe

Rien van de Weygaert[1]

ABSTRACT Galaxy redshift surveys have uncovered the existence of a salient and pervasive foamlike pattern in the distribution of galaxies on scales of a few up to more than a hundred Megaparsec. The significance of this frothy morphology of cosmic structure has been underlined by the results of computer simulations. These suggest the observed cellular patterns to be a prominent and natural aspect of cosmic structure formation for a large variety of scenarios within the context of the gravitational instability theory of cosmic structure formation.

We stress the importance of stochastic geometry as a branch of mathematical statistics particularly suited to model and investigate nontrivial spatial patterns. One of its key concepts, Voronoi tessellations, represents a versatile and flexible mathematical model for foamlike patterns. Based on a seemingly simple definition, Voronoi tessellations define a wealthy stochastic network of interconnected anisotropic components, each of which can be identified with the various structural elements of the cosmic galaxy distribution. The usefulness of Voronoi tessellations is underlined by the fact that they appear to represent a natural asymptotic situation for a range of gravitational instability scenarios of structure formation in which void-like regions are prominent.

Here we describe results of an ongoing thorough investigation of a variety of aspects of cosmologically relevant spatial distributions and statistics within the framework of Voronoi tessellations. Particularly enticing is the recent finding of a profound scaling of both clustering strength and clustering extent for the distribution of tessellation nodes, suggestive for the clustering properties of galaxy clusters. This is strongly suggestive of a hitherto unexpected fundamental and profound property of foamlike geometries. In a sense, cellular networks may be the source of an intrinsic "geometrically biased" clustering.

[1]Kapteyn Institute, University of Groningen

12.1 Introduction

Macroscopic patterns in nature are often due the collective action of basic, often even simple, physical processes. These may yield a surprising array of complex and genuinely unique physical manifestations. The macroscopic organization into complex spatial patterns is one of the most striking. The rich morphology of such systems and patterns represents a major source of information on the underlying physics. This has made them the subject of a major and promising area of inquiry. However, most such studies still reside in a relatively youthful state of development, hampered by the fact that appropriate mathematical machinery for investigating and solidly characterizing the geometrical intricacies of the observed morphologies is not yet firmly in place.

In an astronomical context one of the most salient geometrically complex patterns is that of the foamlike distribution of galaxies, revealed by a variety of systematic and extensive galaxy redshift surveys. Over the two past decades, these galaxy mapping efforts have gradually established the frothy morphology as a universal aspect of the spatial organization of matter in the Univers. Comprising features on a typical scale of tens of Megaparsec, it offers a direct link to the matter distribution in the primordial Universe. The cosmic web is therefore bound to contain a wealth of information on the cosmic structure formation process. It will therefore represent a key to unravelling one of the most pressing enigmas in modern astrophysics, the rise of the wealth and variety of structure in the present-day Universe from a almost perfectly smooth, virtually featureless, pristine cosmos.

However, a lack of straightforward quantitative measures of such patterns has yet prevented a proper interpretation, or indeed identification, of all relevant pieces of information. Quantitative analysis of matter distribution has been largely restricted to first order galaxy clustering measures, useful in evaluating gross statistical properties of the matter distribution but inept for characterizing the intricate foamlike morphologies observed on Megaparsec scales.

Here we will address the meaning and interpretation of the cellular morphology of the cosmic matter distribution. Prominent as it is, its assessment rarely exceeds mere qualitative terminology, seriously impeding the potential exploitation of its content of significant information. One of the most serious omissions concerns a proper appreciation and understanding of the physical and statistical repercussions of the nontrivial cellular geometry. This propelled us to focus on this important aspect, for which we were impelled to invoke ideas and concepts from the relevant field of mathematics, stochastic geometry. Particularly fruitful has been our application and investigation of Voronoi tessellations, a central concept in this mathematical branch addressing the systematics of geometrical entities in a stochastic setting. The phenomenological similarity of Voronoi foams to the cellular morphology seen in the galaxy distribution justifies further exploration of

its virtues as a model for cosmic structure. In the following we will indicate that such similarity is a consequence of the tendency of gravity to shape and evolve structure emerging from a random distribution of tiny density deviations into a network of anisotropically contracting features. Its application gets solidly underpinned by a thorough assessment of the implications for spatial clustering, vindicating the close resemblance of Voronoi foams to the frothy patterns in the observed reality. It is within the context of testing its spatial statistical properties that unexpected profound 'scaling' symmetries were uncovered, shedding new light on the issue of "biased" spatial clustering.

12.2 Patterns in the Galaxy Distribution: the Cosmic Foam

One of the most striking examples of a physical system displaying a salient geometrical morphology, and the largest in terms of sheer size, is the Universe as a whole. The past few decades have revealed that on scales of a few up to more than a hundred Megaparsec, galaxies conglomerate into intriguing cellular or foamlike patterns that pervade throughout the observable cosmos. A dramatic illustration is the map of the 2dF Galaxy Redshift Survey and the newest results of the SDSS survey (see contribution M. Strauss). The recently published map of the distribution of more than 150,000 galaxies in a narrow region on the sky yielded by the 2dF – two-degree field – redshift survey. Instead of a homogenous distribution, we recognize a sponge-like arrangement, with galaxies aggregating in filaments, walls and nodes on the periphery of giant voids.

This frothy geometry of the Megaparsec Universe is evidently one of the most prominent aspects of the cosmic fabric, outlined by galaxies populating huge *filamentary* and *wall-like* structures, the sizes of the most conspicuous one frequently exceeding $100\mathrm{h}^{-1}$ Mpc. The closest and best studied of these massive anisotropic matter concentrations can be identified with known supercluster complexes, enormous structures comprising one or more rich clusters of galaxies and a plethora of more modestly sized clumps of galaxies. A prominent and representative nearby specimen is the Perseus-Pisces supercluster, a $5h^{-1}$ wide ridge of at least $50h^{-1}$ Mpc length, possibly extending out to a total length of $140h^{-1}$ Mpc. In addition to the presence of such huge filaments the galaxy distribution also contains vast planar assemblies. A striking example of is the *Great Wall*, a huge planar assembly of galaxies with dimensions that are estimated to be of the order of $60h^{-1} \times 170h^{-1} \times 5h^{-1}$ Mpc (Geller & Huchra 1989). Within and around these anisotropic features we find a variety of density condensations, ranging from modest groups of a few galaxies up to massive compact *galaxy clusters*. The latter stand out as the most massive fully collapsed

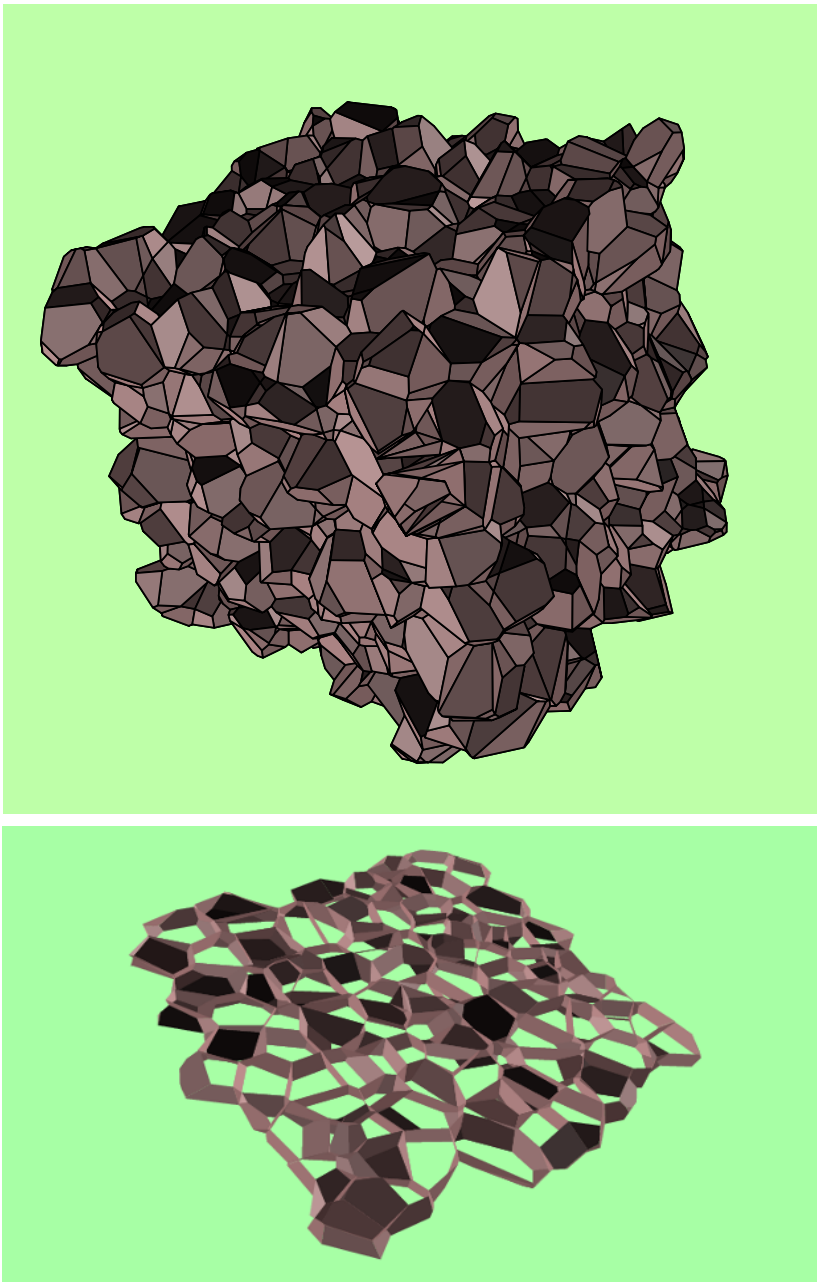


FIGURE 12.1. A full 3-D tessellation comprising 1000 Voronoi cells/polyhedra generated by 1000 Poissonian distributed nuclei. Courtesy: Jacco Dankers

and virialized objects in the Universe. In nearby representatives like the Virgo and Coma cluster typically more than a thousand galaxies have been identified within a radius of a mere $1.5h^{-1}$ Mpc around the core. They may be regarded as a particular population of cosmic structure beacons as they typically concentrate near the interstices of the cosmic web, *nodes* forming a recognizable tracer of the cosmic matter distribution out to vast distances (e.g. Borgani & Guzzo 2001). Complementing this cosmic inventory leads to the existence of large *voids*, enormous regions with sizes in the range of $20 - 50h^{-1}$ Mpc that are practically devoid of any galaxy, usually roundish in shape. The earliest recognized one, the Boötes void (Kirshner et al. 1981, 1987), a conspicuous almost completely empty spherical region with a diameter of around $60h^{-1}$Mpc, is still regarded as the canonic example. The role of voids as key ingredients of the cosmic matter distribution has since been convincingly vindicated in various extensive redshift surveys, up to the recent results produced by 2dF redshift survey and the Sloan redshift surveys.

Of utmost significance for our inquiry into the issue of cosmic structure formation is the fact that the prominent structural components of the galaxy distribution – clusters, filaments, walls and voids – are not merely randomly and independently scattered features. On the contrary, they have arranged themselves in a seemingly highly organized and structured fashion, the *cosmic foam*. They are woven into an intriguing *foamlike* tapestry that permeates the whole of the explored Universe. Voids are generically associated with surrounding density enhancements. In the galaxy distribution they represent both contrasting as well as complementary components ingredients, the vast under-populated regions, (the *voids*), being surrounded by *walls* and *filaments*. At the intersections of the latter we often find the most prominent density enhancements in our universe, the *clusters* of galaxies.

12.3 Gravitational Foam Formation and Bubble Dynamics.

Foamlike patterns have not only been confined to the real world. Equally important has been the finding that foamlike patterns do occur quite naturally in a vast range of structure formation scenarios within the context of the generic framework of gravitational instability theory. Prodded by the steep increase in computing power and the corresponding proliferation of ever more sophisticated and extensive simulation software, a large range of computer models of the structure formation process have produced telling images of similar foamlike morphologies. They reveal an evolution proceeding through stages characterized by matter accumulation in structures with a pronounced cellular morphology.

The generally accepted theoretical framework for the formation of struc-

ture is that of gravitational instability. The formation and moulding of structure is ascribed to the gravitational growth of tiny initial density- and velocity deviations from the global cosmic density and expansion. An important aspect of the gravitational formation process is its inclination to progress via stages in which the cosmic matter distribution settles in striking anisotropic patterns. Aspherical overdensities, on any scale and in any scenario, will contract such that they become increasingly anisotropic, as long as virialization has not yet set in. At first they turn into a flattened 'pancake', possibly followed by contraction into an elongated filament. Such evolutionary stages precede the final stage in which a virialized object, e.g. a galaxy or cluster, will emerge. This tendency to collapse anisotropically finds its origin in the intrinsic primordial flattening of the overdensity, augmented by the anisotropy of the gravitational force field induced by the external matter distribution, i.e. by tidal forces. Naturally, the induced anisotropic collapse has been the major agent in shaping the cosmic foam-like geometry.

Inspired by early computer calculations, Icke (1984) pointed out that for the understanding of the formation of the large coherent patterns pervading the Universe it may be more worthwhile to direct attention to the complementary evolution of underdense regions. By contrast to the overdense features, the low-density regions start to take up a larger and larger part of the volume of the Universe. Icke (1984) then made the interesting observation that the arguments for the dynamics and evolution of slightly anisotropic – e.g. ellipsoidal – primordial overdensities are equally valid when considering the evolution of *low*-density regions. The most important difference is that the sense of the final effect is reversed. The continuously stronger anisotropy of the force field in collapsing ellipsoidal leads to the characteristics tendency for *slight initial asphericities to get amplified during the collapse*, the major internal mechanism for the formation of the observed filaments in the galaxy distribution. By contrast, a void is effectively a region of negative density in a uniform background. Therefore, they will expand as the overdense regions collapse, while *slight asphericities decrease as the voids become larger*. This can be readily appreciated from the fact that with respect to an equally deep spherical underdensity, an ellipsoidal void has a decreased rate of expansion along the longest axis of the ellipsoid and an increased rate of expansion along the shortest axis. Together with the implied *Hubble-type velocity field*, voids will thus behave like low-density 'super-Hubble' expanding patches in the Universe. To describe this behaviour we coined the term "Bubble Theorem" (Icke 1984).

Evidently, we have to be aware of the serious limitations of the ellipsoidal model. It grossly oversimplifies in disregarding important aspects like the presence of substructure in and the immediate vicinity of peaks and dips in the primordial density field. Still, it is interesting to realize that in many respects the homogeneous model is a better approximation for underdense regions than it is for overdense ones. Voids expand and get drained, and

the interior of a (proto)void rapidly flattens out, which renders the validity of the approximation accordingly better. Such behaviour was clearly demonstrated in circumstances of voids embedded in a full complex general cosmic density field (see e.g. Van de Weygaert & van Kampen 1993, their Fig. 16). Their systematic study also showed how voids in general will evolve towards a state in which they become genuine *"Superhubble Bubbles"*.

In realistic circumstances, expanding voids will sooner or later encounter their peers or run into dense surroundings. The volume of space available to a void for expansion is therefore restricted. Voids will also be influenced by the external cosmic mass distribution, and substructure may represent an additional non-negligible factor within the void's history. In general, we deal with a complex situation of a field of expanding voids and collapsing peaks, of voids and peaks over a whole range of sizes and masses, expanding at different rates and at various stages of dynamical development. For the purpose of our geometric viewpoint, the crucial question is whether it is possible to identify some characteristic and simplifying elements within such a complex. Indeed, simulations of void evolution (e.g. Dubinski et al. 1993) represent a suggestive illustration of a hierarchical process akin to the *void hierarchy* seen in realistic simulations (e.g. Van de Weygaert 1991b). It shows the maturing of small-scale voids until their boundaries would reach a shell-crossing catastrophe, after which they merge and dissolve into a larger embedding void. This process gets continuously repeated as the larger parent voids in turn dissolve into yet larger voids. For a primordial Gaussian density field, corresponding analytical calculations (Sheth & Van de Weygaert 2002) then yield a void size distribution (broadly) peaked around a characteristic void size.

A bold leap then brings us to a geometrically interesting situation. Taking the voids as the dominant dynamical component of the Universe, and following the "Bubble Theorem", we may think of the large scale structure as a close packing of spherically expanding regions. Then, approximating a peaked void distribution by one of a single scale, we end up with a situation in which the matter distribution in the large scale Universe is set up by matter being swept up in the bisecting interstices between spheres of equal expansion rate. This *ASYMPTOTIC* description of the cosmic clustering process leads to a geometrical configuration that is one of the main concepts in the field of stochastic geometry: *VORONOI TESSELLATIONS.*

12.4 Voronoi Tessellations: the Geometric Concept

A Voronoi tessellation of a set of nuclei is a space-filling network of polyhedral cells, each of which delimits that part of space that is closer to its nucleus than to any of the other nuclei. In three dimensions a Voronoi foam consists of a packing of Voronoi cells, each cell being a convex polyhedron

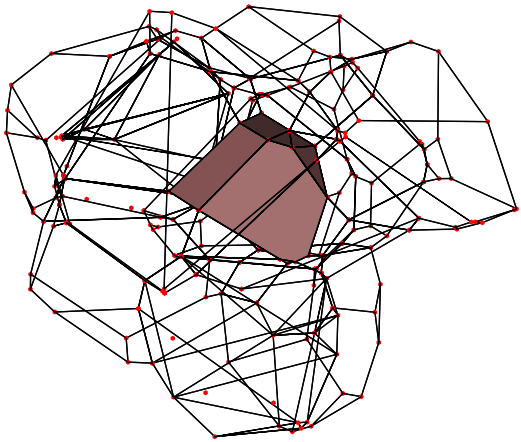


FIGURE 12.2. Wireframe illustration of interrelation between various Voronoi tessellation elements. The central "Voronoi cell" is surrounded by its wire-frame depicted "contiguous" Voronoi neighbours. The boundaries of the cells are the polygonal "Voronoi walls". The wire edges represent the Voronoi edges. The "Voronoi vertices", indicated by red dots, are located at each of the 2 tips of a Voronoi edge, each of them located at the centre of the circumsphere of a corresponding set of four nuclei. Courtesy: Jacco Dankers.

enclosed by the bisecting planes between the nuclei and their neighbours. A Voronoi foam consists of four geometrically distinct elements: the polyhedral cells (*voids*), their walls (*pancakes*), edges (*filaments*) where three walls intersect, and nodes (*clusters*) where four filaments come together.

Formally, each Voronoi region Π_i is the set of points which is nearer to nucleus i than to any of the other nuclei j in a set Φ of nuclei $\{x_i\}$ in d-dimensional space $\Re^d$, or a finite region thereof, $\Pi_i = \{\vec{x} | d(\vec{x}, \vec{x}_i) < d(\vec{x}, \vec{x}_j)\ ,\ \forall\, j \neq i\}$, where $\vec{x}_j$ are the position vectors of the nuclei in Φ, and $d(\vec{x}, \vec{y})$ the Euclidian distance between $\vec{x}$ and $\vec{y}$ (evidently, one can extend the concept to any arbitrary distance measure). From this basic definition, we can directly infer that each Voronoi region Π_i is the intersection of the open half-spaces bounded by the perpendicular bisectors (bisecting planes in 3-D) of the line segments joining the nucleus i and any of the the other nuclei. This implies a Voronoi region Π_i to be a convex polyhedron (or polygon when in 2-D), a *Voronoi polyhedron*. The complete set of $\{\Pi_i\}$ constitute a space-filling tessellation of mutually disjunct cells in d-dimensional space $\Re^d$, the *Voronoi tessellation* $\mathcal{V}(\Phi)$ relative to Φ. A good impression of the morphology of a complete Voronoi tessellation can be seen in figure 1, a tessellation of 1000 cells generated by a Poisson distribution of 1000 nuclei in a cubic box.

Taking the three-dimensional tessellation as the archetypical representa-

tion of structures in the physical world, we can identify four constituent *elements* in the tessellation, intimately related aspects of the full Voronoi tessellation. In addition to (1) the polyhedral *Voronoi cells* Π_i these are (2) the polygonal *Voronoi walls* Σ_{ij} outlining the surface of the Voronoi cells, (3) the one-dimensional *Voronoi edges* Λ_{ijk} defining the rim of both the Voronoi walls and the Voronoi cells, and finally (4) the *Voronoi vertices* V_{ijkl} which mark the limits of edges, walls and cells. While each Voronoi cell is defined by one individual nucleus in the complete set of nuclei Φ, each of the polygonal Voronoi walls Σ_{ij} is defined by two nuclei i and j, consisting of points $\vec{x}$ having equal distance to i and j. The Voronoi wall Σ_{ij} is the subregion of the full bisecting plane of i and j which consists of all points $\vec{x}$ closer to both i and j than other nuclei in Φ. In analogy to the definition of a Voronoi wall, a Voronoi edge Λ_{ijk} is a subregion of the equidistant line defined by three nuclei i, j and k, the subregion consisting of all points $\vec{x}$ closer to i, j and k than to any of the other nuclei in Φ. Evidently, it is part of the perimeter of three walls as well, Σ_{ij}, Σ_{ik} and Σ_{jk}. Pursuing this enumeration, Voronoi vertices V_{ijkl} are defined by four nuclei, i, j, k and l, being the one point equidistant to these four nuclei and closer to them than to any of the other nuclei belonging to Π_i. Note that this implies that the circumscribing sphere defined by the four nuclei does not contain any other nuclei. To appreciate the interrelation between these different geometric aspects, figure 2 lifts out one particular Voronoi cell from a clump of a dozen Voronoi cells. The central cell is the one with its polygonal Voronoi walls surface-shaded, while the wire-frame representation of the surrounding Voronoi cells reveals the Voronoi edges defining their outline and the corresponding vertices as red dots. Notice, how the distribution of vertices, generated by the stochastic point process of nuclei, is in turn a new and uniquely defined point process, that of the *vertices* !!!

12.5 Voronoi Tessellations: the Cosmological Context

In the cosmological context *Voronoi Tessellations* represent the *Asymptotic Frame* for the ultimate matter distribution distribution in any cosmic structure formation scenario, the skeleton delineating the destination of the matter migration streams involved in the gradual buildup of cosmic structures. The premise is that some primordial cosmic process generated a density fluctuation field. In this random density field we can identify a collection of regions where the density is slightly less than average or, rather, the peaks in the primordial gravitational potential perturbation field. As we have seen, these regions are the seeds of the voids. These underdense patches become "expansion centres" from which matter flows away until it runs into its surroundings and encounters similar material flowing out of adjacent voids. Notice that the dependence on the specific structure forma-

tion scenario at hand is entering via the spatial distribution of the sites of the density dips in the primordial density field, whose statistical properites are fully determined by the spectrum of primordial density fluctuations.

Matter will collect at the interstices between the expanding voids. In the asymptotic limit of the corresponding excess Hubble parameter being the same in all voids, these interstices are the bisecting planes, perpendicular bisecting the axes connecting the expansion centres. For any given set of expansion centres, or *nuclei*, the arrangement of these planes define a unique process for the partitioning of space, a *Voronoi tessellation* (Voronoi 1908, see Fig. 1 and 2). A particular realisation of this process (i.e. a specific subdivision of N-space according to the Voronoi tessellation) may be called a *Voronoi foam* (Icke & Van de Weygaert 1987). Within such a cellular framework the interior of each "*VORONOI CELL*" is considered to be a void region. The planes forming the surfaces of the cells are identified with the "*WALLS*" in the galaxy distribution (see e.g. Geller & Huchra 1989). The "*EDGES*" delineating the rim of each wall are to be identified with the filaments in the galaxy distribution. In general, what is usually denoted as a flattened "supercluster" or cosmic "wall" will comprise an assembly of various connecting walls in the Voronoi foam. The elongated "superclusters" or "filaments" will usually consist of a few coupled edges (Fig. 3 clearly illustrates this for the Voronoi kinematic model). Finally, the most outstanding structural elements are the "*VERTICES*", tracing the surface of each wall, outlining the polygonal structure of each wall and limiting the ends of each edge. They correspond to the very dense compact nodes within the cosmic network, amongst which the rich virialised Abell clusters form the most massive representatives.

Cosmologically, the great virtue of the Voronoi foam is that it provides a conceptually simple model for a cellular or foamlike distribution of galaxies, whose ease and versatility of construction makes it an ideal tool for statistical studies. By using such geometrically constructed models one is not restricted by the resolution or number of particles. A cellular structure can be generated over a part of space beyond the reach of any N-body experiment. Even though the model does not and cannot address the galaxy distribution on small scales, it is nevertheless a useful prescription for the spatial distribution of the walls and filaments themselves. This makes the Voronoi model particularly suited for studying the properties of galaxy clustering in cellular structures on very large scales, for example in very deep pencil beam surveys, and for studying the clustering of clusters in these models.

12.6 Voronoi Galaxy Distributions

Having established the cosmological context for Voronoi tessellations in the form of, approximate and asymptotic, skeletal template for the large-

scale mass distribution we set about to generate the corresponding matter distributions. Matter is supposed to aggregate in and around the various geometrical elements of the cosmic frame, such as the walls, the filaments and the vertices.

It is the stochastic yet non-Poissonian geometrical distribution of the walls, filaments and clusters embedded in the cosmic framework which generates the large-scale clustering properties of matter and the related galaxy populations. The small-scale distribution of galaxies, i.e. the distribution within the various components of the cosmic skeleton, will involve the complicated details of highly nonlinear small-scale interactions of the gravitating matter. N-body simulations are preferred for treating that problem. For our purposes, we take the route of complementing the large-scale cellular distribution induced by Voronoi patterns by a user-specified small-scale distribution of galaxies. Ideally, well-defined and elaborate physical models would fill in this aspect. A more practical alternative approach involves the generation of either tailor-made purely heuristic "galaxy" distributions in and around the various elements of a Voronoi tessellation (e.g. pure uniform distributions). Alternatively, we can generate distributions that more closely resemble the outcome of dynamical simulations, and represent an idealized and asymptotic description thereof. Such a model is the *kinematic model* defined by Van de Weygaert & Icke (1989).

Particular emphasis should be put on that fact that this Voronoi strategy has the unique and fundamental feature of studying galaxy distributions around geometrical features that themselves have some distinct and well-defined stochastic spatial distribution. The galaxies are residing in walls, filaments and vertices which are distributed themselves as an integral component of the Voronoi spatial network. Their distribution is not a pure random, but instead one in which these components themselves are spatially strongly correlated, connecting into coherent "super"structures !!! This background frame of spatially clustered geometrical elements not only determines the overall clustering properties of its galaxy population, it also represents and distinguishes it from from less physically motivated stochastic toy models (e.g. the double Poisson process).

12.6.1 Voronoi galaxy distributions: the Kinematic Model

The kinematic Voronoi model is based on the notion that when matter streams out of the voids towards the Voronoi skeleton, cell walls form when material from one void encounters that from an adjacent one. In the original "pancake picture" of Zel'dovich and collaborators, it was gaseous dissipation fixating the pancakes (walls), automatically leading to a cellular galaxy distribution. But also when the matter is collisionless, the walls may be hold together by their own self-gravity. Accordingly, the structure formation scenario of the kinematic model proceeds as follows. Within a void, the mean distance between galaxies increases uniformly in the course

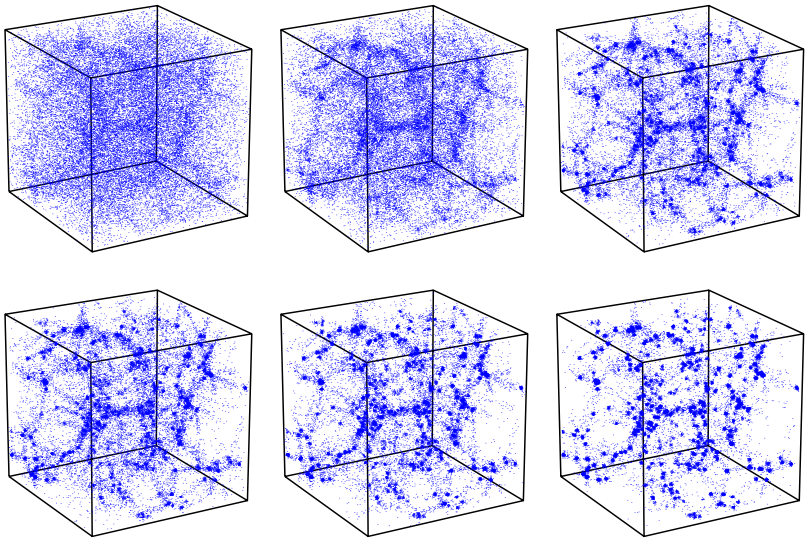


FIGURE 12.3. A sequel of consecutive timesteps within the kinematic Voronoi cell formation process. The depicted boxes have a size of $100h^{-1}$Mpc. Within these cubic volumes some 64 Voronoi cells with a typical size of $25h^{-1}$Mpc delineate the cosmic framework around which some 32000 galaxies have aggregated (corresponding roughly to the number density of galaxies yielded by a Schechter luminosity function with parameters according to Efstathiou, Ellis & Peterson 1988), where we restricted ourselves to galaxies brighter than $M_{gal} = -17.0$. In the full "simulation box" of $200h^{-1}$Mpc, this amounts to 268,235 galaxies.

of time. When a galaxy tries to enter an adjacent cell, the gravity of the wall, aided and abetted by dissipational processes, will slow its motion down. On the average, this amounts to the disappearance of its velocity component perpendicular to the cell wall. Thereafter, the galaxy continues to move within the wall, until it tries to enter the next cell; it then loses its velocity component towards that cell, so that the galaxy continues along a filament. Finally, it comes to rest in a node, as soon as it tries to enter a fourth neighbouring void. Of course the full physical picture is expected to differ considerably in the very dense, highly nonlinear regions of the network, around the filaments and clusters. Nonetheless, the Voronoi kinematic model produces a structural morphology containing the revelant characteristics of the cosmic foam, both the one seen in large redshift surveys as the one found in the many computer model N-body simulations.

The evolutionary progression within our Voronoi kinematic scheme, from an almost featureless random distribution, via a wall-like and filamentary morphology towards a distribution in which matter ultimately aggregates into conspicuous compact cluster-like clumps can be readily appreciated from the sequence of 6 cubic 3-D particle distributions in Figure 3. The steadily increasing contrast of the various structural features is accompanied by a gradual shift in topological nature of the distribution. The

virtually uniform particle distribution at the beginning (upper lefthand frame) ultimately unfolds into the highly clumped distribution in the lower righthand frame. At first only a faint imprint of density enhancements and depressions can be discerned. In the subsequent first stage of nonlinear evolution we see a development of the matter distribution towards a wall-dominated foam. The contrast of the walls with respect to the general field population is rather moderate (see e.g. second frame), and most obviously discernable by tracing the sites where the walls intersect and the galaxy density is slightly enhanced. The ensuing frames depict the gradual progression via a wall-like through a filamentary towards an ultimate cluster-dominated matter distribution. By then nearly all matter has streamed into the nodal sites of the cellular network. The initially almost hesitant rise of the clusters quickly turns into a strong and incessant growth towards their appearance as dense and compact features which ultimately stand out as the sole dominating element in the cosmic matter distribution (bottom righthand frame).

12.7 Superclustering: the clustering of clusters

Maps of the spatial distribution of clusters of galaxies show that clusters themselves are not Poissonian distributed, but turn out to be highly clustered (see e.g. Bahcall 1988). They aggregate to form huge supercluster complexes. For the sake of clarity, it is worthwhile to notice that such superclusters represent moderate density enhancements on a scale of tens of Megaparsec, typically in the order of a few times the average. They are still co-expanding with the Hubble flow, be it at a slightly decelerated rate, and are certainly not to be compared with collapsed, let alone virialized, identifiable physical entities like clusters.

The *first* characteristic of superclustering is the finding that the clustering of clusters is considerably more pronounced than that of galaxies. According to most studies the two-point correlation function $\xi_{cc}(r)$ of clusters is consistent with it being a scaled version of the power-law galaxy-galaxy correlation function, $\xi_{cc}(r) = (r_\circ/r)^\gamma$. While most agree on the same slope $\gamma \approx 1.8$ and a correlation amplitude that is significantly higher than that for the galaxy-galaxy correlation function, the estimates for the exact amplitude differ considerably from a factor $\simeq 10 - 25$. The original value found for the "clustering length" r_o for rich $R \geq 1$ Abell clusters was $r_\mathrm{o} \approx 25h^{-1}$Mpc (Bahcall & Soneira 1983), up to a scale of $100h^{-1}$ Mpc (Bahcall 1988). Later work favoured more moderate values in the order of $15 - 20h^{-1}$Mpc (e.g. Sutherland 1988, Dalton et al. 1992, Peacock & West 1992).

A related *second* characteristic of superclustering is that the differences in estimates of r_o are at least partly related to the specific selection of clusters, i.e. the applied definition of clusters. Studies dealt with cluster samples of

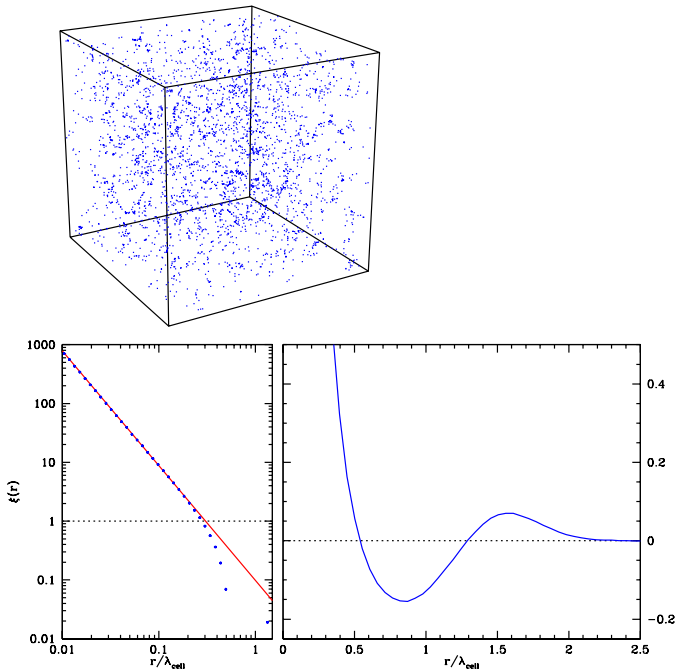


FIGURE 12.4. Two-point correlation function analysis of a selection of galaxies in a Voronoi kinematic model realization. Top frame: depiction of a galaxy sample in a $150h^{-1}$Mpc box at the present cosmic epoch $\sigma(8h^{-1}\text{Mpc} \approx 1$. Note the cellular morphology of walls and filaments with a few conspicuously dense cluster "nodes". Bottom left: a log-log plot of the $\xi(r)$, with distance r in units of the basic cellsize $\lambda_{\rm cell}$. The power-law character of ξ up to $r \sim 0.5\lambda_c$ is evident. Bottom right: a lin-lin plot of ξ showing ringing behaviour out to scales $r \sim 2\lambda_{\rm cell}$. From: Van de Weygaert 2002.

rich $R \geq 1$ Abell clusters, others also included poorer clusters, or employed a physically well-founded criterion on the basis of X-ray emission. On the basis of such analyses we find a trend of an increasing clustering strength as the clusters in the sample become more rich ($\approx$ massive). On the basis of the first related studies, Szalay & Schramm (1985) even put forward the (daring) suggestion that samples of clusters selected on richness would display a 'fractal' clustering behaviour, in which the clustering scale $r_{\rm o}$ would scale linearly with the typical scale L of the cluster catalogue. This typical scale $L(R)$ is then the mean separation between the clusters of richness higher than R: $\xi_{cc}(r) = \beta\ (L(r)/r)^{\gamma}$ where $L(R) = n^{-1/3}$. While the exact scaling of $L(r)$ with mean number density n is questionable, observations follow the qualitative trend of a monotonously increasing $L(R)$. It also seems to adhere to the increasing level of clustering that selections of more massive clusters appear to display in large-scope N-body simulations (e.g. Colberg 1998), given some telling detailed differences.

A *third* aspect of superclustering, one that often escapes emphasis but which we feel is important to focus attention on, is the issue of the spatial range over which clusters show positive correlations, the "coherence" scale of cluster clustering. Currently there is ample evidence that $\xi_{cc}(r)$ extends out considerably further than the galaxy-galaxy correlation ξ_{gg}, possibly out to $50h^{-1} - 100h^{-1}$Mpc. This is not in line with conventional presumption that the stronger level of cluster clustering is due to the more clustered locations of the (proto)cluster peaks in the primordial density field with respect to those of (proto)galaxy peaks. According to this conventional "peak bias" scheme we should not find significant non-zero cluster-cluster correlations on scales where the galaxies no longer show any significant clustering. If indeed ξ_{gg} is negligible on these large scales, explaining the large scale cluster-cluster clustering may be posing more complications than a simple interpretation would suggest.

12.8 Superclustering: the Voronoi Vertex Distribution

In the Voronoi description vertices are identified with the clusters of galaxies, a straightforward geometric identification without need to invoke additional descriptions. Like genuine clusters, these vertices then act as the condensed and compact complexes located at the interstices in the cosmic framework. The immediate and highly significant consequence is that – for a given Voronoi foam realization – the spatial distribution of clusters is fully and uniquely determined. The study of the clustering of these vertices can therefore be done without any further assumptions, fully set by the geometry of the tessellation. When doing this, we basically use the fact that *the Voronoi node distribution is a topological invariant* in co-moving coordinates, and does not depend on the way in which the walls, filaments, and nodes are populated with galaxies. The statistics of the nodes should therefore provide a robust measure of the Voronoi properties.

A first inspection of the spatial distribution of Voronoi vertices (Fig. 4, top frame) immediately reveals that it is not a simple random Poisson distribution. The full spatial distribution of Voronoi vertices in the $150h^{-1}$Mpc cubic volume involves a substantial degree of clustering, a clustering which is even more strongly borne out by the distribution of vertices in a thin slice through the box (bottom lefthand frame) and equally well reflected in the sky distribution (bottom righthand frame). The impression of strong clustering, on scales smaller than or of the order of the cellsize $\lambda_{\rm C}$, is most evidently expressed by the corresponding two-point correlation function $\xi(r)$ (Fig. 4, log-log plot lefthand frame, lin-lin plot in the righthand frame). Not only can we discern a clear positive signal but – surprising at the time of its finding on the basis of similar computer experiments (van de Weygaert & Icke 1989) – out to a distance of at least $r \approx 1/4\,\lambda_{\rm c}$ the correlation function

appears to be an almost perfect power-law,

$$\xi_{vv}(r) = \left(\frac{r_{\rm o}}{r}\right)^{\gamma}\,; \qquad\qquad \gamma = 1.95; \qquad r_{\rm o} \approx 0.3\,\lambda_{\rm c}\,. \tag{12.1}$$

The solid line in the log-log diagram in Fig. 4 represents the power-law with these parameters, the slope $\gamma \approx 1.95$ and "clustering length" $r_{\rm o} \approx 0.3\,\lambda_{\rm c}$. (the solid line represents the power-law with these parameters). Beyond this range, the power-law behaviour breaks down and following a gradual decline the correlation function rapidly falls off to a zero value once distances are of the order of (half) the cellsize. Assessing the behaviour of $\xi(r)$ in a linear-linear plot, we get a better idea of its behaviour around the zeropoint "correlation length" $r_{\rm a} \approx 0.5\lambda_c$ (bottom righthand frame fig. 4). Beyond $r_{\rm a}$ the distribution of Voronoi vertices is practically uniform. Its only noteworthy behaviour is the gradually declining and alternating quasi-periodic ringing between positive and negative values similar to that we also recognized in the "galaxy" distribution, a vague echo of the cellular patterns which the vertices trace out. Finally, beyond $r \approx 2\lambda_c$ any noticeable correlation seems to be absent.

The above 2pt correlation function of Voronoi vertices is a surprisingly good and solid match to the observed world. It sheds an alternative view on the power-law clustering with power law $\gamma \approx 2$ found in the cluster distribution. Also, the observed cluster clustering length $r_{\rm o} \approx 20h^{-1}\,$Mpc can be explained within the context of a cellular model, suggesting a cellsize of $\lambda_{\rm c} \approx 70h^{-1}\,$Mpc as the basic scale of the cosmic foam.

On the other hand, the latter also reveals a complication. The suggested cell scale is surely well in excess of the $25h^{-1} - 35h^{-1}\,$Mpc size of the voids in the galaxy distribution. In addition, it appears to point to an internal inconsistency within the Voronoi concept. We saw above that if we tie the observed galaxy-galaxy correlation to the clustering of objects in the walls and filaments of the same tessellation framework, it suggests a cellsize $\lambda_c \approx 25h^{-1}$Mpc. This would conflict with the cellsize that would correspond to a good fit of the Voronoi vertex clustering to cluster clustering. The solution to this dilemma lead to an intriguing finding (for a complete description of this result see Van de Weygaert 2001).

12.8.1 Biased Voronoi Vertex Selections

We first observe that the vertex correlation function of eqn. (2) concerns the full sample of vertices, irrespective of any possible selection effects based on one or more relevant physical aspects. In reality, it will be almost inevitable to invoke some sort of biasing through the definition criteria of the involved catalogue of clusters. Interpreting the Voronoi model in its quality of asymptotic approximation to the galaxy distribution, its vertices will automatically comprise a range of "masses".

Brushing crudely over the details of the temporal evolution, we may assign each Voronoi vertex a "mass" estimate by equating that to the total

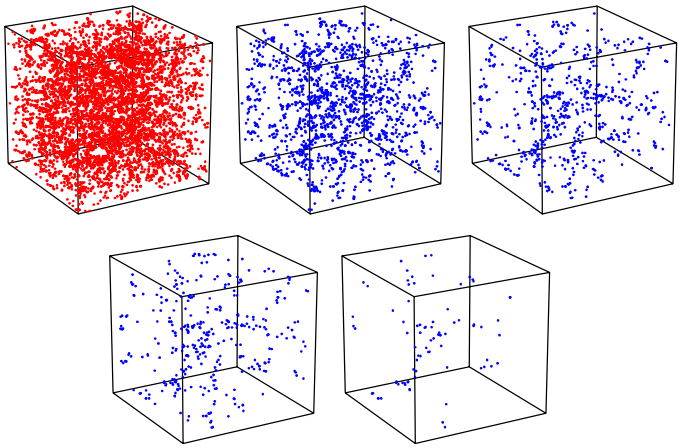


FIGURE 12.5. Selections of vertices from a full sample of vertices. Depicted are the (100%) full sample (top left), and subsamples of the 25%, 10%, 5% and 1% most massive vertices (top centre, top right, bottom left, bottom right). Note how the richer vertices appear to highlight ever more pronounced a filamentary superstructure running from the left box wall to the box centre. From: Van de Weygaert 2002a.

amount of matter ultimately will flow towards that vertex. When we use the "Voronoi streaming model" as a reasonable description of the clustering process, it is reasonably straightforward if cumbersome to calculate the "mass" or "richness" $\mathcal{M}_{\rm V}$ of each Voronoi vertex by pure geometric means. Evidently, vertices surrounding large cells are expected to be more massive. The details, turn out to be challengingly complex, as it concerns the (purely geometric) calculation of the volume of a non-convex polyhedron centered on the Voronoi vertex. The related nuclei are the ones that supply the Voronoi vertex with inflowing matter.

To get an impression of the resulting selected vertex sets, Figure 5 shows 5 times the same box of $250h^{-1}$Mpc size, each with a specific subset of the full vertex distribution (top lefthand cube). In the box we set up a realization of a Voronoi foam comprising 1000 cells with an average size of $25h^{-1}$Mpc. From the full vertex distribution we selected the ones whose "richness" $\mathcal{M}_{\rm V}$ exceeds some specified lower limit. The depicted vertex subsets correspond to progressively higher lower mass limits, such that 100%, 25%, 10%, 5% and 1% most massive vertices are included (from top lefthand to bottom righthand). The impression is not the one we would get if the subsamples would be mere random diluted subsamples from the full vertex sample. On the contrary, we get the definite impression of a growing coherence scale !!! Correcting for a possibly deceiving influence of the dilute sampling, and sampling an equal number of vertices from each "selected" sample only considerably strengthens this impression. There is an intrinsic effect in changing clustering properties as a function of (mass-

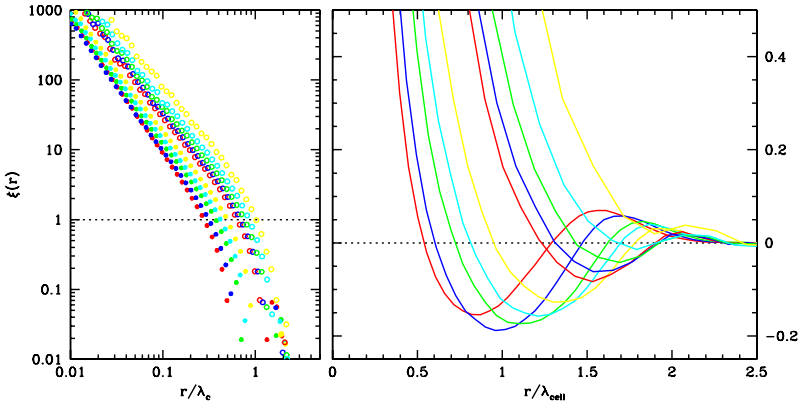


FIGURE 12.6. Scaling of the two-point correlation function of Voronoi vertices, for a variety of subsamples selected on the basis of "richness", ranging from samples with the complete population of vertices down to subsamples containing the 2.5% most massive vertices. Left: log-log plot of $\xi(r)$ against r/λ_c, with λ_c the basic tessellation cellsize ($\equiv$ intranucleus distance). Notice the upward shift of $\xi(r)$ for subsamples with more massive vertices. Right: lin-lin plot of $\xi(r)$ against r/λ_c. Notice the striking rightward shift of the "beating" pattern as richness of the sample increases. From: Van de Weygaert 2002a.

defined) cluster sample.

12.8.2 Vertex clustering: Geometric Biasing?

All in all, Fig. 5 provides ample testimony of a profound largely hidden large-scale pattern in foamlike networks, a hithero entirely unsuspected large-scale coherence over a range exceeding many cellsizes.

To quantify the impression given by the distribution of the biased vertex selections, we analyzed the two-point correlation function for each vertex sample. We computed $\xi(r)$ for samples ranging From the complete sample down to the ones merely containing the 2.5% most massive ones. As the average distance $\lambda_v(R) = n(R)^{-1/3}$ between the sample vertices increases monotonously with rising subsample richness, in the following we will frequently use the parameter λ_v for characterizing the richness of the sample, ranging from $\lambda_v \approx 0.5\lambda_c$ up to $\lambda_v \approx 1.5\lambda_c$ for vertex samples comprising all vertices up to samples with the 10% most massive vertices (the basic cellsize λ_c functions objective distance unit).

The surprising finding is that all subsamples of Voronoi vertices do retain a two-point correlation function displaying the same qualitative behaviour as the $\xi_{vv}(r)$ for the full unbiased vertex sample (Fig 6). Out to a certain range it invariably behaves like a power-law (lefthand frame), while beyond that range the correlation functions all show the decaying oscillatory behaviour that already has been encountered in the case of the full sample.

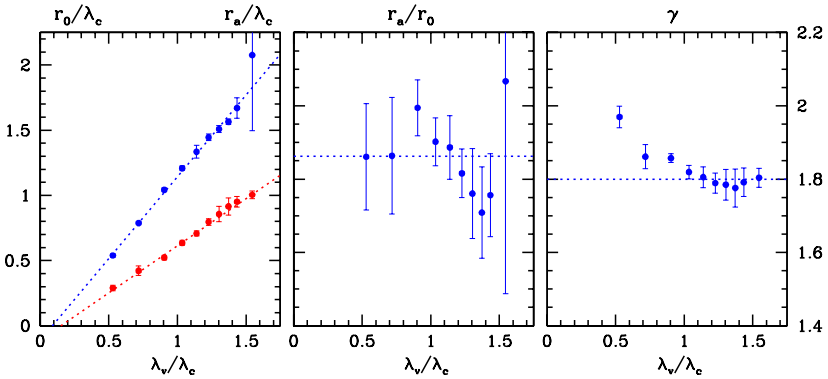


FIGURE 12.7. Scaling of Voronoi vertex two-point correlation function parameters for vertex subsamples over a range of "richness"/"mass". Left: the clustering length r_0 (red, $\xi(r_0) \equiv 1.0$) and the correlation (coherence) length r_a (blue, $\xi(r_a) \equiv 0$) as a function of average spatial separation between vertices in (mass) selected subsample, λ_v/λ_c. Centre: the ratio between clustering length r_0 and coherence length r_a as function of subsample intravertex distance λ_v/λ_c. Right: the power-law slope γ as function of λ_v/λ_c.

While all vertex $\xi_v v(r)$ convincingly confirm the impression of clustered point distributions, merely by the fact that it is rather straightforward to disentangle the various superposed two-point correlation functions we can immediately infer significant systematic differences.

First observation is that the amplitude of the correlation functions increases monotonously with rising vertex sample richness. Expressing the amplitude in terms of the "clustering length" r_o and plotting this against the λ_v between the sample vertices (both in units of λ_c), a striking almost perfectly linear relation is resulting (Fig. 7, lefthand frame, lower line). In other words, almost out of the blue, the "fractal" clustering scaling description of Szalay & Schramm (1985) appears to be stealthily hidden within foamy geometries. Although in the asymptotic Voronoi model we may be partially beset by the fact that we use an asymptotic measure for the vertex "mass" – the total amount of mass that ultimately would settle in the nodes of the cosmic foam – it may have disclosed that ultimately it reflects the foamy structured spatial matter distribution. Overall, the scaling of the clustering strength explains the impression of the increasingly compact clumpiness seen in the "biased" vertex distributions in Fig. 5. Summarizing, we can conclude that the foamy geometry is the ultimate ground for the observed amplified levels of cluster clustering.

A *second* significant observation is that the lin-lin large-scale behaviour of the ξ_{vv} seems to extend to larger and larger distances as the sample richness is increasing. The oscillatory behaviour is systematically shifting outward for the richer vertex samples, which reflects the fact that clustering patterns extend increasingly outward. Even though the basic cellular

pattern had a characteristic scale of only λ_c, the sample of the 5% richest nodes apparently seem to set up coherent patterns extending at least 2 to 3 times larger. This is clearly borne out by the earlier shown related point distributions (Fig. 5). Foamlike geometries seemingly induce coherent structures significantly larger than their basic size !!! This may hint at another tantalizing link between the galaxy and the cluster distribution. To elucidate this behaviour further, in Fig. 7 (lefthand frame, higher line) we also plotted the "correlation (coherence) scale" $r_{\rm a}$ versus the average sample vertex distance λ_v. And yet again, as in the case of $r_{\rm o}$, we find an almost perfectly linear relation !!!

Combining the behaviour of $r_{\rm o}$ and $r_{\rm a}$ we therefore find a remarkable 'self-similar' scaling behaviour, in which the ratio of correlation versus clustering length is virtually constant for all vertex samples, $r_{\rm a}/r_{\rm o} \approx 1.86$ (see Fig. 7, central frame). *Foamlike networks appear to induce a clustering in which richer objects not only cluster more strongly, but also further out !!!*

A *final* interesting detail on the vertex clustering scaling behaviour is that a slight and interesting trend in the behaviour of power-law slope. The richer samples correspond to a tilting of of the slope. Interestingly, borne out by the lower righthand frame in Fig. 7, we see a gradual change from a slope $\gamma \approx 1.95$ for the full sample, to a robust $\gamma \approx 1.8$ for the selected samples.

12.9 Conclusions: Bias, Cosmic Geometry and Self-Similarity

The uncovered systematic trends of vertex clustering have uncovered a hidden 'self-similar' clustering of vertices. This may be appreciated best from studying a particular realization of such behaviour (see Fig. 8)

The above results form a tantalizing indication for the existence of self-similar clustering behaviour in spatial patterns with a cellular or foamlike morphology. It might hint at an intriguing and intimate relationship between the cosmic foamlike geometry and a variety of aspects of the spatial distribution of galaxies and clusters. One important implication is that with clusters residing at a subset of nodes in the cosmic cellular framework, a configuration certainly reminiscent of the observed reality, it would explain why the level of clustering of clusters of galaxies becomes stronger as it concerns samples of more massive clusters. In addition, it would successfully reproduce positive clustering of clusters over scales substantially exceeding the characteristic scale of voids and other elements of the cosmic foam. At these Megaparsec scales there is a close kinship between the measured galaxy-galaxy two-point correlation function and the foamlike morphology of the galaxy distribution. In other words, the cosmic geometry apparently implies a 'geometrical biasing" effect, qualitatively different from the more conventional "peak biasing" picture (Kaiser 1984).

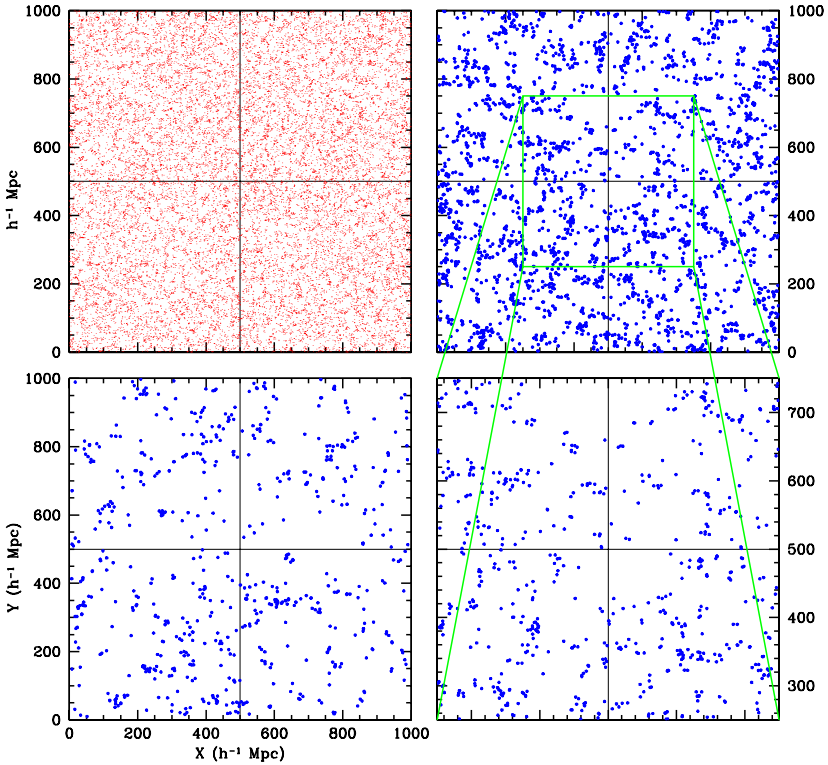


FIGURE 12.8. A depiction of the meaning of 'self-similarity' in the vertex distribution. Out of a full sample of vertices (top left) in a central slice, (top right) the 20.0% richest vertices. Similarly, (bottom left) the 2.5% richest vertices. When lifting the central $1/8^{th}$ region out of the 20% vertex subsample in the (top right-hand) frame and sizing it up to the same scale as the full box, we observe the similarity in point process between the resulting (bottom righthand) distribution and that of the 2.5% subsample (bottom lefthand). Self-similarity in pure form !

Acknowledgments: I would like to thank J. Dankers for his help in Voronoi plotting with Geomview. Writing this contribution, fond memories emerged of the many years over which the encouragement by Vincent Icke and Bernard Jones paved my way on the path through the world of Voronoi.

12.10 References

[1] Bahcall, N.A., 1988, ARA&A, 26, 631
[2] Bahcall, N.A., Soneira, R., 1983, ApJ, 270, 20
[3] Borgani, S., Guzzo, L., 2001, Nature, 409, 39
[4] Borgani, S., Guzzo, L., 2001, Nature, 409, 39
[5] Colberg, J., 1998, Ph.D. thesis, Ludwig-Maximilian Univ. München

[6] Dalton, G.B., Efstathiou, G., Maddox, S.J., Sutherland, W.J., 1992, ApJ, 390, L1
[7] De Lapparent, V., Geller, M.J., Huchra, J.P., 1986, ApJ, 302, L1
[8] Dubinski, J., Da Costa, L.N., Goldwirth, D.S., Lecar, M. Piran, T., 1993, ApJ, 410, 458
[9] Efstathiou, G., Ellis, R.S., Peterson, B.A., 1988, MNRAS, 232, 431
[10] Geller, M.J., Huchra, J., 1989, Science, 246, 897
[11] Icke, V., 1972, Ph.D. Thesis, University Leiden
[12] Icke, V., 1984, MNRAS, 206, 1P
[13] Icke, V., van de Weygaert, R., 1987, A&A, 184, 16
[14] Kaiser, N., 1984, ApJ, 284, L9
[15] Kirshner, R.P., Oemler, A., Schechter, P.L., Shectman, S.A., 1981, ApJ, 248, L57; 1987, ApJ, 314, 493
[16] Peacock, J.A., West, M.J., 1992, 259, 494
[17] Sheth, R.K., van de Weygaert, R., 2002, in prep.
[18] SubbaRao, M.U., Szalay, A.S., 1992, ApJ, 391, 483
[19] Sutherland, W., 1988, MNRAS, 234, 159
[20] Szalay, A.S., Schramm, D.N., 1985, Nature, 314, 718
[21] Voronoi, G., 1908, J. reine angew. Math., 134, 198
[22] van de Weygaert, R., 1991a, MNRAS, 249, 159
[23] van de Weygaert, R., 1991b, Ph.D. Thesis, University Leiden
[24] van de Weygaert, R., 1994, A&A, 283, 361
[25] van de Weygaert, R., 2002, A&A, to be submitted
[26] van de Weygaert, R., Icke, V., 1989, A&A, 213, 1
[27] van de Weygaert, R., van Kampen, E., 1993, MNRAS, 263, 481

13

Statistics and the Cosmic Microwave Background

Andrew H. Jaffe[1]

ABSTRACT We discuss the statistics of fluctuations in the Cosmic Microwave Background, and the statistical analysis of CMB experiments. Using Bayesian techniques, we proceed from the time-ordered data through maps of the sky, to power spectra, and to cosmological parameters. We discuss computational problems encountered along the way, and review recent results.
This paper is followed by a commentary by the Pittsburgh Institute for Computational Astrostatistics.

13.1 Introduction

The Cosmic Microwave Background (CMB) is made up of photons that last interacted with ordinary matter when the Universe was 100,000 years old and had a temperature, T, corresponding to $kT \simeq 1$ eV, where k is Boltzmann's constant and eV are units of energy. At this epoch, the protons and electrons that had been kept ionized by the high temperature were able to form neutral hydrogen for the first time. Prior to this, the charged proton/electron plasma was opaque to photons; thereafter the Universe was transparent. Hence, the CMB photons we see today have been streaming freely for the subsequent 15 billion years, redshifting by a factor of 1,000 to the microwave band, only to be captured finally in one of the several detectors we have designed to do just that. Starting with Penzias and Wilson's 1967 radio telescope, through the COBE satellite [5] launched in the late 80s, today's MAXIMA[13, 19], BOOMERANG[8, 3] and DASI[21, 12, 28] experiments, the just-launched MAP satellite[23], and the Planck satellite[27], planned for 2007, we observe the CMB with increasing sensitivity and higher resolution. The results are a two-dimensional snapshot of the Universe at this epoch of "Last Scattering" or "Recombination" filtered through the physics of the baryons, electrons, photons and Dark Matter making up the Universe. As there have been many fine reviews of these physical processes and what we can hope to learn from

[1]Department of Astronomy, University of California Berkeley

them about the Universe as a whole, I will simply commend the interested reader to them (e.g., [17, 15]), here concentrating on statistical issues. Much of this material is necessarily review of various more technical references (e.g., [3, 32]).

13.2 The statistics of CMB anisotropies

Unlike many other areas of astronomy, here we are concerned with an underlying physical phenomenon that is itself statistical in nature, rather than deterministic. That is, we are not interested in the long run in the details of the temperature distribution, but rather in its overall statistical properties. Within the inflationary paradigm of structure formation, perturbations to the otherwise-smooth matter density are laid down via a quantum-mechanical mechanism; these three-dimensional perturbations are described by a power spectrum, $P(k) \propto k^{n_s}$ (possibly with small corrections) and in most inflationary models with a Gaussian distribution, so that the power spectrum is the *only* information needed to describe them. Because they are extremely small (fractionally less than 10^{-5} at the time of last scattering), we can use linear perturbation theory to determine the impact on the CMB. Any linear transformation of a Gaussian field is another Gaussian field, and hence the CMB fluctuations are themselves described by a 2d power spectrum, C_ℓ, where ℓ is spherical harmonic wavenumber. We start with the temperature pattern on the sky, $\Delta T(\hat{\mathbf{x}})/T = [T(\hat{\mathbf{x}}) - \bar{T}]/\bar{T}$, where $\bar{T}$ is the average temperature and $\hat{\mathbf{x}}$ is a unit vector, and expand this in spherical harmonic multipoles:

$$\frac{\Delta T}{T}(\hat{\mathbf{x}}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\hat{\mathbf{x}}) \tag{13.1}$$

Under the assumptions of Gaussianity and an isotropic distribution on the sky, we can treat the components $a_{\ell m}$ as if they were drawn from a multivariate (but uncorrelated) Gaussian distribution with variance

$$\langle a_{\ell m} a_{\ell' m'} \rangle = C_\ell \delta_{\ell,\ell'} \delta_{m,-m'} \ . \tag{13.2}$$

Then, our task will be to determine C_ℓ from an actual noisy realization of some part of the sky.

13.3 The Bayesian paradigm

We will start with a statement of *Bayes' Theorem*:

$$P(\theta|DI) = \frac{P(\theta|I)P(D|\theta I)}{P(D|I)}, \tag{13.3}$$

where $P(a|b)$ is the probability (or density) for a given b. The parameters of the theory we are testing are θ, the data is D, and the "background information," is I. We mix "propositions" like I, with parameter values, like θ. $P(\theta|I)$ is the *prior*, $P(D|\theta I)$ is the *likelihood*, and

$$P(D|I) = \int d\theta \; P(\theta|I)P(D|\theta I) \tag{13.4}$$

is a normalization factor, occasionally referred to as the *evidence*.

We will in the end wish to report some limits on the parameters, otherwise known as "credible regions." These are defined as

$$P(\theta_{\rm min} < \theta < \theta_{\rm max}|DI) \equiv \int_{\theta_{\rm min}}^{\theta_{\rm max}} P(\theta|DI)\; d\theta = C\;. \tag{13.5}$$

That is, the probability that the parameter is within the given region is C.

In CMB experiments, the data we start with is a timestream,

$$d_t = A_{tp}T_p + n_t \tag{13.6}$$

where d_t is the data taken at time $t = 1 \ldots N_t$, T_p is the sky temperature at pixel $p = 1 \ldots N_p$ with center located at position $\hat{\mathbf{x}}_p$, n_t is the value of the noise (instrumental and otherwise) at t, and finally A_{tp} is the matrix operator converting the temperature on the sky labeled by positions p to that observed at time t (so $A_{tp} = 1$ if pixel p is observed at time t, and 0 otherwise). We will take T_p to be already smeared by the effects of beam and pixel: $T_p = \int d^2x B(\hat{\mathbf{x}}_p, \hat{\mathbf{y}})S(\hat{\mathbf{y}})$, where B gives the response of the beam at position $\hat{\mathbf{x}}_\mathbf{p}$ from a signal at $\hat{\mathbf{y}}$ and S is the underlying temperature on the sky. In the following, we will freely mix matrix notation and the summation convention: $AT \equiv (AT)_t \equiv A_{tp}T_p \equiv \sum_p A_{tp}T_p$.

What are the parameters, θ, in which we are interested? The most obvious appear directly in Eq. 13.6: the underlying CMB sky, T_p. But we can also ask about the power spectrum, C_ℓ, which is responsible for correlations in T_p between different positions, or even the cosmological parameters underlying the spectrum. It is most efficient to ask each of these questions in turn, reducing the amount of data at each step. There is nothing to stop us from calculating $P(\Omega|d_tI)$, and finding the value of the density parameter Ω directly from the timestream. We will see that this would give us the same answer as calculating it from the power spectrum: $P(\Omega|d_tI) = P(\Omega|C_\ell I)$.

13.3.1 From the timestream to a map

We will take the noise to be described by a Gaussian with correlation function

$$\langle n_t n_{t'} \rangle = N_{tt'}. \tag{13.7}$$

We will further take the noise to be stationary, at least over short periods of time, so that $N_{tt'} = N(|t - t'|)$. In practice the noise needs to be solved

for iteratively[11, 30]. Here, we will assume that $N(t)$ is known exactly. Given these definitions, we can write down the likelihood function for T, which we will call the *map*:

$$P(d|TI) = |2\pi N|^{-1/2} \exp\left[-\frac{1}{2}(d - AT)^\dagger N^{-1}(d - AT)\right] . \qquad (13.8)$$

(Note that d and T refer to the full vectors d_t and T_p.) We can now ask, what is the most probable map, $\widehat{T}_p$, given the data? To do this we must specify a prior, which we shall take to be uniform: $P(T|I) = \text{const}$.

Our problem then becomes simple least-squares, albeit with large dimensionality and complicated correlations. By completing the square, we can rewrite the likelihood as

$$P(d_t|T_pI) \propto P(\widehat{T}|TI) = |2\pi C_N|^{-1/2} \exp\left[-\frac{1}{2}(\widehat{T} - T)^\dagger C_N^{-1}(\widehat{T} - T)\right] . \qquad (13.9)$$

where

$$\widehat{T} = (A^\dagger N^{-1} A)^{-1} A^\dagger N^{-1} d \qquad \text{and} \qquad C_N = (A^\dagger N^{-1} A)^{-1} . \qquad (13.10)$$

The posterior distribution for T is just proportional to Eq. 13.9: the underlying map, T_p, is distributed around $\widehat{T}_p$ as a Gaussian with correlation matrix $C_{Npp'}$.

We also see that irrespective of the form of the prior, the likelihood can be written as a function of $\widehat{T}$ and C_N, rather than d and N; $\widehat{T}$ and C_N are *sufficient statistics*. If we retain this information, we can throw out the original timestream data for any further inference we might wish to make from the data.

As an important aside, we note that the calculation of the maximum-likelihood map and its covariance matrix requires $O(N_{\rm pix}^3)$ operations (the map-making itself can be reduced to $O(N_{\rm pix}^2)$, but the correlation matrix is required for further operations). This becomes suitable for supercomputers at the current $N_{\rm pix} \simeq 50,000$ of MAXIMA and BOOMERANG. A parallel implementation of the full calculation exists in the MADCAP package[22], as do various implementations of $O(N_{\rm pix}^2)$ map-making[9].

We can assign a more informative prior distribution for the sky temperature. If we assume that the temperature itself is distributed as a zero-mean Gaussian with some covariance matrix $C_{Tpp'} = \langle T_p T_{p'} \rangle$, we can combine the two Gaussian distributions by the usual complete-the-square mechanics, and find that the posterior for T is once again a Gaussian, now with mean $T^W = C_T(C_T + C_N)^{-1}\widehat{T}$ and variance $C_W = C_T(C_T + C_N)^{-1}C_T$. This is just the *Wiener filter* which can also be derived on minimum-variance grounds. Note that a particular prior *power spectrum*, C_ℓ, defines a particular prior C_T, and hence a particular Wiener filter.

In Figure 13.1 we show an example, the map made from the MAXIMA-1 data[19].

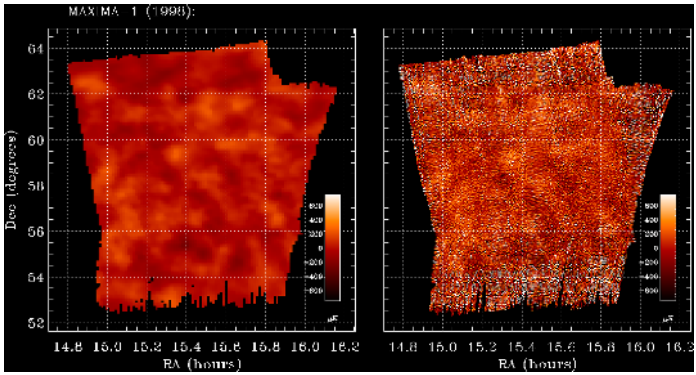


FIGURE 13.1. Maps, $\widehat{T}$, made from the MAXIMA data. The left panel shows a Wiener filtered 5 arcminute-square pixel map (using a Best-fit power spectrum to define C_T), and the right shows the 3 arcminute-square-pixel maximum-likelihood map from [19].

13.3.2 From maps to power spectra

How do we then determine the power spectrum of our data? For the data to have some power spectrum is to say that we can describe the underlying data as being drawn from a distribution with a variance given by

$$\begin{aligned} C_{Tpp'} &\equiv \langle T_p T_{p'} \rangle = \sum_{\ell m, \ell' m'} B_\ell^2 Y_{\ell m}(\hat{\mathbf{x}}_p) Y_{\ell' m'}(\hat{\mathbf{x}}_{p'}) \langle a_{\ell m} a_{\ell' m'} \rangle \\ &= \sum_\ell \frac{2\ell+1}{4\pi} B_\ell^2 C_\ell P_\ell \left(\hat{\mathbf{x}}_p \hat{\mathbf{x}}_{p'}\right) \end{aligned} \tag{13.11}$$

where the $\hat{\mathbf{x}}_p$ is the position of pixel p, B_ℓ is the spherical harmonic transform of the beam and pixelization function (see [35] for details and a full description of complications associated with asymmetric beams). The $a_{\ell m}$ are the spherical harmonic components of T, which we have eliminated using the definition of the power spectrum, Eq. 13.2 above, and the addition formula for spherical harmonics. The P_ℓ are the Legendre polynomials, for integer $\ell = 0, 1, 2, \ldots$, although we usually concentrate on $\ell \geq 2$ as the lower multipoles arise from different physical mechanisms. Beam-smearing cuts off our observations at some maximum ℓ and the physically processes themselves usually take $C_\ell \to 0$ smoothly for ℓ more than about a thousand.

We can use this information to write the *joint* likelihood for the underlying map and the power spectrum. First, we assign a prior for T based on Eq. 13.11. If we only have the mean and variance of T, the maximum entropy prior (and hence in some sense the least informative prior) is a Gaussian distribution, giving

$$P(T|C_\ell I) = |2\pi C_T|^{-1/2} \exp\left[-\frac{1}{2} T^\dagger C_T^{-1} T\right]. \tag{13.12}$$

The posterior is thus

$$P(T, C_\ell|\widehat{T}) \propto P(C_\ell|I)P(T|C_\ell I)P(T|\widehat{T}I). \tag{13.13}$$

This requires the specification of a prior for C_ℓ, but we can defer that decision until later. We first marginalize over T, which takes on the role of a nuisance parameter. We can perform this integral by (once again) completing the square, giving

$$P(C_\ell|\widehat{T}) \propto P(C_\ell|I)|2\pi\left(C_T + C_N\right)|^{-1/2} \exp\left[-\frac{1}{2}\widehat{T}^\dagger\left(C_T + C_N\right)^{-1} T\right]. \tag{13.14}$$

As we would expect, $\widehat{T}$ is distributed as a zero-mean Gaussian with variance $C_T + C_N$; equivalently, it is the sum of two independent zero-mean Gaussian-distributed quantities with variances C_T and C_N.

The question now becomes a computational one: given $\widehat{T}$ and C_N, how do we characterize this as a function of C_ℓ? Unlike when solving for $\widehat{T}$ itself, we cannot do this analytically. Because we *can* calculate derivatives of the likelihood function, we use a modified form of the Newton-Raphson method to find where $dP(C_\ell)/dC_\ell = 0$ and the curvature around that point.[2, 31] Experience shows that the likelihood space is well-structured, with a single maximum, so this procedure is sufficient.

As in the map-making procedure, this calculation unfortunately scales overall as $O(N_{\rm pix}^3)$, making it difficult even for current experiments, and effectively impossible for high-resolution full-sky experiments such as MAP and Planck. The MADCAP package[22] contains a parallel implementation of the likelihood maximization.

Note that we traditionally bin the power spectrum in ℓ. We assume that C_ℓ has a particular shape in some bins, and estimate the amplitude. Because we measure a finite amount of sky, overly-fine bins would be oversampled (as in fourier-analysis on the plane, where you can only get information for frequency intervals $> O(1/\text{length})$. Put another way, narrower bins would be very strongly correlated. Of course, these correlations would be encoded in the likelihood function, but the calculation also scales with the number of bins, another reason to keep this down to a reasonable number.

In Figure 13.2 we show the results of this and related procedures performed on the map of Figure 13.1, as well as data from BOOMERANG[3] and DASI[12]. The error bars are typically given by the inverse curvature of the posterior, but that is numerically very similar to the marginalized likelihood, or in fact any other standard measure of 1$sigma$ uncertainty.

13.3.3 Alternate approaches

Because of the $O(N_{\rm pix}^3)$ scaling of this C_ℓ estimation algorithm, other approaches have been suggested. One abandons the map as the input data,

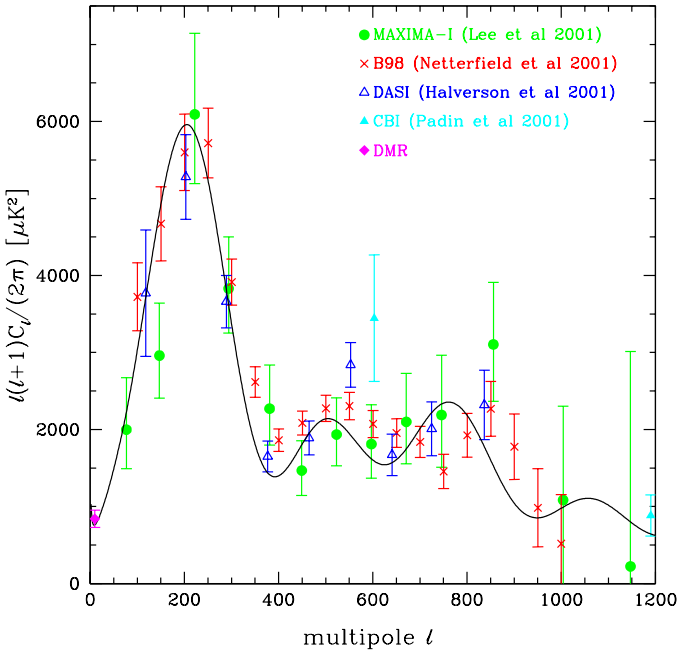


FIGURE 13.2. MAXIMA[19], BOOMERANG[3], DASI[12] and COBE/DMR[5] power spectra. The MAXIMA, CBI and DASI spectra were calculated with variants of the maximum-likelihood method described here; the BOOMERANG spectrum was calculated using a monte-carlo method[2], with modifications to approximate the maximum likelihood. The smooth curve was chosen from the space of cosmologicals discussed here to fit only a subset of the data, but remains a good fit to the entire data.

using instead quadratic combinations of the data, in particular the squared spherical-harmonic components—i.e., the naive power spectrum of the map, sometimes referred to as *pseudo-C_ℓs*.[34] In some simple cases, one can exactly calculate the likelihood function for these quadratics as a function of C_ℓ, in analogy to Eq. 13.14. The use of this approach with real, complicated data has yet to be investigated.

Another speedup takes advantage of the notion that one can smooth a map to investigate power at large scales (low ℓ), and conversely consider small sub-maps for small-scale (high ℓ) information. This has recently been formalized in the context of an approximation to the iterative Newton-Raphson likelihood maximization.[10]

Yet another possibility involves taking advantage of the structure of the noise and signal correlations in certain experimental configurations. The MAP satellite is expected to have noise correlations that are uncorrelated

and approximately azimuthally symmetric. The noise and signal correlations are then both highly structured in the spherical harmonic basis, and this fact can be used to provide matrix preconditioners for operations involved in likelihood maximization.[25]

13.3.4 From power spectra to parameters

Until now, we have left off the prior probability for the power spectrum. However, within the context of adiabatic inflationary models, we can write $C_\ell = C_\ell(\Omega_i, n_s, h, \tau_C, \ldots) \equiv C_\ell(\theta)$, where now θ represents the (7-10 or so) cosmological parameters we wish to determine. Thus, we again defer assigning a prior for the cosmological parameters themselves, just writing $P(C_\ell|\theta I) = \delta\left[C_\ell - C_\ell(\theta)\right]$. But we do have one problem: above, we determined the location of the peak of the likelihood as a function of C_ℓ, and the curvature around that peak, but nothing else about the shape. In particular, this shape is not well-described as that of a Gaussian. With enough computing power at our disposal, we could just calculate the value of the likelihood directly using Eq. 13.14; the $O(N_{\rm pix}^3)$ scaling rears its head, and this quickly becomes prohibitive.

However, the likelihood is well-approximated as a Gaussian in $\ln(C_\ell + x_\ell)$, where x_ℓ is related to the noise properties of the experiment[4]. Hence, once we have found the peak of the likelihood, the curvature at the peak, and this x, we can use simple χ^2 techniques. (Note that this *ansatz* describes the likelihood as a function of the theoretical spectrum, C_ℓ, the quantity of interest to Bayesians. It does *not* describe the likelihood as a function of the *data* (or any statistic of the data) which would be of interest in a frequentist analysis; see other contributions to this volume.)

Now, however, we are finally forced to confront the problem of assigning a prior probability to our cosmological parameters. This is complicated by several factors:

1. For computational reasons, and because we do not fully understand the offset log-normal *ansatz* in the presence of strong correlations, we bin the power spectrum. That is, we calculate the most probable amplitude of the power spectrum in some band, assuming some known shape. Since we are in general not comparing to theories with the same shape that we have assumed, there is subtlety in calculating the likelihood function. This is addressed through the use of "filter functions" with a formalism developed in Ref. [18].

2. The dependence of C_ℓ upon the parameters is highly nonlinear.

3. For most of the parameters, there is no natural measure to define a non-informative prior. Moreover, because the parameters enter the cosmological physics in different ways and different combinations in different problems, we cannot make a simple prior choice such as

using uniform priors in all cases. (For example, we could choose to parameterize in terms of the densities Ω_i, along with the Hubble parameter, h. Alternately, we could parameterize in terms of physical densities $\propto \Omega_i h^2$ which control the physics.

4. There are several approximate degeneracies in the parameter space. That is, there are loci of parameters that give practically indistinguishable spectra.

5. Because of this, we cannot define a compact subspace of the full parameter space for which the likelihood goes to nearly-zero at the boundary. Hence, the results will always depend on the parameter volume over which we choose to calculate models to compare to data.

6. Even if we wish to use informative priors for the cosmological parameters, different experiments measure different parameters, and indeed different experiments disagree.

Given all of these issues, the most practical advice is simply to be sure to enumerate the explicit and implicit priors used. Moreover, it is important to check that the results are not *too* strongly dependent on the form of the prior, or at least that the dependence is physically understood.

There are other considerations when reporting such results. If we are interested in a single parameter, it is traditional to marginalize over all others. In this case, however, we are interested in the parameters both together and separately. That is, we would like to know what value the CMB gives for $\Omega_B h^2$ (say) and so it may be appropriate to marginalize over the other parameters. However, because the parameter space is quite large, we would also like to know where in terms of the other parameters the marginalized distribution picks up most of its mass. In problems with a more simply-structured likelihood space, this is accomplished by just reporting the likelihood maximum and some version of the covariance matrix around the maximum. In this case, the likelihood is not well-fit by a Gaussian (especially when the aforementioned degeneracies show their presence) and our intuition may be misguided. As the data improve we will indeed hone in on at least the non-degenerate parameters; we have seen this in the past few years as we have passed to the latest vintages of data.[20, 16]

For example, we find that somewhat generically, current C_ℓ data can be well fit by models in two very different regimes. One is a "standard" model with reasonable parameters, but another has several of the parameters which control the location and relative height of the peaks changed considerably to "unphysical" values. However, adding a very simple prior requiring $h > 0.45$ eliminates this unphysical regime.

Similarly, it is well known that the CMB is sensitive largely to the overall curvature, and thus to the total mass density, of the Universe, but not to the way in which it is apportioned among matter and a cosmological constant.

We cannot make any a priori obvious cuts on the parameter space to break this degeneracy. Hence, in Figure 13.3, we see the likelihood function in two dimensions (marginalized over all other parameters) for a subset of CMB data.

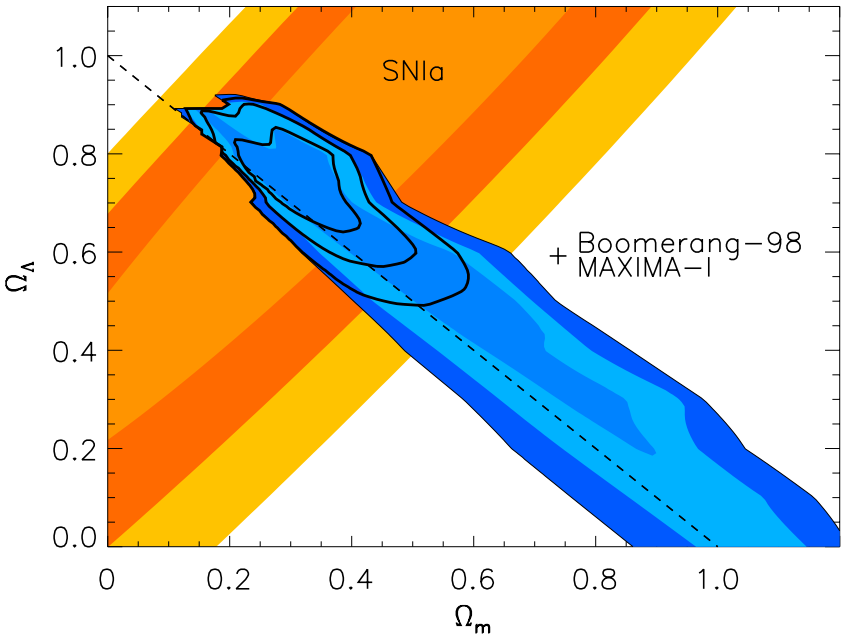


FIGURE 13.3. Likelihood in the Ω_m, Ω_Λ plane from the combined COBE/DMR, MAXIMA and BOOMERANG data as of late 2000. Blue contours along the $\Omega_m + \Omega_\Lambda = 1$ line are from the CMB alone, perpendicular orange contours are from an 'orthogonal' dataset of Supernovae distances[29, 26], and the heavy contours are for the combination of the two. Contours are 1-, 2- and 3-sigma as defined by the equivalent likelihood ratio for a 2-d Gaussian. From ref. [16]

13.3.5 Non-Gaussianity?

So far, we have used a Gaussian distribution to describe both the distribution of noise and of the signal. Perhaps foremost, we do this for simplicity: we can write down all of the above equations! Physically, a Gaussian arises when the "Central Limit Theorem" obtains: when we are concerned with something like the sum of very many small contributions. This holds for many sources of instrumental noise. It is also appropriate for the quantum-mechanical fluctuations produced in most models of inflation.

The specificity of the Gaussian distribution has led to much worry that our methods may be incorrect if the "actual signal isn't distributed as a Gaussian." Within the Bayesian paradigm, the Gaussian form for the likelihood arises when we assign a Gaussian prior to both the noise (Eq. 13.8) and the signal (Eq. 13.12). A Gaussian has the property that it is the *maximum-entropy* distribution given a known mean and variance. Hence, as long as our signal is described by a variance like Eq. 13.2, the Gaussian assumption is, in fact, the most conservative assignment that can be made. (Note, however, that the particular Gaussian we have chosen is *not* completely general: we require by Eq. 13.2 that the different $a_{\ell m}$ at a given ℓ have the same variance—i.e., isotropy.)

Conversely, if we somehow knew that the distribution had a particular non-Gaussian distribution (as predicted by, for example, certain classes of inflationary models [6, 33, 7]), we could use that instead of Eq. 13.12, although it may not be possible in that case to marginalize analytically over the map in the joint distribution of C_ℓ and the map (Eq. 13.13).

13.4 Alternatives: frequentist measures

The community so far has taken a largely Bayesian approach to the analysis of CMB data. Philosophical issues aside, there are alternatives, the so-called "frequentist" or "orthodox" approach, some of whose aspectes were explored at this meeting in the contributions of Schafer & Stark and Wasserman et al. In the former work, they attempt to characterize the 'distance' between cosmological models in terms of the ability of CMB data to discriminate between them. This arises as a problem because of the highly nonlinear relationship between the physical cosmological variables of interest and the measurable quantity, the power spectrum of CMB fluctuations. Such a characterization will prove useful not only in frequentist analyses of CMB data but also in any use of CMB power spectra which require a greater understanding of the mapping between parameters and spectra.

In the following, we wish to comment further on the Bayesian and Frequentist approaches to the CMB data analysis problem in general. Without caricaturing it too much, we can summarize the frequentist approach as follows. Just as in the Bayesian approach, we start with the likelihood function. Then, we choose an "estimator", some function of the data, chosen to somehow represent an estimate of the parameter we wish to determine. We then calculate the sampling distribution of this estimator, under the assumption of some fixed value of the theoretical parameters. If the likelihood is $P(d|\theta)$, and our estimator is $\hat{\theta}(d)$, we need $P(\hat{\theta}|\theta)$. If the estimator is some simple function of the data, then we can just use the usual transformations $P(x)dx = P(y)dy$ and do this analytically, otherwise we can perform Monte Carlo sampling of $P(d|\theta)$. Armed with this distribution, we

define a confidence interval in the usual way. To do this, we will need

$$P(\theta_{\rm min} < \hat{\theta} < \theta_{\rm max}|\theta I) \equiv \int_{\theta_{\rm min}}^{\theta_{\rm max}} P(\hat{\theta}|\theta I)\, d\hat{\theta} = C\ . \tag{13.15}$$

(cf. Eq. 13.5.) The art of frequentist statistics is in the choice of the estimator. Often, it is chosen to be unbiased, $\int d\hat{\theta}\ \hat{\theta} P(\hat{\theta}|\theta) = \theta$, and have some appropriately small or minimum variance under the same distribution.

It is worth belaboring the point that these intervals are a priori completely distinct from Bayesian intervals. These intervals say that, if you repeated the experiment many times, each time drawing from the same sampling distribution, in some fraction C of the trials you would get the answer within the stated limits.

This is in contrast to the Bayesian credible region, although the form is similar. Even in the simplest case, we are dealing with two very different functions: the Bayesian uses $P(\theta|\hat{\theta}I)$ for fixed data, $\hat{\theta}$, whereas the frequentist uses $P(\hat{\theta}|\theta I)$ for fixed θ. Even for a uniform prior when these functions are proportional to one another, the two approaches are concerned with it as a function of different variables! Of course, we whet our teeth on problems in which θ and $\hat{\theta}$ appear symmetrically in the likelihood—estimation with linear, Gaussian models. In this case, then, the frequentist and Bayesian results are agreement, but in general they will not be so. In particular, even if these correspondances do obtain (at least approximately), they do not help us understand other aspects of the frequentist distribution — for example, the offset-lognormal ansatz of Sec. 13.3.4 above applies to the likelihood as a function of C_ℓ, not as a function of some estimator $\hat{C}_\ell$.

13.4.1 Monte-Carlo power spectra

Nonetheless, intuition and longstanding practice suggest that such frequentist measures have a place. Indeed, for the estimation of Power Spectra in particular, there is a deeper reason to use them, even within the Bayesian paradigm. Consider the very simplest spectrum estimation problem, an all-sky experiment with uniform noise, and a pixel scale negligible compared to that of the sky signal. Then, there are exact correspondences between the (uniform prior) Bayesian and Frequentist results: The Bayesian maximum-likelihood is the same as the frequentist mean, and the "error bars" as calculated from the Bayesian curvature are the same as the Frequentist variance. [These correspondences are not strictly true if the noise is comparable to the signal, since the frequentist mean and variance are calculated for (signal + noise) > 0 rather than for signal > 0.] These are well-known to hold asymptotically, but this is a case in which they hold for finite data as well.

Unfortunately, these correspondences do not remain exact for realistic experiments. Nonetheless, experience has thus far shown that we can indeed

extract useful approximate Bayesian information from Monte Carlo power spectra[2, 3]; this is an ongoing area of research, especially as the era of experiments with $N_{\rm pix} \gg 100,000$ approaches. In the regime to be probed by MAP and Planck, with millions of pixels covering the whole sky, some alternative to the brute-force matrix manipulation will be necessary. For a full-sky experiment, we can take advantage of fast spherical-harmonic transforms to speed up some of the Monte Carlo calculations[2].

13.4.2 Cosmological parameter estimates

We must nonetheless take some care in interpreting such frequentist results. In particular, how do we use them in the next step of the calculation? If we take them to be approximations to the Bayesian solution, then the procedure is clear, assuming, at least, that we also have access to parameters like x_ℓ (Sec. 13.3.4 above) in order to compute the posterior distribution. However, a fully-fledged frequentist analysis in this case is somewhat more involved, since there clearly are no good sufficient statistics for the cosmological parameters given the data.

One can still imagine an entirely frequentist algorithm for calculating not just the power spectrum, but the cosmological parameters themselves, directly from the data. In practice, it is found that the limits are not very different from the Bayesian intervals[1], although this is a subject of ongoing research, and not immune to technical problems.

13.5 Conclusions

The cosmic microwave background has become one of the primary tools for exploring the early Universe. The simple, linear physics describing the phenomena make it relatively straightforward to connect the measurement process to the underlying cosmological phenomena. Conversely, the highly accurate measurements of the cosmological parameters that this data will allow requires that we understand our measurement and analysis procedure in great detail. This brings to the fore both computational issues in manipulating highly-correlated multivariate distributions and philosophical issues regardin the underlying analysis methods.

Acknowledgments The author would like to especially thank Lloyd Knox, Dmitri Pogosyan, Dick Bond, Julian Borrill, Pedro Ferreira and Radek Stompor for their work and innumerable discussions, as well as the whole MAXIMA, BOOMERANG and COMBAT teams. Portions of this work were supported by NASA LTSA Grant NAG5-6552 and by NSF KDI grant 9872979.

13.6 References

[1] M. Abroe et al, Frequentist Estimation of Cosmological Parameters from the MAXIMA-I data.

[2] J. R. Bond, A. H. Jaffe, and L. Knox, Phys Rev D **57**, 2117 (1998).

[3] J. R. Bond, R. G. Crittenden, A. H. Jaffe, and L. E. Knox, Computers in Science and Engineering **1**, 21 (1999).

[4] J. R. Bond, A. H. Jaffe, and L. Knox, Astrophys J **533**, 19 (2000).

[5] http://space.gsfc.nasa.gov/astro/cobe/.

[6] C. R. Contaldi, R. Bean, and J. Magueijo, Physics Letters B **468**, 189 (1999).

[7] C. R. Contaldi and J. . Magueijo, Phys Rev D **63**, 3512+ (2001).

[8] P. de Bernardis *et al.*, Nature **404**, 955 (2000).

[9] O. Doré, R. Teyssier, F. R. Bouchet, D. Vibert, and S. Prunet, Astro Astrophys **374**, 358 (2001).

[10] O. Dore, L. Knox, and A. Peel, CMB Power Spectrum Estimation via Hierarchical Decomposition, submitted to Phys. Rev. D, (2001).

[11] P. G. Ferreira and A. H. Jaffe, Mon Not Royal Astro Soc **312**, 89 (2000).

[12] N. W. Halverson *et al.*, DASI First Results: A Measurement of the Cosmic Microwave Background Angular Power Spectrum, 2001, astro-ph/0104489.

[13] S. Hanany *et al.*, Astrophys J Lett **545**, L5 (2000).

[14] E. Hivon *et al.*, MASTER of the CMB Anisotropy Power Spectrum: A Fast Method for Statistical Analysis of Large and Complex CMB Data Sets Astrophys J **567**, 2 (2001)

[15] W. Hu, http://background.uchicago.edu/.

[16] A. H. Jaffe *et al.*, Physical Review Letters **86**, 3475 (2001).

[17] M. Kamionkowski and A. Kosowsky, Ann. Rev. Nucl. Part. Sci. **49**, 77 (1999).

[18] L. E. Knox, Phys Rev D **60**, 103516 (1999).

[19] A. T. Lee *et al.*, A High Spatial Resolution Analysis of the MAXIMA-1 Cosmic Microwave Background Anisotropy Data, Astrophys J **561**, L1 (2001),

[20] A. E. Lange *et al.*, Phys Rev D **63**, 411+ (2001).

[21] E. M. Leitch *et al.*, Experiment Design and First Season Observations with the Degree Angular Scale Interferometer, Astrophys J **568**, 28 (2001) astro-ph/0104488.

[22] http://www.nersc.gov/ borrill/cmb/madcap.html.

[23] http://map.gsfc.nasa.gov/.

[24] C. B. Netterfield *et al.*, A measurement by BOOMERANG of multiple peaks in the angular power spectrum of the cosmic microwave background, 2001, astro-ph/0104460.

[25] S. P. Oh, D. N. Spergel, and G. Hinshaw, Astrophys J **510**, 551 (1999).

[26] S. Perlmutter et al, Astrophys J **517**, 565 (1999).

[27] http://astro.estec.esa.nl/SA-general/Projects/Planck/.

[28] C. Pryke *et al.*, Cosmological Parameter Extraction from the First Season of Observations with DASI, Astrophys J **568**, 46 (2001) astro-ph/0104490.

[29] A. Riess et al, Astron. J. **116**, 1009 (1998).

[30] R. Stompor *et al.*, Making Maps Of The Cosmic Microwave Background: The MAXIMA Example, 2001, astro-ph/0106451.

[31] M. Tegmark, Phys Rev D **55**, 5895 (1997).

[32] M. Tegmark, Phys Rev D **56**, 4514 (1997).

[33] L. Verde, L. Wang, A. F. Heavens, and M. Kamionkowski, Mon Not Royal Astro Soc **313**, 141 (2000).

[34] B. D. Wandelt, E. Hivon, and K. M. Gorski, Phys Rev D **64**, 083003 (2001).

[35] J. H. P. Wu *et al.*, Astrophys J Suppl **132**, 1 (2001).

Commentary by The Pittsburgh Institute for Computational Astrostatistics[2]

Andrew Jaffe has given a nice summary of challenges in analyzing the Cosmic Microwave Background (CMB). Jaffe seems to prefer a Bayesian analysis though he notes that such an analysis does have some problems. In our discussion we review the statistical model, we highlight some challenges and we suggest a new approach.

13.7 The Statistical Model

A simplified version of the problem that Jaffe states is as follows. We observe

$$\begin{aligned} T &\sim \text{Normal}(0, C_T) \\ d|T &\sim \text{Normal}(AT, N) \end{aligned}$$

where T is the vector of unobserved temperatures and d is the vector of observed data. In this simplified form, the matrices N and A are assumed known. The matrix $C_T = C_T(\omega)$ contains unknown parameters $\omega = (\omega_1, \omega_2 \ldots)$. We use ω_ℓ where Jaffe uses C_ℓ to avoid confusion with C_T and C_N (defined below).

The least squares estimate of T is $\hat{T} = (A^T N^{-1} A)^{-1} A^T N^{-1} d$ with variance $C_N = (A^T N^{-1} A)^{-1}$. We may then re-express the model as

$$\begin{aligned} T &\sim \text{Normal}(0, C_T) \\ \hat{T}|T &\sim \text{Normal}(T, C_N). \end{aligned} \tag{13.16}$$

The marginal distribution of $\hat{T}$ is $d \sim \text{Normal}(0, C_N + C_T)$. The likelihood function is

$$\mathcal{L}(\omega) \propto \frac{1}{|C_N + C_T(\omega)|^{1/2}} \exp\left\{ -\frac{1}{2} \hat{T}^T (C_T(\omega) + C_N)^{-1} \hat{T} \right\}. \tag{13.17}$$

A more direct route to the likelihood is to note that, from (13.16), $d \sim \text{Normal}(0, A^T C_T(\omega) A + N)$ and thus $\mathcal{L}(\omega) = f(d|\omega)$ which is identical to (13.17).

The likelihood $\mathcal{L}(\omega)$ depends on parameters $\omega = (\omega_1, \omega_2, \ldots,)$ which are, essentially, the values of the true power spectrum at each multipole moment. For a variety of reasons, evaluating $\mathcal{L}(\omega)$ directly is hard. Instead,

[2]The members of PICA, in reverse order of seniority, are: Woncheol Jang, Chris Miller, Andy Connolly, Jeff Schneider, Chris Genovese, Bob Nichol, Andrew Moore and Larry Wasserman.

one extracts point estimates $\hat{\omega}_\ell$ of the parameters for subset of values of ℓ. Apparently, $\hat{\omega}_\ell$ is something like the maximum likelihood estimate obtained using either a profile or marginal likelihood. At least approximately, one has

$$\hat{\omega}_\ell = \omega_\ell + \epsilon_\ell \tag{13.18}$$

where $\epsilon_\ell \sim \text{Normal}(0, \sigma_\ell^2)$ and the $\epsilon_\ell' s$ are approximately uncorrelated. We then have the approximate likelihood

$$\hat{\mathcal{L}}(\omega) \propto \exp\left\{-\frac{1}{2}\sum_\ell \frac{(\omega_\ell - \hat{\omega}_\ell)^2}{\sigma_\ell^2}\right\}.$$

Each cosmological parameters κ can be viewed as a nonlinear function of the $\omega_\ell' s$. Thus we may write $\kappa = U(\omega)$ for some function U. The Bayesian approach is to place a prior on ω which, together with the likelihood yields a posterior $\pi(\omega|d)$. ¿From the posterior, we may compute the marginal posterior $\pi(\kappa|d)$ for any quantity of interest κ.

The conceptual simplicity of the Bayesian approach is appealing. Jaffe notes, however, that there are some complications. We now discuss some of these complications.

13.8 Identifiability

Jaffe points out that there are "... approximate degeneracies in the parameter space." In statistical parlance, we say that the model is under-identified or that some parameters are non-identifiable. Basically, this means that the data are not highly informative about all the parameters. For example, if $X \sim \text{Normal}(a+b, 1)$ then we can estimate $\mu = a+b$ but we cannot estimate a and b separately. In the Bayesian framework, we could still put a prior on (a, b) and find marginal posteriors for a and b. But lack of identifiability is a warning flag that standard Bayesian or even likelihood methods may not be satisfactory. The lack of identifiability implies that the posterior will be highly sensitive to the prior. Further complications occur when we integrate out many parameters as we now explain.

13.9 Dangers of Integrating out Nuisance Parameters

Inferring a parameter of interest in the presence of nuisance parameters is conceptually simple in the Bayesian approach. One merely integrates out the nuisance parameters. But there are dangers. Integrating out many parameters can lead to a posterior distribution with strange properties. In

particular, the 95 per cent posterior interval may contain the true value of the parameter with very low frequency. Put another way, the posterior may be badly biased. Here is an extreme example due to Stein (1959). Observe independent observations $X_i \sim N(\theta_i, 1)$ $i = 1, \ldots, n$ and suppose we want to estimate $\psi = \sum_i \theta_i^2$. Suppose we use a flat prior on $\theta = (\theta_1, \ldots, \theta_n)$. Let $A = [a, \infty)$ where a is defined by $P(Z > a) = 1 - \alpha$ and Z has a non-central χ^2 distribution with n degrees of freedom and non-centrality parameter $\sum_i X_i^2$. It can be shown that A is a $1 - \alpha$ posterior interval, i.e. $P(\psi \in A | X_1, \ldots, X_n) = 1 - \alpha$. However, Stein showed that $P(\psi \in A | \theta) \approx 0$ so the interval will rarely contain the true value in the frequency sense.

13.10 An Approach Based on Nonparametric Regression

In our contribution to this volume, we took a different approach to the problem. We review the main idea here. Let $f(\ell)$ denote the true power spectrum at multipole moment ℓ. We can write (13.18) as

$$Y_\ell = f(\ell) + \epsilon_\ell$$

where $Y_\ell = \hat{\omega}_\ell$. Written this way, we see that this is really a regression problem. Our approach is to nonparametrically estimate the regression f and find a nonparametric $1 - \alpha$ confidence set C_n for f. Then we express cosmological parameters as functions of f: $\kappa = U(f)$. A confidence interval for κ is given by

$$\left(\min_{f \in C_n} U(f), \ \max_{f \in C_n} U(f) \right).$$

If the parameter κ is under-identified this will show up automatically as a wide confidence interval. Moreover, the intervals have correct frequency coverage, simultaneously over all parameters of interest. This approach sidesteps many of the problems, gives correct confidence intervals and avoids any need for integration.

13.11 Reference

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.*, **30**, 877-880.

14

Inference in Microwave Cosmology: A Frequentist Perspective

Chad M. Schafer and Philip B. Stark

ABSTRACT Estimating cosmological parameters using measurements of the Cosmic Microwave Background (CMB) is scientifically important and computationally and statistically challenging. Bayesian methods and blends of Bayesian and frequentist ideas are common in cosmology. Constructing purely frequentist confidence intervals raises questions about the probability that the intervals falsely contain incorrect values. A computable bound on this *false coverage probability* can help find optimal confidence intervals. This paper is followed by a commentary by astronomy Andrew H. Jaffe.

14.1 The Problem

Key cosmological parameters are related to tiny temperature fluctuations among photons released during the last scattering, when the universe had cooled enough for photons to travel freely. These photons form the Cosmic Microwave Background (CMB). Many cosmological models treat the observed CMB temperature anisotropy as a realization of a random $n-$vector X that has a Gaussian distribution with mean zero and covariance matrix Σ_θ, where θ is the vector of cosmological parameters. For example, in the initial analysis of the MAXIMA data [1] θ consisted of six numerical parameters, $\theta = (\Omega, \Omega_\Lambda, \Omega_b h^2, \Omega_c h^2, n_s, \tau_c)$. Henceforth here the *parameter space* $\Theta \subset \mathcal{R}^p$ is the collection of feasible models.

For any $\theta \in \Theta$, the matrix Σ_θ is a linear combination of known matrices, but the mapping from Θ into the vector of weights is highly nonlinear, and does not have a closed-form expression. This makes it computationally challenging to find the distribution of a statistic for different values of θ, which is at the heart of frequentist approaches to estimation. Moreover, two points θ, θ' in Θ may differ greatly in the value of one or more of their components, but still yield covariance matrices $\Sigma_\theta, \Sigma_{\theta'}$ that are "close" in the sense that the L_1-distance between the two probability distributions is small.

14.2 False Coverage Probability

Write $\theta = (\theta_1, \theta_2, \ldots, \theta_p) \in \Re^p$, and consider estimating θ_1, treating the other components of θ as nuisance parameters. An interval estimator for θ_1 is a function $\mathcal{C}$ that maps the space of possible observations into a set of real numbers. The *false coverage probability* $\gamma_{\mathcal{C}}(\theta, a) = P_\theta(\mathcal{C}(X) \ni a)$ is the probability that the interval includes (covers) a when θ is the truth—a fundamental measure of accuracy. A $1-\alpha$ confidence interval for θ_1 must have $\gamma_{\mathcal{C}}(\theta, \theta_1) \geq 1-\alpha$ whatever be $\theta \in \Theta$. Subject to that coverage constraint, it is desirable that $\mathcal{C}$ minimize $\gamma_{\mathcal{C}}(\theta, a)$ for all θ and all $a \neq \theta_1$, but such uniformity is rarely possible.

If $\mathcal{C}$ is a $1-\alpha$ confidence interval estimator, then

$$\gamma_{\mathcal{C}}(\theta', a) \geq \sup_{\{\theta \in \Theta:\, \theta_1 = a\}} \left(1 - \frac{1}{2}\Delta_1(\theta', \theta) - \alpha\right),$$

where Δ_1 is the L_1-*distance* between the probability distribution for X when the cosmological parameter vector equals θ and the probability distribution for X when the cosmological parameter vector equals θ'. The *affinity* between θ' and θ,

$$\rho(\theta', \theta) = \int \left(f_{\theta'}(x)\, f_\theta(x)\right)^{1/2} dx,$$

can be easier to calculate than the L_1-distance. The L_1-distance and the affinity are related [2]:

$$\gamma_{\mathcal{C}}(\theta', a) \geq \sup_{\{\theta \in \Theta:\, \theta_1 = a\}} \left(1 - \left[1 - \rho^2(\theta', \theta)\right]^{1/2} - \alpha\right).$$

For the Gaussian case at hand,

$$\rho(\theta', \theta) = \frac{2^{n/2} \left|\Sigma_\theta^{-1} + \Sigma_{\theta'}^{-1}\right|^{-1/2}}{\left|\Sigma_\theta\right|^{1/4} \left|\Sigma_{\theta'}\right|^{1/4}}.$$

This lets one bound the false coverage probability through the pixel covariance matrix Σ_θ, a natural representation of the cosmology. Currently, computing the affinity is tractable only for small experiments, but better algorithms might allow large experiments to be analyzed.

14.3 References

[1] Jaffe, A., et. al. (2001) "Cosmology from MAXIMA-1, BOOMERANG & COBE/DMR CMB Observations," *Phys. Rev. Lett.,* **86**, 3475-3479, astro-ph/000733

[2] LeCam, L.M. and Yang, G.L. (1990) *Asymptotics in Statistics: Some Basic Concepts.* First Edition. New York: Springer-Verlag.

Commentary by Andrew H. Jaffe

14.4 Alternatives: frequentist measures

The community so far has taken a largely Bayesian perspective to the analysis of CMB data. Philosophical issues aside, there are alternatives, the so-called "frequentist" or "orthodox" approach, some of whose aspects are explored in this contribution by Schafer and Stark. In this work, they attempt to characterize the 'distance' between cosmological models in terms of the ability of CMB data to discriminate between them. This arises as a problem because of the highly nonlinear relationship between the physical cosmological variables of interest and the measurable quantity, the power spectrum of CMB fluctuations. Such a characterization will prove useful not only in frequentist analyses of CMB data but also in any use of CMB power spectra which require a greater understanding of the mapping between parameters and spectra.

In the following, we wish to comment further on the Bayesian and Frequentist approaches to the CMB data analysis problem in general. Without caricaturing it too much, we can summarize the frequentist approach as follows. Just as in the Bayesian approach, we start with the likelihood function. Then, we choose an "estimator", some function of the data, chosen to somehow represent an estimate of the parameter we wish to determine. We then calculate the sampling distribution of this estimator, under the assumption of some fixed value of the theoretical parameters. If the likelihood is $P(d|\theta)$, and our estimator is $\hat{\theta}(d)$, we need $P(\hat{\theta}|\theta)$. If the estimator is some simple function of the data, then we can just use $P(x)dx = P(y)dy$ and do this analytically, otherwise we can perform Monte Carlo sampling of $P(d|\theta)$. Armed with this distribution, we define a confidence interval in the usual way. To do this, we will need

$$P(\theta_{\rm min} < \hat{\theta} < \theta_{\rm max}|\theta I) \equiv \int_{\theta_{\rm min}}^{\theta_{\rm max}} P(\hat{\theta}|\theta I)\, d\hat{\theta} = C\,. \qquad (14.1)$$

The art of frequentist statistics is in the choice of the estimator. Often, it is chosen to be unbiased, $\int d\hat{\theta}\ \hat{\theta} P(\hat{\theta}|\theta) = \theta$, and have some appropriately small or minimum variance under the same distribution.

It is worth belaboring the point that these intervals are a priori completely distinct from Bayesian intervals. These intervals say that, if you repeated the experiment many times, each time drawing from the same sampling distribution, in some fraction C of the trials you would get the answer within the stated limits.

This is in contrast to the Bayesian credible region, although the form is similar. Even in the simplest case, we are dealing with two very different functions: the Bayesian uses $P(\theta|\hat{\theta}I)$ for fixed data, $\hat{\theta}$, whereas the frequen-

tist uses $P(\hat{\theta}|\theta I)$ for fixed θ. Even for a uniform prior when these functions are proportional to one another, the two approaches are concerned with it as a function of different variables! Of course, we whet our teeth on problems in which θ and $\hat{\theta}$ appear symmetrically in the likelihood—estimation with linear, Gaussian models. In this case, then, the frequentist and Bayesian results are agreement, but in general they will not be so.

14.4.1 Monte-Carlo power spectra

Nonetheless, intuition and longstanding practice suggest that such frequentist measures have a place. Indeed, for the estimation of Power Spectra in particular, there is a deeper reason to use them, even within the Bayesian paradigm. Consider the very simplest spectrum estimation problem, an all-sky experiment with uniform noise, and a pixel scale negligible compared to that of the sky signal. Then, there are exact correspondences between the (uniform prior) Bayesian and Frequentist results: The Bayesian maximum-likelihood is the same as the frequentist mean, and the "error bars" as calculated from the Bayesian curvature are the same as the Frequentist variance. [These correspondences are not strictly true if the noise is comparable to the signal, since the frequentist mean and variance are calculated for (signal + noise) > 0 rather than for signal > 0.] These are well-known to hold asymptotically, but this is a case in which they hold for finite data as well.

Unfortunately, these correspondences do not remain exact for realistic experiments. Nonetheless, experience has thus far shown that we can indeed extract useful approximate Bayesian information from Monte Carlo power spectra [2, 3]; this is an ongoing area of research, especially as the era of experiments with $N_{\rm pix} \gg 100,000$ approaches. In the regime to be probed by MAP and Planck, with millions of pixels covering the whole sky, some alternative to the brute-force matrix manipulation will be necessary. For a full-sky experiment, we can take advantage of fast spherical-harmonic transforms to speed up some of the Monte Carlo calculations [2].

14.4.2 Frequentist parameter estimates

Finally, one can imagine an *entirely* frequentist algorithm for calculating not just the power spectrum, but the cosmological parameters themselves, directly from the data. In practice, it is found that the limits are not very different from the Bayesian intervals [1], although this is a subject of ongoing research, and not immune to technical problems.

14.5 References

[1] M. Abroe et al., Frequentist Estimation of Cosmological Parameters from the MAXIMA-I data.

[2] E. Hivon *et al.*, MASTER of the CMB Anisotropy Power Spectrum: A Fast Method for Statistical Analysis of Large and Complex CMB Data Sets Astrophys J **567**, 2 (2001)

[3] C. B. Netterfield *et al.*, A measurement by BOOMERANG of multiple peaks in the angular power spectrum of the cosmic microwave background, 2001, astro-ph/0104460.

This page intentionally left blank

15

Nonparametric Inference in Astrophysics

The Pittsburgh Institute for Computational Astrostatistics (PICA)[1]

ABSTRACT We discuss nonparametric density estimation and regression for astrophysics problems. In particular, we show how to compute nonparametric confidence intervals for the location and size of peaks of a function. We illustrate these ideas with recent data on the Cosmic Microwave Background. We also briefly discuss nonparametric Bayesian inference.
This paper is followed by commentaries by astronomers Michael A. Strauss and Jeffrey D. Scargle, and a rejoinder by the authors.

15.1 Nonparametric inference

The explosion of data in astrophysics provides unique opportunities and challenges. The challenges are mainly in data storage and manipulation. The opportunities arise from the fact that large sample sizes make nonparametric statistical methods very effective. Nonparametric methods are statistical techniques that make as few assumptions as possible about the process that generated the data. Such methods are inherently more flexible than more traditional parametric methods that impose rigid and often unrealistic assumptions. With large sample sizes, nonparametric methods make it possible to find subtle effects which might otherwise be obscured by the assumptions built into parametric methods. We begin by discussing two prototypical astrostatistics problems.

PROBLEM 1. DENSITY ESTIMATION. Let $X_1, \ldots, X_n$ denote the positions of n galaxies in a galaxy survey. Let $f(x)dx$ denote the probability of finding a galaxy in a small volume around x. The function f is a *probability density function*, satisfying $f(x) \geq 0$ and $\int f(x)dx = 1$. We regard $X_1, \ldots, X_n$ as n random draws from f. Our goal is to estimate $f(x)$ from the data $(X_1, \ldots, X_n)$ while making as few assumptions about f as possible. Figure 15.1 shows redshifts from a pencil beam from the Sloan Digital

[1]The members of PICA, in reverse order of seniority, are: Woncheol Jang, Chris Miller, Andy Connolly, Jeff Schneider, Chris Genovese, Bob Nichol, Andrew Moore and Larry Wasserman.

Sky Survey. The figure shows several nonparametric density estimates that will be described in more detail in Section 3. The structure in the data is evident only if we smooth the data by just the right amount (lower left plot).[2]

PROBLEM 2. REGRESSION. Figures 15.2 and 15.3 show cosmic microwave background (CMB) data from BOOMERaNG (Netterfield et al. 2001), Maxima (Lee et al. 2001) and DASI (Halverson 2001). The data consist of n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. Here, X_i is multipole moment and Y_i is the estimated power spectrum of the temperature fluctuations. If $f(x)$ denotes the true power spectrum then

$$Y_i = f(X_i) + \epsilon_i$$

where ϵ_i is a random error with mean 0. This is the standard *regression model*. We call Y the *response* variable and X the *covariate*. Other commonly used names for X include *predictor* and *independent* variable. The function f is called the *regression function*. The goal in nonparametric regression is to estimate f making only minimal smoothness assumptions about f.

The main messages of this paper are: (1) with large data sets one can estimate a function f *nonparametrically*, that is, without assuming that f follows some given functional form; (2) one can use the data to estimate the optimal amount of smoothing; (3) one can derive confidence sets for f as well as confidence sets for interesting features of f. The latter point is very important and is an example of where rigorous statistical methods are a necessity; the usual confidence intervals of the form "estimate plus or minus error" will not suffice.

The outline of this paper is as follows. Section 2 discusses some conceptual issues. Section 3 discusses kernel density estimation. Section 4 discusses nonparametric regression. Section 5 explains something that might be less familiar to astrophysicists, namely, nonparametric estimation via shrinkage. Section 6 discusses nonparametric confidence intervals. In Section 7 we briefly discuss nonparametric Bayesian inference. We make some concluding remarks in Section 8. Other examples of nonparametric methods in the astronomical literature can be found in Merritt (1997) and Merritt & Tremblay (1994).

Notation: We denote the mean of a random quantity X by $E(X)$, often written as $\langle X \rangle$ in physics. The variance of X is denoted by $\sigma^2 \equiv Var(X) = E(X - E(X))^2$. A random variable X has a Normal (or Gaussian) distribution with mean μ and variance σ^2, denoted by $X \sim N(\mu, \sigma^2)$, if

$$Pr(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx.$$

[2]The data involve selection bias since we can only observe brighter objects for larger redshifts. However, the sampling is fairly complete out to about $z = 0.2$.

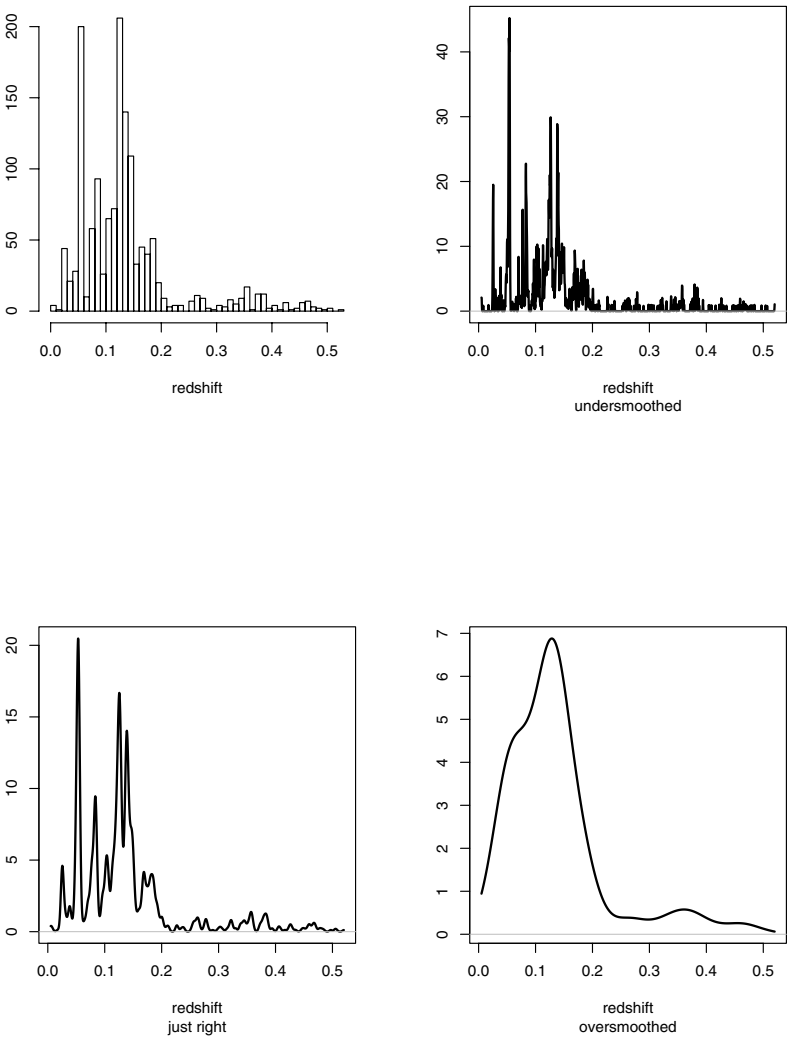


FIGURE 15.1. Redshift data. Histogram and three kernel density estimates based on three different bandwidths. The bandwidth for the estimate in the lower left panel was estimated from the data using cross-validation.

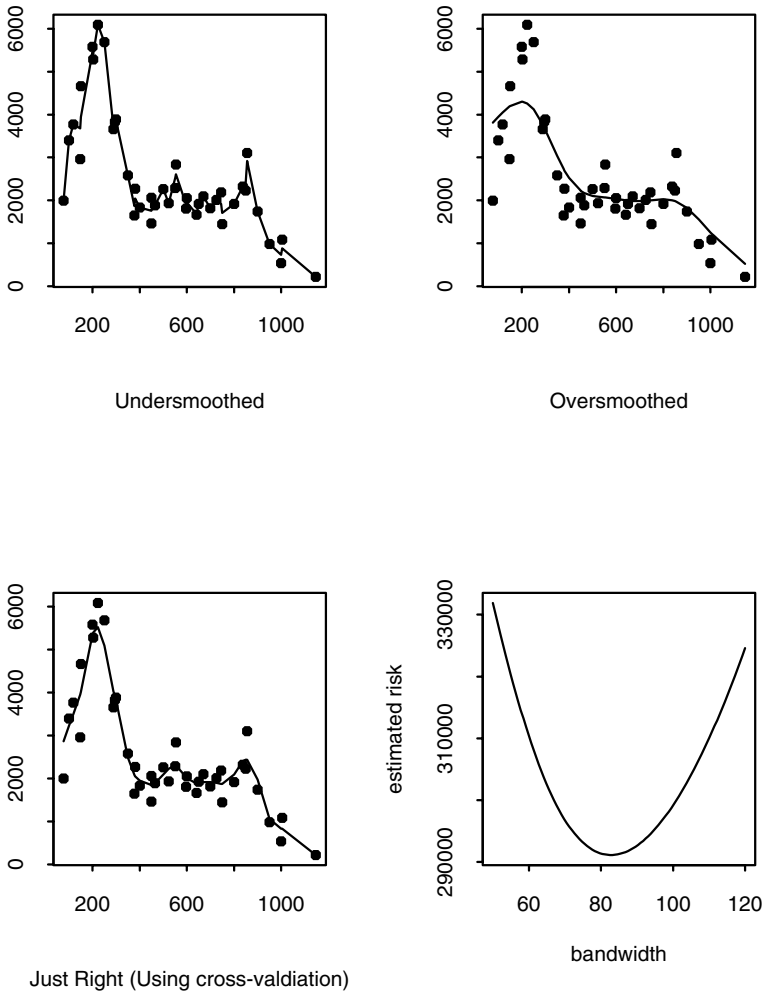


FIGURE 15.2. CMB data. Section 4 explains the methods. The first fit is undersmoothed, the second is oversmoothed and the third is based on cross-validation. The last panel shows the estimated risk versus the bandwidth of the smoother. The data are from BOOMERaNG, Maxima and DASI.

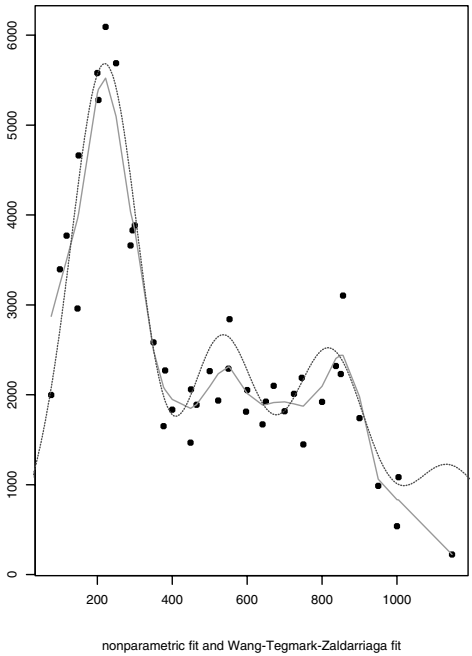


FIGURE 15.3. Best nonparametric fit together with parametric fit from Wang, Tegmark and Zaldarriaga (2001).

We use $\hat{f}$ to denote an estimate of a function f.

15.2 Some conceptual issues

15.2.1 The Bias-Variance tradeoff

In any nonparametric problem, we need to find methods that produce estimates $\hat{f}$ of the unknown function f. Obviously, we would like $\hat{f}$ to be close to f. We will measure closeness with squared error:

$$L(f, \hat{f}) = \int (f(x) - \hat{f}(x))^2 dx.$$

The average value of the error is called the *risk* or *mean squared error* (MSE) and is denoted by:

$$R(f, \hat{f}) = E_f \left[L(f, \hat{f}) \right].$$

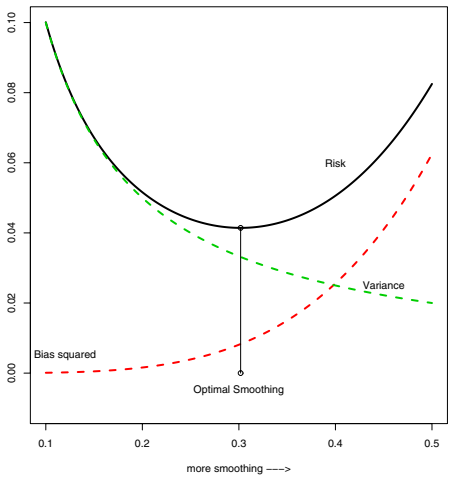


FIGURE 15.4. The Bias-Variance tradeoff. The bias increases and the variance decreases with the amount of smoothing. The optimal amount of smoothing, indicated by the vertical line, minimizes the risk = bias^2 + variance.

A simple calculation shows that

$$R(f, \hat{f}) = \int \text{Bias}_x^2 \; dx + \int \text{Var}_x \; dx$$

where $\text{Bias}_x = E[\hat{f}(x)] - f(x)$ is the bias of $\hat{f}(x)$ and $\text{Var}_x = Var[\hat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$ is the variance of $\hat{f}(x)$. In words:

$$\text{RISK} = \text{BIAS}^2 + \text{VARIANCE}.$$

Every nonparametric method involves some sort of data-smoothing. The difficult task in nonparametric inference is to determine how much smoothing to do. When the data are over-smoothed, the bias term is large and the variance is small. When the data are under-smoothed the opposite is true; see Figure 15.4. This is called the *bias-variance tradeoff*. Minimizing risk corresponds to balancing bias and variance.

15.2.2 Nonparametric confidence sets

Let f be the function of interest, for example, the true power spectrum in the CMB example. Assume that $f \in \mathcal{F}$ where $\mathcal{F}$ is some very large class of functions. A valid (large sample) $1-\alpha$ confidence set C_n is a set $C_n \subset \mathcal{F}$ such that

$$\liminf_{n\to\infty} \inf_{f\in\mathcal{F}} Pr(f \in C_n) \geq 1-\alpha$$

where n is sample size. In words, C_n traps the true function f with probability approximately $1-\alpha$ (or greater). In parametric models, confidence intervals take the form $\hat{\theta} \pm 2\,\mathrm{se}$ where $\hat{\theta}$ is an estimate of a parameter θ and se is the standard error of the estimate $\hat{\theta}$. Bayesian interval estimates take essentially the same form. Nonparametric confidence sets are derived in a different way as we shall explain later in the paper.

If prior information is available on f then it can be included by restricting C_n. For example, if it is thought that f has at most three peaks and two dips, we replace C_n with $C_n \cap \mathcal{I}$ where $\mathcal{I}$ is the set of functions with no more than three peaks and two dips.

Having constructed the confidence set we are then in a position to give confidence intervals for features of interest. We express features as functions of f, written $T(f)$. For example, $T(f)$ might denote the location of the first peak in f. Then

$$\left(\inf_{f\in C_n} T(f),\ \sup_{f\in C_n} T(f) \right)$$

is a $1-\alpha$ confidence interval for the feature $T(f)$. In fact, we can construct valid, simultaneous confidence intervals for many features of interest this way, once we have C_n. In section 6, we report such intervals for the CMB data.

Let us dispel a common criticism about confidence intervals. An oft cited but useless interpretation of a 95 per cent confidence interval is: if we repeated the experiment many times, the interval would contain the true value 95 per cent of the time. This interpretation leads many researchers to find confidence sets to be irrelevant since the repetitions are hypothetical. The correct interpretation is: if the method for constructing C_n is used on a stream of (unrelated) scientific problems, we will trap the true value 95 per cent of the time. The latter interpretation is correct and is more scientifically useful than the former.

15.2.3 Where is the likelihood?

The likelihood function, which is a familiar centerpiece of statistical inference in parametric problems, is notably absent in most nonparametric methods. It is possible to define a likelihood and even perform Bayesian inference in nonparametric problems. But for the most part, likelihood and

Bayesian methods have serious drawbacks in nonparametric settings. See section 7 for more discussion on this point.

15.3 Kernel density estimation

We now turn to problem 1, density estimation. Let us start this section with its conclusion: the choice of kernel (smoothing filter) is relatively unimportant; the choice of bandwidth (smoothing parameter) is crucial; the optimal bandwidth can be estimated from the data. Let us now explain what this means.

Let $X_1, \ldots, X_n$ denote the observed data, a sample from f. The most commonly used density estimator is the *kernel density estimator* defined by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

where K is called the *kernel* and h is called the *bandwidth*. This amounts to placing a smoothed out lump of mass of size $1/n$ over each data point X_i. Excellent references on kernel density estimation include Silverman (1986) and Scott (1992).

The kernel is usually assumed to be a smooth function satisfying $K(x) \geq 0$, $\int xK(x)dx = 0$ and $\tau \equiv \int x^2 K(x)dx > 0$. A fact that is well known in statistics but appears to be less known in astrophysics is that the choice of kernel K is not crucial. The optimal kernel that minimizes risk (for large samples) is called the Epanechnikov kernel $K(x) = .75(1 - x^2/5)/\sqrt{5}$ for $|x| < \sqrt{5}$. But the estimates using another other smooth kernel are usually numerically indistinguishable. This observation is confirmed by theoretical calculations which show that the risk is very insensitive to the choice of kernel. In this paper we use the Gaussian kernel $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$.

What does matter is the choice of bandwidth h which controls the amount of smoothing. Figure 15.1 shows the density estimate with four different bandwidths. Here we see how sensitive the estimate $\hat{f}$ is to the choice of h. Small bandwidths give very rough estimates while larger bandwidths give smoother estimates. Statistical theory tells us that, in one dimensional problems,

$$\begin{aligned} R(f, \hat{f}) &= \text{BIAS}^2 + \text{VARIANCE} \\ &\approx \frac{1}{4} h^4 c_1 A(f) + \frac{c_2}{nh} \end{aligned}$$

where $c_1 = \int x^2 K(x)dx$, $c_2 = \int K(x)^2 dx$ and $A(f) = \int (f''(x))^2 dx$. The risk is minimized by taking the bandwidth equal to

$$h_* = c_1^{-2/5} c_2^{1/5} A(f)^{-1/5} n^{-1/5}.$$

This is informative because it tells us that the best bandwidth decreases at rate $n^{-1/5}$ and leads to risk of order $O(n^{-4/5})$. Generally, one cannot find a nonparametric estimator that converges faster than $O(n^{-4/5})$. This rate is close to the rate of parametric estimators, namely, $O(n^{-1})$. The difference between these rates is the price we pay for being nonparametric.

The expression for h_* depends on the unknown density f which makes the result of little practical use. We need a data-based method for choosing h. The most common method for choosing a bandwidth h from the data is *cross-validation*. The idea is as follows.

We would like to choose h to minimize the squared error $\int(f(x) - \hat{f}(x))^2 dz = \int \hat{f}^2(x)dz - 2\int \hat{f}(x)f(x)dx + \int f^2(x)dx$. Since $\int f^2(x)dx$ does not depend on h, this corresponds to minimizing

$$J(h) = \int \hat{f}^2(x)dz - 2\int \hat{f}(x)f(x)dx.$$

It can be shown that

$$\hat{J}(h) = \int \hat{f}^2(x)dz - 2\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i).$$

is an unbiased estimate of $E[J(h)]$, where $\hat{f}_{-i}$ is the "leave-one-out" estimate obtained by omitting X_i. Some algebra shows that

$$\hat{J}(h) \approx \frac{1}{hn^2}\sum_i \sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0) \tag{15.1}$$

where $K^*(x) = K^{(2)}(x) - 2K(x)$ and $K^{(2)}$ is the convolution of K with itself. Hence, it is not actually necessary to compute $\hat{f}_{-i}$. We choose the bandwidth $\hat{h}$ that minimizes $\hat{J}(h)$. The lower left panel of Figure 15.1 was based on cross-validation. An important observation for large data bases is that (15.1) can be computed quickly using the fast Fourier transform; see Silverman (1986, p 61-66).

15.4 Nonparametric kernel regression

Returning to the regression problem, consider pairs of points $(X_1, Y_1), \ldots, (X_n, Y_n)$ related by

$$Y_i = f(X_i) + \epsilon_i.$$

The kernel method for density estimation also works for regression. The estimate $\hat{f}$ is a weighted average of the points near x: $\hat{f}(x) = \sum_{i=1}^n w_i Y_i$ where the weights are given by $w_i \propto K\left(\frac{x - X_i}{h}\right)$. This estimator is called the Nadaraya-Watson estimator. Figure 15.2 shows that estimator for the CMB data. Note the extreme dependence on the bandwidth h.

Once again, we use cross-validation to choose the bandwidth h. The risk is estimated by

$$\hat{J}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{-i}(X_i))^2.$$

The first three panels in Figure 15.2 show the regression data with different bandwidths. The second plot is based on the cross-validation bandwidth. The final plot shows the estimated risk $\hat{J}(h)$ from cross validation. Figure 15.3 compares the nonparametric fit with the fit by Wang, Tegmark and Zaldarriaga (2001).

Given the small sample size and the fact that we have completely ignored the cosmological models (as well as differential error on each data point) the nonparametric fit does a remarkable job. It "confirms," nonparametrically, the existence of three peaks, their approximate positions and approximate heights. Actually, the degree to which the fit confirms the three peaks requires confidence statements that we discuss in section 6.

15.5 Smoothing by shrinking

There is another approach to nonparametric estimation based on expanding f into an orthogonal series. The idea is to estimate the coefficients of the series and then "shrink" these estimates towards 0. The operation of shrinking is akin to smoothing. These methods have certain advantages over kernel smoothers. First, the problem of estimating the bandwidth is replaced with the problem of choosing the amount of shrinkage which is, arguably, supported by better statistical theory than the former. Second, it is easier to construct valid confidence sets for f in this framework. Third, in some problems one can choose the basis in a well-informed way which will lead to improved estimators. For example, Donoho and Johnstone (1994, 1995) and Johnstone (this volume) show that wavelet bases can be used to great advantage in certain problems.

Suppose we observe $Y_i = f(x_i) + \epsilon_i$ where, for simplicity, we assume that $x_1 = 1/n, x_2 = 2/n, \ldots, x_n = 1$. Further suppose that $\epsilon_i \sim N(0, \sigma^2)$. Let $\phi_1, \phi_2, \ldots$ be an orthonormal basis for $[0, 1]$:

$$\int_0^1 \phi_j^2(x)dx = 1 \quad \text{and} \quad \int_0^1 \phi_i(x)\phi_j(x)dx = 0 \text{ when } i \neq j.$$

For illustration, we consider the cosine basis: $\phi_1(x) \equiv 1$, $\phi_2(x) = \sqrt{2}cos(\pi x)$, $\phi_2(x) = \sqrt{2}cos(2\pi x), \ldots$. Expand f in this basis: $f(x) \sim \sum_{j=1}^{\infty}\beta_j\phi_j(x) \approx \sum_{j=1}^{n}\beta_j\phi_j(x)$. Estimating f then amounts to estimating the β_j's. Let $Z_j = n^{-1/2}\sum_{i=1}^{n} Y_i\phi_j(i/n)$. It can be shown that $Z_j \approx N\left(\theta_j, \sigma^2\right), \quad j = 1, \ldots, n$ where $\theta_j = \sqrt{n}\beta_j$. Once we have estimates $\hat{\theta}_j$, we set $\hat{\beta}_j = n^{-1/2}\hat{\theta}_j$ and $\hat{f}(x) = \sum_{j=1}^{n}\hat{\beta}_j\phi_j(x)$.

How do we estimate $\theta = (\theta_1, \ldots, \theta_n)$ from $Z = (Z_1, \ldots, Z_n)$? A crude estimate is $\hat{\theta}_j = Z_j$, $j = 1, \ldots, n$. This leads to a very noisy (unsmoothed) estimate of f. Better estimates can be found by using *shrinkage* estimators. The idea – which goes back to James and Stein (1961) and Stein (1981) – is to estimate θ by shrinking the vector Z closer to the origin. A major discovery in mathematical statistics was that careful shrinkage leads to estimates with much smaller risk. Following Beran (2000) we consider shrinkage estimators of the form $\hat{\theta} = (\alpha_1 Z_1, \alpha_2 Z_2, \ldots, \alpha_n Z_n)$ where $1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq 0$ which forces more shrinkage for higher frequency cosine terms.

Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ and let $R(\alpha)$ denote the risk of $\hat{\theta}$ using shrinkage vector α. An estimate of $R(\alpha)$, called *Stein's unbiased risk estimate* (SURE), is

$$\hat{R}(\alpha) = \sum_j \left[\hat{\sigma}^2 \alpha_j^2 + (Z_j^2 - \hat{\sigma}^2)(1 - \alpha_j)^2\right]$$

where σ^2 has been estimated by $\hat{\sigma}^2 = \frac{1}{k}\sum_{i=n-k+1}^{n} Z_i^2$ with $k < n$. Using appropriate numerical techniques, we minimize $\hat{R}(\alpha)$ subject to the monotonicity constraint. The minimizer is denoted by $\hat{\alpha}$ and the final estimate is $\hat{\theta} = (\hat{\alpha}_1 Z_1, \hat{\alpha}_2 Z_2, \ldots, \hat{\alpha}_n Z_n)$. Beran (2000) shows that the estimator obtained this way has some important optimality properties. Beran calls this approach REACT (Risk Estimation, Adaptation, and Coordinate Transformation). The estimated function $\hat{f}$ turns out to be similar to the kernel estimator; due to space limitations we omit the plot.

15.6 Confidence sets

When estimating a scalar quantity θ with an estimator $\hat{\theta}$, it is common to summarize the uncertainty for the estimate by reporting $\hat{\theta} \pm 2se$ where $\text{se} \approx \sqrt{Var(\hat{\theta})}$ is the *standard error* of the estimator. Under certain *regularity conditions*, this interval is a 95 per cent confidence interval, that is,

$$Pr\left(\hat{\theta} - 2\,\text{se} \;\leq\; \theta \;\leq\; \hat{\theta} + 2\,\text{se}\right) \approx .95.$$

This follows because, under the conditions alluded to above, $\hat{\theta} \approx N(\theta, \text{se}^2)$.

But the "plus or minus 2 standard errors" rule fails in nonparametric inference. Consider estimating a density $f(x)$ at a single point x with a kernel density estimator. It turns out that

$$\hat{f}(x) \approx N\left(f(x) + \text{bias}, \frac{c_2 f(x)}{nh}\right) \tag{15.2}$$

where

$$\text{bias} = \frac{1}{2}h^2 f''(x) c_1 \tag{15.3}$$

is the bias, $c_1 = \int x^2 K(x)dx$ and $c_2 = \int K^2(x)dx$. The estimated standard error is

$$\text{se} = \left\{ \frac{c_2 \hat{f}(x)}{nh} \right\}^{1/2}. \tag{15.4}$$

Observe from (15.2) that $(\hat{f}(x) - f(x))/se \approx N(\text{bias/se}, 1)$. If use the "estimate plus/minus 2 se" rule then

$$\begin{aligned} Pr\left(\hat{f}(x) - 2\,\text{se} \leq f(x) \leq \hat{f}(x) + 2\,\text{se}\right) &= Pr\left(-2 \leq \frac{\hat{f}(x) - f(x)}{\text{se}} \leq 2\right) \\ &\approx Pr\left(-2 \leq N\left(\frac{\text{bias}}{\text{se}}, 1\right) \leq 2\right). \end{aligned}$$

If bias/se $\to 0$ then this becomes $Pr(-2 < N(0,1) < 2) \approx .95$. As we explained in Section 2, the optimal bandwidth is of the form $h = cn^{-1/5}$. If you plug $h = cn^{-1/5}$ this into (15.3) and(15.4) you will see that bias/se does not tend to 0. The confidence interval will have coverage less than .95. In summary, "estimate plus/minus 2 standard errors" is not appropriate in nonparametric inference. There are a variety of ways to deal with this problem. One is to use kernels with a suboptimal bandwidth. This undersmooths the estimate resulting in a reduction of bias.

Another approach is based on the REACT method (Beran and Dumbgen, 1998). We construct a confidence set C_n for the vector of function values at the observed data, $\mathbf{f}_n = (f(X_1), \ldots, f(X_n))$. The confidence set C_n satisfies: for any $c > 0$,

$$\limsup_{n\to\infty} \sup_{||\mathbf{f}_n|| \leq c} |Pr(\mathbf{f}_n \in C_n) - (1-\alpha)| \to 0$$

where $||a|| = \sqrt{n^{-1}\sum_i a_i^2}$. The supremum is important: it means that the accuracy of the coverage probability does not depend on the true (unknown) function.

The confidence set, expressed in terms of the coefficients θ, is

$$C_n = \left\{ \theta : \; n^{-1} \sum_j (\theta_j - \hat{\theta}_j)^2 \leq \hat{R}_r + n^{-1/2} \hat{\tau} z_\alpha \right\}$$

where z_α is such that $P(Z > z_\alpha) = \alpha$ where $Z \sim N(0,1)$ and $\hat{\tau}$ is a quantity computed from the data whose formula we omit here. Finally, the confidence set for f is

$$\mathcal{D}_n = \left\{ f : \; f = \sum_j \beta_j \phi_j : \; \beta_j = n^{-1/2}\theta_j, \; \theta \in \mathcal{C}_n \right\}.$$

Let us return to the CMB example. We constructed a 95 per cent confidence set C_n, then we searched over C_n and found the possible number, location and heights of the peaks. We restricted the search to functions with no more than three peaks and two dips as it was deemed unlikely that the true power spectrum would have more than three peaks. Curves with one or two peaks cannot be ruled out at the 95 per cent level. The confidence intervals, restricted to three peak models, are as follows.

Peak	Location	Height
1	(118,300)	(4361,8055)
2	(377,650)	(1822,4798)
3	(597,900)	(1839,4683)

The 95 per cent confidence interval for the ratio of the height of the second peak divided by the height of the first peak is $(.21, 1.4)$. The 95 per cent confidence interval for the ratio of the height of the third peak divided by the height of the second peak is $(.22, 2.82)$. Not surprisingly, the intervals are broad because the data set is small. In a further work by our group (Miller et al 2001) we investigate the improvements in measurement error that are needed to get more precise confidence sets.

15.7 Nonparametric Bayes

There seems to be great interest in Bayesian methods in astrophysics. The reader might wonder if it is possible to perform nonparametric Bayesian inference. The answer is, sort of.

Consider estimating a density f assumed to belong to some large class of functions such as $\mathcal{F} = \{f : \int (f''(x))^2 dx \leq C\}$. The "parameter" is the function f and the likelihood function is $\mathcal{L}_n(f) = \prod_{i=1}^n f(X_i)$. Maximizing the likelihood leads to the absurd density estimate that puts infinite spikes on each data point. It is possible to put a prior π over $\mathcal{F}$. The posterior distribution on $\mathcal{F}$ is well defined and Bayes theorem still holds:

$$Pr(f \in C \mid X_1, \ldots, X_n) = \frac{\int_C \mathcal{L}_n(f) d\pi(f)}{\int_{\mathcal{F}} \mathcal{L}_n(f) d\pi(f)}.$$

Lest this seem somewhat abstract, take note that much recent work in statistics lately has led to methods for computing this posterior.

However, there is a problem. The parameter space $\mathcal{F}$ is infinite dimensional and, in such cases, the prior π is extremely influential. The result is that the posterior may concentrate around the true function very slowly. Worse, the 95 per cent Bayesian credible sets will contain the true function with very low frequency. In many cases the frequency coverage probability of the Bayesian 95 per cent credible set is near 0! Since high dimensional parametric models behave like nonparametric models, these remarks should

give us pause before casually applying Bayesian methods to parametric models with many parameters.

The results that make these comments precise are fairly technical. The interested reader is referred to Diaconis and Freedman (1986), Barron, Schervish and Wasserman (1999), Ghosal, Ghosh and van der Vaart (2000), Freedman (2000), Zhao (2000) and Shen and Wasserman (2001). The bottom line: in nonparametric problems Bayesian inference is an interesting research area but is not (yet?) a practical tool.

15.8 Conclusion

Nonparametric methods are at their best when the sample size is large. The amount and quality of astrophysics data have increased dramatically in the last few years. For this reason, we believe that nonparametric methods will play an increasingly important role in astrophysics. We have tried to illustrate some of the key ideas and methods here. But we have really only touched on a few main points. We hope through our continued interdisciplinary collaboration and through others like it elsewhere, that the development of nonparametric techniques in astrophysics will continue in the future.

15.9 References

Barron, A., Schervish, M. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics.* **27**, 536-561.

Beran, R. (2000). REACT Scatterplot Smoothers: Superefficiency Through Basis Economy. *Journal of the American Statistical Association.*

Beran, R. and Dumbgen, L. (1998). Modulation Estimators and Confidence Sets. *The Annals of Statistics*, 26, 1826-1856.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14**, 1-67.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81, 425-455.

Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90, 1200-1224.

Freedman, D. A. (1999). On the Bernstein-Von Mises theorem with infinite dimensional parameters. *The Annals of Statistics*, **27**, 1119-1140.

Ghosal, S., Ghosh, J.K. and van der Vaart, A. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, **28**, 500-531.

Halverson, N.W. et al. (2002), *The Astrophysical Journal*, **568**, 38-45, astro-ph/0104489

James, W. and Stein, C. (1961). Estimation with quadratic loss. In: *Proc. Fourth Berkeley Symp. Math. Statist. Prob., 1, (J. Neyman, ed.)* University of California Press, Berkeley. pp. 361–380.

Lee, A.T., et al. (2001) *Astrophysical Journal*, **561**, L1

Merritt, David, and Tremblay, Benoit (1994). Nonparametric estimation of density profiles. The Astronomical Journal, vol. 108, no. 2, p. 514-537

Merritt, David (1997). Recovering Velocity Distributions Via Penalized Likelihood. Astronomical Journal v.114, p. 228-237.

Miller, C., Nichol, R., Genovese, C., Wasserman, L. (2002). A nonparametric analysis of the cosmic microwave background power spectrum. Astrophysical Journal, V.565, p. L67-L70.

Netterfield, C.B., et al. (2001) astro-ph/0104460.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, New York.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. To appear: *The Annals of Statistics.*

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman-Hall: New York.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, **9**, 1135-1151.

Wang, X., Tegmark, M., and Zaldarriaga, M. (2001). To appear in Physcial Review D. astro-ph/0105091.

Zhao, L. H. (2000). Bayesian aspects of some nonparametric priors. *The Annals of Statistics*, **28**, 532-552.

Commentary by Michael A. Strauss[3]

I enjoyed this paper a lot; this is one statistics talk where as an astronomer, I could immediately see application to problems that I tackle every day. I would point out that for the problems used to illustrate the talk, it was not always clear what the scientific question that was being addressed actually was, and therefore the statistical methods used were not necessarily optimal. For example, the example of the large-scale distribution of galaxies was given to show how one can choose an optimal filter. In fact, astronomers are interested in the structures on a *range* of scales. As Figure 15.2 in my contribution to these proceedings makes clear, there is a great deal of information for a variety of smoothing lengths, all of which is useful in trying to come to a physical understanding of galaxy clustering (see also the contributions by V. Martínez and R. van de Weygaert). One should also keep in mind that the galaxy distribution data become noisier as one goes further out (galaxies further away are fainter than those closer in), and astronomers have used methods like the Wiener filter and its variants to come up with optimal smoothing of the data.

I was quite impressed by the demonstration of techniques for demonstrating the validity of certain features in the data (such as the third bump in the power spectrum of the Cosmic Microwave Data) without fitting explicit models; that is quite an important advance. Nevertheless, it is worth emphasizing that the fitting of physical models to data continues to have its place in analyses of these data; it is these fits which allow us to constrain cosmological models directly from the CMB observations.

Finally, let me echo one of the more important messages of this paper, namely that the *shape* of one's filter is not nearly as important as its *width*. This is a non-trivial, and sometimes non-intuitive fact, but understanding it makes life quite a bit simpler for astronomers when faced with a bewildering variety of different filtering techniques for their data.

[3]Princeton University Observatory

Commentary by Jeffrey D. Scargle[4]

15.10 Nonparametric Inference

The excellent overview by the Pittsburgh Institute for Computational Astrostatistics (PICA) group (Jang, Miller, Connolly, Schneider, Genovese, Nichol, Moore, and Wasserman) begins with the statement

> "Nonparametric methods are statistical techniques that make as few assumptions as possible about the process that generated the data. Such methods are inherently more flexible than more traditional parametric methods that impose rigid and often unrealistic assumptions."

The informality of this definition is warranted by the fact that the terms *parametric* and *nonparametric* are used somewhat loosely, and in different ways in a variety of contexts.

A few additional comments may help astronomers. Parametric methods typically use models in specific functional forms containing one or more parameters. Example: a Gaussian form for a distribution, where the parameters are the mean and variance. In contrast, *nonparametric* methods use generic models in which the number of parameters depends on the number of data points (Rissanen 1989). Example: Fourier and wavelet representations, in which the number of coefficients in is equal to the number of data points. Polynomial fitting is another example, where the number of parameters is one of the parameters of the problem; in such cases determining the optimum order of the model is often the hardest part of the problem.

Paradoxically, then, nonparametric methods do not avoid the use of parameters. The distinction is between generic models and those with specialized, explicit forms – not the use of parameters. Unfortunately, through long usage we are stuck with this misleading terminology.

15.11 Smoothing and the Bias-Variance Tradeoff

PICA gives a very clear picture of nonparametric inference, emphasizing the trade-off between bias and variance of estimators. They discuss density estimation and regression, both of which can be viewed as determination of an unknown function f of some independent variable, say X, using noisy observations Y:

$$Y = f(X) + \epsilon \ , \tag{15.5}$$

[4]NASA Ames Research Center

where ϵ is the noise, here assumed additive. They treat the estimation procedure as a smoothing of the observations. The key point is that the optimum amount of smoothing, which is unknown, can be determined by finding the smoothing parameter that minimizes the *mean squared error*.

This analysis implicitly *assumes that the function f is relatively smooth*, so that the roughness of the samples Y is due to observational noise. Removing noise and smoothing are thus viewed as essentially synonymous. This viewpoint is expressed in the statement (PICA §2):

> "Every nonparametric method involves some sort of data-smoothing."

In astronomical and other exploratory data analysis where one is searching for a signal of unknown smoothness, this assumption may not be desirable. That wavelet denoising (Iain Johnstone's paper in these proceedings) and Bayesian Blocks (my paper in these proceedings) can detect structure on any scale, as long as it is supported by the data, are counterexamples to PICA's statement above.

That this is a real issue is exemplified by the extremely short spike found within a gamma-ray burst (Scargle, Norris, and Bonnell 1998). This very real and interesting $\approx 100\mu$-sec-scale feature would have been completely lost in any histogram with bin size indicated by the $\approx$ second to millisecond time scale suggested by other bursts – and not contradicted by the appearance of the raw data for this one. More to the point, it would be washed away by almost any known smoothing technique.

15.12 Nonparametric Bayesian Methods

PICA espouses a rather pessimistic view of Bayesian methods for nonparametric inference (§7). I disagree.

First, the well-studied *smoothness priors* for f allow one to express in a precise way assumptions of the kind discussed above. Second, consider the discussion in §7 that takes the model space to be functions with square-integrable second derivatives, and concludes:

> "Maximizing the likelihood leads to the absurd density estimate that puts infinite spikes on each data point."

There are several problems here. One is that the usefulness of a density estimate depends on how you look at it and what you are going to do with it. Consider the cumulative distribution function (CDF)

$$F = \int_{-\infty}^{x} f(x')dx'. \tag{15.6}$$

The CDF corresponding to placing δ-functions in f at each datum has some nice features, and is far from absurd. F could be quite useful for computing estimates of other quantities, such as moments of the distribution.

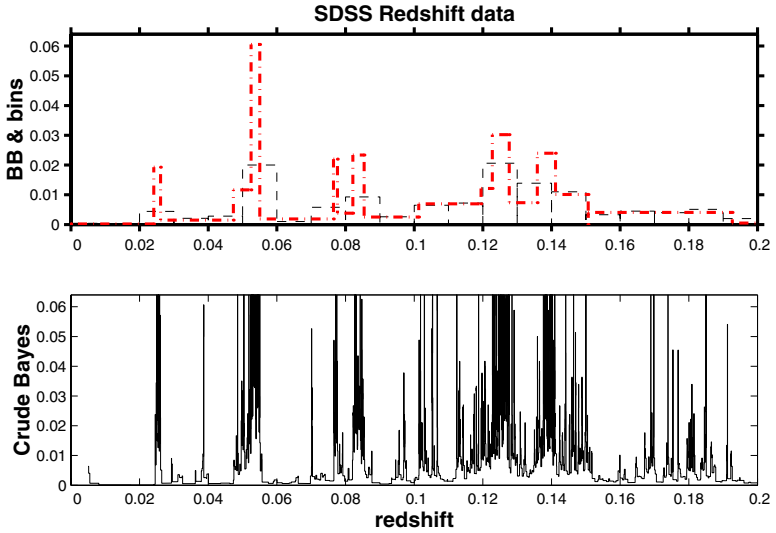


FIGURE 15.5. PICA's SDSS Redshift data. Top: Bayesian Blocks (solid lines) compared to the PICA binning of the data (dashed lines). Bottom: Crude local density estimator; the vertical scale is as in the top panel, so many spikes are off-scale.

A more fundamental problem has to do with the specification of the model space, which can be seen by studying the redshift data from the Sloan Digital Sky Survey discussed by PICA and replotted here in Figure 15.5 with an expanded redshift scale.

Astronomical data is always discrete in nature. Not only are the data recorded with finite precision, but usually there is an inherent quantization imposed by the measurement apparatus. Hence use of results based on continuous variables is dangerous.

Of course, the PICA argument has a discrete version, in which f is allowed to have arbitrary values at the discrete grid of allowed values of X. It would be more reasonable to model f as piecewise constant on intervals centered on the data points – i.e. assign to x_n the interval from $\frac{1}{2}(x_{n-1}+x_n)$ to $\frac{1}{2}(x_n+x_{n+1})$. The local estimate obtained by maximizing the likelihood for each such interval separately, is shown in the bottom panel of Figure 15.5. While this could be described as *undersmoothed*, it is definitely not an absurd estimate of the true density.

A more complete Bayesian solution, based on a model space consisting of functions piece-wise constant on arbitrary intervals (Scargle, 1998, 2001), is plotted in the top panel, superimposed on an evenly spaced binned histogram more or less the same as the upper-left panel of PICA's Figure 15.1. Within the assumptions of this model, this algorithm aims at finding the optimum such piecewise constant representation. No smoothness crite-

rion has been imposed, the goal being to provide an estimate expressing all of the structure, regardless of scale, that is supported by the data. One accordingly avoids the dependence of the solution on the smoothing parameter (since there is none!), or on the sizes and locations of preselected bins.

The rather sharp features present in the Bayesian Block solution are consistent with the emerging picture of the Universe consisting of sheets of galaxies. Of course we are handicapped by not knowing what the true density is. Studies with simulated data of known properties are indicated.

Piecewise constant representations, although surprisingly useful (my and Alanna Connors's papers in these proceedings), are restrictive. A fully Bayesian solution, with carefully applied smoothness priors, awaits future work.

I thank the PICA group for providing me the data shown here.

15.13 References

Rissanen, J., 1989, *Stochastic Complexity and Statistical Inquiry*, Singapore: World Scientific.

Scargle, J., 1998, Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, A New Method to Analyze Structure in Photon Counting Data, *Astrophysical Journal*, **504**, p. 405-418, see astro-ph/9711233

Scargle, J. D., Norris, J. P., and Bonnell, J. T., 1998, Attributes of GRB Pulses: Bayesian Blocks Analysis of TTE Data; a Microburst in GRB920229, Gamma-Ray Bursts: 4th Huntsville Symposium, September, 1997, Eds Meegan, C. A., Preece, R.D., and Koshut, T., American Institute of Physics (AIP conference proceedings; 428), p. 181.

Scargle, J. D., 2001, Bayesian Blocks: Divide and Conquer, MCMC, and Cell Coalescence Approaches, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering,* 19th International Workshop, Boise, Idaho, 2-5 August, 1999. Eds. Josh Rychert, Gary Erickson and Ray Smith, AIP Conference Proceedings, Vol. 567, p. 245-256

Rejoinder by PICA

We thank Michael Strauss and Jeff Scargle for their comments. We completely agree that the parametric/nonparametric terminology is less than satisfactory and we are grateful that he added further clarification on this points.

We wrote that every nonparametric method involves smoothing and Jeff is correct to point out that this is not true. First, as Jeff notes, the cumulative distribution function can be estimated by the empirical distribution functions that put mass $1/n$ at each data point. This involves no smoothing at all and the resulting estimator is consistent. But for estimating regression functions and density functions, the raw data cannot be used 'as is'. Perhaps we used the word smoothing too loosely. We meant that some processing step inevitably involves some sort of bias-variance tradeoff. We refer to this process broadly as smoothing.

Wavelet denoising and Bayesian Blocks do require smoothing in this broader sense. For example, wavelet denoising involves setting a threshold for the wavelet coefficients. Larger thresholds lead to more bias and less variance and smaller thresholds lead to more variance and less bias. The reason wavelets detect finer structure is because they form an unconditional basis for a larger class of function speces than a Fourier basis, and because of the non-linear nature of the thresholding rule.

Similarly, a smoothness assumption is implicitly build into the prior for Bayesian Blocks. If the prior does not induce a smoothness restriction than the posterior will not be consistent (Barron et al. 1999).

We did not mean to sound so pessibistic about nonparametric Bayesian methods. Currently, we know of no nonparametric Bayesian methods whose 95 percent confidence sets actually contain the true function approximately 95 percent of the time in the frequency sense. But we hope one will be found and we encourage Jeff and others to keep developing such tools.

Barron, A., Schervish, M. & Wasserman, L., 1999. The consistency of posterior distributions in nonparametric problems, *Annals of Statistics*, 14, 536-561.

This page intentionally left blank

16

Random Forests: Finding Quasars

Leo Breiman[1], Michael Last and John Rice

ABSTRACT
The automatic classification of objects from catalogues or other sources of data is a common statistical problem in many astronomical surveys. We describe an effective method, Random Forests, in which votes for class membership are polled from a large random ensemble of tree classifiers. This procedure is illustrated by the problem of identifying quasars from the FIRST survey.
This paper is followed by a commentary by astronomer Eric D. Feigelson.

16.1 Introduction

The automatic classification of objects from catalogues is a common statistical problem encountered in many surveys. From a list of values of variables (e.g. color, magnitude) associated with an object, it is desired to identify the object's type (e.g. star, galaxy). In the last section of this paper, we discuss an example in which we classify objects as quasars or non-quasars using the combined results of a radio survey and an optical survey. Such classification helps guide the choice of which objects to follow up with relatively expensive spectroscopic measurements.

The last five years of research in the Machine Learning field has produced classification methods with significantly higher accuracies than previous methods. There have been two lines of productive research. One estimates the border between classes by increasing the dimensionality of the input predictor space. The classifiers produced by this method are called Support Vector Machines [Vapnik(1995)], [Vapnik(1998)].

The other creates a varied ensemble of classifiers, lets each classifier vote for the class it favors, and then outputs the classification that has the plurality of votes. The most accurate classifier of this type is called Random Forests [Breiman 1999], abbreviated RF. We will describe the construction of RF, and compare its performance with single CART trees. RF can also quantify which variables are important to the class and this procedure is described as well.

[1]Department of Statistics, University of California, Berkeley

16.2 Construction of Random Forests (RF)

Recall the steps in constructing an ordinary CART tree: A node is a subset of the data. The root node contains all data. At each node, search through all variables to find the best split into two children nodes. Split all the way down and then prune the tree up to get minimal test set error.

The construction of RF differs:

1. The root node contains a bootstrap sample from the original data. A different bootstrap sample is drawn for each tree to be grown.
2. An integer K is fixed, K is much smaller than the number of variables. K is the only parameter that needs to be specified. The default is the square root of number of variables.
3. At each node, K of the variables are selected at random. Only these variables are searched through for the best split. The largest tree possible is grown and is not pruned.
4. The forest consists of N trees. To classify a new object having coordinates x, put x down each of the N trees. Each tree gives a classification for x.
5. The forest chooses that classification having the most votes out of the N votes cast

Code for random forests is publicly available. [2]

16.3 Accuracy of RF Compared to CART

Accuracy of single trees (CART) to random forests is compared using data sets from the UCI repository (ftp.ics.uci.edu/pub/MachineLearningDatabases).

For the five smaller data sets above the line, the test set error was estimated by leaving out a random 10% of the data, then running CART and the forest on the other 90%.

The left-out 10% was run down the tree and the forest and the error on this 10% computed for both. This was repeated 100 times and the errors averaged. The larger data sets below the line came with a separate test set.

[2] http://www.stat.Berkeley.EDU/users/breiman/

TABLE 16.1. **Data Set Descriptions**

Data Set	Training	Test	Variables	Classes
cancer	699	-	9	2
ionosphere	351	-	34	2
diabetes	768	-	8	2
glass	214	-	9	6
soybean	683	-	35	19
letters	15,000	5000	16	26
satellite	4,435	2000	36	6
shuttle	43,500	14,500	9	7
DNA	2,000	1,186	60	3
digit	7,291	2,007	256	10

TABLE 16.2. **Test Set Misclassification Error (%)**

Data Set	Forest	Single Tree
breast cancer	2.9	5.9
ionosphere	5.5	11.2
diabetes	24.2	25.3
glass	22.0	30.4
soybean	5.7	8.6
letters	3.4	12.4
satellite	8.6	14.8
shuttle	7.0	62.0
DNA	3.9	6.2
digit	6.2	17.1

The reductions in test set error are dramatic–often over 50%, and almost always, over 30%. RF achieves state-of-the-art accuracy and on the synthetic data sets it has been tested on, where the lowest possible error rate can be analytically computed, gets close to this lower limit.

16.4 RF Byproducts

A wealth of information can be obtained in a single run of Random Forests, including test set error rate and variable importance. This information comes from using the "out-of-bag" cases in the training set that have been left out of the bootstrapped training set.

Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the k-th tree.

Test Set Error Rate: Put each case left out in the construction of the k-th tree down the k-th tree to get a classification. In this way, a test set classification is gotten for each case in about one third of the trees. Let the final test set classification of the forest be the class having the most votes. Compare this classification with the classification given in the data to get an estimate of the test set error.

Variable Importance: To estimate the importance of variable #4: In the left out cases for the k-th tree, randomly permute all values of variable #4. Put these new covariate values down the tree and get classifications. Proceed as though computing a new test set error. The amount by which this new error exceeds the original test set error is defined as the importance of variable #4.

16.5 Application: Automatic Identification of Quasars

In [White et al.(2000)] decision trees were used to automatically identify quasars, combining information from the FIRST survey and from POSS-I plates. The aim was to construct a radio-selected sample of optically bright quasars, and in particular to bridge the gap between radio-loud and radio-quiet quasars. Continuing that effort with an enlarged set of data, we trained classifiers on 2127 objects (1366 quasars) identified from their spectra.

The following variables were used in constructing the classifiers:

1. The result of a star/galaxy classifier for the red plate
2. Another star/galaxy classifier for the red plate
3. Red magnitude

4. A star/galaxy classifier for the blue plate
5. Another star/galaxy classifier for the blue plate
6. Blue magnitude
7. Color (blue magnitude minus red magnitude)
8. Separation between radio and optical sources in arcseconds
9. Another estimate of separation between radio and optical sources
10. Radio peak flux
11. Radio integrated flux

On the basis of their spectra, the objects were classified in the following categories:

1. A: Narrow line Active Galactic Nucleus
2. B: BL Lac (a kind of blazar)
3. G: Galaxy without emission lines
4. H: H/II star forming galaxy
5. Q: Quasar
6. S: Star

The task was thus to use the measurements of the variables listed above to automatically classify objects into these categories and in particular to discriminate quasars from other types of objects. As would be expected, there is a substantial amount of information available from color and magnitude, as shown in Figure 16.1.

An automatic classifier carves up the 11 dimensional space defined by the variables into regions corresponding to different types of objects. This is illustrated in Figure 16.2 which shows a projection of the data onto a plane determined by several of the variables. The figure indicates that one should be able to achieve fairly good separation.

When objects were classified as either quasars or non-quasars, random forests had misclassification rate of 14.3%. For baseline comparison, a standard classification tree had an error rate of 19.7%. A support vector machine had an error rate of 13.9%, comparable to that of random forests. It might be thought that basically only color and magnitude are informative, but this is not the case: when only these variables are used as classifiers, the error rate is 19.2%. When the categories of blazars and quasars were merged so that one does not try to distinguish between them, the error rate for random forests dropped to 10.5% (Examination of figures like those above makes it clear that it is quite difficult to discriminate quasars from blazars.)

Figure 16.3 shows that the misidentified quasars tended to be bluer and brighter. The quasar fraction increases for fainter objects (because the number of quasars per square degree rises very rapidly as we go fainter), which makes fainter samples easier to classify.

Variables were scored for importance, as discussed above: Color is thus

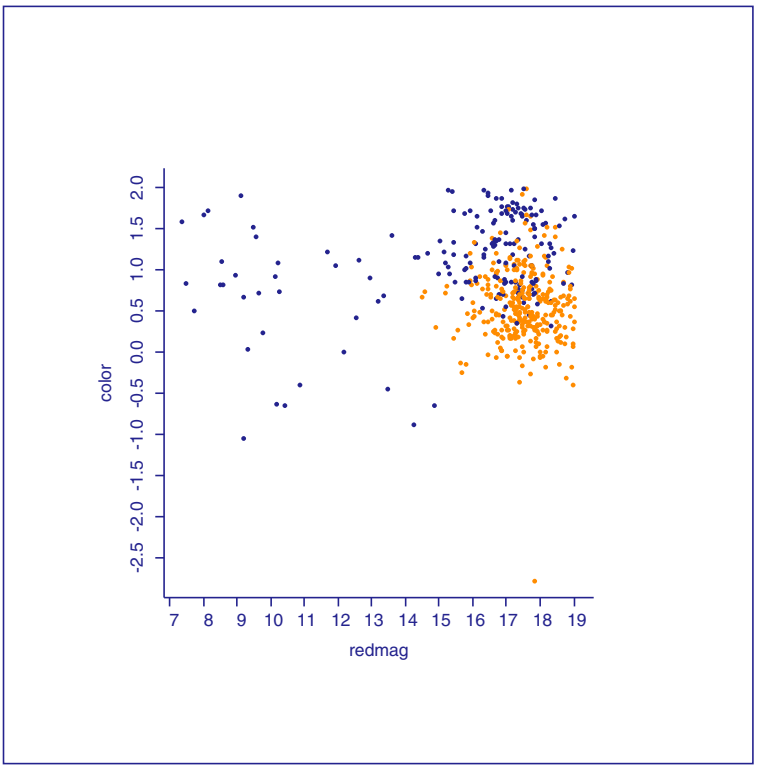


FIGURE 16.1. Color versus red magnitude. Quasars are lighter.

TABLE 16.3. **Variable importance**

Variable	Importance
red star/gal classifier 1	.99
red star/gal classifier 2	.33
red magnitude	4.95
blue star/gal classifier 1	.33
blue star/gal classifier 2	2.64
blue magnitude	.9
color	33.0
radio-optical separation 1	5.61
radio-optical separation 2	1.53
radio peak flux	1.98
radio integrated flux	.33

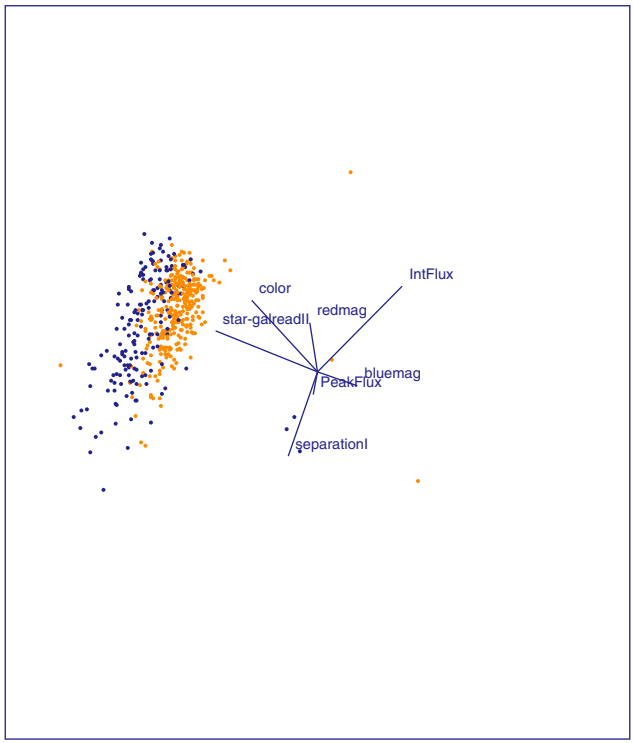


FIGURE 16.2. A projection of the data. Quasars are lighter.

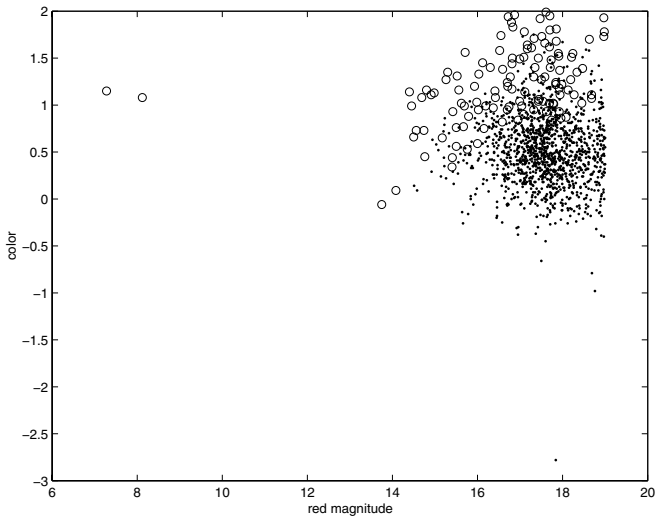


FIGURE 16.3. Errors in quasar identification. Misidentified quasars are shown as circles.

TABLE 16.4. **Confusion matrix**

True Class	Assigned Class					
	A	B	G	H	Q	S
A	36	5	15	26	9	7
B	4	10	2	2	5	0
G	10	1	1	9	90	0
H	59	9	28	159	40	18
Q	36	85	3	59	1292	58
S	1	2	2	8	20	88

by far the most important variable for determining the classification, but as we have seen above adding variables other than color and magnitude increases predictive performance.

Random forests produce an estimate of the probability, $P(Q)$, that an object is a quasar. Examination of the results shows that these probabilities are "calibrated," i.e. of those objects for which the estimated probability of being a quasar is 90%, about 90% are in fact quasars, etc. The decision of whether to follow up an object with spectroscopic observations can thus be guided by its probability $P(Q)$.

An object can be declared to be a quasar if $P(Q) > p$, for a given p (in the results quoted above we used the threshold $p = 0.5$). Varying p produces a tradeoff between two types of errors – false positives (calling an object a quasar when it is not) and false negatives (failing to identify a real quasar as such). Equivalently we can define completeness as the fraction of actual quasars included and reliability as the fraction selected that are in fact quasars and view completeness and reliability as functions of p, as shown in Figure 16.4. From this figure we see that completeness of 90% can be achieved with about 87% reliability.

The classification errors can be examined to find those for which misidentified quasars were badly misidentified, i.e. $P(Q)$ is small. The quasar fraction increases for fainter objects (because the number of quasars per square degree rises very rapidly as we go fainter), which makes fainter samples easier to classify. You can see this effect in your plot that shows the misclassified objects (as large colored dots) in a plot of color vs. red magnitude– misclassifications are much more common for quasars brighter than 16th magnitude. Also, bluer quasars tend to be more likely to be misidentified.

More ambitiously, we attempted to classify each object into each of the categories above, not merely as quasar or non-quasar. The results are shown in the following "confusion matrix," shown in Table 16.5. The columns give the true classes and the rows give the guessed classes. Thus 59 of the 1366 quasars were misidentified as H, etc. It is interesting that the completeness-reliability curve for classifying quasars when attempting to

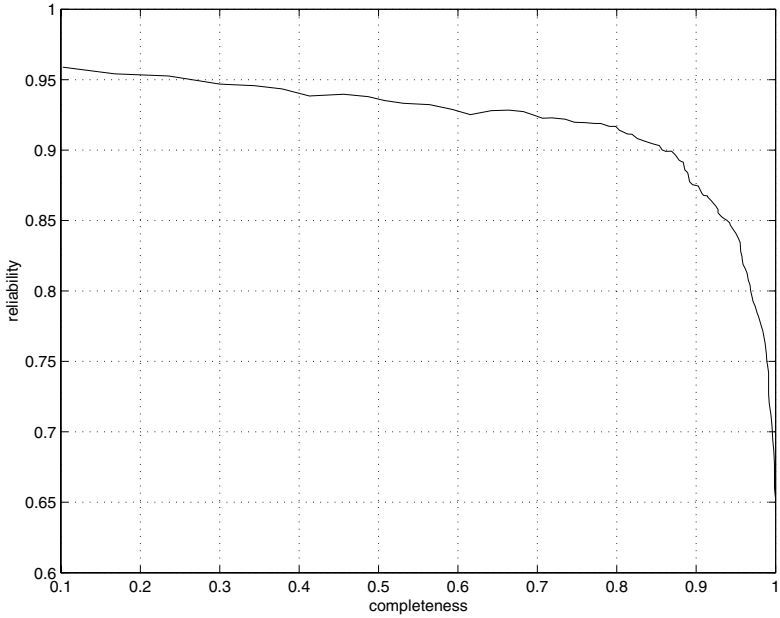


FIGURE 16.4. Completeness-reliability curve

identify all objects is virtually identical to that when quasars are merely discriminated from all other objects, so that little is lost in being more ambitious.

16.6 References

[Vapnik(1995)] Vapnik, V. 1995, *The Nature of Statistical Learning Theory*, Springer

[Vapnik(1998)] Vapnik, V. 1998, *Statistical Learning Theory*, Springer

[Breiman 1999] L. Breiman 1999, 'Random Forests – Random Features', Univ. of California Berkeley, Dept. of Statistics, Tech. Rept. 567

[White et al.(2000)] White, R. L., Becker, R. H., Gregg, M. D., Laurent-Muehleisen, S. A., Brotherton, M. S., Impey, C. D., Petry, C. E., Foltz, C. B., Helfand, D. J., McMahon, R. G. & Cabenela, J. E. 2000, 'The FIRST bright quasar survey. II. 60 nights and 1200 spectra later', Astrophs. J. Suppl, 126, 133-207

Commentary by Eric D. Feigelson[3]

A typical astronomical problem in multivariate classification requires the separation of objects into distinct classes (or outlier status) from a database with N rows and p columns/properties. Common difficulties include:

1. Number of clusters may be unknown
2. Shape of clusters in $p-$space unknown (not multinormal?)
3. Redundancy of variables present but not understood
4. Reliability of derived classification unknown
5. Heteroscedastic (i.e. different for each object) measurement errors with known variances are available for many variables
6. Many variables can suffer censoring (i.e. upper limits due to nondetections)

Many existing statistical methodologies often treat the first four problems but rarely confront the last two.

I list here just a few of the myriad examples of such problems:

- Quasars *vs.* stars from SDSS photometry ($N \sim 10^8,\ p \sim 10$)
- Galaxies *vs.* stars on POSS plates ($N \sim 10^9,\ p \sim 10$)
- Morphology of radio galaxies from FIRST ($N \sim 10^6,\ p \sim ?$)
- Morphology of galaxies from ESO/HST ($N \sim 10^5,\ p \sim 10 - 20$)
- Dusty stars in Milky Way from IRAS ($N \sim 10^5,\ p = 6$)
- Spectral classification of stars ($N \sim 10^4,\ p = 10^2$)
- Gamma-ray bursts from BATSE ($N \sim 10^3,\ p \sim 10$)
- Ty1/Ty2/BAL quasars from optical spectra ($N \sim 10^3,\ p \sim 10^2$)

Astronomers approach such problems from a variety of directions. From an examination of the astronomical literature (http://absads.harvard.edu/-abstract_service.html) and adding my own brief impressions, the methods in order of most to least frequently used are:

Neural networks (~ 150 studies, 1990–)
Often effective but 'black box' results: not easily reproduced and poor insights into important variables

Principal component analysis (~ 80 studies, 1980–)
Often inappropriate for this purpose, as the astronomer typically astrophysical relationships between properties within a class *after* classification. For very-high dimensional datasets or those with redundant variables, it is useful to perform classification on principal components rather than the original dataset.

[3]Department of Astronomy & Astrophysics, Pennsylvania State University

Bayesian classifiers ($\sim$ 30 studies, 1985–)
Often effective but may be difficult to reproduce. *AutoClass* is commonly used.

Decision trees ($\sim$ 20 studies, 1994–)
Often effective, computationally simple. Oblique decision trees recommended by White (1997) in the *Statistical Challenges in Modern Astronomy II* conference.

CART Classification and Regression Trees developed by Breiman (1984). (0 studies)
The absence of CART in the astronomical literature, and rarity of decision tree methods in general, is quite remarkable given then very heavy usage in other fields.

The reported capabilities of Professor Breiman's latest development of the decision tree approach, Random Forests, appear to be exceptional Breiman (1999). He says it gives a classification of a multivariate dataset which does not overfit the data, is not highly sensitive to noise, and gives accurate classifications for known data. In addition to the classification, it indicates which variables are most important, provides proximity measures and measures of outlyingness for existing and future future data points, and gives density estimation (a smoothed p-dimensional surface).

A major disadvantage of the method is its computationally intensive Monte Carlo approach. Astronomers are not daunted by numerically intensive calculations, although this precludes application to very large datasets as envisioned, for example, by the forthcoming Virtual Observatories (Brunner, Djorgovski & Szalay 2001). Many standard methods are not used by astronomers because they suffer a cultural aversion to commercial statistical packages. But Prof. Breiman has deftly circumvented this problem with Random Forests by providing on-line Fortran-77 and R freeware (see http://www.stat.berkeley.edu/users/breiman). Astronomers should know that R, based on S-Plus, is a powerful, UNIX-friendly, public domain, statistical computing environment available at htp://www.R-project.org.

A final important message from this paper is that astronomers are generally ignorant of the developments in decision tree methodology during the past several years such as arcing, bagging and boosting. Few of us read the relevant machine learning and neural computation literature. There is little doubt in my mind that Random Forests and similar methods should be applied to many astronomical problems.

16.7 References

[Breiman(1984)] L. Breiman, *Classification and regression trees*, Belmont CA:Wadsworth

[Breiman(1999)] Breiman, L. 1999, Tech. Rpt. 567, UC Berkeley Statistics. Available on-line at http://www.stat.berkeley.edu/tech-reports

[Brunner et al.(2001)] Brunner, R. J., Djorgovski, S. G., & Szalay, A. S. (eds.) 2001, *Virtual Observatories of the Future*, San Francisco:Astron. Soc. Pacific Conf. Ser. 225,

[White(1997)] White, R. L. 1997, in *Statistical Challenges in Modern Astronomy II*, New York:Springer, 135

17

Interactive and Dynamic Graphics for Data Analysis: A Case Study On Quasar Data

Dianne Cook[1]

ABSTRACT This paper describes the use of interactive and dynamic statistical graphics for a classification task of separating quasars from non-quasars, using measurements on red and blue plates, radio and optical values. Multivariate plotting techniques used are the scatterplot matrix, parallel coordinate plot and tours, an extension of 3D rotation to arbitrary dimensional rotation.
This paper is followed by a commentary by statistician Fionn D. Murtagh.

17.1 Introduction

This paper gives a brief introduction to interactive and dynamic graphics methods that may be useful in studying astronomical measurements. The reader can interpret it as a short literature review of existing graphics methodology, focusing on how some of these methods can be applied to gain insight about an astronomical data set: quasars. One word of caution (or preemptive apology) to the reader: the analysis is done by a statistician with little knowledge of astronomy. It should be used as a potential guide to the use of general graphical tools for astronomical data and not a definitive explanation of the study of quasars.

17.2 Methods

In statistics we typically think of data as meaning information that has been processed into a table or a list. The simplest format of data is a matrix where columns correspond to variables, and rows correspond to replications or objects on which the variables are measured. Variables may be generated by recoding raw measurements into quantitative values. The goal of data visualization is to examine the abstract relationships between variables or columns of the table, in order to quantify the joint distribution between the variables. The number of variables in arbitrary. In the general field of

[1]Department of Statistics, Iowa State University

computer-generated visualizations emphasis is placed on generating graphics of 2D or 3D objects. This is not sufficient for data visualization because data rarely comes with just 2 or 3 variables. Hence a critical aspect of data visualization is the use of abstract graphics such as histograms, scatterplots, time series. There are several good references for these techniques (see for example Cleveland (1983), Tufte (1983), Wilkinson (1999)).

Graphics provides a complement to numerical techniques. The primary advantage is that they can reveal overlooked structure in data due the the ability of the human eye to recognize complex structure. For example, graphics are traditionally used to assess goodness-of-fit of models, to check for such structure as unmodeled non-linear relationships. Graphics are very helpful in uncovering the unexpected in exploratory data analysis, or data mining.

The current state of fast computing allows for a high level of interaction and rapid redrawing of plots. Much of the current research in data visualization is producing direct manipulation of graphical elements and dynamic graphics. There are many examples of these environments (for example, see Cleveland et al (1988), Swayne et al (1998), Buja et al (1996), Carr et al (1996), Cook et al (1996), Hofmann et al (2000), Buja et al (1991), Sutherland et al (2000)).

17.3 Example

The data used here comes from a study on quasars (Breiman et al, 2001). There are 2101 cases with 12 variables:

```
star_class_red = star/gal classifier for red plate
star_class_red2 = another star/gal classifier for red plate
red_magnitude = red magnitude
star_class_blue = star/gal classifier for blue plate
star_class_blue2 = another star/gal classifier for blue plate
blue_magnitude = blue magnitude
color = blue mag minus red mag
sep_radio_optic = separation between radio and optical in arcsec
sep_radio_optic2 = another estimate of separation between
                   radio and optical in arcsec
radio_peak_flux = radio peak flux
radio_integ_flux = radio integrated flux
spectral_type = spectral type (H, Q, A, S, B, G, Unknown)
where the values for spectral type
  5=Q: quasar
  2=B: BL Lacs (a kind of blazer)
  1=A: narrow-line AGN
  4=H: H II/star forming galaxies
  6=S: stars
  3=G: galaxies without emission lines
```

`0=unknown`

Our understanding is that we are interested in distinguishing between quasars (spectral type 5) and non-quasars, based on the other variables. To approach this visually we use color to code categorical class information, that is, points corresponding to quasars differently from the other points (we use red solid circles for quasars and navy blue crosses for non-quasars). Then we use the typical array of multivariate plotting techniques to examine the quantitative variables.

Figure 17.1 shows univariate textured dot plots (Tukey et al, 1990) for variables where there appears to be some difference between quasars and non-quasars: red magnitude, blue magnitude, color, sep_radio_optic, radio peak flux, radio integrated flux. In each case the horizontal axis displays the indicator variable for quasar or not, and these values are jittered, randomly spread in a horizontal space. Most of the variables have highly skewed distributions. Normally one might consider transforming the skewed variables to spread the values more evenly. This is often as useful in visual displays, as it is essential in statistical analysis. However in the presence of cluster structure, as we have here, transforming to spread values out does not assist in finding differences between the two classes. The variables' red magnitude and blue magnitude seem to display the most difference in values for quasars as opposed to non-quasars. But the difference is mostly in the variation of the measurements: quasars have more concentrated higher values of red magnitude than non-quasars. The exception is that there are two points classed as quasars which have unusually low values. Probing these values indicates they correspond to cases 1247 and 1365. Blue magnitude has a similar pattern between quasars and non-quasars, and the same two unusual cases. The variable color shows some distributional difference between quasars and non-quasars. Quasars have a more "normal" or "bell-shaped" shaped distribution, whereas non-quasars have a more skewed distribution. Both classes are centered around similar values and have similar ranges, with quasars perhaps having slightly higher average values. So it is difficult to distinguish the two groups. There is very little difference between the two classes on sep_radio_optic. The two radio flux variables have more spread in the quasar class than in non-quasars.

Figure 17.2 shows a scatterplot matrix display of 4 of the seemingly important variables, red and blue magnitudes, color and and sep_radio_optic. A scatterplot matrix is a multi-layout plot displaying pairwise views of the included subset of variables in a convenient matrix format. Red and blue magnitudes are strongly linearly related so it may be that it's not useful to include both variables to obtain the best classification. Red magnitude and color are not linearly related, and it looks like there is a small increase in the difference between the two classes when these two variables are used: the difference occurs at high values on both variables. In contrast the plot of red magnitude against sep_radio_optic shows little advantage to including the additional variable: the difference between the classes on this

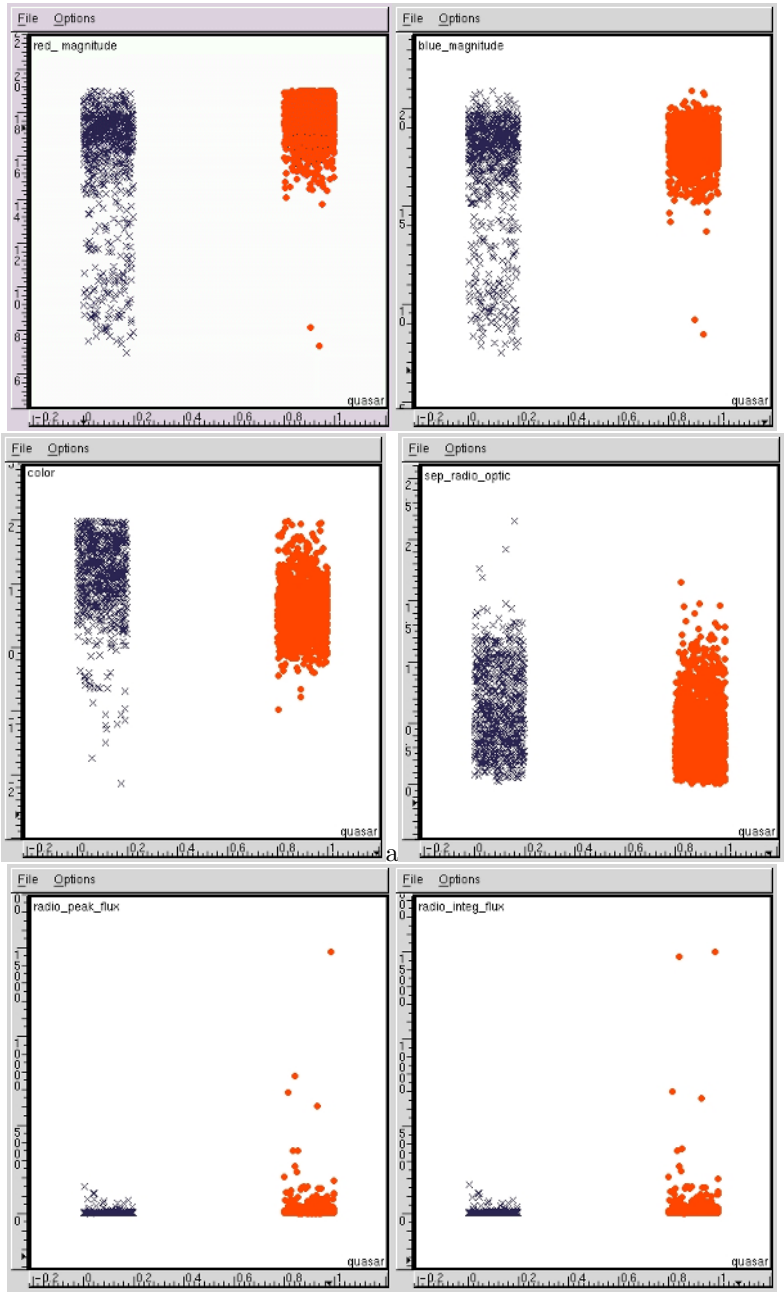


FIGURE 17.1. Univariate plots of quasar data: red magnitude, blue magnitude, color, sep_radio_optic, radio peak flux, radio integrated flux.

second variable is due to a few high values. Again this is corroborated with the plot of color and sep_radio_optic: little is gained in separating the two classes by including the second variable.

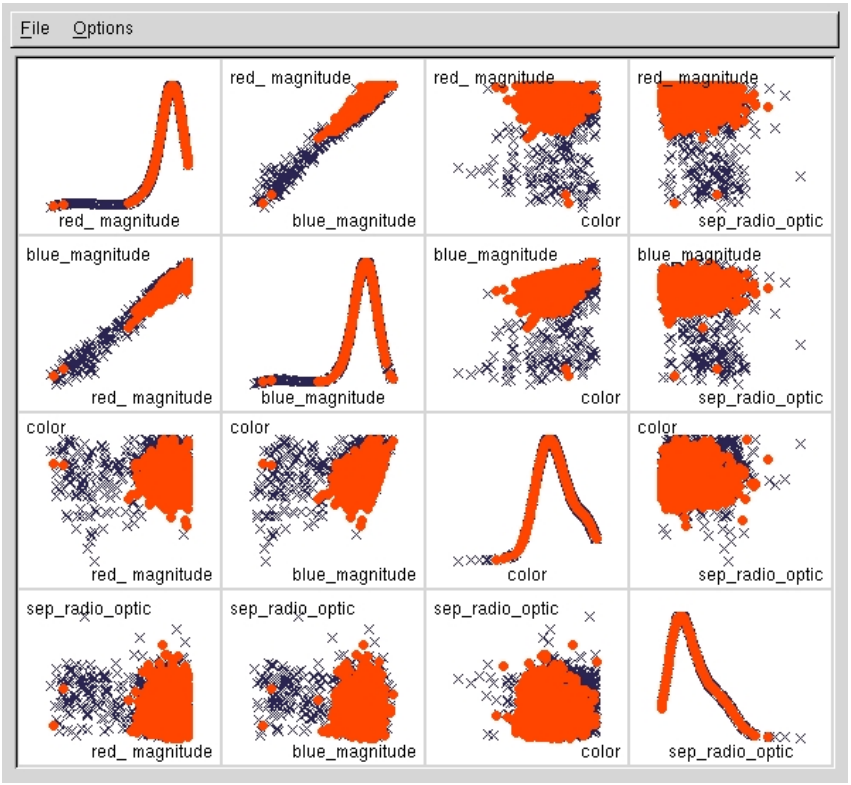


FIGURE 17.2. Scatterplot matrix of four variables in the quasar data: red magnitude, blue magnitude, color, sep_radio_optic.

Figure 17.3 shows a parallel coordinate plot (Inselberg, 1985; Wegman, 1990) of 5 of the seemingly important variables: red magnitude, blue magnitude, color, sep_radio_optic, radio peak flux. A parallel coordinate plot displays the variable axes parallel to each other rather than orthogonal to each other. Values for a particular row (case) in the data are connected by lines, so each line trace corresponds to a single case. Line traces corresponding to quasars are colored red and non-quasars are navy blue. The values on the axes are also represented with the solid circles for quasars and crosses for non-quasars. In reading parallel coordinate plots, for classification tasks like this one, look for several patterns: separations between points on the

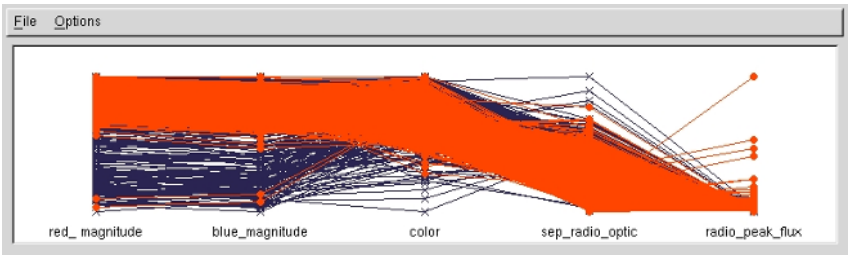


FIGURE 17.3. Parallel coordinate plot of five variables in the quasar data: red magnitude, blue magnitude, color, sep_radio_optic, radio peak flux.

axes, or lines between axes, crossing of lines between the axes. In the parallel coordinate plot of the quasar data we see mostly parallel traces for both quasars and non-quasars for the first 3 variables, with quasar traces higher and more concentrated in comparison to non-quasar traces. Between color and sep_radio_optic there is some crossing of lines and a block of non-quasar traces can be seen higher than non-quasars mid-way between the two variable axes. This corresponds to a distributional difference between quasars and non-quasars that can be seen in the pairwise scatterplot (Figure 17.2): the shape of non-quasars is boxy, or square in the high values of these two variables as opposed to quasar rounded shape. This means it would be possible to construct a non-linear boundary that carves off a group of non-quasars in the high values for these two variables. There is also crossing of lines between sep_radio_optic and radio peak flux as which corresponds to the high values of non-quasar sep_radio_optic dropping to have low values on radio peak flux, and the reverse pattern for some quasars. In general to extract more information from a parallel coordinate plot it would be necessary to permute the order of the axes.

Figure 17.4 shows views from a tour of the data. Tours are dynamic graphics which extend 3D rotation to arbitrary dimensional rotations (for example, see Asimov (1985), Cook et al (1995), Cook et al (1997), Buja et al (1997)). There are several different types of algorithms for generating the rotations through high-dimensional space. Here we make use of a manual tour, for the most part, and a grand (random) tour. All tours are based on taking projections of the variable axes. Here we use 2D projections of the 11D space. The manual tour allows the user to manually rotate a variable into and out of an existing projection. We start with a projection of two variables: (top left) 3 (red magnitude) and 7 (color). Then we manually rotate in additional variables into this view to examine the effect of separating the quasars from non-quasars. The top right plot shows the effect of rotating variable 8 (sep_radio_optic) into the projection. Some advantage is gained from this combination of variables in that there is a marginally improved difference between quasars and non-quasars. The bottom left plot shows the effect of rotating a fourth variable, radio peak flux into the pro-

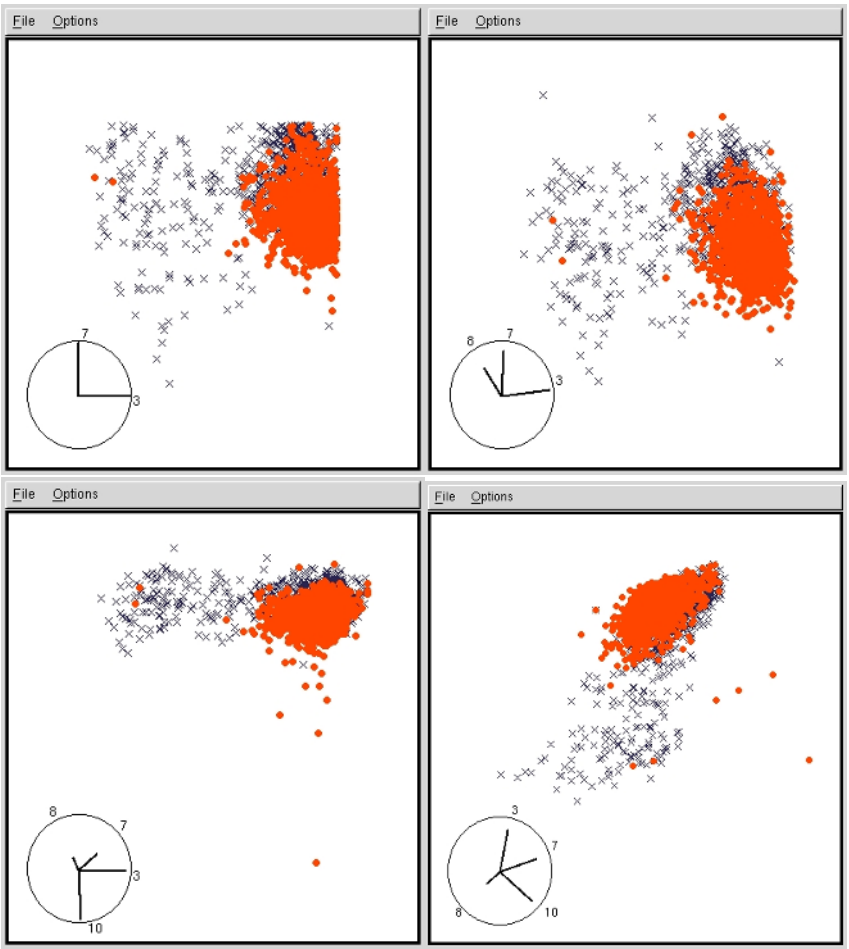


FIGURE 17.4. Tour views of quasar data: (top left) red magnitude (3) vs color (7), (top right) sep_radio_optic (8) is manually rotated into the projection, (bottom left) radio peak flux (10) is rotated into the projection, and (bottom right) another projection of these four variables.

jection: the few cases with extreme values on this variable are now visible, but perhaps little is gained in separating the two classes. The bottom right view shows a random combination of these 4 variables which reveals some difference between the two classes.

In summary, based on the available selection of 11 variables it doesn't appear that quasars are clearly distinct from non-quasars. Rather the quasar class is embedded within the non-quasar class. If the goal is to classify new observations the best that could be done would be to confidently state that a new observation was *not* a quasar given particular combinations of red magnitude, color and sep_radio_optic. It is also not a particularly high-dimensional problem, in that little additional classification power is gained by using more than 3 of the 11 variables.

17.4 Conclusion

In this paper we have naively worked through an astronomical data set collected to classify quasars from non-quasars to demonstrate the use of statistical graphics techniques in the analysis process. Several recently developed graphical techniques were included. There is new work on extending the current methods to extremely large datasets in progress, with reports at `http://www.public.iastate.edu/~dicook/Limn/index.html`.

Acknowledgements: The ideas on methods are joint work with Andreas Buja and Deborah F. Swayne, from AT&T Labs. The software used in this paper is `ggobi`. It is freely available from the web site `http://www.ggobi.org`. GGobi is a joint effort with Deborah F. Swayne and Andreas Buja and additionally Duncan Temple Lang from Lucent Bell Labs, and Heike Hofmann, AT&T Labs.

Asimov, A. 1985, The Grand Tour: A Tool for Viewing Multidimensional Data, SIAM Journal on Scientific and Statistical Computing, 6(1):128-143.

Breiman, L. et al, 2001, Random Forests: Finding Quasars, this volume.

Buja, A. et al, 1991, Computing and Graphics in Statistics, Springer-Verlag: New York, NY.

Buja, A., et al, 1996, Interactive High-Dimensional Data Visualization, Journal of Computational and Graphical Statistics, 5(1):78-99.

Buja, A. et al, 1997, Dynamic Projections in High-Dimensional Visualization: Theory and Computational Methods, AT&T Technical Report, Florham Park, NJ `http://www.research.att.com/~andreas/#dataviz/papers/dynamic-projections.ps.gz`

Carr, D. B., et al, 1996, ExplorN: Design Considerations Past and Present, Center for Computational Statistics, George Mason University, Technical Report No 129.

Cleveland, W. S. 1993 Visualizing Data, Hobart Press: Summit, NJ.

Cleveland, W. S. et al (ed), 1988, Dynamic Graphics for Statistics, Wadsworth: Monterey, CA.

Cook, D. et al, 1995, Grand Tour and Projection Pursuit, Journal of Computational and Graphical Statistics, 4(3):155-172.

Cook D., et al, 1996, Dynamic Graphics in a GIS: Exploring and Analyzing Multivariate Spatial Data using Linked Software, Computational Statistics: Special Issue on Computer Aided Analyses of Spatial Data, 11(4):467-480.

Cook, D. et al, 1997, Manual Controls For High-Dimensional Data Projections, Journal of Computational and Graphical Statistics, 6(4):464-480.

Hofmann, H., et al, 2000, Visualizing Association Rules with Interactive Mosaic Plots, Proceedings of the 6th International Conference of SigKDD, 227-235.

Inselberg, A. 1985, The Plane with Parallel Coordinates, The Visual Computer, 1:69-91.

Sutherland, P. et al 2000, Orca: A Visualization Toolkit for High-Dimensional Data, Journal of Computational and Graphical Statistics, 9(3):509-529.

Swayne, D. F., et al 1998, XGobi: Interactive Dynamic Graphics in the X Window System, Journal of Computational and Graphical Statistics, 7(1):113-130.

Tufte, E. R. 1983, The Visual Display of Quantitative Information, Graphics Press: Cheshire, CT.

Tukey, J., et al, 1990, Strips Displaying Empirical Distributions: I. Textured Dot Strips, Bellcore Technical Memorandum.

Wegman, E., 1990, Hyperdimensional Data Analysis Using Parallel Coordinates, Journal of American Statistical Association, 85:664-675.

Wilkinson, L. 1999, The Grammar of Graphics, Springer-Verlag: New York, NY.

Commentary by Fionn D. Murtagh[2]

A person with much influence in the area of interactive and dynamic statistical graphics, as described by Dianne Cook, is John Tukey. Tukey had a long career at Bell Labs and Princeton University, and died on July 26, 2000. He was a major figure, who contributed to the English language (and other languages besides!) words such as "software", and "bit". He was also instrumental in developing the Fast Fourier Transform algorithm,

[2]School of Computer Science, Queen's University Belfast

robust statistics, and of course exploratory data analysis.

The PRIM-9 system to which he contributed was an early 1970s hardware and software platform for interactive display of statistical data. It was developed at the Stanford Linear Accelerator Center (SLAC). A video of Tukey talking about PRIM-9 is available from Bell Labs (James, 1998). With acknowledgement to the Bell Labs/Lucent Video Library, I now show this video each year in the multivariate data analysis course I teach in the doctoral program at Strasbourg Observatory.

Dianne Cook also has a number of tapes in the Bell Labs/Lucent Video Library. There is much that is very exciting in this collection, not least on the XGobi system, the precursor of GGobi discussed by Dianne. From PRIM-9 to GGobi is a journey over three decades, and the progress has been immense.

GGobi (GGobi, 2001) offers many nice features. It is in the public domain, available for Linux and Windows, it supports an XML data input format, it is compatible with the R statistical language and environment (Leisch, 2001) – which in turn makes it compatible with the closely-related S-Plus language and environment – and it interfaces to Postgres and MySQL database management systems.

The work surveyed by Dr Cook is visually exciting. It occupies a center stage slot, in taking data analysis and interpretation forward. As we have seen with PRIM-9, even three decades ago it was known that such software and/or hardware environments are necessary for data interpretation. The natural home of Dianne Cook's work lies midway between human factors and human computer interaction, on the one hand, and statistical inference and modeling, on the other. Current evolution in astronomy towards the virtual observatory, and towards the astronomical data grid, both ensure that interactive statistical graphics will remain very central to astronomical data analysis and interpretation.

To borrow a phrase from an earlier publication by Dianne Cook, this work is all about "calibration of one's eyes", as one tackles the problem of seeing in high-dimensional data spaces.

1. GGobi Data Visualization System, 2001. http://www.ggobi.org

2. D.A. James, "The Statistical Graphics Section's Video Lending Library", Graphics Section Library, Bell Labs, Lucent Technologies, Murray Hill, NJ, 1998.
http://cm.bell-labs.com/cm/ms/departments/sia/video-library/libry.html

3. F. Leisch, "The Comprehensive R Archive Network", 2001.
http://cran.r-project.org

18

Computational AstroStatistics: Fast and Efficient Tools for Analysing Huge Astronomical Data Sources

Robert C. Nichol[1], S. Chong, A. J. Connolly, S. Davies, C. Genovese, A. M. Hopkins, C. J. Miller, A. W. Moore, D. Pelleg, G. T. Richards, J. Schneider, I. Szapudi, and L. Wasserman

ABSTRACT I present here a review of past and present multi–disciplinary research of the Pittsburgh Computational AstroStatistics[2] (PiCA) group. This group is dedicated to developing fast and efficient statistical algorithms for analysing huge astronomical data sources. I begin with a short review of multi–resolutional *k*d-trees which are the building blocks for many of our algorithms. For example, quick range queries and fast N–point correlation functions. I will present new results from the use of Mixture Models (Connolly et al. 2000) in density estimation of multi–color data from the Sloan Digital Sky Survey (SDSS). Specifically, the selection of quasars and the automated identification of X–ray sources. I will also present a brief overview of the False Discovery Rate (FDR) procedure (Miller et al. 2001a) and show how it has been used in the detection of "Baryon Wiggles" in the local galaxy power spectrum and source identification in radio data. Finally, I will look forward to new research on an automated Bayes Network anomaly detector and the possible use of the Locally Linear Embedding algorithm (LLE; Roweis & Saul 2000) for spectral classification of SDSS spectra.

This paper is followed by a commentary by statisticians Fionn D. Murtagh and Dianne Cook.

[1]Department of Physics, Carnegie Mellon University

[2]See http://www.picagroup.org for a full list of PiCA members and our latest papers, research and software

18.1 Introduction

In this paper, I present an update on the past and present work of the Pittsburgh Computational AstroStatistics (PiCA) group; a multi–disciplinary group of researchers from Computer Science, Statistics, and Astrophysics dedicated to developing fast and efficient algorithms for the analysis of huge astronomical datasets (see Nichol et al. 2000 a previous review of our work). The work presented by Larry Wasserman in this volume is part of the PiCA group research but is not discussed herein for obvious reasons.

The motivation for this work is two–fold. First, the quantity of data being collected is increasing rapidly and we stand on the threshold of the so-called "data flood". By the end of this decade, we will have collected petabytes of astronomical data *e.g.* LSST & Planck. The sheer size and dimensionality of these datasets will restrict our ability to navigate and analyse these huge data sources and we will need new techniques to help us. The proposed "Virtual Observatory" (VO; see papers by Alex Szalay and George Djorgovski in this volume) is designed to address the issues of management, distribution and manipulation of such huge, multi–dimensional astronomical datasets. In this paper, we focus on the need for new analysis algorithms since an N^2 or N^3 algorithm – where N is the number of data points – will no work any longer.

Second, we are entering the realm of high precision astrophysics where the need to make measurements with higher and higher accuracy will increase (see recent review by Turner 2001). In cosmology, for example, the next decade will see the drive to measure the cosmological parameters to an accuracy of a few percent as well as confidently map the distribution of mass in both the local and distant universe. The drive for higher precision will greatly benefit from new statistical tools like those discussed herein and by others in this volume. In general, these new statistical techniques are computationally intense – *e.g.* the non-parametric techniques discussed by Larry Wasserman (this volume) – and therefore, to gain their potential, we will need to develop fast and efficient implementations of such algorithms. In this paper, I present some examples of such implementations.

In Section 18.2, I present a brief review of multi–resolutional KD–trees which are at the heart of much of our technology. In Section 18.3, I provide some examples of how such trees can speed–up simple counting queries. In Section 18.4, I will review Mixture Models and their use in Astrophysics. In Section 18.5, I will quickly present a new statistical tool called False Discovery Rate (FDR) and show two recent applications of this technique. In Section 18.6, I will outline our new work on a Bayes Network anomaly detector, while in Section 18.7, I present initial results from our research of algorithms for mapping high dimensional spaces.

18.2 Multi–Resolutional KD–trees

A multi–resolutional KD–tree (*k*d-tree) is a way of organizing a set of data-points in k-dimensional space in such a way that once built, whenever a query arrives requesting a list of all points in a neighborhood, the query can be answered quickly without needing to scan every single point.

The root node of the *k*d-tree owns all the data points. Each non-leaf-node has two children, defined by a splitting dimension $n.\textsc{SplitDim}$ and a splitting value $n.\textsc{SplitValue}$. The two children divide their parent's data points between them, with the left child owning those data points that are strictly less than the splitting value in the splitting dimension, and the right child owning the remainder of the parent's data points.

*k*d-trees are usually constructed top-down, beginning with the full set of points and then splitting in the center of the widest dimension. It has been shown that this splitting criteria – instead of, say, splitting at the median of the widest dimension – produces a more balanced tree which is thus closer to obtaining the desired $O(log\, N)$ performance (see Moore 1991). This produces two child nodes, each with a distinct set of points. This is then repeated recursively on each of the two child nodes.

A node is declared to be a leaf, and is left unsplit, if the widest dimension of its bounding box is $\leq$ some threshold, MinBoxWidth. A node is also left unsplit if it denotes fewer than some threshold number of points, $r_{\min}$. A leaf node has no children, but instead contains a list of k-dimensional vectors: the actual data-points contained in that leaf. The values MinBoxWidth $= 0$ and $r_{\min} = 1$ would cause the largest *k*d-tree structure because all leaf nodes would denote single data points. In practice, we set MinBoxWidth to 1% of the range of the data point components and $r_{\min}$ to around 10. The tree size and construction thus cost considerably less than these bounds because in dense regions, tiny leaf nodes are able to summarize dozens of data points. The operations needed in tree-building are computationally trivial and therefore, the overhead in constructing the tree is negligible. Also, once a tree is built it can be re-used for many different analysis operations.

18.3 Example Uses of *k*d-trees

18.3.1 Range Counting and Cached Sufficient Statistics

One of the most common queries made in Astronomy is: how many objects are within 1 arcminute (or distance r) of a given position. As discussed below, such a query can be performed very quickly using a *k*d-tree.

The key to the speed of such a query is the decorations of the *k*d-tree with extra information which we refer to as *cached sufficient statistics* (see Moore & Lee 1998). Specifically, we can store for each node the bounding

box of all the points it contains (call this box n.BoundBox). The implication of this is that every node must contain two new k dimensional vectors to represent the lower and upper limits of each dimension of the bounding box. The range search operation takes two inputs. The first is a k-dimensional vector $\mathbf{q}$ called the *query point*. The second is a separation distance s_{hi}. The operation returns the complete set of points in the kd-tree that lie within distance s_{hi} of $\mathbf{q}$. Also, we can store n.NumPoints, which is the number of points contained in each node. Furthermore, we also store the centroid of all points in a node and their covariance matrix.

Once we have n.NumPoints and n.BoundBox, it is trivial to write an operation that exactly counts the number of data-points within some range without explicitly visiting all the data-points. .

- **RangeCount**$(n, \mathbf{q}, s_{hi})$
  Returns an integer: the number of points that are both inside the n and also within distance s_{hi} of $\mathbf{q}$.
- Let MinDist := the closest distance from $\mathbf{q}$ to n.BoundBox.
- If MinDist $\geq s_{hi}$ then it is impossible that any point in n can be within range of the query. So simply return 0.
- Let MaxDist := the furthest distance from $\mathbf{q}$ to n.BoundBox.
- If MaxDist $\leq s_{hi}$ then every point in n must be within range of the query. So simply return n.NumPoints.
- Else, if n is a leaf node, we must iterate through all the data-points in its leaf list. Start a counter at zero. For each point, find if it is within distance s_{hi} of $\mathbf{q}$. If so, increment the counter by one. Return the count once the full list has been scanned.
- Else, n is not a leaf node. Then:
  - Let $C_{\rm left}$:= **RangeCount**$(n$.Left$, query, s_{hi})$
  - Let $C_{\rm right}$:= **RangeCount**$(n$.Right$, query, s_{hi})$
  - Return $C_{\rm left} + C_{\rm right}$.

18.3.2 Fast N–point Correlation Functions

N–point correlation functions have a rich history in Astrophysics and have been extensively used to characterize the large–scale distribution of matter in the Universe. Moreover, higher–order correlation functions will become critically important in this new era of high precision cosmology as they are important tests of biasing and gaussianity (see Szapudi et al. 2001).

N–point correlation functions are however, computationally intensive to compute especially for large databases and high values of N. We have used

a dual *k*d-tree approach to help solve this problem and provide substantial speed–ups for calculating the N–point correlation functions (see Moore et al. 2000 & 2001). We note here that substantial speed–ups can also be achieved by binning the data into cells and performing the calculation directly on that grid. This is fine for separations larger than the grid size but fails as one approaches the resolution of the bin size. Our method is equivalent to an "all–pairs" calculation *i.e.* if one had visited all possible pairs of points in the dataset and binned them appropriately.

For more details on our N–point correlation function code, the reader is referred to Moore et al. (2000 & 2001) as well as our website `http://www.-autonlab.org/`. We note here that the tree structures discussed herein are optimal for relatively low dimensional spaces (*e.g.* a few tens of dimensions) and other tree structures like Ball–trees and AD–trees are better for higher–dimensional spaces (see Moore & Lee 1998)

18.4 Using Mixture Models in Astrophysics

In Connolly et al. (2000), we presented the use of Mixture Models of Gaussians to model the probability density function of multi–dimensional astronomical data. The reader is referred to Connolly et al. (2000) for a detailed review of Mixture Models including our fast implementation of the algorithm based upon the *k*d-tree technology discussed above. In this section, we provide two recent applications of this technology to the Sloan Digital Sky Survey (SDSS).

18.4.1 Finding X–ray Sources

Even after years of hard work, the number of detected X–ray sources with an optical identification remains small. For example, the WGACAT[3], SHARC (Romer et al. 2000) & RASS (Voges et al. 1999) X–ray catalogs, which contain hundreds of thousands of X–rays sources, are still mostly unidentified. This is due to the laborious nature of the optical follow–up.

This will hopefully change soon primarily due to new optical surveys of the sky and the approaching VO era which will provide new, automated tools to assist the user. As a pilot study, we are using the SDSS data and the mixture model algorithm to help automate the optical identification of X–ray sources.

This is achieved as follows. We first obtain photometric multi–color data (u',g',r',i',z') within 15 arcseconds of 7300 WGACAT and SHARC sources within the boundaries of the SDSS EDR data (see Stoughton et al. 2001). This results in 377 matches between an SDSS and X–ray source.

[3] http://wgacat.gsfc.nasa.gov

Using these data, we cluster the sources in 4D color–space and thus determine the probability density function for these sources (the best fit mixture model contains 33 gaussians). This pdf is then used to determine the likelihood of any new source being an X–ray source. We plan to extend this work to include further optical and X–ray information *e.g.* the optical morphology and the ratio of the optical and X–ray fluxes (see Stocke et al. 1991). This will facilitate a robust and automatic identification for a large number of X–ray sources presently lying undiscovered in catalogs like WGACAT. A preliminary version of this system is available at `http://ranger.phys.cmu.edu/users/xray/`.

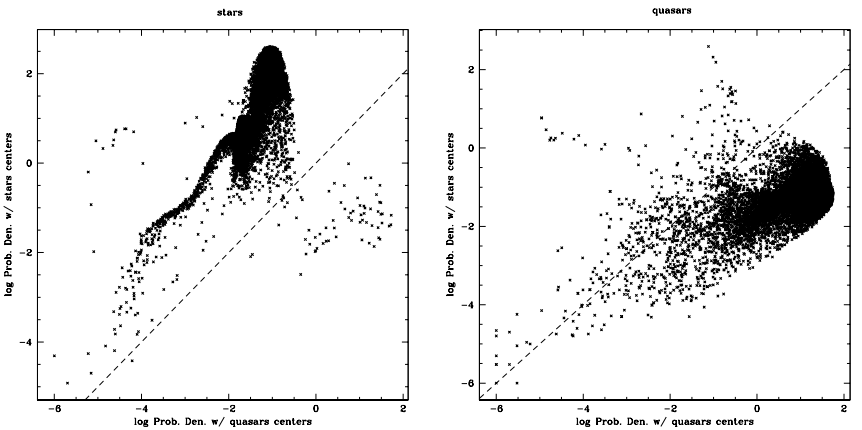


Figure 1: Relative Likelihood of a SDSS source being a star or quasar based on their observed colors. The 45 degree separation line is shown.

18.4.2 Quasar Target Selection

We have also begun to use the Mixture Model algorithm to help in the selection of quasars in multi-color space. In Figure 1, we show a preliminary implementation of such an algorithm using the SDSS data. Here, we have clustered 8833 spectroscopically–confirmed SDSS quasars and 9999 SDSS stars (selected to be point–like objects) in 4D color–space ($u'-g'$,$g'-r'$,$r'-i'$,$i'-z'$) to obtain two pdf's; one for quasars and the other for stars. Then given a new SDSS source with measured colors, one can easily compute the relative likelihood that it is a star or quasar. As illustrated in Fig. 1, we can achieve a high success rate with 96% of the quasars having a quasar probability density larger than stellar probability densities and 99% of the stars having a stellar probability density higher than quasar probability densities *i.e.* the dashed lines in these figures.

We plan several major improvements to this technique. This includes *i)* the addition of other parameters like star–galaxy separation probability,

magnitudes, radio and X–ray fluxes *etc.*; *ii)* the use of synthetic quasar and star SDSS colors to ensure we are not biasing ourselves since the observed data clearly includes the survey selection function; *iii)* increased testing using significantly more spectroscopic and photometric data from the SDSS.

In addition, these applications of the mixture model algorithm have highlighted the need for improvements to the core technology, specifically the need for the algorithm to incorporate observational errors on the data points being clustered to obtain the pdf's. This is traditionally ignored in such computer science orientated algorithms but is vital when analysing real astronomical data. We also need to develop and improve the visualization of the mixture model. At present, this is woefully inadequate and is beginning to hinder our ability to quickly interpret the results of our mixture model. These improvements to the algorithm will require new computer science and statistical research.

18.5 False Discovery Rate

In a recent paper by Miller et al. (2001a), we introduced the False Discovery Rate (FDR) to the astronomical community. This is a new statistical procedure for performing multiple hypothesis tests on data and has three key advantages over more traditional methods like a "3–sigma" threshold or the Bonferroni method: *i)* It has a higher probability of correctly detecting real deviations between the model and the data; *ii* it controls a scientifically relevant quantity – the average fraction of false discoveries over the total number of discoveries; *iii)* it can be trivially adapted to handle correlated data.

We have recently used FDR to solve two astronomical problems. The first is the detection of the acoustic oscillations ("Baryon Wiggles") in the power spectrum of matter in the local universe (see Miller et al. 2001a,b,c for the full details of this discovery). In Figure 2, we show our detection of the "Baryon Wiggles" along with a comparison of our work with the recently released CMB Balloon data (MAXIMA & BOOMERANG). The agreement between these two measurements is impressive and it is re-assuring that our detection of the "Baryon Wiggles" is fully consistent with the CMB at a $z \sim 1000$. In summary, the FDR procedure is a less conservative procedure than the more traditional multiple hypothesis testing methodologies (like "2 sigma" thresholding) commonly used in Astronomy. This has allowed us to detect the "Baryon Wiggles" in the local universe with much fewer data. This illustrates the power of using new statistical tools in this era of high precision cosmology as we attempt to extract the maximum amount of information from these future surveys.

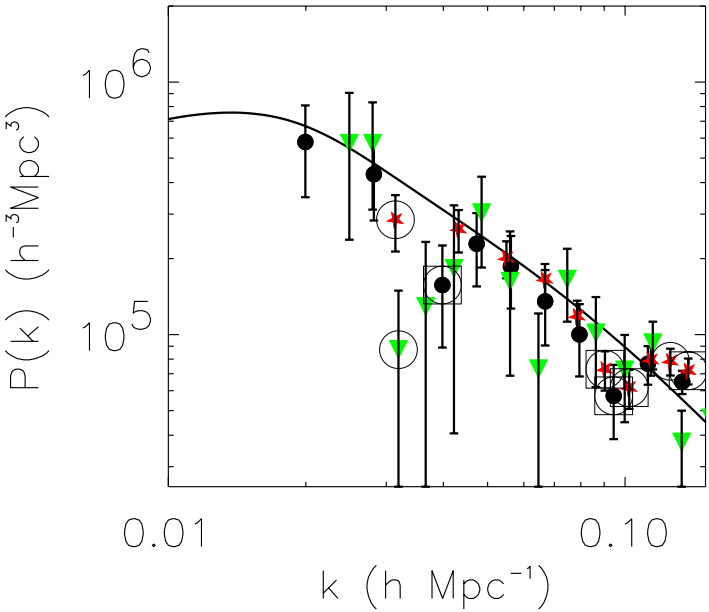


Figure 2: Figure 5 taken from Miller et al. (2001a). The figure shows the amplitude–shifted power spectra for the three samples of uncorrelated data (see Miller et al. 2001b for details). The points highlighted with a circle denote rejections with $\alpha = 0.25$ (e.g. a quarter of the rejections may be mistakes). The points highlighted by squares are for $\alpha = 0.10$ (e.g. a tenth of the rejections may be mistakes). The analysis utilizes our best-fit model with the baryon wiggles removed as the null hypothesis. By controlling the false discovery rate, we can say with statistical confidence that the two "valleys" are detected as features in the power spectra.

A second application of FDR is given in Hopkins et al. (2001) as part of a new source detection algorithm for radio data. Specifically, Hopkins et al. (2001) use FDR to determine which pixels in their radio telescope images are consistent with sky noise or are part of a source. Traditionally, this is done by apply a "5 sigma" threshold which, as discussed by Hopkins et al. (2001) and Miller et al. (2001a), is a very conservative test. Hopkins et al. (2001) compare the FDR method with Imsad and Sextractor (two traditional methods of detecting sources in imaging data) and find it is significantly better than these methods in detecting more, real sources without increasing the false source detection rate.

18.6 Bayes Network Anomaly Detector

Bayes Networks are a popular method for representing joint probability distributions over many variables. The Bayes Nets have the advantage that instead of using a single joint probability function (which can be prohibitive since it may require a large number of parameters to fit the data), they factor the distribution into a smaller number of conditional probability functions for only a subset of the important variables. In practical terms, Bayes Nets have two limitations. They are computationally slow to learn and traditionally only work for discrete data. We have tackled both of these issues using a new implementation of Bayes Networks called *Mix-Nets* (see Davies & Moore 2000) which uses the mixture model of Gaussians to fit the data quickly over different subsets of the domain variables which can then be combined into a coherent joint probability model for the entire domain. Once learned, the Bayes Net offers the ability to isolate sources with a low likelihood of being produced by the model and this identifies those sources as anomalies. Moreover, the Bayes Net provides the variables in the joint probability model which cause this source to be anomalous.

We have used this technology to search for anomalies within the SDSS photometric archive. Specifically, we have used 1.5 million SDSS detected sources, each with 25 variables (magnitudes, sizes and shape parameters in all 5 of the SDSS passbands), to build a Bayes Network. We derive the overall probability of each source (using the learned network) and rank the sources by this probability. The bottom 1000 sources are flagged as anomalies and visually inspected as they are unusual objects, within the data, based on the joint probability model of these 25 attributes.

One of the major problems with this present approach is the existence of errors within the data. At present, the most unusual objects are diffraction spikes (around stars), asteroids and de–blending errors. This is understandable since these errors have unphysical colors and shapes making them gross outliers to many of the joint conditional probability distributions.

We plan to tackle this problem – which is an issue of productivity – using an iterative loop where the scientist helps the Bayes Network focus on the interesting astronomical anomalies. First, we will initially learn the Bayes Network with all attributes and all data points of interest. The scientist will be presented the bottom 1000 sources (the anomalies with the lowest probability) and will interactively highlight obvious errors (like those mentioned above). As the Bayes Networks also stores the conditional probabilities that caused this anomaly, we can use this information to suppress further examples of such an error when we re–learn the Bayes Network *i.e.* if diffraction spikes are always "long" and "red" we can use that information to ignore further examples of this error. After a few iterations, we should have interactively suppressed obvious errors based on this feed–back loop and the scientist will be presented with a higher percentage of physical anomalies. This is research in progress.

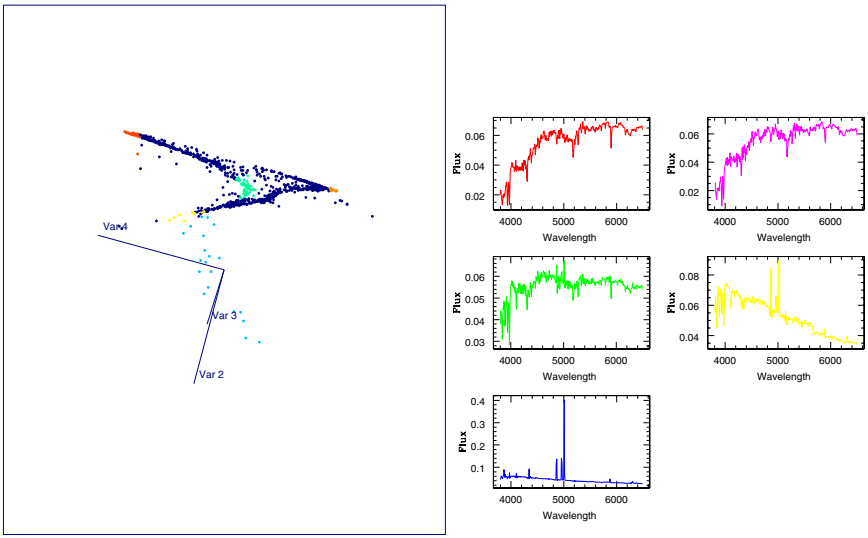


Figure 3: The LLE algorithm is applied to a sample of 500 galaxy spectra (each of 2000 wavelength elements) in order to determine if galaxy spectra occupy a lower dimensional subspace (*i.e.* if strong correlations are present between the individual spectra). Using LLE to compress this 500x2000 space down to a 3 dimensional subspace (see left panel for the distribution of the coefficients for the 500 spectra in this 3D space). We find that the position of a galaxy within this subspace is directly correlated with its spectral type (or, mean age of the galaxy). The right panel shows the typical spectra associated with those points highlighted in the left panel. This simple example demonstrates how new computational techniques might enable a radical compression in the dimensionality of physical data sets.

18.7 Very High Dimensional Data

The next generation of astronomical data will contain many thousands of dimensions. This presents a new paradigm for data analysis techniques since present algorithms and tools do not scale–up into such regimes. The handling of very high dimensional data is an active research area in computer science and statistics *e.g.* Isomap (Tenenbaum, de Silva & Langford 2000) and LLE (Roweis & Saul 2000). In Fig. 3, we show the power of such algorithms through the use of LLE to non–parametrically study the classification of SDSS spectra.

18.8 Conclusions

In this paper, I have outlined an array of fast and efficient statistical algorithms we are developing as part of the Pittsburgh Computational AstroStatistics (PiCA) Group. This is a balanced, multi–disciplinary research effort where all parties gain substantially from this cross–discipline collaboration. For example, the fast algorithms enable new astrophysics to be done and conceived (N–point functions), while the astrophysics problems drive new computer science and statistics *e.g.* the incorporation of errors into Bayes Networks and Mixture Models as well as new statistical theory in extending FDR to slightly correlated data. Therefore, it is a rich collaboration with many possibilities to simulate new and cutting–edge research in computer science, statistics and astrophysics. This work is funded in part through the NSF KDI and ITR programs and the NASA AISRP program and makes use of SDSS data (see www.sdss.org). We acknowledge Don York for carefully reading this manuscript.

18.9 References

Connolly, A. J., et al. 2000, AJ (submitted), see astro-ph/0008187

Davies, S., Moore, A. W., 2000, *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*

Hopkins A. M., et al. 2002, AJ, 123, 1086

Miller, C. J., et al. 2001a, AJ, 122, 3492

Miller, C. J., Nichol, R. C., Batuski, D.J., 2001b, ApJ, 555, 68

Miller, C. J., Nichol, R. C., Batuski, D.J., 2001c, Science, 292, 2302

Moore, A. W., 1991, Ph.D. Thesis, University of Cambridge

Moore, A. W., Lee, M. S., *Volume 8 of Journal of Artificial Intelligence Research*

Moore, A. W., et al., 2000, *Proceedings of MPA/MPE/ESO Conference "Mining the Sky"*, 71. astro-ph/0012333

Nichol, R.C., 2000, *Proceedings from "Virtual Observatories of the Future"*, 265. astro-ph/0007404

Romer, A. K., 2000, ApJS, 126, 209

Roweis, S., Saul, L. K., 2000, Science, 290, 5500

Stocke, J. T., et al. 1991, ApJS, 76, 813

Stoughton, C., et al. in preparation

Szapudi, I., et al. 2001, ApJ, 548, 115

Tenenbaum, J. B., de Silva V., Langford, J. C., 2000, Science, 290, 5500

Turner, M. S., 2001, PASP, 113, 653. astro-ph/0102057

Voges, W., et al. 1999, A&A, 349, 389

Commentary by Fionn D. Murtagh

Computational efficiency in Nichol et al. comes from use of the kd-tree or multidimensional binary search tree. This generalization of the binary search tree was developed a quarter of a century ago. Some important references include the following.

1. J.H. Friedman, J.L. Bentley and R.A. Finkel, "An algorithm for finding best matches in logarithmic expected time", ACM Transactions on Mathematical Software, 3, 209–226, 1977. The feature space dimensionality used here is from 2 to 8.

2. J.L. Bentley and J.H. Friedman, "Fast algorithms for constructing minimal spanning trees in coordinate spaces", IEEE Transactions on Computers, C-27, 97–105, 1978.

3. A.J. Broder, "Strategies for incremental nearest neighbor search", Pattern Recognition, 23, 171-178, 1990. Lisp code is contained in this article.

4. An approach to using the multidimensional binary search tree in higher dimensional spaces was by means of batching the features. Then search is based on a rule such as the following: Take the left subtree if some one of the node-defining features is present. This approach lends itself to boolean data, e.g. keyword presence-absence data in an information retrieval context. It was used in C.M. Eastman and S.F. Weiss, "Tree structures for high dimensionality nearest neighbor searching", Information Systems, 7, 115-122, 1982.

5. More recent related work includes B. Thiesson, C. Meek and D. Heckerman, "Accelerating EM for large databases", Microsoft Research technical report, 1999.

6. Work on external memory algorithms, when data sets are too large to fit in memory and therefore must be stored in slower external memory, includes of course kd-trees. See J.S. Vitter, "External memory algorithms and data structures", ACM Computing Surveys, 2001, in press.

Bob colorfully described the recent developments in the kd-tree area as "decorating [the kd-tree] with cache statistics". Needed are covariances and centroids for mixture model clustering, and bounding box information for range queries.

For efficient clustering, or searching, or other analysis of large data collections, many other approaches have been developed over the last decades. Stochastic approximation is used, for example, to carry out eigen-reduction. This is dealt with under the heading of "direct reading algorithms" in L. Lebart, A. Morineau and K.M. Warwick, Multivariate Descriptive Statistical Analysis, Wiley, 1984. In neural networks, analogous work is characterized as on-line or real-time.

A review of much state of the art work on the processing of large data sets is to be found in J. Abello, P.M. Pardalos and M.G.C. Resende, Handbook of Massive Data Sets, Kluwer, 2001. This book has approximately 1250 pages, – a massive book to deal with massive data set problems!

On the different topic of mixture models in astronomy, the work of S. Mukherjee, E.D. Feigelson, G.J. Babu, F. Murtagh, C. Fraley and A. Raftery, "Three types of gamma ray bursts", The Astrophysical Journal, 508, 314-327, 1998, was a first application in astronomy of the use of a Bayes factor approach to answer the question: how many mixtures are appropriate?

In conclusion, Bob Nichol in presenting this work is to be thanked for the range of important astronomy-motivated computing and statistical problems raised.

Commentary by Dianne Cook

This paper describes the variety of research directions being conducted at Carnegie Mellon University. This is a powerful cross-disciplinary team.

It would be helpful to have more detailed references to web site and literature for multi-resolutional KD-trees. A google search produces these links: `http://hissa.nist.gov/dads/HTML/kdtree.html` has definitions, `http://www.rolemaker.dk/nonRoleMaker/uni/algogem/kdtree.htm` has

a 2D java applet demo, http://www-hpcc.astro.washington.edu/faculty/marios/papers/perform/node2.html an application to astrophysical data. KD-trees appear to be useful and may be useful for visualization work for constructing indices for fast linking between multiple views.

For high-dimensional problems, there are several other approaches to reducing dimensionality than those mentioned in the paper. Projection pursuit (Huber, 1985) is a technique developed in statistics that finds interesting projections of the p variables. Principal component analysis can be considered a subset of projection pursuit. Independent component analysis (http://www.cnl.
salk.edu/∼tony/ica.html) from the signal processing community is also strongly related to projection pursuit. Other methods include local principal components analysis (http://lib.stat.cmu.edu/general/
xnavigation), sliced inverse regression (Li, 1991), and prosection views (Furnas et al, 1994).

18.10 References

Furnas, G. et al, 1994, Prosection Views: Dimensional Inference Through Sections and Projections, Journal of Computational and Graphical Statistics, 3(4):323-385.

Li, K-C., 1991, Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414):316-342.

19

Clustering in High-Dimensional Data Spaces

Fionn D. Murtagh[1]

ABSTRACT By high-dimensional we mean dimensionality of the same order as the number of objects or observations to cluster, and the latter in the range of thousands upwards. Bellman's "curse of dimensionality" applies to many widely-used data analysis methods in high-dimensional spaces. One way to address this problem is by array permuting methods, involving row/column reordering. Such methods are closely related to dimensionality reduction methods such as principal components analysis. An imposed order on an array is beneficial not only for visualization but also for use of a vast range of image processing methods. For example, clustering becomes in this context image feature detection.

19.1 Introduction

Bellman's (1961) [1] "curse of dimensionality" refers to the exponential growth of hypervolume as a function of dimensionality. Many problems become tougher as the dimensionality increases. Nowhere is this more evident than in problems related to search and clustering.

In [2] (see also [3]), a constant computational time or $O(1)$ approach to cluster analysis was described. The computational complexity was, as is usual, defined in terms of the number of observations. This work related to problem spaces of dimensionality 2, with generalization possible to 3-dimensional spaces [4]. Byers and Raftery [5] proposed another very competitive approach.

It may be helpful to distinguish this work from clustering understood as mixture distribution modeling. A characterization follows which will describe the broad picture. Banfield and Raftery [6] discuss algorithms for optimal cluster modeling and fitting. On the other hand, the work on clustering of Murtagh and Starck [2], and Byers and Raftery [5], is based on noise modeling. Mixture modeling and cluster modeling are essentially signal modeling. Given that observed data can be considered as a mixture of signal and of noise, one can approach data analysis from either of two

[1]School of Computer Science, Queen's University Belfast

perspectives: accurately model the signal, as in mixture modeling, with perhaps noise components included in the mixture; or accurately model the noise.

The latter lends itself well to the problem representation to be described in this article. In general, it lends itself well to situations when we consider data arrays as images. We will next look at when and how we can do this.

19.2 Matrix sequencing

We take our input object-attribute data, e.g. document-term or hyperlink array, as a 2-dimensional image. In general, an array is a mapping from the Cartesian product of observation set, I, and attribute set, J, onto the reals, $f : I \times J \longrightarrow \mathbb{R}$, while an image (single frame) is generally defined for discrete spatial intervals X and Y, $f : X \times Y \longrightarrow \mathbb{R}$. A table or array differs from a 2-dimensional image, however, in one major respect. There is an order relation defined on the row- and column-dimensions in the case of the image. To achieve invariance we must induce an analogous ordering relation on the observation and variable dimensions of our data table.

A natural way to do this is to seek to optimize contiguous placement of large (or nonzero) data table entries. Note that array row and column permutation to achieve such an optimal or suboptimal result leaves intact each value x_{ij}. We simply have row and column, i and j, in different locations at output compared to input. Methods for achieving such block clustering of data arrays include combinatorial optimization ([7, 8, 9]) and iterative methods ([10, 11]). In an information retrieval context, a simulated annealing approach was also used in [12]. Further references and discussion of these methods can be found in [13, 14, 15]. Treating the results of such methods as an image for visualization purposes is a very common practice (e.g. [16]).

We now describe briefly two algorithms which work well in practice.

Moments Method [10] Given a matrix, $a(i, j)$, for $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$. Define row moments as $m(i) = (\sum_j ja(i, j))/(\sum_j a(i, j))$. Permute rows in order of nondecreasing row moments. Define column moments analogously. Permute columns in order of nondecreasing column moments. Reiterate until convergence.

This algorithm results (usually) in large matrix entries being repositioned close to the diagonal. An optimal result cannot be guaranteed.

Bond Energy Algorithm [7] Permute matrix rows and columns such that a criterion, BEA $= \sum_{i,j} a(i, j)(a(i - 1, j) + a(i + 1, j) + a(i, j - 1) + a(i, j + 1))$ is maximized.

An algorithm to implement the BEA is as follows: Position a row arbitrarily. Place the next row such that the contribution to the BEA

criterion is maximized. Place the row following that such that the new contribution to the BEA is maximized. Continue until all rows have been positioned. Then do analogously for columns. No further convergence is required in this case. This algorithm is a particular use of the traveling salesperson problem, TSP, which is widely used in scheduling. In view of the arbitrary initial choice of row or column, and more particularly in view of the greedy algorithm solution, this is a suboptimal algorithm.

Matrix reordering rests on (i) permuting the rows and columns of an incidence array to some standard form, and then data analysis for us in this context involves (ii) treating the permuted array as an image, analyzed subsequently by some appropriate analysis method.

19.2.1 Matrix permutation and singular value decomposition

Dimensionality reduction methods, including principal components analysis (suitable for quantitative data), correspondence analysis (suitable for qualitative data), classical multidimensional scaling, and others, is based on singular value decomposition. It holds:

$$AU = \Lambda U$$

where we have the following. A is derived from the given data – in the case of principal components analysis, this is a correlation matrix, or a variance/covariance matrix, or a sums of squares and cross products matrix.

Zha et al. [17] formalize the reordering problem as the constructing of a sparse rectangular matrix

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}$$

so that W_{11} and W_{22} are relatively denser than W_{12} and W_{21}. Permuting rows and columns according to projections onto principal axes achieves this pattern for W. Proceeding recursively (subject to a feasibility cut-off), we can further increase near-diagonal density at the expense of off-diagonal sparseness.

19.2.2 Lerman's theorem for ultrametric matrices

As is well-known, a geometric space has an induced metric. The Euclidean metric is widely used. The Euclidean metric ($L = 2$) is just one of the Minkowski metrics, with others including the Hamming or city-block metric ($L = 1$), and the Chebyshev metric ($L = \infty$):

$$d_p(x, y) = \sqrt[p]{\sum_j \mid x_j - y_j \mid^p} \qquad p \geq 1.$$

A metric satisfies the property of triangular inequality $d(x, y) \leq d(y, z) + d(z, y)$. The ultrametric inequality is a more restrictive condition: $d(x, y) \leq \max(d(y, z), d(z, y))$. Consider a classification hierarchy defined on the object set. We can represent the tree with the objects at the base, and the embedded clusters extending upwards. If one defines the distance between objects as the lowest level in the tree in which the two objects first find themselves associated with the same cluster, then the resulting distance is an ultrametric one. Inducing a tree on an object-set is the transforming of a metric space into an ultrametric one.

Ultrametric distance matrices can be represented, subject to an appropriate ordering of objects, with quite particular relations between values as we move away from the diagonal. Lerman [18] discusses ultrametric spaces in detail. Lerman's Theorem 2 ([18], p. 45) describes properties of ultrametric distance matrices. The result we are most interested in is in regard to matrix reordering: an order can be found such that array elements are necessarily non-increasing as we move away from the diagonal, and row and column array elements have a number of such inequality properties.

Lerman's Theorem for the Form of Ultrametric Matrices: An $n \times n$ matrix of positive reals, symmetric with respect to the diagonal, is a matrix of distances associated with an ultrametric on the object-set iff a permutation can be applied to the matrix such that the matrix has the following form:

1. Beyond the diagonal term equaling 0, values in the same row are non-decreasing.
2. For each index k, if (condition b1) $d(k, k+1) = d(k, k+2) = \ldots = d(k, k+l+1)$ then (implication b2) $d(k+1, j) \leq d(k, j)$ for $k+1 < j \leq k+l+1$ and (implication b3) $d(k+1, j) = d(k, j)$ for $j > k+l+1$.

Therefore $l \geq 0$ is the length of the section starting, beyond the principal diagonal, the interval of columns containing equal values in row k.

We will exemplify Lerman's theorem using the Fisher iris data. The iris data of Anderson used by Fisher [19] is a very widely-used benchmark dataset. The data consists of 3 varieties of iris flower, each providing 50 samples. There are measurements on 4 variables, petal and sepal length and breadth. The data matrix is therefore one of dimensions 150×4.

To derive ultrametric distances, we took the Fisher iris data, in its original 150×4 form. We constructed a complete link hierarchical clustering, using the Euclidean distance between the observation vectors. We read off the 150×150 ultrametric distances (ranks were used, rather than agglomeration criterion values) from this dendrogram. Fig. 1 (left) shows this ultrametric matrix. (The greyscale values have been histogram-equalized for better contrast.) When we reorder the rows and columns (the matrix is symmetric of course) in accordance with the ordering of singletons used by the dendrogram representation we get the visualization shown in Fig. 1 (right). Again contrast stretching through histogram-equalization was used.

Note that the origin is in the lower left, i.e., following the image convention rather than the matrix convention.

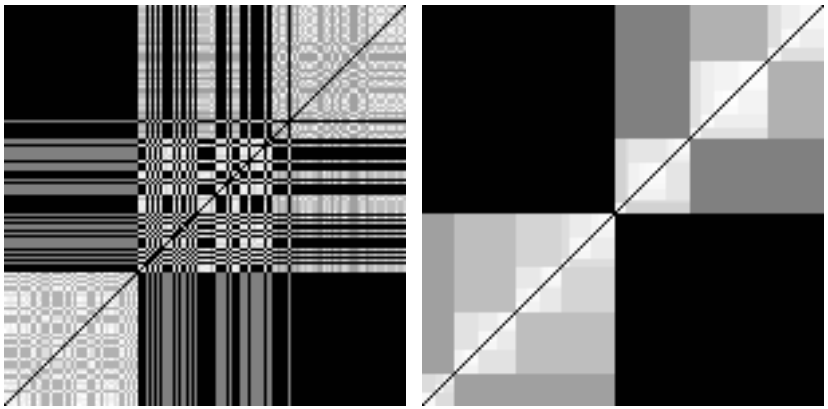


FIGURE 19.1. Left: ultrametric matrix of 150 observations, in given order – clusters 1, 2 and 3 correspond to sequence numbers 1–50, 51–100, 101–150. Right: ultrametric matrix of these same observations, with the rows and columns permuted in accordance with a non-crossover representation of the associated dendrogram.

19.2.3 Permuting large sparse arrays

A few comments on the computational aspects of array permuting methods when the array is very large and very sparse follow [20]. Gathering larger (or nonzero) array elements to the diagonal can be viewed in terms of minimizing the envelope of nonzero values relative to the diagonal. This can be formulated and solved in purely symbolic terms by reordering vertices in a suitable graph representation of the matrix. A widely-used method for symmetric sparse matrices is the Reverse Cuthill-McKee (RCM) method.

The complexity of the RCM method for ordering rows or columns is proportional to the product of the maximum degree of any vertex in the graph represented by the array and the total number of edges (nonzeroes in the matrix). For hypertext matrices with small maximum degree, the method would be extremely fast. The strength of the method is its low time complexity but it does suffer from certain drawbacks. The heuristic for finding the starting vertex is influenced by the initial numbering of vertices and so the quality of the reordering can vary slightly for the same problem for different initial numberings. Next, the overall method does not accommodate dense rows (e.g., a common link used in every document), and if a row has a significantly large number of nonzeroes it might be best to process it separately; i.e., extract the dense rows, reorder the remaining matrix and augment it by the dense rows (or common links) numbered last.

One alternative approach is based on linear algebra, making use of the extremely sparse incidence data which one is usually dealing with. The execution time required by RCM may well require at least two orders of magnitude (i.e., 100 times) less execution time compared to such methods. However such methods, including for example sparse array implementations of correspondence analysis, appear to be more competitive with respect to bandwidth (and envelope) reduction at the increased computational cost.

Elapsed CPU times for a range of arrays are given in [20], and as an indication show performances between 0.025 to 3.18 seconds for permuting a 4000×400 array.

19.3 Incidence data and image models

Consider co-occurrence data, or document-term dependence data. Contiguity of links, or of data values in general, is important if we take the 2-way data array as a 2-dimensional image. It is precisely this issue which distinguishes a data array from an image: in the latter data type, the rows and columns are permutation invariant.

We can define permutation invariance by some appropriate means. We can use the output of some matrix permuting method, such as the bond energy algorithm [7] or a permuting method related to singular value decomposition [20].

The non-uniqueness of such solutions is not unduly important in this article and will not be discussed in detail. However we must justify our approach since it does rely on an array permutation method selected by the user. The resulting non-unique solution is acceptable because our ultimate goals are related to data visualization and exploratory data analysis. Our problem-solving approach is *unsupervised* rather than *supervised*, to use terms which are central in pattern recognition. We seek *an* interpretation of our data, rather than *the* unique interpretation. Of course, the unsupervised data analysis may well precede or be otherwise very closely coupled to supervised analysis (discriminant analysis, statistical estimation, exact database match, etc.) in practice.

19.4 Clustering of document-term data

Experiments were carried out on a set of bibliographical data – documents in the literature crossed by user-assigned index terms. This bibliographic data is from the journal *Astronomy and Astrophysics*. It is used currently to provide a cluster-based graphical user interface to further information on these articles, and in many cases (if one's library subscribes to the journal) to the full online articles themselves. This document map can be accessed

at URL http://cdsweb.u-strasbg.fr/Abstract.html. Further information on the construction and maintenance of these document maps is available in [21, 22]. We looked at a set of such bibliography relating to 6885 articles published in *Astronomy and Astrophysics* between 1994 and early 1999. A sample of the first 3 records is as follows.

```
1994A&A...284L...1I 102 167
1994A&A...284L...5W 4 5 14 16 52 69
1994A&A...284L...9M 29
```

A 19-character unique identifier (the *bibcode*) is followed by the sequence numbers of the index terms. There are 269 of the latter. They are specified by the author(s) and examples will be seen below in this section. The experiments to follow were based on the first 512 documents in order to facilitate presentation of results. Fig. 2 shows the 512×269 incidence array used. We investigated the row and column permuting of this array, based on the ordering of projections on the principal component, but limited clustering was brought about. This was due to the paucity of index term "overlap" properties in this dataset, i.e. the relatively limited numbers of index terms shared by any given pair of documents. For this reason, we elected to base subsequent work on the contingency table. Fig. 3 shows this.

A principal components analysis of the 512×269 dataset is dominated by the $O(m^3)$, $m = 269$ diagonalization requirement. Calculating the principal component projections for the rows takes linear (in document space) time. We used the order of principal component projections to provide a standard permutation of rows and columns of the document contingency table. The resulting permuted contingency table is shown in Fig. 4.

We can interpret clusters on the basis of their most highly associated index terms. This in turn relates to the ordering of index terms on the first principal component axis in this case. Applying an arbitrary cut-off to principal component projections, we find the index terms most associated with the two ends of the first principal component as follows (first three shown):

```
stars:circumstellar matter
X-rays:stars
stars:abundances
```

The other extremity of the first principal component axis is associated with the following index terms (limited to three):

```
galaxies:redshifts
galaxies:luminosity function,mass function
galaxies:compact
```

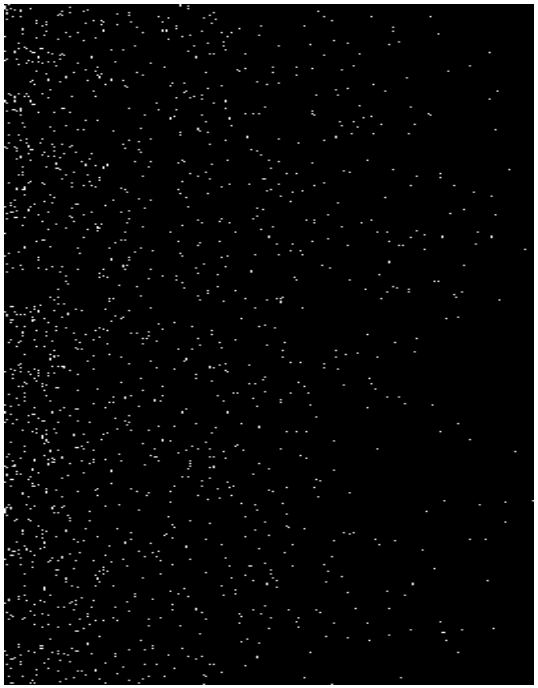


FIGURE 19.2. Rows: 512 documents, columns: 269 index terms.

The distinction is clear – between stars, and stellar topics of inquiry, on the one hand, and interstellar matter (ISM) and galaxies, i.e. topics in cosmology, on the other hand.

19.5 Application to hypertext

From the Concise Columbia Encyclopedia (1989 2nd ed., online version) a set of data relating to 12025 encyclopedia entries and to 9778 cross-references or links was used. We took the first 1203×977 values, based on the correspondence analysis reordering. About the lower half of this array was close to diagonal, and was therefore relatively straightforward to analyze. (The clusters in fact formed a one-dimensional ordering or seriation, and therefore were particularly easy to process.) The upper part of the array was more dispersed and this is what we analyzed using our method. Fig. 5 shows this 500×450 array.

This part of the encyclopedia data was filtered in wavelet transform space using a Poisson noise model. [23] contains further details of the procedure followed. The result is shown in Fig. 6. The first relatively long "horizontal bar" was selected – it corresponds to column index (link) 1733 = `geological era`. The corresponding row indices (articles) are, in sequence:

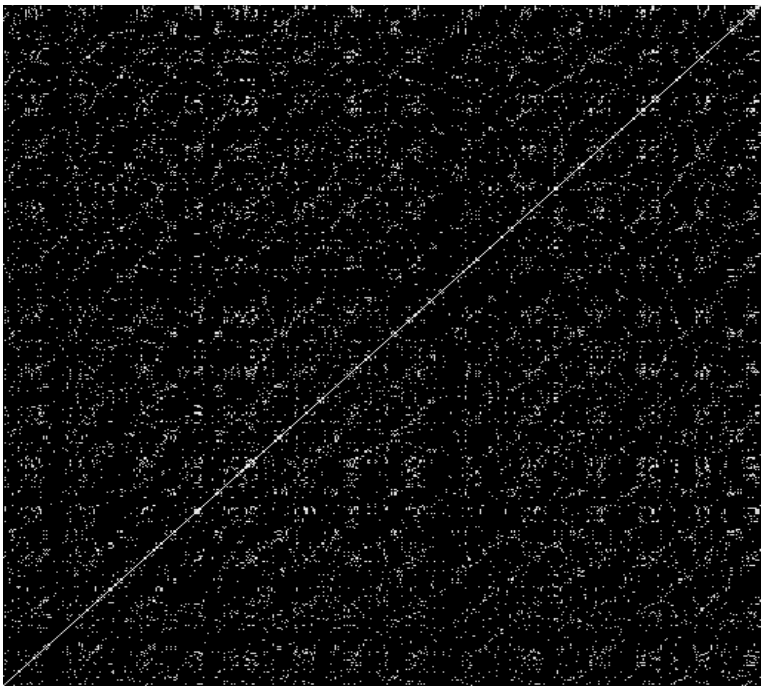


FIGURE 19.3. Contingency table of 512 documents.

SILURIAN PERIOD, PLEISTOCENE EPOCH, HOLOCENE EPOCH, PRECAMBRIAN TIME, CARBONIFEROUS PERIOD, OLIGOCENE-EPOCH, ORDOVICIAN PERIOD, TRIASSIC PERIOD, CENOZOIC ERA, PALEOCENE EPOCH, MIOCENE EPOCH, DEVONIAN PERIOD, PALEOZOIC ERA, JURASSIC PERIOD, MESOZOIC ERA, CAMBRIAN PERIOD, PLIOCENE EPOCH, CRETACEOUS PERIOD

19.6 Conclusion

The methodology developed here is fast and effective. It is based on the convergence of a number of technologies: (i) data visualization techniques; (ii) data matrix permuting techniques; and (iii) appropriate image analysis methods, if feasible of linear computational cost. We have discussed its use for large incidence arrays. We introduced noise modeling of such data, and showed how noise filtering can be used to provide as output a set of significant clusters in the data. Such clusters may be overlapping. Further development of this work would be to investigate hierarchical clusters, possibly overlapping, derived from the multiple scales.

We have also discussed this innovative methodology using a number of different datasets. It is clearly related to other well-established data analysis

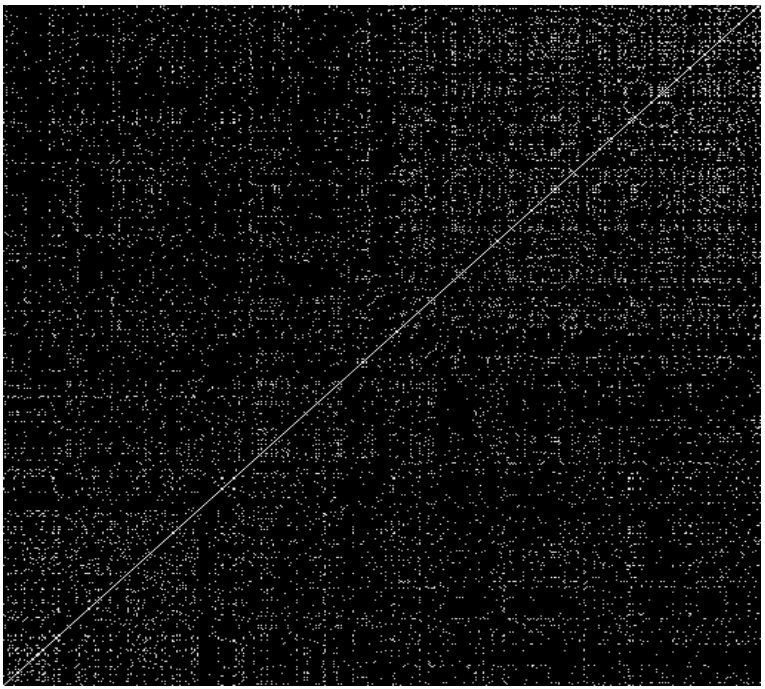


FIGURE 19.4. Row/column-permuted contingency table of 512 documents, based on projections onto first principal component.

methods, such as seriation (one-dimensional ordering of observations), and nonparametric density estimation (the wavelet transform can be viewed as performing such density estimation).

We can note also the potential use of our new methodology for use in graphical user interfaces. The Kohonen self-organizing feature map, by now quite widely used for support of clickable user interfaces ([21, 22]), presents a map of the documents (say), but not as explicitly of the associated index terms. Our maps cater equally for both documents and index terms. Furthermore, the way is open to the exploration of what can be offered by recent developments in client-server based image storage and delivery (see some discussion in Chapter 7 of [3]) e.g. progressive transmission and foveation (i.e. progressive transmission in a local region) strategies. This perspective opens up onto a line of inquiry which could be characterized as *multiple resolution information storage, access and retrieval*. This is particularly relevant in the context of current international initiatives on the computational and data Grid infrastructure of the future.

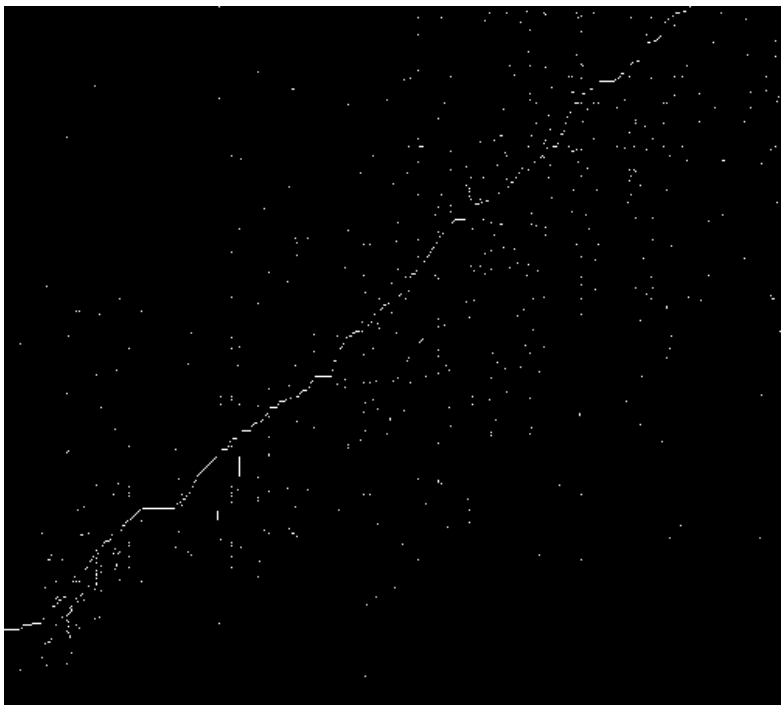


FIGURE 19.5. Part (500×450) of original encyclopedia incidence data array.

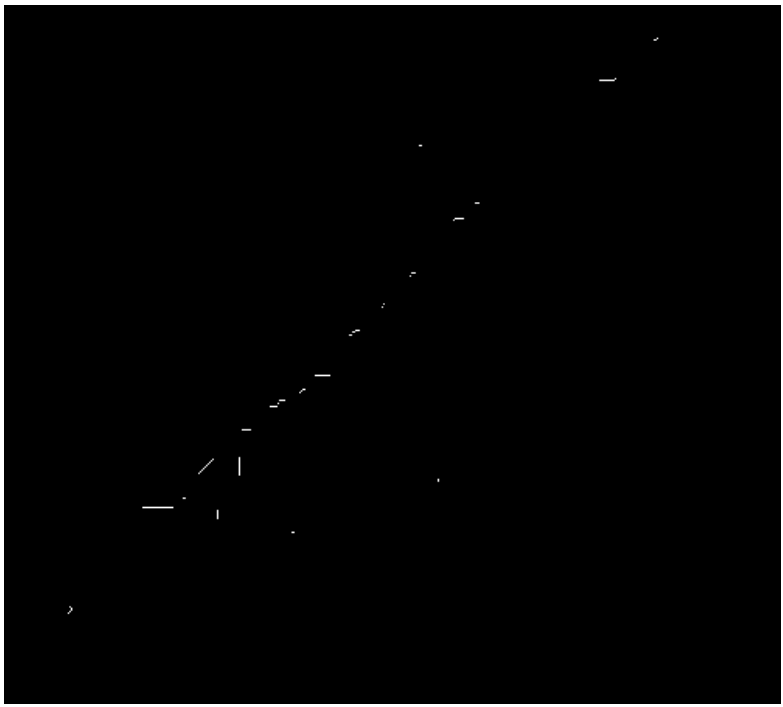


FIGURE 19.6. End-product of the filtering of the array shown in the previous Figure.

19.7 References

[1] Bellman, R. (1961) *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton.

[2] Murtagh, F. and Starck, J.L. (1998) "Pattern clustering based on noise modeling in wavelet space", *Pattern Recognition*, **31**, 847–855.

[3] Starck, J.L., Murtagh, F. and Bijaoui, A. (1998) *Image and Data Analysis: The Multiscale Approach*. Cambridge University Press, Cambridge.

[4] Chereul, E., Crézé, M. and Bienaymé, O. (1997) "3D wavelet transform analysis of Hipparcos data", in Maccarone, M.C., Murtagh, F., Kurtz, M. and Bijaoui, A. (eds.). *Advanced Techniques and Methods for Astronomical Information Handling*, Observatoire de la Côte d'Azur, Nice, France, 41–48.

[5] S. Byers and A.E. Raftery (1996) "Nearest neighbor clutter removal for estimating features in spatial point processes", Technical Report 305, Department of Statistics, University of Washington.

[6] Banfield, J.D. and Raftery, A.E. (1993) "Model-based Gaussian and non-Gaussian clustering", *Biometrics*, **49**, 803–821.

[7] McCormick, W.T., Schweitzer, P.J. and White, T.J. (1972) Problem decomposition and data reorganization by a clustering technique, *Operations Research*, **20**, 993–1009.

[8] Lenstra, J.K. (1974) "Clustering a data array and the traveling-salesman problem", *Operations Research*, **22**, 413–414.

[9] Doyle, J. (1988) "Classification by ordering a (sparse) matrix: a simulated annealing approach", *Applied Mathematical Modelling*, **12**, 86–94.

[10] Deutsch, S.B. and Martin, J.J. (1971) "An ordering algorithm for analysis of data arrays", *Operations Research*, **19**, 1350–1362.

[11] Streng, R. (1991) "Classification and seriation by iterative reordering of a data matrix", in Bock, H.-H. and Ihm, P. (eds.). *Classification, Data Analysis and Knowledge Organization Models and Methods with Applications*, Springer-Verlag, Berlin, pp. 121–130.

[12] Packer, C.V. (1989) "Applying row-column permutation to matrix representations of large citation networks", *Information Processing and Management*, **25**, 307–314.

[13] Murtagh, F. (1985) *Multidimensional Clustering Algorithms*. Physica-Verlag, Würzburg.

[14] March, S.T. (1983) "Techniques for structuring database records", *Computing Surveys*, **15**, 45–79.

[15] Arabie, P., Schleutermann, S., Dawes, J. and Hubert, L. (1988) "Marketing applications of sequencing and partitioning of nonsymmetric and/or two-mode matrices", in Gaul, W. and Schader, M. (eds.), *Data, Expert Knowledge and Decisions*. Springer-Verlag, Berlin, pp. 215–224.

[16] Gale, N., W.C. Halperin and Costanzo, C.M. (1984) "Unclassed matrix shading and optimal ordering in hierarchical cluster analysis", *Journal of Classification*, **1**, 75–92.

[17] Hongyuan Zha, Ding, C., Ming Gu, Xiaofeng He and Simon, H. (2001), "Bipartite graph partitioning and data clustering", preprint.

[18] Lerman, I.C. (1981) Classification et Analyse Ordinale des Données. Dunod, Paris.

[19] Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179–188.

[20] Berry, M.W., Hendrickson, B. and Raghavan, P. (1996) Sparse matrix reordering schemes for browsing hypertext, in *Proceedings of the AMS-SIAM Summer Seminar on Mathematics of Numerical Analysis: Real Number Algorithms*, Park City, UT, July 17 – August 11, 1995. Published in *Lectures in Applied Mathematics (LAM) Vol. 32: The Mathematics of Numerical Analysis*, Renegar, J., Shub, M. and Smale, S. (eds.). American Mathematical Society, pp. 99–123. (http://www.cs.utk.edu/∼berry/order/index.html).

[21] Poincot, P., Lesteven, S. and Murtagh, F. (1998), "A spatial user interface to the astronomical literature", *Astronomy and Astrophysics Supplement Series*, **130**, 183–191.

[22] Poincot, P., Lesteven, S. and Murtagh, F. (2000), "Maps of information spaces: assessments from astronomy", *Journal of the American Society for Information Science*. **51**, 1081–1089.

[23] Murtagh, F., Starck, J.L. and Berry, M. (2000), "Overcoming the curse of dimensionality in clustering by means of the wavelet transform", *The Computer Journal*, **43**, 107-120, 2000.

20

Advanced Tools for Astronomical Time Series and Image Analysis

Jeffrey D. Scargle[1]

ABSTRACT The algorithms described here, which I have developed for applications in X-ray and γ-ray astronomy, will hopefully be of use in other ways, perhaps aiding in the exploration of modern astronomy's data cornucopia. The goal is to describe principled approaches to some ubiquitous problems, such as detection and characterization of periodic and aperiodic signals, estimation of time delays between multiple time series, and source detection in noisy images with noisy backgrounds. The latter problem is related to detection of clusters in data spaces of various dimensions. A goal of this work is to achieve a unifying view of several related topics: signal detection and characterization, cluster identification, classification, density estimation, and multivariate regression. In addition to being useful for analysis of data from space-based and ground-based missions, these algorithms may be a basis for a future automatic science discovery facility, and in turn provide analysis tools for the Virtual Observatory. This chapter has ties to those by Larry Bretthorst, Tom Loredo, Alanna Connors, Fionn Murtagh, Jim Berger, David van Dyk, Vicent Martínez & Enn Saar.
The paper is followed by commentaries by Thomas J. Loredo and Peter E. Freeman.

"The unconscious goal of the scientific philosopher is the automation of science."
Irving John Good, *The Estimation of Probabilities*, 1965

"Automate or die."
Silicon Valley Billboard, June, 2001

20.1 Statistical challenges in modern astronomy

One of the most important statistical challenges in science today is the effective analysis of data from NASA's observational astronomy programs. The work discussed here is meant to provide algorithms of general applicability in the framework of automated science analysis. It is hoped that

[1]NASA Ames Research Center

they will be useful in addressing various challenges in astronomy – such as mining information from the Sloan Digital Sky Survey (see presentation by Michael Strauss) and other cosmological datasets (presentations by Vincent Martinez and Enn Saar, and A. H. Jaffe).

Automated processing already plays a large role in astronomical data analysis, and will be increasingly important as astronomy progresses into the Century of Data. How far along the path to the final scientific output can automatic processing be taken? I feel artificially intelligent data analysis will soon become surprisingly practical. See (Glymour *et al.* 1997, Glymour and Cooper 1999, Heckerman 1997, and Shalizi and Crutchfield 1999) for modern approaches to automatic analysis of data.

20.2 Periodic signals

Definitive presentations of the modern approach to detection of a sinusoidal signal in the presence of noise appear in (Bretthorst, 1988, 2001). A key result that we will need for future reference is that the posterior probability density for the frequency ω of a single component is

$$P(\omega) \propto e^{\frac{C(\omega)}{\sigma^2}} \ , \tag{20.1}$$

[Bretthorst, 1988, Eq. (2.7)] where $C(\omega)$ is the ordinary *Schuster periodogram*, and σ is the variance of the noise, here assumed known. This equation shows that the periodogram is a sufficient statistic for this problem, and contains all information needed to compute frequency estimates and their uncertainties. (Bretthorst 2001) shows that the Lomb-Scargle periodogram serves the same role for unevenly spaced data.

The situation just described is an instructive case study in the relation between the frequentist approach employing a *statistic* and the Bayesian computation of a *posterior distribution*:

- As initially introduced, the periodogram is an *ad hoc* frequentist statistic. Since it is the inner product of a sinusoid and the data, it is reasonable that the periodogram will be large at frequencies at which a harmonic signal is present, small otherwise. But otherwise it is "pulled out of a hat" – an interesting quantity offered with minimal motivation, no justification[2] for preferring it over other possibilities, and only an indirect connection to detection probabilities.
- The Bayesian approach, so eloquently expounded in (Bretthorst, 1988), computes directly and straightforwardly the probability of sinusoidal

[2]Of course the periodogram's statistical behavior more or less validates its choice, after the fact. Indeed, the reason for constructing a modified periodogram for unevenly spaced data was to make its statistical behavior the same simple behavior shown by the Schuster periodogram for even spacing (Scargle 1982, 1989).

signal being present. It devolves that the resulting expression contains the periodogram, nicely clarifying its meaning – but this was by no means guaranteed.

Which of these two approaches is more satisfying is a matter of some debate.

20.3 Time delays and scaling

One often wants to determine the *lag* between two time series. That is, we picture the process generating the second time series as a delayed and possibly scaled version of that generating the first, and we wish to estimate the value of the delay. The approach here follows closely Bretthorst's, mentioned in §20.2. Only results are given here; see (Scargle 2001b) for details.

Assume that the underlying process is a signal, S, superimposed on a background, B. Take as given the two background rates, B_X and B_Y. We seek to characterize the signals that rise above these backgrounds. In some applications the backgrounds should be treated as unknown nuisance parameters, assigned a prior probability distribution, and then marginalized. In one case of special interest (gamma-ray bursts), the background levels are well determined by other data, and can properly be fixed at known constant values. Even here the ideal procedure is to represent this extrinsic data with a prior distribution for the background and marginalize it.

The complete model, expressing delay and scaling between the two signals, is:

$$X_m^{Model} = S_m + B_X \tag{20.2}$$

$$Y_m^{Model} = aS_{m-\tau} + B_Y \ , \tag{20.3}$$

where the lag is τ, and the Y-signal is an overall factor a times the X-signal.

For TTE data, m is measured is quantized units – here called time *ticks*, as defined by the electronics of the data acquisition system – and the above equations give the probability of a photon being detecting during tick m. The observed values, X_m, Y_m have values 1 or 0, depending on whether or not an event was recorded at tick m. After the usual procedure of writing down the likelihoods and marginalizing[3] the signal amplitudes, we find the posterior probability density for τ and a is

$$G_{total}(\tau, a) = G_0(a)\ e^{\frac{\gamma_{X,Y}(\tau)}{\Sigma^2}} \tag{20.4}$$

where

$$\gamma_{X,Y}(\tau) = \sum_{m=1}^{M} X_{m+\tau} Y_m \tag{20.5}$$

[3]That is, integrating out.

is the cross-correlation function of X and Y, and M is the length of the observation interval in ticks. This function arises naturally in the development, and is not introduced in an *ad hoc* manner. It can be readily and rapidly computed using the fast Fourier transform, representing X and Y as arrays of zeros punctuated by unit amplitude δ-functions at the values of m at which photons were detected. The coefficients G_0 and Σ (given in Scargle 2001b) depend on a and the backgrounds, but not on τ. The posterior for evenly binned count data (Scargle 2001b), at least for the case where the variances are independent of time, has exactly the same form.

Note that Eq. (20.4) has a clear similarity to the probability density for ω quoted above, Eq. (20.1) in §20.2. **The cross correlation function, γ is a sufficient statistic for lags, just as the periodogram is for frequencies.** The maximum likelihood value of the lag is just the value of τ that maximizes the cross-correlation function, so the main added feature is the ability to compute the full distributions of τ and a.

20.4 Signal structure: Segmentation yields structure

Now turn to the problem of detecting and characterizing signal structure, from time series data. This section described a very practical representation of time-domain structures corrupted by observational noise[4], namely *partitioning of the data space into subsets in which the signal is assumed constant.*

20.4.1 Data

We consider data consisting of signal measurements, corrupted by noise, blurring, or other instrumental effects. These measurements may be in spaces of one dimension (*e.g.*, time series, energy spectra, *etc.*), two dimensions (images), or more (galaxy redshift/position catalogs).

I distinguish three types of measurement. The first is *event data*[5]. One measures positions of discrete points in the data space under consideration. Examples from the Compton Gamma Ray Observatory are time-tagged photon data from BATSE and sky-image data from EGRET, consisting of lists of photon positions, energies and times. While the usual coordi-

[4] An important point, often leading to confusion, is that *noise* in astronomy has two quite distinct meanings: random observational errors, and random variability intrinsic to the source. The latter, part of the signal, is often just what one is studying, whereas observational noise is a corruption, to be eliminated as much as possible.

[5] This term is appropriate to the context of 1D time series; *point data* is used in the context of 2D images.

nate representation of such points uses real numbers, in practice the corresponding infinite accuracy or resolution is not achievable. The coordinate is quantized in some small unit. In time series from high energy astrophysics, *e.g.*, the points are the times of detection of individual photons, and the corresponding quantum is the resolution of the spacecraft clock, typically somewhere in the range of microseconds to milliseconds.

In the second type of measurement, the entire observation interval (or area, or volume) is partitioned into pre-specified bins (or pixels, or cells), and one records the number of events in each. Event data can be converted to this mode, by adopting a set of bins and counting the points that fall in each bin. This process discards information, diminishes the resolution to that of the bins, and makes the results dependent on the sizes and locations of the bins.

The third type of sequential measurement does not involve explicit counting of events, but some other measurement of a quantity at a set of times or points in space. Here the statistical distribution of the observational errors is not tied to the Poisson distribution, as for the other two types, but can in principle be anything – most commonly normal (Gaussian). The values of the independent variable can be points, intervals, or defined by a spread-out sampling function. For example, spatial power spectra of cosmic microwave background measurements are typically reported in terms of window functions with various shapes; Bharat Ratra and Tarun Souradeep maintain a WWW site (`http://www.phys.ksu.edu/~tarun/CMBwindows/wincomb/wincomb_tf.html`) that gives details for many CMB experiments.

20.4.2 The model

A key step in any likelihood analysis is definition of a model representing the underlying process (*i.e.*, the true signal) and the corruption process obscuring the true signal. We must compute the probability that the observed data would be obtained, given the model and its parameters. This function, called the *likelihood*, depends on the data mode, the sampling process, and the nature of the signal, the noise and other corruption processes.

A big advantage of point data is that they are efficiently described by a single, very simple model. The *Poisson process* is appropriate whenever the events are independent of each other. By this is meant that the occurrence of one event does not change the probability of any others. A common example of dependence is *dead time* in time series data: each photon is followed by an interval in which the detection of a second photon is inhibited. See (Stoyan, Kendall and Mecke 1995) for an excellent discussion of point processes in general, Poisson point processes in particular, and a number of ways that real world data can depart from being Poisson.

Independence implies that the probability an event will occur in any element of data space is proportional to the volume of that element. The

proportionality coefficient is the local event rate, often called the *Poisson parameter* λ. It need not be constant, but can vary in an arbitrary way over the data space.

If λ varies randomly, the process is said to be a *Cox process*, or more descriptively a *doubly-stochastic process*. In such cases it is important to distinguish the two random processes at play. (The usual assumption is that these two are independent of each other.) Keeping this distinction clearly in mind, one can show that events occurring at two different locations are independent[6], even if the event rates at the two points are strongly correlated.

It is remarkable that the seemingly highly special Poisson model is in reality quite general – and surprisingly appropriate for most astronomical processes. All that is required is that the events are independent of each other, and their rate is described by an unrestricted function of position in the data space. Even dependences can be accounted for by incorporating them into the likelihood.

This function representing λ's dependence on location can be either parametric or nonparametric[7]. Since we do not want to impose an explicit signal shape, we use a nonparametric model, namely *piecewise constant* functions. This very convenient model class has the following properties:

- nonparametric
- general: capable of representing any reasonable signal
- simple, easy to compute: rate constant on finite intervals
- useful, *i.e.* easy computation of physically significant properties:
  - pulse peak times, widths, rise times, and decay times
  - pulse amplitudes
  - background level
- extendible to 2D and higher data spaces
- data adaptive, *i.e.* can respond to local features

This representation is also useful in domains such as classification, cluster detection, regression, and density estimation. One can think of it as implementing density estimation with blocks taking on the role of bins.

[6] *I.e.*, their joint probability is the product of the individual probabilities.

[7] Somewhat paradoxically, the number of parameters of a nonparametric model depends on the number of data points (Rissanen 1989). Examples are polynomial fitting, Fourier analysis, and wavelets. The basic idea is that one is really representing the structure in terms of elementary basis functions, whose number depends on how much information is present – rather than fitting a predefined shape to the signal.

Importantly, bin locations and sizes are determined by the data, through the condition that the blocks represent the statistically significant variations in the signal.

Note that we don't really assert that the underlying physical process has a rate that changes in this blocky, discontinuous way. The true signal is no doubt relatively smooth. We represent it as piecewise constant in the same spirit as step-function approximations of a smooth curve. The idea is not that this representation is exact in some limit (often the justification for blocky models; *cf.* wavelet theory, especially the innovative ideas of Donoho 1994a,b), but simply that the blockiness reflects the statistical uncertainty of the data.

One could consider more accurate, *e.g.* piecewise *linear*, representations. But if continuity is imposed, the number of free parameters is almost the same as for piecewise constant models. For the most part the added accuracy is illusory and merely serves to complicate model interpretation.

Another issue has to do with what use the model will be put to. Often we are not really interested in the true shape itself, but in more generic information. For example, in the study of impulsive phenomena, such as Gamma ray bursts, one is interested in rise times, decay times, and other pulse properties. Since there are convenient ways to estimate these parameters directly from the blocks, our seemingly crude representation may adequately encode all usable shape information.

20.4.3 Algorithms

Three algorithms for implementing this Bayesian approach to modeling time series have been described elsewhere (Scargle 1998, 2001a), so only a brief sketch will be given here. The basic component of the model, called a *block* and denoted B_i, comprises a time interval of length T_i and ascribes the N_i data points within this interval to a Poisson process with event rate λ_i. The posterior for this model is

$$P(\ \mathrm{B}_i) = \Phi(N_i, T_i) = \frac{\Gamma(N_i+1)\Gamma(T_i-N_i+1)}{\Gamma(T_i+2)} = \frac{N_i!(T_i-N_i)!}{(T_i+1)!}\ . \quad (20.6)$$

Note that λ_i does not appear, since it has been marginalized. $P(\ \mathrm{B}_i)$ depends on only the size of the block and the number of data points in it. The posterior for the whole model is just $\prod_{i=1} P(\ \mathrm{B}_i)$, where i ranges over all elements of the partition.

Broadly, the three approaches are:

- **Divide and Conquer**: model comparison specifies the optimum *changepoint* at which to subdivide the interval; apply iteratively to all sub-intervals
- **Markov Chain Monte Carlo** (MCMC): sum the posterior probability by expeditiously exploring changepoint space

- **Cell Coalescence**: start from an ultra-fine representation assigning one block to each datum; merge block pairs based on model comparison

The first and last can be thought of as *top-down* and *bottom-up* approaches, respectively. Consider two adjacent intervals, described by N_1, V_1 and N_2, V_2. The corresponding *Bayes merge factor* is computed using Eq. (20.6) to give the ratio of posteriors for the two regions merged and not merged, respectively:

$$\frac{P(\text{Merged})}{P(\text{Not Merged})} = \frac{\Phi(N_1 + N_2, V_1 + V_2)}{\Phi(N_1, V_1)\Phi(N_2, V_2)} \ . \tag{20.7}$$

In both cases one iterates until subdivision or merge operations no longer improve the posterior probability of the model. They are *greedy* algorithms, meaning that they choose the greatest gain possible at each step of the numerical optimization. This is sometimes called *myopic optimization* – a "take what you can now, with no regard for the future" strategy. On termination, the result may be a local optimum, perhaps a good approximate solution – but not guaranteed to be the global optimum. Cell Coalescence is easily generalizable to higher dimensions, as we will soon see.

MCMC (*e.g.*, Gilks, Richardson and Spiegelhalter 1996) is the most rigorous approach, as it solves for all changepoints simultaneously. Convergence of MCMC algorithms is a subtle issue.

20.5 High dimensional structure: Cluster analysis and classification

Cluster analysis in data spaces of higher dimension faces many vexing problems (Backer 1995, Gordon 1999), including determination of the number of clusters, a bewildering variety of proposed methods, loss of information due to restricted data modes, incorporation of prior information, nuisance parameters, and *post facto* validation of clusters. The Bayesian approach deals effectively with all of these issues. This section sketches an extension of the cell coalescence version of Bayesian Blocks to higher dimensions. The posterior in Eq. (20.6) applies unchanged in a space of any dimension, and the principles of the algorithm are identical to those in 1D.

Happily use of the *Voronoi tessellation (Okabe, Boots, Sugihara, and Chiu 2000)*[8] unravels the only real complication that arises in higher dimension – namely the geometry. The Voronoi tessellation partitions the

[8]Due to their importance in computer graphics, fast algorithms yielding the unique Voronoi cell partition of a space of arbitrary dimension are readily available. MatLab (© The MathWorks, Inc.), *e.g.*, has one that represents the resulting data structures in a form very convenient for present purposes.

data space into cells: cell i is that region of the data space closer to datum i than to any other datum.

The Voronoi tessellation is an excellent representation of the data. It contains all relevant information in the raw data. It reduces the search space from the hugely infinite space of all possible partitions to the quite finite space of all possible Voronoi cell subsets which form a partition. It provides a simple generalization of the notion of adjacent intervals: blocks containing cells that touch at one or more points. And it even provides a crude but effective density estimation right off the bat, through the relation that the local point density is the reciprocal of the volume of the Voronoi cell.

The greedy cell coalescence algorithm collects Voronoi cells into larger and larger blocks by iteratively merging the pair of blocks with the largest *merge factor* from Eq. (20.7). In many applications it is both required and efficient to permit only blocks touching each other to merge. The iteration halts if the maximum merge factor falls below 1, at which point the data space has typically been partitioned into blocks much fewer in number than the original data points. Each block has a density equal to the number of data points in it divided by its volume. Then, if desired, high-density blocks adjacent to each other can be collected into *clusters*.

A slightly more detailed discussion of this work in progress is in (Scargle 2001c).

I am greatly indebted to Larry Bretthorst, Alanna Connors, Ayman Farahat, Karl Young, Tom Loredo, Jay Norris, Peter Cheeseman, and Peter Sturrock for comments and suggestions.

References

Backer, E. 1995, *Computer-assisted Reasoning in Cluster Analysis*, Prentice Hall, New York

Bretthorst, G. L. (1988), *Bayesian Spectrum Analysis and Parameter Estimation*, Lecture Notes in Statistics, Springer-Verlag. Available (legally) by downloading from `http://bayes.wustl.edu/`.

Bretthorst, G. L. (2001), "Frequency Estimation And Generalized Lomb-Scargle Periodograms," in this volume, and other papers on his www site `http://bayes.wustl.edu/glb/bib.html`

Donoho, D.L., (1994a), "Smooth Wavelet Decompositions with Blocky Coefficient Kernels," in *Recent Advances in Wavelet Analysis*, L Schumaker and G. Webb, eds., Academic Press, pp. 259-308.

Donoho, D. L. (1994b), "On Minimum Entropy Segmentation," in *Wavelets: Theory, Algorithms, and Applications*, ed. Chui, C.K., Montefusco, L., and Pucciio, L., Academic Press: New York, pp. 233-269.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., eds., (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. 1997, "Statistical

Themes and Lessons for Data Mining," in *Data Mining and Knowledge Discovery*, Vol. **1**, p. 11

Glymour, C., and Cooper, G. 1999, *Computation, Causation and Discovery*, MIT/AAAI.

Gordon, A. D.(1999), Classification, 2nd Edition, Monographs on Statistics and Applied Probability 82, Chapman & Hall/CRC, New York

Heckerman, D. 1997, "Bayesian Networks for Data Mining," in *Data Mining and Knowledge Discovery*, Vol. **1**, p. 79

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N., 2000, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd Edition, John Wiley & Sons: New York.

Rissanen, Jorma, 1989, *Stochastic Complexity and Statistical Inquiry*, Singapore: World Scientific.

Scargle, J. 1982, "Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data," 1982, *Astrophysical Journal*, **263**, pp. 835-853.

Scargle, J. 1989, "Studies in Astronomical Time Series Analysis. III. Fourier Transforms, Autocorrelation and Cross-correlation Functions of Unevenly Spaced Data," 1989, *Astrophysical Journal*, **343**, pp. 874-887.

Scargle, J. "Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, A New Method to Analyze Structure in Photon Counting Data," 1998, *Astrophysical Journal*, **504**, p. 405-418, September 1, 1998. `http://xxx.lanl.gov/abs/astro-ph/9711233`

Scargle, J. (2001a), "Bayesian Blocks: Divide and Conquer, MCMC, and Cell Coalescence Approaches," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering,* 19th International Workshop, Boise, Idaho, 2-5 August, 1999. Eds. Josh Rychert, Gary Erickson and Ray Smith, AIP Conference Proceedings, Vol. 567, p. 245-256.

Scargle, J. (2001b), "Bayesian Estimation of Time Series Lags and Structure," Contribution to **Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2001)**, held at Johns Hopkins University, Baltimore, MD USA on August 4-9, 2001.

Scargle, J. D. (2001c), "Bayesian Blocks in Two or More Dimensions: Image Segmentation and Cluster Analysis," Contribution to **Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2001)**, held at Johns Hopkins University, Baltimore, MD USA on August 4-9, 2001.

Shalizi, C. and Crutchfield, J. 1999, "Computational Mechanics: Pattern and Prediction, Structure and Simplicity," http://www.santafe.edu/sfi/publications/Abstracts/99-07-044abs.html

Stoyan, D., Kendall, W. S., and Mecke, J. (1995), *Stochastic Geometry and its Applications,* 2nd edition, John Wiley & Sons: New York

Commentary by Thomas J. Loredo[9]

Scargle's contribution provides both algorithms that are of immediate practical use, and much food for thought to inspire future research. In this commentary, I focus on his segmented Poisson process approach to modeling signal structure, known colloquially as "Bayes blocks" (BB). I first point out an important astrostatistical problem on which his work could have immediate impact, and then discuss the rationale and generality of the BB approach to modeling signal structure. I offer some brief comments on Scargle's lag estimation work in my commentary on Bretthorst's paper.

20.6 A mundane application

Scargle forsees wide application of his BB approach to modeling signal structure, spanning many disciplines and data sets of varying dimension. Many of these applications will require significant innovation in computational tools for implementing BB. Here I describe a more mundane application for which current algorithms are likely adequate. Despite being mundane, the application is of broad applicability and significant importance; successful application of BB to this problem would be very valuable.

The problem concerns accounting for background rates in binned counting data, such as that produced by X-ray spectrometers. Typically, astrophysical or space-based sources of background dominate the background rate, and so must be measured as part of one's observation of an interesting source. One points the instrument off-source to measure the background, and on-source to measure the background plus source; joint analysis of these data allow one to infer the source spectrum.

Underlying most current methods for analysis of "on/off" binned spectra is a bin-by-bin background model. Off-source, the expected number of counts in bin i is modeled as $\bar{n}_i = b_i T_{\rm off}$, where b_i is the background rate in the bin, and $T_{\rm off}$ is the off-source data interval (which may have units of time, time $\times$ area, etc.). On-source, the expected number of counts is $\bar{n}_i = [b_i + s_i(\theta)]T_{\rm on}$, where $s_i(\theta)$ is the source rate in bin i according to some spectrum model with parameters θ, and $T_{\rm on}$ is the on-source interval. This model underlies both the traditional background-subtracted χ^2 approach to fitting, and more recent Bayesian approaches that rigorously account for the Poisson nature of the data.

The problem with the model is its presumption of no connection between the background rates in adjacent bins. This assumption was probably adequate in the past when bin widths were large, so adjacent bins might in fact be dominated by different features of the background spectrum. But

[9]Department of Astronomy, Cornell University

with improved instrument resolution and telemetry bandwidth, bin widths are becoming ever finer, and the assumption is sacrificing precision. A simple calculation (left to the reader!) illustrates the problem: Let $s_i = s$, a constant, and consider, say, 5 or 10 bins of data. Generate the data with a *constant* background rate, and infer s with standard methods. Then infer s again, using a model with $b_i = b$, a constant. The latter model effectively pools the background data from all the bins, leading to a more accurate background estimate, and thus a more accurate s estimate. In real data, the problem is that we do not know in advance which bins to pool together. Scargle's cell coalescence BB algorithm seems tailor-made to this purpose, and could significantly improve the precision of inferences made with such data without requiring complicated parametric modeling of the background.

20.7 Rationale

Scargle anticipates application of the BB approach to disciplines as diverse as X-ray astronomy and galaxy clustering. Here I raise a few issues regarding justification of the approach in various domains, answering some questions but leaving others open for future research.

A stumbling block for some potential BB users may be its Poisson process foundations. In particular, in a Poisson model counts in disjoint regions are independent, yet typical models for phenomena such as galaxy clustering use tools such as correlation functions that show counts to be correlated. Does this rule out BB for such processes? Perhaps. But it may be a surprise to some readers that the answer isn't simply "yes." Answering the question reveals some features of Poisson processes that many astronomers are unaware of; the answer presented here also shows how a Bayesian look at the process offers a particularly clear insight into the question.

The key to this question is noting that the probability distribution for counts in disjoint regions is independent for a Poisson process *when one conditions on the underlying intensity parameter(s)* (Scargle's λ). When the intensity is unknown, the (unconditioned) joint distribution for the counts can exhibit correlations. This is perhaps obvious conceptually once stated; if the underlying rate over a region is constant but of unknown magnitude, then obviously the number of counts I expect in one part of the region will depend on what I have observed elsewhere. To show this mathematically, Bayesian probability theory is especially appropriate. Let n_1 and n_2 denote the number of counts in two disjoint regions, and suppose the expected number in each region is the same and given by λ. Then conditional on λ, the joint distribution is

$$p(n_1, n_2|\lambda, M) = \frac{\lambda^{n_1} e^{-\lambda}}{n_1!} \times \frac{\lambda^{n_2} e^{-\lambda}}{n_2!}, \tag{20.8}$$

where M denotes the Poisson modeling assumptions. This is the product of two independent Poisson distributions. But now suppose that λ is unknown. For concreteness, describe the uncertainty in λ with an exponential prior density, $p(\lambda|M) \propto \exp(-\lambda/\lambda_0)$ (so larger λ_0 corresponds to greater uncertainty about λ). The *unconditioned* joint distribution is then

$$p(n_1, n_2|M) = \int d\lambda\, p(\lambda|M)\, p(n_1, n_2|\lambda, M) \tag{20.9}$$

$$= \frac{1}{\lambda_0(2+\lambda_0^{-1})^{n_1+n_2+1}} \frac{(n_1+n_2)!}{n_1!n_2!}. \tag{20.10}$$

This is clearly not a product of independent distributions. One can use Bayes's theorem to find the conditional distribution for n_2 given n_1,

$$p(n_2|n_1, M) = \frac{(1+\lambda_0^{-1})^{n_1+1}}{(2+\lambda_0^{-1})^{n_1+n_2+1}} \frac{(n_1+n_2)!}{n_1!n_2!}. \tag{20.11}$$

This distribution peaks at $n_2 \approx n_1$ for large λ_0, just the type of correlation one might expect.

This simple exercise shows that a Poisson model can account for correlated counting structure, and thus provides motivation for broad use of the BB approach. But can one account for *any* kind of correlation within the Poisson framework? Here the answer appears to be "no." To explore this, we again study the two-bin case, but let the expected values for n_1 and n_2 be given by two separate parameters, λ_1 and λ_2. The most general joint distribution for the counts within the Poisson framework can then be written

$$p(n_1, n_2|M) = \int d\lambda_1 \int d\lambda_1 \frac{\lambda_1^{n_1}\lambda_2^{n_2}}{n_1!n_2!} e^{-(\lambda_1+\lambda_2)}\, p(\lambda_1, \lambda_2|M), \tag{20.12}$$

where the final factor is a joint prior density for the Poisson intensity parameters. Now change variables so we can separately focus on the total intensity and the "shape" (relative bin-to-bin variation). Replace λ_1 and λ_2 with the total amplitude $\lambda = \lambda_1 + \lambda_2$ and the fractions $f_i = \lambda_i/\lambda$, constrained by definition so $f_1 + f_2 = 1$. Also, let $N = n_1 + n_2$. Rewrite the prior as $p(\lambda_1, \lambda_2|M) = h(\lambda)g(f_1, f_2|\lambda)\delta(1 - f_1 - f_2)$, with h the prior for the amplitude and $g\delta$ the (conditional) prior for the shape. Then using equation (20.12) one can show that the joint distribution for n_1 and n_2 conditioned on N can be written

$$p(n_1, n_2|N, M) = \int df_1 \int df_2 \frac{N!}{n_1!n_2!} f_1^{n_1} f_2^{n_2} G_N(f_1, f_1)\delta(1 - f_1 - f_2), \tag{20.13}$$

where $G_N(f_1, f_1)\delta(1 - f_1 - f_2)$ is just the joint distribution for the shape parameters given N, $p(f_1, f_2|N, M)$, and of course $n_1 + n_2 = N$ throughout (so $p(n_1, n_2|N, M)$ is really 1-dimensional). This exercise is useful because

joint distributions of the form of equation (20.13) arise both in statistics (in studies of exchangeable sequences) and in statistical mechanics, and are thus well-studied. For the Poisson model to be completely general, it must be capable of producing any possible $p(n_1, n_2|N, M)$. However, in a study of the N-representability problem in statistical mechanics, Jaynes showed that some such joint distributions can be expressed in the form of equation (20.13) only if one allows $G_N(f_1, f_1)$ to be negative (Jaynes 1986). This is not possible in the Poisson framework, so some forms of correlation cannot be accurately modeled with Poisson processes.

A more general issue is the adequacy of piecewise-constant (PC) modeling when one knows the underlying process is continuous. That is, a priori one knows the model is certainly false. Some comfort can perhaps be found in the realization that, on some level, this is probably true of *all* statistical models, yet many seem to succeed regardless, presumably because modeling errors are small compared with uncertainties due to limited data. But the discrepancy between the model and reality seems especially stark with PC models (I say this despite having helped introduce such models into astrostatistics in work with Phil Gregory).

Fortunately, the adequacy of similar histogram and changepoint models has been well-studied in statistics, and some interesting theorems offer solace. Since the underlying "true model" is unknown, the theorems require use of measure theory to work in the infinite-dimensional spaces required for nonparametric modeling and are challenging for nonexperts. Lavine (1994) provides a concise and readable summary of some key results, which he summarizes as showing that "good priors are those that are approximately right for most densities; parametric priors [e.g., histograms] are often good enough." Unfortunately, the theorems all invoke some limiting process (e.g., proof of consistency, i.e., the right result when the number of data tends to ∞). Theorems that provide more practical criteria for model adequacy are desirable. Also, one should note that results in nonparametric statistics can depend on one's choice of distance measure between distributions in infinite-dimensional spaces, and there is some controversy over the appropriate measure. Finally, the theorems all concern *estimation*; but one is also interested in *detection*. This is where I am personally most concerned. A signal with smooth but varying structure may require many PC segments to model, but Bayesian model comparison penalizes models according to the number of parameters, so one may pay a high penalty to model smoothly varying signals, high enough to prevent detection. For the same data, an inherently smooth model with fewer parameters might succeed. More work is needed on this issue; continuous segmented models studied in the Bayesian literature (e.g., piecewise linear or quadratic splines) may help circumvent any problems while retaining some of the virtues of the discontinuous PC basis.

In summary, Scargle's BB approach has greater generality than may first be apparent; combined with its conceptual simplicity and straightforward

algorithms, this should help motivate its wider use. But it also has some limitations. Scargle's intriguing presentation will hopefully motivate both applications of this new tool and generalizations using other underlying point processes and segment types.

References

E. T. Jaynes. Some applications and extensions of the de finetti representation theorem. In P. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno de Finetti*, pages 31–42. Elsevier Science Publishers, Amsterdam, 1986.

M. Lavine. Discussion. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 5*, pages 250–251. Oxford University Press, 1996.

Commentary by Peter E. Freeman[10]

This paper provides an excellent overview of "algorithms of general applicability," showing how relatively simple Bayesian-based methods can be used to perform an array of fundamental analyses. I will not comment upon the details of the algorithms themselves,[11] but rather use this space as a bully pulpit from which to exhort meeting participants and interested readers to follow the path that Scargle is currently following, by creating advanced tools that typical astronomers will understand and more importantly *will want to use.*

It sounds simple, but I must sound a cautionary note: winning the hearts and minds of typical astronomers, whose knowledge of statistics may extend no further than a perusal of Bevington,[12] will not necessarily be an easy task. In my current position as a scientist and programmer developing tools for the Chandra Interactive Analysis of Observations (CIAO) software package (`http://cxc.harvard.edu/ciao/`) I have observed firsthand that comfort and speed, rather than rigor, dictates how many analy-

[10] Harvard-Smithsonian Center for Astrophysics

[11] I will mention, however, that in addition to the work of Bretthorst on the detection of sinusoidal signals, readers should be aware of the Bayesian-based Gregory & Loredo method (ApJ 398:146 1992 and Gregory's brief article in this volume) which tests for the presence of stepwise periodic signals, an elegant algorithm that I have used in my own work and plan to code in C/C++ for general use.

[12] P. R. Bevington & D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences* (New York: McGraw-Hill, 2nd edition updated from the first 1969 edition) provided the first, incomplete statistics education for many astronomers, including me.

ses are performed. This is true *even when users know better*. For instance, many astronomers continue to group data in adjacent bins of counts spectra and use the χ^2 statistic in fits to these data, rather than use the Poisson likelihood, because this is what they were taught to do and/or they are most comfortable using reduced χ^2 as a measure of model acceptability. And many continue to subtract observed background (off) data from source (on) data before performing spectral fits, rather than fitting the background and source data with separate models simultaneously, because the former is faster and "it won't make any difference in the final results anyway."

Thus creating a better algorithm is but the first step. You must strive to make the user interface as simple as possible, and to produce more than enough documentation. This is more than just how to use your tool—you should provide examples, threads, test cases, anything to make the learning curve less steep. You should strive to make it fast (although with the constant improvement in computing power, this is becoming less important). And you should be prepared for rejection, as some astronomers will still resist what is new and exotic. But many others, especially graduate students and post-docs, will come around, and astronomy as a whole will benefit.

I will end this sermon by mentioning that the fight for the hearts and minds of astronomers would be considerably easier if someone (or some group) were to write a basic applied astrostatistics text. In 2001, I gave two lectures on the use of statistics in X-ray astronomy at the First X-ray Astronomy School,[13] and afterwards it became quite clear that the students had learned much from these lectures, and that were hungry to learn more. A textbook would help these and future astronomy students immensely.

[13]Organized by Keith Arnaud and the High Energy Astrophysics Science Archive Research Center at NASA-Goddard Space Flight Center; see `http://heasarc.gsfc.nasa.gov/docs/xrayschool/index.html`.

21

Frequency Estimation and Generalized Lomb-Scargle Periodograms

G. Larry Bretthorst[1]

ABSTRACT Using Bayesian probability theory we demonstrate that the Lomb-Scargle periodogram may be generalized in a straightforward manner to nonuniformly nonsimultaneously sampled quadrature data when the sinusoid has arbitrary amplitude modulation. This generalized Lomb-Scargle periodogram is the sufficient statistic for single frequency estimation in a wide class of problems ranging from stationary frequency estimation in real uniformly sampled data, to frequency estimation for a single sinusoid having exponential, Gaussian, or arbitrary amplitude modulation. In addition we define the bandwidth of a nonuniformly sampled data set and show that the phenomenon of aliases exists in both uniformly and nonuniformly sampled data and that the phenomenon has the same cause in both types of data. Finally, we show that nonuniform sampling does not affect the accuracy of the frequency estimates; although it may affect the accuracy of the amplitude estimates.
This paper is followed by a commentary by Thomas J. Loredo.

21.1 Introduction

The problem of estimating the frequency or period of a sinusoid arises in an a host of contexts in the sciences. For example, in nuclear magnetic resonance (NMR) the signals are sinusoidal with exponential decay. In meteorology, temperature data obviously fluctuate sinusoidally on a daily and yearly basis. In astrophysics, the period of variable stars may be on the order of days to years with nonstationary nonsinusoidal oscillations about a trend. The data gathered by the different sciences are almost as varied as phenomena being observed. In NMR, the quadrature data are almost always uniformly sampled (a quadrature measurement is one in which a measurement of both the real and imaginary components of the sinusoids has been made). In astrophysics, the data may be magnitudes or velocities sampled at unevenly spaced intervals.

[1]School of Medicine, Washington University

The standard way to deal with such data is to compute a discrete Fourier transform of the data and then view the transform as an absorption spectrum, a power spectrum, a Schuster periodogram (Schuster 1905), or a Lomb-Scargle periodogram (Lomb 1976, and Scargle 1982 and 1989). See Priestley (1981) and Marple (1987) for reviews of classical spectral estimation techniques. The problem with all such techniques is that they have not be derived from any single set of unifying principles that indicate the optimal way to estimate the period. In this paper we change that by using Bayesian probability theory to deriving the discrete Fourier transform, the power spectrum, the weighted power spectrum, the Schuster periodogram and the Lomb-Scargle periodogram as special cases of a generalized Lomb-Scargle periodogram, and show that the generalized Lomb-Scargle periodogram is a sufficient statistic for single frequency estimation. (A sufficient statistic summarizes all of the information in the data relevant to the question being asked.)

In Bayesian probability theory, there are two basic rules for manipulating probabilities: the product rule and the sum rule. All other rules may be derived from these. If A, B, and C stand for three arbitrary hypotheses, then the product rule states

$$P(AB|C) = P(A|C)P(B|AC), \tag{21.1}$$

where $P(AB|C)$ is the joint probability that "A and B are true given that C is true," $P(A|C)$ is the probability that "A is true given C is true," and $P(B|AC)$ is the probability that "B is true given that both A and C are true."

In Aristotelian logic, the hypothesis "A and B" is the same as "B and A," so the numerical value assigned to the probabilities for these hypotheses must be the same. The order may be rearranged in the product rule, Eq. (21.1), to obtain:

$$P(BA|C) = P(B|C)P(A|BC), \tag{21.2}$$

which may be combined with Eq. (21.1) to obtain a seemingly trivial result

$$P(A|BC) = \frac{P(A|C)P(B|AC)}{P(B|C)}. \tag{21.3}$$

This is Bayes' theorem. It is named after Rev. Thomas Bayes, an 18th century mathematician who derived a special case of this theorem. Bayes' calculations were published in 1763, two years after his death. This theorem, as generalized by Laplace, is the basic starting point for inference problems using probability theory as logic.

The second rule of probability theory, the sum rule, relates to the probability for "A or B." The operation "or" is indicated by a + inside a probability symbol. The sum rule states that given three hypotheses A, B,

and C, the probability for "A or B given C" is

$$P(A+B|C) = P(A|C) + P(B|C) - P(AB|C). \tag{21.4}$$

If the hypotheses A and B are mutually exclusive, that is the probability $P(AB|C)$ is zero, the sum rule becomes:

$$P(A+B|C) = P(A|C) + P(B|C). \tag{21.5}$$

The sum rule is especially useful because it allows one to investigate an interesting hypothesis while removing an uninteresting or nuisance hypothesis from consideration.

To illustrate how to use the sum rule to eliminate nuisance hypotheses, suppose D stands for the data, f the hypothesis "the frequency of a sinusoidal oscillation was f," and B the hypothesis "the amplitude of the sinusoid was B." Now suppose one wishes to compute the probability for the frequency given the data, $P(f|D)$, but the amplitude B is present and must be dealt with. The way to proceed is to compute the joint probability for the frequency and the amplitude given the data, and then use the sum rule to eliminate the amplitude from consideration. Suppose, for argument's sake, the amplitude B could take on only one of two mutually exclusive values $B \in \{B_1, B_2\}$. If one computes the probability for the frequency and (B_1 or B_2) given the data one has

$$P(f|D) \equiv P(f[B_1 + B_2]|D) = P(fB_1|D) + P(fB_2|D). \tag{21.6}$$

This probability distribution summarizes all of the information in the data relevant to the estimation of the frequency f. The probability $P(f|D)$ is called the marginal probability for the frequency f given the data D.

The marginal probability $P(f|D)$ does not depend on the amplitudes at all. To see this, the product rule is applied to the right-hand side of Eq. (21.6) to obtain

$$P(f|D) = P(B_1|D)P(f|B_1D) \; + \; P(B_2|D)P(f|B_2D) \tag{21.7}$$

but

$$P(B_1|D) + P(B_2|D) = 1 \tag{21.8}$$

because the hypotheses are exhaustive. So the probability for the frequency f is a weighted average of the probability for the frequency given that one knows the various amplitudes. The weights are just the probability that each of the amplitudes is the correct one. Of course, the amplitude could take on more than two values; for example if $B \in \{B_1, \cdots, B_m\}$, then the marginal probability distribution becomes

$$P(f|D) = \sum_{j=1}^{m} P(fB_j|D), \tag{21.9}$$

provided the amplitudes are mutually exclusive and exhaustive. In many problems, the hypotheses B could take on a continuum of values, but *as long as only one value of B is realized when the data were taken* the sum rule becomes

$$P(f|D) = \int dB P(fB|D). \tag{21.10}$$

Note that the B inside the probability symbols refers to the hypothesis; while the B appearing outside of the probability symbols is a number or index. A notation could be developed to stress this distinction, but in most cases the meaning is apparent from the context.

21.2 Frequency estimation: The generalized Lomb-Scargle periodogram

The problem addressed is the estimation of the frequency, f, of a sinusoid having known arbitrary amplitude modulation given nonuniformly nonsimultaneously sampled quadrature data. To apply Bayesian probability theory to any problem one must relate the parameter of interest, here the frequency, to the measured data. This is done through a model. For a sinusoid having arbitrary amplitude modulation, the frequency may be related to the real data by

$$d_R(t_i) = A\cos(2\pi f t_i - \theta)Z(t_i) + B\sin(2\pi f t_i - \theta)Z(t_i) + n_R(t_i) \tag{21.11}$$

where $d_R(t_i)$ denotes the real data measured at time t_i, A and B are the cosine and sine amplitudes of the sinusoid, and $n_R(t_i)$ denotes the noise at time t_i. Following Lomb's example, θ will be defined in such a way as to make the cosine and sine functions orthogonal on the discretely sampled times. The function $Z(t_i)$ specifies the amplitude modulation of the sinusoid; $Z(t)$ could be an exponential, a Gaussian, or any other function appropriate to the signal being modeled. If $Z(t)$ is a function of any parameters, these parameters are presumed known; for example, if $Z(t)$ is a decaying exponential, then we assume the decay rate constant is known. Of course, in any Bayesian analysis we could turn our attention to the parameters in $Z(f)$ and estimate them, but for this problem we will consider them as known and suppress them from the notation.

In a quadrature data set one also has a measurement of the imaginary or quadrature part of the signal. The imaginary data are 90° out of phase with the real data. Here this means that model for the imaginary data is 90° out of phase with the model for the real data:

$$d_I(t'_j) = -A\sin(2\pi f t'_j - \theta)Z(t'_j) + B\cos(2\pi f t'_j - \theta)Z(t'_j) + n_I(t'_j). \tag{21.12}$$

We have labeled the times at which the imaginary data were acquired with a prime superscript to distinguish them from the times at which the real

data were acquired and we have added a subscript, I, to several quantities to indicate that these quantities refer to the imaginary part of the signal. The total number of data values will be designated as $N = N_R + N_I$, where N_R and N_I are the number of data values in the real and imaginary channels respectively; N_R and N_I need not be equal and can be zero.

In Bayesian probability theory all of the information in the data relevant to the problem being solved is summarized in a probability density function. For the problem of estimating the frequency, this probability is denoted as $P(f|D_R D_I I)$, where this should be read as the posterior probability for the frequency f given the real data D_R, the imaginary data D_I and the prior information I. In this probability all of the arguments are hypotheses. For example f stands for a hypotheses of the form "at the time the data were take the value of the frequency was f." Thus probability theory is ranking a whole series of models, one for each value of f, and the width of the posterior probability is a natural measure of how uncertain one is of the frequency. The hypotheses I appearing in $P(f|D_R D_I I)$ refers to all of our prior information—including the model equations—and does not refer to the imaginary data; rather it refers to the general background information that goes into making this a well posed problem.

Using the sum rule of probability theory, Eq. (21.5), the posterior probability for the frequency is computed from the joint posterior probability for all of the parameters:

$$P(f|DI) = \int dA dB d\sigma P(fAB\sigma|D_R D_I I) \tag{21.13}$$

where σ is the standard deviation of the Gaussian noise prior probability. The right-hand side of this equation may be factored using Bayes' theorem, Eq. (21.3), and the product rule, Eq. (21.1); one obtains

$$\begin{aligned} P(f|D_R D_I I) &\propto \int dA dB d\sigma P(f|I) P(A|I) P(B|I) P(\sigma|I) \\ &\times P(D_R|fAB\sigma I) P(D_I|fAB\sigma I) \end{aligned} \tag{21.14}$$

where we have assumed logical independence of the parameters, and that the standard deviation of the noise prior probability is the same for both the real and imaginary data; i.e., our prior information indicate that real and imaginary data have the same noise levels.

If we assign uniform prior probabilities to $P(f|I)$, $P(A|I)$, $P(B|I)$, a Jeffreys' prior $(1/\sigma)$ to $P(\sigma|I)$, and assign the two likelihoods using Gaussian noise prior probabilities, one obtains:

$$P(f|DI) \propto \int_{-\infty}^{\infty} dA \int_{-\infty}^{\infty} dB \int_{0}^{\infty} d\sigma \sigma^{-(N+1)} \exp\left\{-\frac{Q}{2\sigma^2}\right\} \tag{21.15}$$

where

$$Q \equiv N\overline{d^2} - 2AR(f) - 2BI(f) + A^2 C(f) + B^2 S(f). \tag{21.16}$$

The mean-square data value, $\overline{d^2}$, is defined as

$$\overline{d^2} = \frac{1}{N}\left[\sum_{i=1}^{N_R} d_R(t_i)^2 + \sum_{j=1}^{N_I} d_I(t'_j)^2\right]. \tag{21.17}$$

The function $R(f)$ is defined as

$$\begin{aligned} R(f) &\equiv \sum_{i=1}^{N_R} d_R(t_i)\cos(2\pi f t_i - \theta)Z(t_i) \\ &- \sum_{j=1}^{N_I} d_I(t'_j)\sin(2\pi f t'_j - \theta)Z(t'_j) \end{aligned} \tag{21.18}$$

which for uniformly sampled data reduces to the real part of a weighted discrete Fourier transform of the complex data. The function $Z(t)$ plays the role of the weight or apodizing function. Similarly, the function $I(f)$ is defined as

$$\begin{aligned} I(f) &\equiv \sum_{i=1}^{N_R} d_R(t_i)\sin(2\pi f t_i - \theta)Z(t_i) \\ &+ \sum_{j=1}^{N_I} d_I(t'_j)\cos(2\pi f t'_j - \theta)Z(t'_j) \end{aligned} \tag{21.19}$$

which for uniformly sampled data reduces to the imaginary part of a weighted discrete Fourier transform of the complex data. The function $C(f)$ is defined as

$$C(f) \equiv \sum_{i=1}^{N_R} \cos^2(2\pi f t_i - \theta)Z(t_i)^2 + \sum_{j=1}^{N_I} \sin^2(2\pi f t'_j - \theta)Z(t'_j)^2 \tag{21.20}$$

and is an effective number of data items in the real data, see Bretthorst 2000 for more on this. Similarly the function $S(f)$ is defined as

$$S(f) \equiv \sum_{i=1}^{N_R} \sin^2(2\pi f t_i - \theta)Z(t_i)^2 + \sum_{j=1}^{N_I} \cos^2(2\pi f t'_j - \theta)Z(t'_j)^2 \tag{21.21}$$

and is the effective number of data items in the imaginary data. Finally, the condition that the cross terms cancel, i.e., that the model functions are orthogonal, is used to determine the value of θ. This condition is given by:

$$\begin{aligned} 0 &= \sum_{i=1}^{N_R} \cos(2\pi f t_i - \theta)\sin(2\pi f t_i - \theta)Z(t_i)^2 \\ &- \sum_{j=1}^{N_I} \sin(2\pi f t'_j - \theta)\cos(2\pi f t'_j - \theta)Z(t'_j)^2. \end{aligned} \tag{21.22}$$

Note that if the data are simultaneously sampled, $t_i = t'_i$, Eq. (21.22) is automatically satisfied, so θ may be defined to be zero. Otherwise, θ is given by

$$\theta = \frac{1}{2}\tan^{-1}\left[\frac{\sum_{i=1}^{N_R}\sin(4\pi f t_i)Z(t_i)^2 - \sum_{j=1}^{N_I}\sin(4\pi f t'_j)Z(t'_j)^2}{\sum_{i=1}^{N_R}\cos(4\pi f t_i)Z(t_i)^2 - \sum_{j=1}^{N_I}\cos(4\pi f t'_j)Z(t'_j)^2}\right]. \quad (21.23)$$

The triple integral in Eq. (21.15) may be evaluated as follows: First, the integrals over the two amplitudes are uncoupled Gaussian quadrature integrals and are easily done. One needs only complete the square in the exponent, and a simple change of variables to evaluate them. The remaining integral over the standard deviation of the noise prior probability may be transformed into a Gamma integral and is also easily evaluated. We do not give the details of these evaluations; rather we simply give the results:

$$P(f|DI) \propto \frac{1}{\sqrt{C(f)S(f)}}\left[N\overline{d^2} - \overline{h^2}\right]^{\frac{2-N}{2}} \quad (21.24)$$

where the sufficient statistic $\overline{h^2}$ is given by

$$\overline{h^2} = \frac{R(f)^2}{C(f)} + \frac{I(f)^2}{S(f)} \quad (21.25)$$

and is a generalization of the Lomb-Scargle periodogram.

The generalized Lomb-Scargle periodogram, Eq. (21.25), has a number of very interesting features. First, when the data are real and the sinusoid is stationary, the sufficient statistic for single frequency estimation is the Lomb-Scargle periodogram; not the Schuster periodogram, i.e., not the power spectrum. Second, when the data are real, but $Z(t)$ is not constant, then Eq. (21.25) generalizes the Lomb-Scargle periodogram in a very straightforward manner to account for the amplitude modulation of the signal. Third, for uniformly sampled quadrature data when the sinusoid is stationary, Eq. (21.25) reduces to a Schuster periodogram or the power spectrum of the data. So while the Schuster periodogram is not a sufficient statistic for frequency estimation in real nonquadrature data, it is a sufficient statistic for quadrature data. Fourth, for uniformly sampled quadrature data when the sinusoid is not stationary, Eq. (21.25) reduced to a weighted power spectrum of the data. Thus the weighted power spectrum is the sufficient statistic for single frequency estimation when the data are quadrature. Fifth, when the quadrature data are nonuniformly but simultaneously sampled, Eq. (21.25) generalizes the weighted power spectrum to account for the nonuniform samples, but otherwise is the exact analogue of a weighted power spectrum. Finally, when the data are nonuniformly and nonsimultaneously sampled, Eq. (21.25) generalizes to a functional form that is formally identical to a Lomb-Scargle periodogram but adapted to an amplitude modulated quadrature detected sinusoid.

21.3 Aliasing

Now that we have finished deriving the generalized Lomb-Scargle periodogram, we would like to investigate some of its properties, in particular the aliasing phenomenon. First, it is easy to show that the parameter θ appearing in the generalized Lomb-Scargle model does not change the location of the peak in the Lomb-Scargle periodogram; fixing θ only changes the estimated phase of the sinusoid. Consequently, fixing θ simplifies the functional form of the sufficient statistic to formally resemble a power spectrum and indeed the generalized Lomb-Scargle periodogram reduces to a power spectrum for simultaneously sampled quadrature data. Now a powers spectrum, i.e,. the discrete Fourier transform, is a periodic function of frequency. The period is called the bandwidth, and the bandwidth is the largest frequency interval free of repeats or aliases. Because the generalized Lomb-Scargle periodogram reduces to a power spectrum under appropriate conditions, the bandwidth of the generalized Lomb-Scargle periodogram is exactly the same as the bandwidth of the discrete Fourier transform. The question we would like to investigate in this section, is what happens to these repeats or aliases when the data are nonuniformly nonsimultaneously sampled? Are the aliases still there? If not, where did they go?

First, the discrete Fourier transform may be defined as

$$\mathcal{F}(f_k) \equiv \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\left\{2\pi \mathbf{i} f_k t_j\right\} \tag{21.26}$$

where the complex data $\mathbf{d}$ is given by $\mathbf{d} \equiv d_R(t_i) + \mathbf{i} d_R(t_i)$. For uniformly sample data the times are given by

$$t_j = j\Delta T \tag{21.27}$$

and if the fast discrete Fourier transform is used to perform this calculation, the frequencies f_k are given by

$$f_k = \frac{k}{N\Delta T} \qquad k \in \left\{-\frac{N}{2}, -\frac{N}{2}+1, \cdots, \frac{N}{2}\right\}. \tag{21.28}$$

The time ΔT is the time interval between data samples and can be used to define the Nyquist critical frequency,

$$f_{N_c} = \pm\frac{1}{2\Delta T}. \tag{21.29}$$

The Nyquist critical frequency may be used to define the bandwidth:

$$\text{bandwidth} \equiv \left(-f_{N_c} \leq f \leq f_{N_c}\right). \tag{21.30}$$

It is the largest frequency interval over which the discrete Fourier transform is not a periodic function of frequency.

To understand why the discrete Fourier transform is a periodic function of frequency, suppose we wish to evaluate the discrete Fourier transform at the frequencies outside the bandwidth:

$$f_k = \frac{k}{N\Delta T}, \qquad k = mN + k', \quad k' \in \left\{-\frac{N}{2}, -\frac{N}{2}+1, \cdots, \frac{N}{2}\right\}. \tag{21.31}$$

The index k' specifies the nonaliased frequency interval of a discrete Fourier transform. The integer m shifts this frequency interval up or down by an integer multiple of the total bandwidth. If $m = 0$, we are in the interval $(-f_{Nc} \le f_k \le f_{Nc})$; if $m = 1$, we are in the interval $(f_{Nc} \le f_k \le 3f_{Nc})$, etc. If we now substitute Eqs. (21.31, and 21.27) into Eq. (21.26), the reason the discrete Fourier transform is periodic becomes readily apparent

$$\begin{aligned}
\mathcal{F}(f_{k'}) &\equiv \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\left\{\frac{2\pi\mathbf{i}(mN+k')j}{N}\right\}, && (21.32)\\
&= \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\{2\pi\mathbf{i}mj\} \exp\left\{\frac{2\pi\mathbf{i}k'j}{N}\right\}, && (21.33)\\
&= \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\left\{\frac{2\pi\mathbf{i}k'j}{N}\right\}, && (21.34)\\
&= \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\{2\pi\mathbf{i}f_{k'}t_j\}. && (21.35)
\end{aligned}$$

In going from Eq. (21.33) to (21.34) a factor, $\exp\{\mathbf{i}(2\pi mj)\}$, was dropped because both m and j are integers, so $(2\pi mj)$ is an integer multiple of 2π, and the complex exponential is one. Aliases occur because the complex exponential canceled leaving behind a discrete Fourier transform on the interval $(-f_{Nc} \le f_k \le f_{Nc})$. The integer m specifies which integer multiple of the bandwidth is being evaluated and will always be an integer no matter how the data are collected. However, the integer j came about because the data were uniformly sampled. If the data had not been uniformly sampled the relationship, $t_j = j\Delta T$, would not hold, the complex exponential would not have cancelled, and aliases would not have been present.

In the present problem, nonuniformly nonsimultaneously sampled data, there is no ΔT such that all of the acquisition times are integer multiples of this time; not if the times are truly sampled randomly. However, all data and times must be recorded to finite accuracy. Consequently, there must be a largest effective dwell time, $\Delta T'$, such that all of the times (both the real and imaginary) must satisfy

$$t_l = k_l \Delta T' \qquad t_l \in \{\text{Real } t_i \text{ or Imaginary } t'_j\} \tag{21.36}$$

where k_l is an integer. The subscript l was added to k to indicate that each of the times t_l requires a different integer k_l to make this relationship true.

Of course, this was also true for uniformly sampled data: its just that for uniformly sampled data the integers were consecutive, $k_l = 0, 1, \cdots N-1$. The effective dwell time is always less than or equal to the smallest time interval between data items, and is the least common denominator for all of the times. Additionally, the effective dwell time is the dwell time at which one would have had to acquire data in order to obtain a uniformly sampled data set with data items at each of the times t_i and t'_j. The effective dwell time, $\Delta T'$, can be used to define a Nyquist critical frequency

$$f_{Nc} = \frac{1}{2\Delta T'}. \tag{21.37}$$

Aliases *must* appear for frequencies outside this bandwidth.

The reason that aliases must appear for frequencies outside this bandwidth can be made apparent in the following way. Suppose we have a hypothetical data set that is sampled at $\Delta T'$. Suppose further, the hypothetical data are zero everywhere except at the times we actually have data, and there the data are equal to the appropriate $d_R(t_i)$ or $d_I(t'_j)$. If we now compute the discrete Fourier transform of this hypothetical data set, then by the analysis done in Eqs. (21.32)-(21.35) the Nyquist critical frequency of this data set is $1/2\Delta T'$ and frequencies outside the bandwidth are aliased. Now look at the definitions of $R(f)$ and $I(f)$, Eqs. (21.18) and (21.19). You will find that these quantities are just the real and imaginary parts of the discrete Fourier transform of our hypothetical data set. The zeros in the hypothetical data cannot contribute to the sums in the discrete Fourier transform: they act only as place holders, and so the only part of the sums that survive are just where we have data. By construction that is just what Eqs. (21.18) and (21.19) are computing. So aliases must appear at frequencies greater than this Nyquist critical frequency. For much more on aliases see Bretthorst 2000.

21.4 Parameter estimates

The generalize Lomb-Scargle periodogram is a sufficient statistic for the estimation of a frequency in nonuniformly nonsimultaneously sample data. However, the frequency is not the only parameter appearing in the model; the model also implicitly contains an amplitude, phase and possible one or more parameters associated with amplitude modulation of the signal. In this section we would like to investigate what happens to the parameter when the data are nonuniformly nonsimultaneously sampled. In particular we would like to know if the parameter estimates change when the data are nonuniformly nonsimultaneously sampled.

In this discussion we are going to estimate the parameters using the data shown in Fig. 21.1(A) and (B). These two data sets contain exactly

FIGURE 21.1. Uniformly and Nonuniformly Sampling

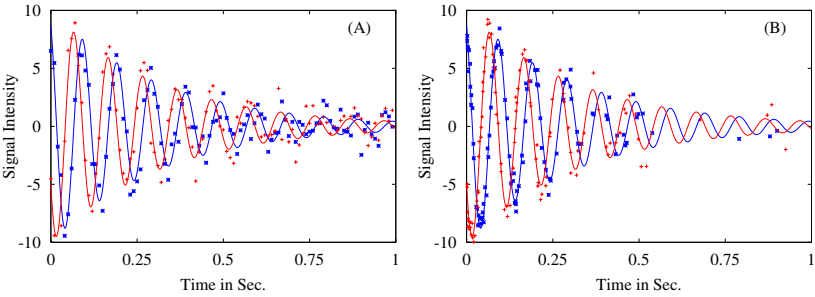


Fig. 21.1. Panel (A) and (B) are simulated data, each data set has exactly the same signal having exactly the same signal-to-noise. The data sets differ only because panel (A) has been uniformly sampled, while (B) has been nonuniformly sampled. Note the nonuniform samples were taken exponentially, thus there are more samples at the beginning of the data and exponentially fewer at the end of the data.

the same signal and have exactly the same signal-to-noise, they differ from each other only in that panel (A) has been uniformly sampled while panel (B) has been randomly sampled. These random samples are distributed exponentially. We mention this only because it will become important later when we consider amplitude estimation. The noise realizations in each data set are different, and this will result in slightly different parameter estimates for each data set.

We will discuss estimation of the frequency, decay rate constant and the amplitude. We will not discuss estimation of the phase and standard deviation of the noise prior probability as these are of less importance. The model we will use is given by

$$d_R(t_i) = A\cos(2\pi f t_i + \phi)\exp\{-\alpha t_i\} \tag{21.38}$$

for the real channel. This model is of the general form of the Lomb-Scargle model, but now we have suppressed the extra phase parameter, as its redundant, we have added an exponential decay rate constant to describe the amplitude modulation, and we have written the model in terms of an amplitude and phase rather than sine and cosine amplitudes.

Markov chain Monte Carlo was used to compute the marginal posterior probability for each parameter. All of the parameters appearing in the model were simulated simultaneously, thus the target distribution of Markov chain Monte Carlo simulation was the joint posterior probability for all the parameters. We targeted the joint posterior probability for all of the parameters for computational convenience; *i.e.,* it was easier to do a single Markov chain Monte Carlo simulation than to do five separate calculations, one for each parameter appearing in the model. Because the probability density functions shown in Fig. 21.2(A), (B) and (D) were formed

FIGURE 21.2. Estimating The Parameters

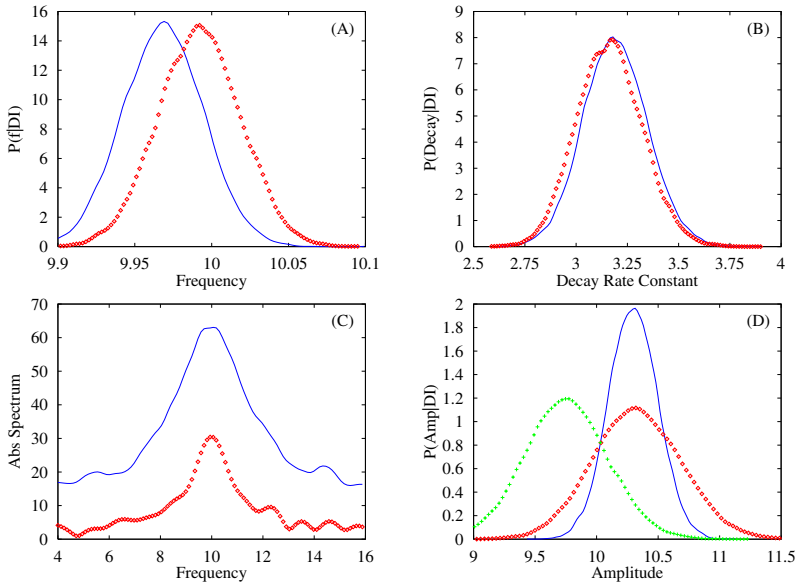


Fig. 21.2. The posterior probability of the parameters was computed for the uniformly and nonuniformly nonsimultaneously sampled data, open characters and solid lines respectively. Panel (A) is the posterior probability for the frequency, (B) the decay rate constant, (D) the amplitude. Panel (C) is the absolute-value spectrum computed for the two data sets. The extra curve in panel (D), the plus signs, was computed from a nonuniformly nonsimultaneously sample data set having uniformly sampled times, see text for details.

by computing a histogram of the Markov chain Monte Carlo samples, there are small, irrelevant, artifacts in these plots that are related to the number of samples drawn from the simulation. For more on Markov chain Monte Carlo methods and how these can be used to implement Bayesian calculations see Neal 1993 and Gilks, et. al. 1996.

The posterior probability for the frequency, decay rate constant, and amplitude are shown in Fig. 21.2(A), (B) and (D) respectively. Each of these plots is the fully normalized marginal posterior probability for the parameter of interest independent of all of the other parameters appearing in the model. Panel (C) contains the absolute-value spectra computed from these two data sets and will be used to compare Fourier transform estimation procedures to the Bayesian calculations. The curves drawn with open characters were computed using the uniformly sampled data shown in Fig. 21.1(A); while the solid lines in these plots were computed from the nonuniformly nonsimultaneously sampled data shown in Fig. 21.1(B).

The marginal posterior probability for the frequency is shown in Fig. 21.2(A). This is the fully normalized marginal posterior probability for the frequency

independent of all of the other parameters, Eq. (21.24). Note that the true frequency, 10 Hz, is well covered by the posterior probability computed from both the uniformly (open characters) and nonuniformly nonsimultaneously (sold line) sampled data. Also note that these distributions are almost identical in height and width. Consequently, both the uniform and nonuniformly nonsimultaneously sampled data have given the same parameter estimates to within the uncertainty in these estimates. Of course the details for each estimated differ, because the noise realizations in each data set differ. Consequently, the frequency estimate is not strongly dependent on the sampling scheme. Indeed this can be derived from the rules of probability theory with the following proviso: the two sampling schemes must cover the *same* total sampling time and must sample the signal in a reasonably dense fashion so that sums may be approximated by integrals. Having said this, we must reemphasize that this is only true for frequency estimates using data having sampling schemes covering the *same* total sampling time; it is not true if the sampling times differ nor is it necessarily true of the other parameters appearing in the model. Indeed one can show that for a given number of data values, the precision of the frequency estimate for a stationary sinusoid is inversely proportional to the total sampling time. Thus, sampling 10 times longer will result in frequency estimates that are 10 times more precise. As noted in Bretthorst 1988 this is equivalent to saying that for frequency estimation data values at the front and back of the data are most important in determining the frequency, because it is in these data that small phase differences are most highly magnified by the time variable.

We have also plotted the absolute-value spectra computed from these two data sets, Fig. 21.2(C). Note that the peaks of these two absolute-value spectra are at essentially the same frequency as the corresponding peaks in panel (A); although they are plotted on differing scales. If the absolute value spectrum is used to estimate the frequency, one would typically use the peak frequency as the estimate, and then claim roughly the half-width-at-half-height as the uncertainty in this estimate. For these two data sets that is about 10 plus or minus 2 Hz. The two fully normalized posterior probabilities shown in panel (A) span a frequency interval of only 0.2 Hz. This frequency interval is roughly 6 standard deviations. Thus the frequency has been estimated to roughly 10 Hz with an uncertainty of $0.2/6 \approx 0.03$ Hz; a 60 fold reduction in the uncertainty in the frequency estimate.

One last note before we begin the discussion of estimating the decay rate constant, we note that all of the details in the wings of the absolute-value spectrum shown in panel (C) are irrelevant to the frequency estimation process. The posterior probability for the frequency has peaked in a region that is very small compared to the scale of these wings, *all of the information about the frequency estimate is contained in a very small region around the single largest peak in the spectrum.* In the discrete Fourier transform,

the presence of multiple peaks may or may not be an indication of multiple resonances. Indeed it is easy to show that the generalize Lomb-Scargle periodogram may have peaks that are related to the sampling scheme. The only way to be certain that multiple resonances are presence, is to postulate a model containing multiple resonances and then compute the posterior probability for the number of resonances.

The marginal posterior probability for the decay rate constant is shown in Fig. 21.2(B). Here we again find that the parameter estimates from both data sets are essentially identical in all of there relevant details. Both probabilities peak at nearly the same value of the decay rate constant, both have nearly the same width, and therefore the same standard deviation; thus like frequency estimates, the estimates for the decay rate constants do not strongly dependent on the sampling scheme. In principle the accuracy of the estimates for the decay rate constants scale with time just like the frequency estimates, of course, with decaying signals this is of little practical importance. Note that the decay rate constant has been estimated to be about 3.2 ± 0.3 Sec.$^{-1}$ at one standard deviation. The true value is 3 Sec.$^{-1}$, so both sampling schemes give reasonable estimates of the decay rate. If one were to try and estimate the decay rate constant from the absolute-values spectrum, the half-width-at-half-height would normally be used, here that is about 2 Sec.$^{-1}$ and no claim about the accuracy of the estimate would be made.

The marginal posterior probability for the amplitude of the sinusoid is shown in Fig. 21.2(D). In this paper we did not directly talk about amplitude estimation (see Bretthorst 1992 for a discussion of this subject), rather we treated the amplitudes of the sine and cosine model functions as nuisance parameters and removed them from the posterior probability for the other parameters. We did this because we wished to explore the relationships between frequency estimation using Bayesian probability theory and the discrete Fourier transform. However, the Markov chain Monte Carlo simulation used Eq. (21.38) as the model for the real data, so it was a trivial matter to compute the posterior probability for the amplitude. If you examine Fig. 21.2(D) you will note that now we do have a difference between the uniform (open characters) and the nonuniformly nonsimultaneously sampled data (solid lines). The amplitude estimates from the nonuniformly nonsimultaneously sampled data are a good factor of 2 more precise than the estimates from the uniformly sampled data. One might think that this is caused by the nonuniform nonsimultaneous sampling and this would be correct, but not for the obvious reasons. If you examine panel (D) you will note that we have plotted a third curve (plus signs). This curve is the posterior probability for the amplitude computed from data with the exact same signal and signal-to-noise ratio, but having times that are nonuniformly nonsimultaneously sampled where the times were generated from a uniform random number generator. We will call this data set the uniform-randomly sampled data. Note that the height and width

of the posterior probabilities computed from both the uniformly and the uniform-randomly sampled data are essentially the same, so by itself the nonuniform nonsimultaneous sampling did not cause the amplitude estimates to improve. The amplitude estimate improved because exponential sampling gathered more data where the signal was large. The accuracy of the amplitude estimate is proportional to the standard deviation of the noise and inversely proportional to square root of the effective number of data values. Because exponential sampling gathered more data where the signal was large, its effective number of data values was larger and so the amplitude estimate improved. In this case, the improvement was about a factor of 2, so the exponential sampling had an effective number of data values that was about a factor of 4 larger than for the uniformly or uniform-randomly sampled data. This fact is also reflected in differing heights of the absolute value spectra plotted in Fig. 21.2(C). The peak height of an absolute value spectrum is proportional to the square root of the effective number of data values. In panel (C) the spectra computed from the uniformly sampled data set, open characters, is roughly a factor of 2 lower than the height of the spectrum computed from the exponentially sampled data set, solid line.

21.5 Summary and conclusions

Probability theory generalizes the Lomb-Scargle periodogram roughly as follows: in uniformly or nonuniformly sampled real data, the sufficient statistic for estimating the frequency of a single stationary sinusoid is the Lomb-Scargle periodogram. When the function $Z(t_i)$ is not a constant, probability theory generalized the Lomb-Scargle periodogram to include this modulation. For a stationary sinusoid, when the data are quadrature simultaneously sampled, probability theory simplifies the Lomb-Scargle periodogram to a Schuster periodogram. When the sinusoid is not stationary, the sufficient statistic becomes a weighted power spectrum where the weighting function is given by $Z(t)$. Finally, when the data are nonuniformly nonsimultaneously sampled, the sufficient statistic is the generalized Lomb-Scargle periodogram.

In a literal sense, probability theory does no such thing as generalize the discrete Fourier transform or the Lomb-Scargle periodogram. Probability theory simply tells one how to analyze a particular problem optimally. For estimation of a sinusoidal frequency, the sufficient statistics turn out to be related to the discrete Fourier transform. This was, for us, a happy coincidence because it enabled us to interpret the results of the analysis in a way that sheds light on the discrete Fourier transform and how it should be used. In the appropriate limits, the discrete Fourier transform power spectrum, the Schuster periodogram, the Lomb-Scargle periodogram and

the generalizations presented in this paper are all optimal frequency estimators for the *single* sinusoidal case. However, when the true signal deviate from this model, for example when there are multiple sinusoids or the data contain a trend, then these statistics are *never* optimal frequency estimators, and there are always other statistics that will improve the resolution of the multiple frequencies or properly account for trend in the data, see Bretthorst 1988, and 2000.

Aliasing is a general phenomenon and exists in both uniformly and nonuniformly nonsimultaneously sampled data for exactly the same reason. It is the fact that all of the times may be expressed as an integer multiple of an effective dwell time that is the cause of aliasing. Two data sets differing only in how precisely the times are recorded generally have different Nyquist critical frequencies.

The analysis in this paper generalized the concept of bandwidth and showed that uniformly simultaneously sampled data have the smallest possible bandwidth. The addition of any nonuniformly nonsimultaneously sampled data always increases the Nyquist critical frequency and thus increases the bandwidth. The Nyquist critical frequency for nonuniformly nonsimultaneously sampled data may be many orders of magnitude greater than that for uniformly simultaneously sampled data having similar acquisition parameters. Consequently, nonuniformly nonsimultaneously sampled data can have tremendous advantages over uniformly sampled data because the critical time is not how fast one can sample data, but how accurately one can vary the acquisition of each data item. This opens up the possibility of measuring very high frequencies with bandwidths much larger than previously possible.

21.6 References

Bretthorst, G. Larry (1988), "Bayesian Spectrum Analysis and Parameter Estimation," in *Lecture Notes in Statistics,* **48**, J. Berger, S. Fienberg, J. Gani, K. Krickenberg, and B. Singer (eds), Springer-Verlag, New York, New York.

Bretthorst, G. Larry, (2000) "Nonuniform Sampling: Aliasing and Bandwidth," *Maximum Entropy and Bayesian Methods*, G. Erickson *ed.,*, Kluwer academic press, the Netherlands.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (1996), "Markov Chain Monte Carlo in Practice," Chapman & Hall, London.

Lomb, N. R. (1976) "Least-Squares Frequency Analysis of Unevenly Spaced Data," *Astrophysical and Space Science*, **39**, pp. 447-462.

Marple, S. L. (1987) *Digital spectral Analysis with applications,* Prentice-Hall , Inc., Englewood Clifts. New Jersey.

Neal, Radford M. (1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods," technical report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, 2 Vols., Academic Press, Inc., Orlando FL, Combined paperback edition with corrections (1983).

Scargle, J. D. (1982) "Studies in Astronomical Time Series Analysis II. Statistical Aspects of Spectral Analysis of Unevenly Sampled Data," *Astrophysical Journal*, **263**, pp. 835-853.

Scargle, J. D. (1989) "Studies in Astronomical Time Series Analysis. III. Fourier Transforms, Autocorrelation and Cross-correlation Functions of Unevenly Spaced Data," *Astrophysical Journal*, **343**, pp. 874-887.

Schuster, A., (1905), "The Periodogram and its Optical Analogy," *Proceedings of the Royal Society of London,* **77**, p. 136.

Commentary by Thomas J. Loredo[2]

The Bayesian and frequentist approaches to statistical inference differ in many ways. Two differences are of special importance for the construction of algorithms. The first concerns the choice of statistic (the function of the data on which to base inferences). In frequentist statistics, specifying a good statistic for a nontrivial problem is a difficult art. In the Bayesian approach, once a hypothesis space is specified, probability theory automatically identifies what functions of the data to use to discriminate between the hypotheses (i.e., the functions that appear in the likelihood). This automatic behavior comes at the cost of having to specify alternative hypotheses (some frequentist calculations can proceed without specifying an alternative to the null hypothesis, e.g., goodness-of-fit tests). Second, the two approaches use the sampling distribution for the data (i.e., the likelihood for the hypotheses) very differently to calculate probabilities associated with inferences. In frequentist calculations, the hypothesis is fixed,

[2]Department of Astronomy, Cornell University

and sums and integrals are calculated in the sample space of hypothetical data. In Bayesian calculations, the data are fixed to the observed values, and sums and integrals are calculated in the hypothesis or parameter space. Consequently, even when the same statistics are used in both approaches, qualitatively different results can be found.

Bretthorst's work provides a rich source of examples of how both of these key differences manifest themselves in real-world calculations. In the limited space of his presentation here, he has emphasized the first difference: how Bayesian probability theory can be used as a "machine" for generating useful statistics. In this commentary I will highlight a few aspects of this difference, but I will dwell on the second difference: how, once the statistic has been identified, adopting the Bayesian approach leads one to use it in ways that can produce results that differ dramatically from those found in frequentist calculations with the same statistic.

21.7 Choosing a statistic

Bretthorst describes a Bayesian calculation that identifies the Lomb-Scargle periodogram as the sufficient statistic for frequency estimation, and then generalizes this, adding to a growing list of Bayesian results that appear almost obvious once stated, but which have somehow escaped notice despite decades of work on time series. These results include the earlier demonstration by Jaynes and Bretthorst that the Schuster periodogram is the sufficient statistic for frequency estimation with uniformly sampled data, as well as Scargle's discovery, reported in these proceedings, that the cross-correlation function is a sufficient statistic for inferring lags between time series. It is worth emphasizing that the mathematics underlying these results is quite simple; these discoveries eluded previous researchers, not because the calculations were difficult, but because *a different conceptual approach was required*. They demonstrate that careful consideration of conceptual issues is not a merely philosophical exercise, concerned only with matters of interpretation, but opens the doors to new results of practical significance.

Of course, periodograms and cross-correlation functions are tools that have been used by time series analysts for many decades. What is new in the Bayesian calculations is a clarified connection between the statistics and time-domain model structure, and a precise "recipe" for using the statistics to calculate probabilities for hypotheses of interest. The former points the way to powerful generalizations. These include statistics developed by Bretthhorst for estimating multiple frequencies and frequency multiplets (which can be resolved even when they are much closer together than the width of a periodogram peak), and for estimating frequencies of decaying sinusoids. Further generalizations follow simply by changing the sinusoid

basis in Bretthorst's calculations. For example, Scargle, Bretthorst, and I have independently developed "Kepler periodogram" approaches for detecting planets in radial velocity and astrometric data by using the periodic functions desribing Keplerian reflex motions as the basis functions.

21.8 Sample space vs. parameter space

The second new aspect of the Bayesian calculations—new recipes for how to use the statistics—is of equally great significance. I believe this aspect, which underlies a truly revolutionary new understanding of periodograms, is little appreciated by astronomers. To highlight it, consider frequency detection and estimation for uniformly sampled time series over a time T, with data d_i at times t_i, for $i = 1$ to N. The periodogram is

$$I(f) = \frac{1}{N}\left[\sum_i d_i \cos(2\pi f t_i)\right]^2 + \frac{1}{N}\left[\sum_i d_i \sin(2\pi f t_i)\right]^2, \qquad (21.39)$$

viewed as a continuous function of the frequency, f. Since there are only N data, there must be at most N "pieces of information" in $I(f)$. Actually, there are $N/2 + 1$ (nearest integer if N is odd) values of $I(f)$ at equally spaced frequencies that determine the entire function. These values can be found using the discrete Fourier transform (DFT) of the data to calculate the power spectrum at $N/2 + 1$ Fourier frequencies, $I_j = I(f_j)$, where $f_j = 2\pi j/T$, with $j = 0$ to $N/2$. The Fourier power spectral density (PSD) familiar to astronomers is I_j, with a possible subtraction of an average term from d_i, and with various normalizations adopted (to simplify its statistical properties). For simplicity, we here call I_j the PSD.

In frequentist analyses, the PSD is typically viewed as *an estimator of the signal's power spectrum*, albeit corrupted by the finite and discrete nature of the data and the presence of noise. The statistical properties of this estimator follow from how the PSD values vary as the values of the N data vary through repeated observation. Attention is focused on the PSD rather than the continuous $I(f)$ because the PSD values are statistically independent under the null hypothesis of a constant (e.g., zero) signal (plus noise). But an unfortunate consequence of there only being $N/2+1$ Fourier frequencies is that the expected behavior of the PSD when a periodic signal is present differs depending on whether the period of the signal lies exactly on or away from a Fourier frequency. This is illustrated in Figure 21.3. Figure 21.3a shows the PSD calculated from data with a signal at a Fourier frequency and a signal-to-noise of 10; Figure 21.3b shows the PSD calculated from similar data, but with the signal frequency midway between two Fourier frequencies. *Spectral leakage* is apparent; when the true frequency is not a Fourier frequency, power "leaks" to neighboring frequencies, reducing the amplitude of the PSD peak, and broadening it. This complicates

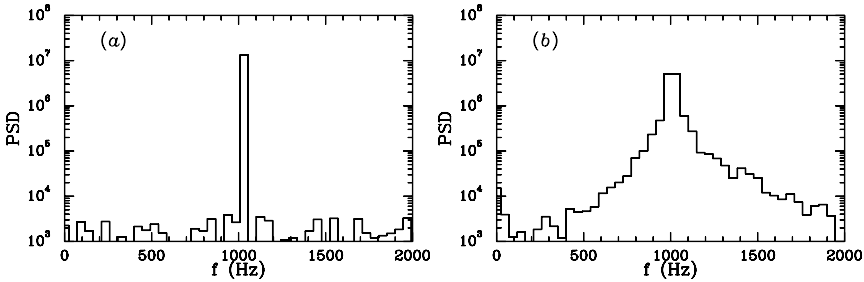


FIGURE 21.3. Leakage in the PSD. (a) The PSD (up to 2 kHz) for data simulated with a weak sinusoidal signal and added Gaussian noise; $S/N = 5$, with 1024 samples at a 48 kHz sampling rate. The sinusoid frequency is at the Fourier frequency nearest 1 kHz (1031.25 Hz). (b) As in (a), but the frequency (1008 Hz) is between two Fourier frequencies.

the interpretation and use of the PSD for both detection and estimation. Conventional remedies for leakage use windowing or tapering of the data (essentially a linear averaging process) to reduce leakage at non-Fourier frequencies, at the expense of spreading the signal power when the signal is at or near a Fourier frequency.

In the Bayesian approach of Bretthorst, one does not address frequency detection and estimation through the intermediary of a spectrum estimator. Instead, one simply calculates the probability that a sinusoidal signal of a specified frequency is present. The continuous periodogram is "handed" to the analyst in the course of this calculation, not as a spectrum estimator, but (roughly) as *the logarithm for the (marginal) posterior probability for the unknown frequency.* Probabilities for hypotheses of interest are found by integrating the exponentiated periodogram over frequency (a parameter space integral). In such calculations, evaluating $I(f)$ *between* Fourier frequencies is important.

Figure 21.4 illustrates some aspects of the Bayesian procedure. Figure 21.4a shows the continuous periodogram for the same data used to produce Figure 21.3a (signal at a Fourier frequency). Dots highlight the values at Fourier frequencies (the values plotted in Fig. 21.3a). Figure 21.4b shows a similar plot, corresponding to Figure 21.3b (signal at a non-Fourier frequency). Although the values at Fourier frequencies exhibit very different behavior in Figures 21.3a and 21.3b, the continuous periodograms are qualitatively very similar for both data sets. The insets in the figures show the marginal posterior distributions for the frequency in each case, found by nonlinear processing of $I(f)$. These distributions are extremely sharp and narrow in both cases, and very accurately pinpoint the correct frequency. The sidelobes and other structure evident in $I(f)$ are exponentially attenuated. Detection probabilities (for determining whether a periodic signal is present), found by integrating the exponentiated periodogram over all

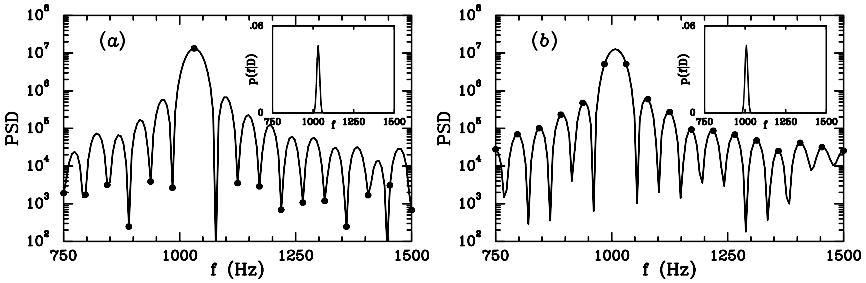


FIGURE 21.4. Details of the continuous periodogram. (a) Periodogram near its peak, for the data used for Fig. 21.3a (at a Fourier frequency); dots show values of the (discrete) PSD. Inset shows the posterior distribution for the frequency found by exponentiating the periodogram scaled by the noise variance. (b) As in (a), but for the data used for Fig. 21.3b (signal frequency between Fourier frequencies).

f, similarly exhibit comparable performance for Fourier and non-Fourier frequencies.

From the Bayesian viewpoint, there is no spectral leakage problem associated with non-Fourier frequencies, which have no special role to play when one is calculating parameter space (rather than sample space) integrals. The continuous periodogram has a complicated shape for *all* possible signal frequencies; the complications are merely hidden in some cases if one examines only Fourier frequencies. The complicated shape of the periodogram results from the finite and discrete nature of the data, but is not viewed as distortion of a "spectrum estimate" due to convolution of with window and sampling functions. Rather, the shape conveys information about how the finite and discrete nature of the data can confuse one's inferences about a single sinusoid when noise is significant (in which case the sidelobes will not be as attenuated as in the examples above), and is similar for data generated by signals at or between Fourier frequencies.

This is but one example of how adopting the Bayesian approach greatly changes how periodograms are used to make inferences. I urge readers intrigued by this brief discussion to further study Bretthorst's book and papers, where more important differences will be found (e.g., use of the periodogram *peak* to infer the noise amplitude, rather than the PSD "background" level; and nonlinear processing of the real and imaginary parts of the DFT to resolve closely spaced frequencies).

This page intentionally left blank

22

Multiscale Methods in Astronomy

Jean-Luc Starck[1]

ABSTRACT Wavelets have been used extensively for several years now in astronomy for many purposes, ranging from data filtering and deconvolution, to star and galaxy detection or cosmic ray removal. We review in this paper a range of methods and applications. A recent method, the ridgelet transform is also described, and we show its interest when the data present anisotropic features.

22.1 Introduction

The wavelet transform has been extensively used in astronomical data analysis during the last ten years. A quick search with ADS shows that around 500 papers contain the keyword "Wavelet" in their abstract, and all astrophysical domains were concerned, from the sun study to the CMB analysis. This large success of the wavelet transform (WT) is due to the fact that astronomical data presents generally complex hierarchical structures, often described as fractals. Using multiscale approaches such as the wavelet transform (WT), an image can be decomposed into components at different scales, and the WT is therefore well-adapted to astronomical data study.

The following section presents the different WT algorithms which can be used. In section 22.3, we discuss how noise, which is always present in astronomical images, is managed. In section 22.4, we review some wavelet based applications. A recent multiscale method, the ridgelet transform, is described in section 22.5, and we show its interest when the data present anisotropic features.

22.2 The Wavelet Transform

There are many WT algorithms [MF98, SMB98]. The (bi-) orthogonal wavelet transform [Mal89], often referred to as the Fast Wavelet Transform

[1]Centre d'Études Atomique, Paris

(FWT), is certainly the most widely used among available discrete wavelet transform algorithms. It is a non-redundant representation of the information. An introduction to this type of transform can be found in [Dau92]. The famous Haar transform belongs to this class. Using the FWT, a signal s can be decomposed by:

$$s(l) = \sum_k c_{J,k}\phi_{J,l}(k) + \sum_k \sum_{j=1}^{J} \psi_{j,l}(k) w_{j,k} \tag{22.1}$$

with $\phi_{j,l}(x) = 2^{-j}\phi(2^{-j}x - l)$ and $\psi_{j,l}(x) = 2^{-j}\psi(2^{-j}x - l)$, where ϕ and ψ are respectively the scaling function and the wavelet function. J is the number of resolutions used in the decomposition, w_j the wavelet (or details) coefficients at scale j, and c_J is a coarse or smooth version of the original signal s.

Another well known algorithm is the à trous wavelet transform. The wavelet transform of an image by this algorithm produces, at each scale j, a set $\{w_j\}$. This has the same number of pixels as the input data set. The original data c_0 can be expressed as the sum of all the wavelet scales and the smoothed array c_p by $c_0 = c_J + \sum_{j=1}^{J} w_j$ and a pixel at position k can be expressed also as the sum of all the wavelet coefficients at this position, plus the smoothed array: $c_{0,k} = c_{J,k} + \sum_{j=1}^{J} w_{j,k}$.

In astronomical images, there are generally no edges, and objects are relatively diffuse. For this reason, an isotropic or symmetric analysis produces better results. This is the reason why the à trous algorithm is often preferred. Furthermore, for the most usual applications (detection, filtering, deconvolution, etc.), undersampling leads to severe artifacts which can be easily avoided by non-orthogonal transforms such the à trous algorithm. For these reasons, the FWT is rarely used in the astronomical domain.

22.3 Significant wavelet coefficients

Astronomical data are always contaminated by a noise, and it is important to detect the wavelet coefficients which are "significant", i.e. the wavelet coefficients which have an absolute value too large to be due to noise. We defined the multiresolution M^D of the data set D by:

$$M^D_{j,k} = \begin{cases} 1 & \text{if } w_{j,k} \text{ is significant} \\ 0 & \text{if } w_{j,k} \text{ is not significant} \end{cases} \tag{22.2}$$

where j is the scale, k the pixel position, and $w_{j,k}$ the wavelet coefficient of D at scale j and at position k. We need now to determine when a wavelet coefficient is significant. For Gaussian noise, it is easy to derive an estimation of the noise standard deviation σ_j at scale j from the noise standard deviation, which can be evaluated with good accuracy in an automated way

[SM98]. To detect the significant wavelet coefficients, it suffices to compare the wavelet coefficients $w_{j,k}$ to a threshold level t_j. t_j is generally taken equal to $k\sigma_j$, and k is chosen between 3 and 5. The value of 3 corresponds to a probability of false detection of 0.27%. If $w_{j,k}$ is small, then it is not significant and could be due to noise. If $w_{j,k}$ is large, it is significant:

$$\begin{array}{l} \text{if } \mid w_{j,k} \mid \geq t_j \quad \text{then } w_{j,k} \text{ is significant} \\ \text{if } \mid w_{j,k} \mid < t_j \quad \text{then } w_{j,k} \text{ is not significant} \end{array} \tag{22.3}$$

Other thresholding approaches have been proposed, like the *universal threshold* [DJ93], or the SURE method [CD95], but they generally do not produce as good results as the k-sigma method.

When the noise is not Gaussian, many strategies have been developed depending on the nature of the noise or directly from simulations [SMB98].

22.4 Wavelet based Methods in Astronomical Data Processing

22.4.1 Filtering

The most used filtering method is the hard thresholding, which consists of setting to 0 all wavelet coefficients which have an absolute value lower than a threshold t_j

$$\tilde{w}_{j,k} = \begin{cases} w_{j,k} & \text{if } w_{j,k} > t_j \\ 0 & \text{otherwise} \end{cases} \tag{22.4}$$

We define the function $\mathcal{T}$ as the function which set to zero all wavelet coefficients outside a given multiresolution support M:

$$\mathcal{F}(M, x) = c_{J,k} + \sum_{j=1}^{J} M_{j,k} w_{j,k} \tag{22.5}$$

where $c_{J,k}$ and $w_{j,k}$ are obtained from the à trous wavelet transform of x. The filtered version $\tilde{s}$ of the input signal s is obtained by $\tilde{s} = \mathcal{F}(M^s, s)$, M^s being the multiresolution support of s. This solution can be refined by the following iterative scheme:

$$\tilde{s}^{n+1}(k) = s^n(k) + \mathcal{F}(M^s, r^n) \tag{22.6}$$

where $r^n = s - \tilde{s}^n$. This algorithm allows us to constraint the residual to have a zero value inside the the multiresolution support of s [SMB98]. For astronomical image filtering, iterating improves significantly the results, especially for the photometry, i.e. the integrated intensity of a source.

22.4.2 Image Deconvolution

Observed data Y in the physical sciences are generally corrupted by noise, which is often additive and which follows in many cases a Gaussian distribution, a Poisson distribution, or a combination of both. Using Bayes' theorem to evaluate the probability of the realization of the original signal X, knowing the data Y, we have

$$Prob(X|Y) = \frac{Prob(Y|X).Prob(X)}{Prob(Y)} \tag{22.7}$$

$Prob(Y|X)$ is the conditional probability of getting the data Y given an original signal X, i.e. it represents the distribution of the noise.

The denominator in equation (22.7) is independent of X and is considered as a constant (stationary noise). $Prob(X)$ is the a priori distribution of the solution X. In the absence of any information on the solution X except its positivity, a possible course of action is to derive the probability of X from its entropy. Several definitions of entropy has been proposed, and the main ones are: i) Burg [Bur78]: $H_b(X) = -\sum_{pixels} \ln(X)$, ii) Frieden [Fri78]: $H_f(X) = -\sum_{pixels} X \ln(X)$, iii) Gull and Skilling [GS91]: $H_g(X) = \sum_{pixels} X - M - X \ln(X|M)$. Each of these entropies can be used, and they correspond to different probability distributions that one can associate with an image [NN86]. It was shown in [NN86] that results vary strongly with the background level, and that these entropy functions produce poor results for negative structures, i.e. structures under the background level, and compact structures in the signal. The Gull and Skilling entropy gives rise to the difficulty of estimating a model. Furthermore it was shown in [BKK94] that the solution is dependent on this choice.

Many studies [BKK94, PS96] have been carried out in order to improve the functional to be minimized. But the question which should be raised is: what is a good entropy measure for signal restoration?

In [SMG98], the benchmark properties for a good "physical" definition of entropy were discussed. Assuming that a signal X is the sum of several components: $X = S + B + N$, where S is the signal of interest, B the background, and N the noise, we proposed that the following criteria should be verified:

1. The information in a flat signal is zero ($S = 0$, $N = 0$ and $B = \text{Cst}$).
2. The amount of information in a signal is independent of the background (i.e., $H(X)$ is independent of B).
3. The amount of information is dependent on the noise (i.e., $H(X)$ is dependent on N). A given signal X does not furnish the same information in the different cases where the noise N is high or small.
4. The entropy must work in the same way for a pixel which has a value $B + \epsilon$, and for a pixel which has a value $B - \epsilon$. $H(X)$ must be a function of the absolute value of S instead of S.

5. The amount of information is dependent on the correlation in the signal. If the signal S presents large features above the noise, it contains a lot of information. By generating a new set of data from S, by randomly taking the pixel values in S, the large features will evidently disappear, and this new signal will contain less information. But the pixel values will be the same as in S.

The Burg and Frieden entropy functions do not verify any of these criteria, and the Skilling one verifies only point 2. Using the wavelet transform, it has been shown [SMG98, SM99, SMQB01] that an entropy function verifying all cited properties can be obtained, which produces very good results.

22.4.3 Interferometric Image Reconstruction

In interferometric imaging, measurements are carried out in Fourier space but the (u, v) plane is not completely covered. The image, called the dirty map, is obtained by a simple inverse Fourier transform of the data and the PSF, called the dirty beam, by an inverse Fourier transform of the (u, v) plane coverage. The presence of secondary lobes in the dirty beam creates very serious artifacts in the dirty map and a deconvolution is necessary. By applying the CLEAN method at each scale of the wavelet transform using the FFT, we can localize significant structures, and an iterative reconstruction algorithm allows solutions to be found which satisfy the positivity constraint, and the constraint of fidelity to measurements (i.e. at each measured $V_m(u, v)$ $+/-$ $\Delta_m(u, v)$, we require that the solution O satisfies $\mid \hat{O}(u, v) - V_m(u, v) \mid < \Delta_m(u, v)$). More details can be found in [SBLP94].

22.4.4 Object detection

Using the à trous algorithm, an image I can be expressed as the sum of all the wavelet scales and the smoothed array c_J by the expression

$$I(k, l) = c_{J,k,l} + \sum_{j=1}^{J} w_{j,k,l}. \tag{22.8}$$

Hence, we have a *multiscale pixel representation*, i.e. each pixel of the input image is associated to a set of pixels of the multiscale transform. A further step is to consider a *multiscale object representation*, which would associate to an object contained in the data, a volume in the multiscale transform. Such a representation obviously depends on the kind of image we need to analyze. A Multiscale Vision Model (MVM) has been developed [BR95] for astronomical data. Using the MVM, an image I can be decomposed, from its wavelet transform, into a set of components:

$$I(k, l) = \sum_{i=1}^{N_o} O_i(k, l) + B(k, l) + N(k, l) \tag{22.9}$$

where N_o is the number of object, O_i are the objects contained in the data (stars galaxies, etc), B is the background image, and N is the noise. Furthermore, it has been shown [SBVM00] that a deconvolution can be introduced in this decomposition, and the set of components verifies:

$$I(k,l) = \sum_{i=1}^{N_o} (P_i * O_i)(k,l) + B(k,l) + N(k,l) \tag{22.10}$$

where P_i is the Point Spread Function (PSF) associated to the object i. We have therefore an elegant solution for the deconvolution with a spatially variant PSF.

22.5 The Ridgelet Transform

The application of the à trous algorithm has lead to impressive results, compared to previous methods, for data restoration and object detection. As it was discussed before, this wavelet transform is well adapted to the analysis of isotropic features. However, all features included in 2D and 3D astronomical data set are not isotropic (filaments, elongated galaxies, planetary images, arclet, ...). The FWT may be better than the à trous for such data set, but still presents some limitations which may impact in some applications. Indeed, if the FWT performs better than the FFT to represent edges in an image, it is still not optimal. There is only a fixed number of directional elements independent of scales, and there is no highly anisotropic elements. For instance, the Haar 2D wavelet transform is optimal to find features with a ratio $length/width = 2$, and a horizontal,vertical, or diagonal orientation. This problem have lead to the development of other multiscale representations, like the ridgelet [CD99] or the curvelet transform [DD00].

The two-dimensional continuous ridgelet transform in $\mathbf{R}^2$ can be defined as follows [CD99]. We pick a smooth univariate function $\psi : \mathbf{R} \to \mathbf{R}$ with sufficient decay and satisfying the admissibility condition

$$\int |\hat{\psi}(\xi)|^2/|\xi|^2 \, d\xi < \infty, \tag{22.11}$$

which holds if, say, ψ has a vanishing mean $\int \psi(t)dt = 0$. We will suppose that ψ is normalized so that $\int |\hat{\psi}(\xi)|^2 \xi^{-2} d\xi = 1$.

For each $a > 0$, each $b \in \mathbf{R}$ and each $\theta \in [0, 2\pi)$, we define the bivariate *ridgelet* $\psi_{a,b,\theta} : \mathbf{R}^2 \to \mathbf{R}^2$ by

$$\psi_{a,b,\theta}(x) = a^{-1/2} \cdot \psi((x_1 \cos\theta + x_2 \sin\theta - b)/a); \tag{22.12}$$

Figure 22.5 (upper left) shows an example ridgelet function. Figure 22.5 upper right, and bottom left shows the the same function after rotation and

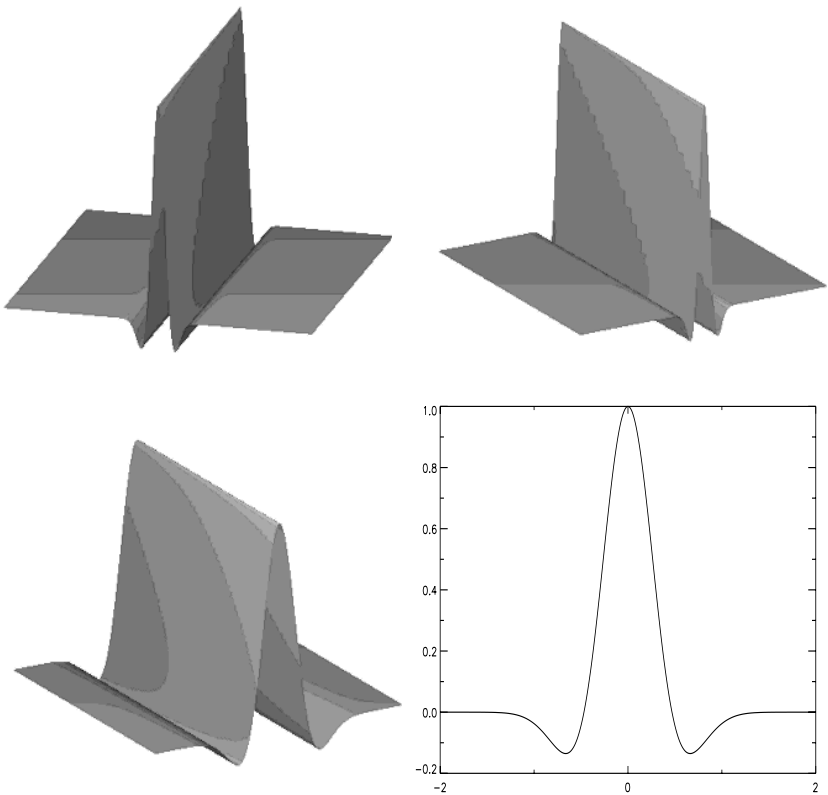


FIGURE 22.1. Example of ridgelet function.

rescaling. This function is constant along lines $x_1 \cos\theta + x_2 \sin\theta = const.$ Transverse to these ridges it is a wavelet (see figure 22.5 bottom right).

Given an integrable bivariate function $f(x)$, we define its ridgelet coefficients by

$$\mathcal{R}_f(a,b,\theta) = \int \overline{\psi}_{a,b,\theta}(x) f(x) dx.$$

We have the exact reconstruction formula

$$f(x) = \int_0^{2\pi} \int_{-\infty}^{\infty} \int_0^{\infty} \mathcal{R}_f(a,b,\theta) \psi_{a,b,\theta}(x) \frac{da}{a^3} db \frac{d\theta}{4\pi} \tag{22.13}$$

valid a.e. for functions which are both integrable and square integrable.

It has been shown [CD99] that the ridgelet transform is precisely the application of a 1-dimensional wavelet transform to the slices of the Radon transform where the angular variable θ is constant and t is varying. More details on the implementation of the digital ridgelet transform can be found in [SCD01].

22.6 Combined Transform

If 2D or 3D astronomical data set may contain anisotropic features, they certainly will also contains isotropic ones. Hence, a perfect multiscale decomposition should benefit of both the à trous algorithm advantages and that of the ridgelet transform as well. More generally, we can imagine that we have $\mathcal{T}_1, ..., \mathcal{T}_{N_t}$ transform operators, each one being optimal to detect one kind of structure. A solution α is obtained by minimizing a functional of the form:

$$J(\alpha) = \| s - \sum_{k=0}^{N_t} \mathcal{T}_k^{-1} \alpha_k \|_2^2 + \lambda \sum_k \| \alpha_k \|_0 \tag{22.14}$$

where s is the original signal, and α_k are the coefficient obtained with the transform $\mathcal{T}_k$.

An algorithm to perform such a minimization has been presented in [Sta01]. It consists in hard thresholding the residual successively on the different bases.

1. Initialize $L_{\max}$ and the number of iterations Ni. For noise filtering, estimate the noise standard deviation σ, and set $L_{\min} = k$. Otherwise, set $\sigma = 1$ and $L_{\min} = 0$.
2. Set $\delta_\lambda = \frac{1}{N_i}(L_{\max} - L_{\min})$, $\lambda = L_{\max}$, and all coefficients α_k to 0.
3. While $\lambda \geq L_{\min}$ do
4. for k = 1, .., N_t do
   - Calculate the residual $R = s - \sum_k \mathcal{T}_k^{-1} \alpha_k$.
   - Calculate the transform $\mathcal{T}_k$ of the residual: $r_k = \mathcal{T}_k R$.
   - For all coefficients $r_{k,i}$ do
     - Update the coefficients: if $\alpha_{k,i} \neq 0$ or $\mid r_{k,i} \mid > \lambda\sigma$ then $\alpha_{k,i} = \alpha_{k,i} + r_{k,i}$.
5. $\lambda = \lambda - \delta_\lambda$, and goto 6.

For an exact representation of the data, k must be set to 0. Choosing $k > 0$ introduces a filtering. If a single transform is used, it corresponds to the standard $k\sigma$ hard thresholding.

22.6.1 Example 1: Simulation

Figure 22.2 illustrates the result in the case where the input image contains only lines and Gaussians. In this experiment, we have initialized $L_{\max}$ to 20, and δ to 2 (10 iterations). Two transform operators were used, the à trous wavelet transform and the ridgelet transform. The first is well adapted to the detection of Gaussian due to the isotropy of the wavelet function [SMB98], while the second is optimal to represent lines [CD99].

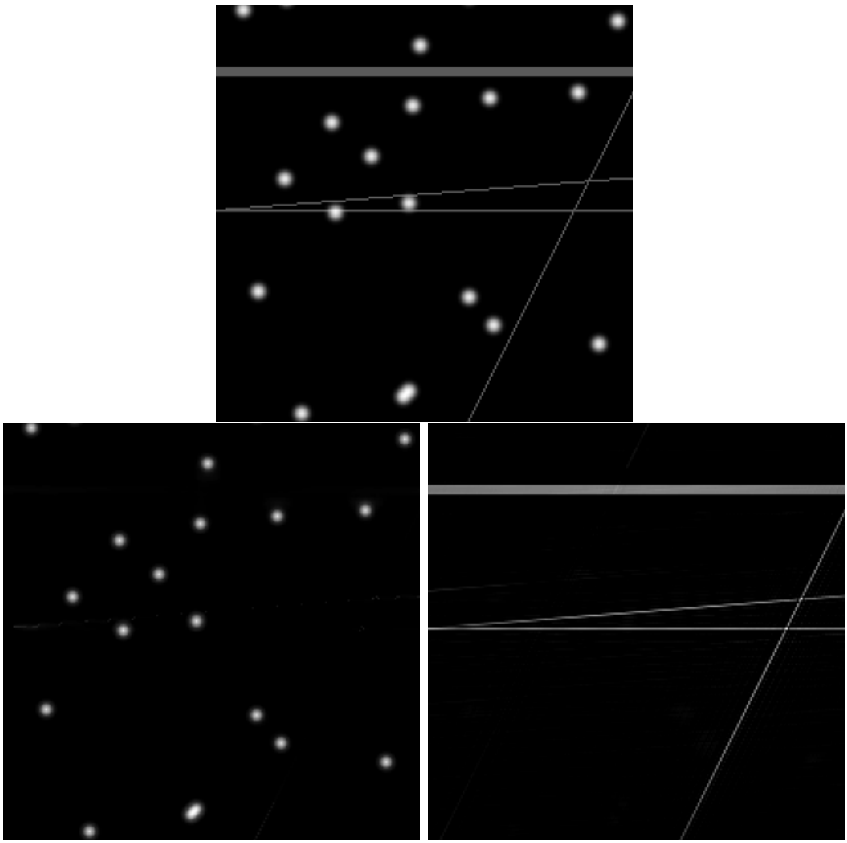


FIGURE 22.2. Top, original image containing lines and gaussians. Botton left, reconstructed image for the à trous wavelet coefficient, bottom right, reconstructed image from the ridgelet coefficients.

Figure 22.2 top, bottom left, and bottom right represents respectively the original image, the reconstructed image from the à trous wavelet coefficient, and the reconstructed image from the ridgelet coefficient. The addition of both reconstructed images reproduces the original one.

22.6.2 Example 2: Elongated - point like object

Figure 22.3 shows the result of a decomposition of a spiral galaxy (NGC2997). This image (figure 22.3 top left) contains many compact structures (stars and HII region), more or less isotropic, and large scale elongated features (NGC2997 spiral part). Compact objects are well represented by isotropic wavelets, and the elongated features are better represented by a ridgelet basis. In order to benefit of the optimal data representation of both transforms, the image has been decomposed on both the à trous wavelet

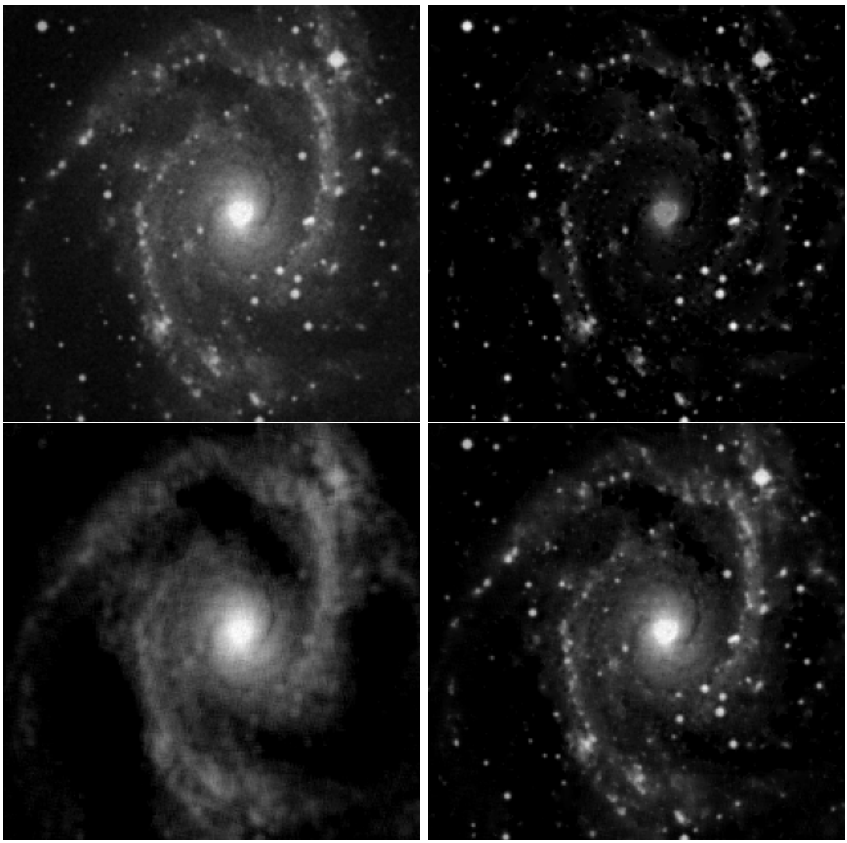


FIGURE 22.3. Top left, galaxy NGC2997, top right reconstructed image from the à trous wavelet coefficients, bottom left, reconstruction from the ridgelet coefficients, and bottom right addition of both reconstructed images.

transform and on the ridgelet transform by using the combined transform method. When the functional is minimized, we get two images, and their coaddition is the filtered version of the original image. The reconstructions from the à trous coefficient, and from the ridgelet the ridgelet coefficient can be seen in figure 22.3 top right and bottom left. The addition of both images is presented in figure 22.3 bottom right.

Acknowledgments

We wish to thank David Donoho and Emmanuel Candès for useful discussions and comments.

22.7 References

[BKK94] Tj.R. Bontekoe, E. Koper, and D.J.M. Kester. Pyramid maximum entropy images of IRAS survey data. *Astronomy and Astrophysics.*, 294:1037–1053, 1994.

[BR95] A. Bijaoui and F. Rué. A multiscale vision model adapted to astronomical images. *Signal Processing*, 46:229–243, 1995.

[Bur78] J.P. Burg. Annual Meeting International Society Exploratory Geophysics, Reprinted in Modern Spectral Analysis, D.G. Childers, ed., IEEE Press, New York, 34–41, 1978.

[CD95] R.R. Coifman and D.L. Donoho. Translation invariant denoising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, pages 125–150, New York, 1995. Springer-Verlag.

[CD99] E.J. Candès and D. Donoho. Ridgelets: the key to high dimensional intermittency? *Phil. trans; R. Soc. Lond. A*, 357:2495–2509, 1999.

[Dau92] I. Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1992.

[DD00] D.L. Donoho and M.R. Duncan. Digital curvelet transform: strategy, implementation and experiments. In H.H. Szu, M. Vetterli, W. Campbell, and J.R. Buss, editors, *Proc. Aerosense 2000, Wavelet Applications VII*, volume 4056, pages 12–29, Bellingham Washington, 2000. SPIE.

[DJ93] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. Technical Report 400, Stanford University, 1993.

[Fri78] B.R. Frieden. *Image Enhancement and Restoration.* Springer-Verlag, Berlin, 1978.

[GS91] S.F. Gull and J. Skilling. *MEMSYS5 Quantified Maximum Entropy User's Manual*, 1991.

[Mal89] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[MF98] S. Mallat and F. Falzon. Analysis of low bit rate image transform coding. *IEEE Transactions on Signal Processing*, 46(4):1027–42, 1998.

[NN86] R. Narayan and R. Nityananda. Maximum entropy image restoration in astronomy. *Ann. Rev. Astron. Astrophys.*, 24:127–170, 1986.

[PS96] E. Pantin and J.L. Starck. Deconvolution of astronomical images using the multiscale maximum entropy method. *Astronomy and Astrophysics, Suppl. Ser.*, 315:575–585, 1996.

[SBLP94] J.L. Starck, A. Bijaoui, B. Lopez, and C. Perrier. Image reconstruction by the wavelet transform applied to aperture synthesis. *Astronomy and Astrophysics*, 283:349–360, 1994.

[SBVM00] J.L. Starck, A. Bijaoui, I. Vatchanov, and F. Murtagh. A combined approach for object detection and deconvolution. *Astronomy and Astrophysics, Suppl. Ser.*, 147:139–149, 2000.

[SCD01] J.L. Starck, E. Candès, and D.L. Donoho. The curvelet transform for image denoising. *IEEE Transactions on Image Processing*, 2001. to appear.

[SM98] J.L. Starck and F. Murtagh. Automatic noise estimation from the multiresolution support. *Publications of the Astronomical Society of the Pacific*, 110(744):193–199, 1998.

[SM99] J.L. Starck and F. Murtagh. Multiscale entropy filtering. *Signal Processing*, 76(2):147–165, 1999.

[SMB98] J.L. Starck, F. Murtagh, and A. Bijaoui. *Image Processing and Data Analysis: The Multiscale Approach*. Cambridge University Press, Cambridge (GB), 1998.

[SMG98] J.L. Starck, F. Murtagh, and R. Gastaud. A new entropy measure based on the wavelet transform and noise modeling. S*pecial Issue on Multirate Systems, Filter Banks, Wavelets, and Applications of IEEE Transactions on CAS II*, 45(8), 1998.

[SMQB01] J.L. Starck, F. Murtagh, P. Querre, and F. Bonnarel. Entropy and astronomical data analysis: Perspectives from multiresolution analysis. *Astronomy and Astrophysics*, 368:730–746, 2001.

[Sta01] J.-L. Starck. Nonlinear multiscale transforms. In T. Barth, T. Chan, and R. Haimes, editors, *Advanced Multiscale and Multiresolution Methods*. Springer-Verlag, 2001.

23

Threshold Selection in Transform Shrinkage

Iain Johnstone[1]

ABSTRACT
The transform shrinkage paradigm is reviewed, of which wavelet denoising is a key example, with a focus on the blockwise approach to processing of transform coefficients. Thresholding approaches are surveyed, with special emphasis is placed on an empirical Bayes approach, which promises to adapt well to the demands of both 'dense' and 'sparse' signals. Since the author has no significant experience with problems in astronomy, discussion and examples (denoising of signals, images and deconvolution) are alas generic.
This paper is followed by a commentary by astronomer Jean-Luc Starck.

23.1 The transform shrinkage paradigm

A familiar strategy in data analysis is to (a) transform the data, via Fourier, wavelet, or some other transform, (b) process the transform coefficients in some way (compression, denoising,...) and finally (c) back transform the processed coefficients to the original domain.

Although much of what we have to say will apply quite generally to transform coefficient processing, for definiteness we begin with a one dimensional signal processing setting, with data

$$y_i = f(t_i) + z_i, \qquad i = 1, \ldots, n, \tag{23.1}$$

observed at $N = 2^J$ equally spaced time points t_i in the presence of additive noise $\{z_i\}$. The goal is to estimate, or reconstruct, f from the data y. The process may then be represented diagrammatically as follows:

$$\begin{array}{ccc} (y_i) & \xrightarrow{W} & ((d_{jk})) \\ & & \downarrow \eta \\ \big(\hat{f}(t_i)\big) & \xleftarrow{W^{-1}} & ((\hat{d}_{jk})) \end{array} \tag{23.2}$$

[1]Department of Statistics, Stanford University

Figure 23.1 illustrates this strategy on an NMR signal - we emphasize here that the values of the (hard) thresholds used are estimated separately at each level j from the data at that level.

23.1.1 General remarks

The thresholding estimate in Figure 1 takes $O(n)$ operations to compute, and is based on simple coordinatewise operations in the transform domain. Yet it demonstrates the possibility of "denoising without smoothing" in that the reconstruction is close to noise-free, without any concomitant broadening of peaks. In statistical terms, one might say that the estimate is automatically spatially adaptive – with more averaging done in regions of low signal variability. Processing methods that are linear in the data, such as Wiener filtering, can not have this property.

A heuristic explanation for the success of the method runs as follows: the wavelet transform produces a representation of the signal that is *sparse* (few large signal coefficients). On the other hand, the (orthogonal) wavelet transform carries white noise into white noise. Consequently, thresholding is a good strategy for harvesting the few large signal coefficients that emerge from the noise.

23.1.2 The transform step

W might be a discrete orthogonal wavelet transform, in which case there are exactly N output coefficients. [For an expanded discussion of wavelet ideas, along with references to the primary literature, we refer throughout to now standard references such as [Mal99] and [Dau92].]

The cascade algorithm (e.g. [Mal99, Ch. 7.3]) begins with $c_J = y$, and at successive steps $j = J-1, ..., L$ applies fixed filters H (high-pass) and G (low-pass) followed by downsampling D, yielding "decimated" vectors

$$c_j = DGc_{j+1}, \qquad d_j = DHc_{j+1},$$

each of length 2^j. The process is stopped at some coarse scale $L \geq 0$, yielding vectors of wavelet coefficients $d_{J-1}, d_{J-2}, ..., d_L$ supplemented by a coarse vector of scaling coefficients c_L. There are $2^{J-1} + 2^{J-2} + \cdots + 2^L + 2^L = 2^J = N$ coefficients in all, so the transform is one-to-one, and in fact the algorithm runs in $O(N)$ time.

If the filter coefficients in H and G are carefully chosen (see e.g. [Mal99, Chapter 7]), the transform is orthogonal, so that the inverse W^{-1} is just the same as the transpose W^t. A further advantage of the orthogonal transform is that it preserves white noise: if the noise z in the time domain is uncorrelated and of constant variance, then the same will be true of the transform coefficients d (both within and across scales).

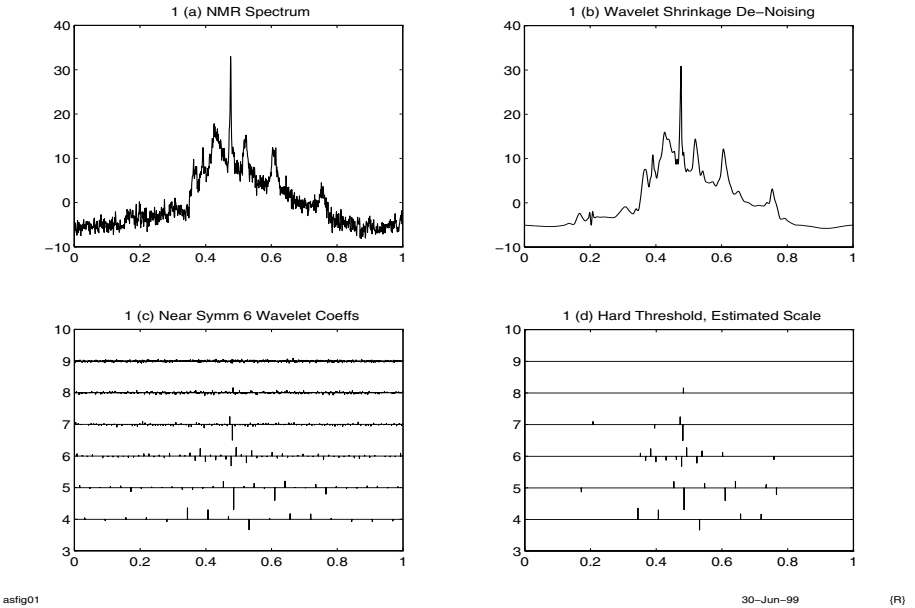


FIGURE 23.1. (a) A noisy NMR signal, $N = 1024$ (from A. Maudsley via C. Raphael) . (c) Discrete orthogonal wavelet transform coefficients, displayed by resolution level $(d_9, d_8, ..., d_4)$ and location within level. Note that the fraction of large coefficients within level decreases as scale becomes finer. (d) result of applying level-dependent hard thresholding (at threshold $\sigma\sqrt{2 \log N}$) to individual coefficients, (b) reconstruction by inverse wavelet transform.

Redundant vs. Non-redundant transform A significant drawback of the orthogonal or non-redundant transform is that it is not translation-invariant: the dyadic decimation process means that different results will be produced if the original signal is shifted, for example by an odd number of time points. A translation-invariant, or stationary, transform can be obtained by not decimating: this *á trous* algorithm (e.g. [Mal99, Sec. 5.5.2]) produces N coefficients at each scale, for a total $N \log_2 N$. The transform is then "redundant", and has slightly greater algorithmic complexity $O(N \log N)$. In addition the transform coefficients, being oversampled, are now correlated. However, the increased quality of the (now shift-invariant) reconstructions is such that this transform is almost always to be preferred in practice. While the theoretical properties of the orthogonal transform coefficients are easier to understand and analyze, it has frequently been found that insights – such as prescriptions for threshold choices – derived from analysis of the orthogonal transform yield even better results when used in conjunction with the stationary transform.

Choice of wavelet filters The orthogonality and multiscale properties of wavelets constrain the choice of the filters H and G severely, as does the desire that the filters have finite length (leading to wavelets of compact support.) Nevertheless, there are still some degrees of freedom, leading to the existence of several families of frequently used filters in addition to the family initially constructed by Daubechies. Among the factors

(a) *support length:* longer filters lead to smoother wavelets, but have more coefficients,

(b) *symmetry:* a real orthonormal wavelet of compact support cannot be exactly symmetric, but the Symmlet family comes close,

(c) *number of vanishing moments:* more lead to better approximation properties for smooth signals.

Space precludes detailed discussion: see [Dau92] or [Mal99, Chapter 7].

23.1.3 Processing Step

A major goal of the transform paradigm is that relatively simple processing should suffice in the transform coefficient domain. In Figure 23.1, for example, hard thresholding $\hat{d}_{jk} = d_{jk} I\{|d_{jk}| \geq \hat{t}_j\}$ is applied to each coefficient, with the threshold $\hat{t}_j$ estimated from the coeffficients d_j at level j. More generally, a block processing strategy might be represented simply

$$\eta = (\eta_j), \qquad \eta_j : d_j \to \hat{d}_j.$$

[Note that in the wavelet setting, these blocks need not necessarily correspond exactly to levels of the discrete wavelet transform - they might be composed of spatially related groups of coefficients within a single level, for instance.]

The focus of this paper will be on simple methods of processing by blocks, so we pause to address some of the attendant pros and cons. In general, the hope is that within a block the data has a degree of homogeneity. For example, within blocks the noise structure may be close to uncorrelated (or even independent, with Gaussian data) with common variance. At the same time, within blocks, some sort of exchangeability assumption about the signal components may be more nearly justified.

A major advantage of treating block processing as a modular subproblem is that solutions can be used in a broad class of transform problems beyond the initial setting of wavelet shrinkage (see the next subsection). A corresponding disadvantage is that we ignore cross block dependence (such as occurs between wavelet coefficients at different scales near the same location). A number of authors have addressed this issue in the wavelet shrinkage setting (e.g. [CNB98, PS00]).

We briefly mention two further advantages of block processing. It will often happen that the noise variance varies between blocks. However, it is often fairly straightforward to estimate the noise variance within a block if it is believed that less than half of the coefficients contain any signal, using the resistant estimate

$$\hat{\sigma}_j^2 = MAD\{d_j\}/0.6745, \tag{23.3}$$

where MAD stands for "median absolute deviation", and the factor 0.6745 calibrates for standard Gaussian noise.

In the noisy signal setting (23.1), if the noise is stationary and correlated, it often happens that the effect of the wavelet transform is decorrelating, so that within scales, the coefficients $k \to d_{jk}$ are nearly uncorrelated, with level dependent variances σ_j^2.

Example Figure 23.2 shows an extract of 2048 data points from a sample generated by physiologist Rick Eisenberg to represent relevant challenges in processing data measuring the picoamp ion currents that flow in single membrane channels. (More details are in [JS97]). The data consists of a step function switching between values 0 ("off") and 1 ("on") at random in the presence of additive, non-white noise of a form known to be representative of laboratory data. Although the signal to noise ratio is low, level-dependent variances of the wavelet coefficients may be estimated using (23.3).

Translation-invariant denoising with the Haar wavelet and hard thresholds $\hat{\sigma}_j\sqrt{2\log n}$ leads to oversmoothing. The use of smaller, appropriately chosen data dependent thresholds leads to a more satisfactory fit (for example, after thresholding the reconstruction to the known values 0 or 1 in the time domain.) How this is done is the subject of Section 23.2.

23.1.4 Block structure in many transform problems

For one-dimensional signals (23.1), we have so far discussed using blocks formed from the 2^j coefficients at each scale of the discrete wavelet trans-

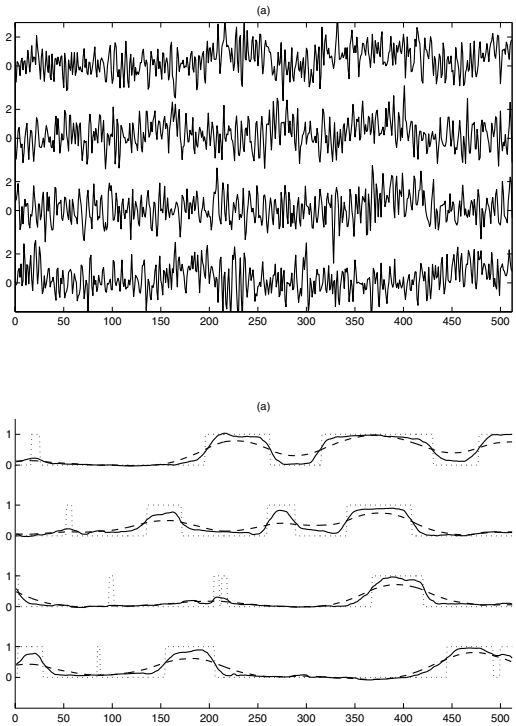


FIGURE 23.2. Ion channel data. Panel (a) sample trace of length 2048. Panel (b) Dotted line: true signal, Dashed line: reconstruction using (translation invariant) thresholding at $\hat{\sigma}_j\sqrt{2\log n}$. Solid line: reconstruction using TI thresholding at data determined thresholds (a combination of SURE and universal).Further details in [JS97].

form. One may wish to use smaller blocks of spatially contiguous coefficients with each level, say of size of order $\log n$ - see for example [CS99, HKP99]. Using the Fourier transform instead, one might form blocks of contiguous frequencies, see [EP84, CT01] where blocks of geometrically, polynomially or logarithmically growing size have been considered.

For images, when using separable wavelet bases (see e.g. [Mal99, Chapter 7.7]), it is natural to treat the horizontal, vertical and diagonal channels as separate blocks within each scale j. Candès and Donoho ([CD99b, CD99a]) have argued that other multiresolution systems based on ridgelets are better adapted to the representation of images: there are natural blocks of coefficients within these representations to which the thresholding methods described here may be applied. Finally, there are other orthonormal systems, such as brushlets ([MC97]) and members of wavelet or cosine packet libraries, within which blocking can be used.

We may consider in addition certain *indirect data* settings, $y = Kf + z$

in which a (linear) operator K acts on the signal or image of interest before (noisy) observations y are taken. Typical examples for K include (fractional) integration, or Radon transformation, or convolution(blurring).

In certain cases, there exists an exact or near diagonalization in some transform domain

$$y_{Jk} = \alpha_J \theta_{Jk} + z_{Jk}, \qquad k \in B_J, J \in \mathcal{J} \tag{23.4}$$

Traditional examples include (a) the Fourier transform, which diagonalizes deconvolution problems and (b) the singular value decomposition (SVD). Exploiting multiresolution ideas, one might have a *wavelet-vaguelette decomposition* (WVD) [Don95], with multiresolution representing systems $\{v_{Jk}\}, \{w_{Jk}\}$ for the domain and range of K respectively, such that $Kv_{Jk} = \alpha_J w_{Jk}$, so that equation (23.4) is the coefficient level representation. Finally, for certain deconvolution problems involving "hyperbolic noise", Kalifa and Mallat ([KM99]) have argued that a "mirror wavelet" basis (actually a particular basis from the wavelet packet library) leads to better reconstructions. In all of these systems, there are obvious ways of blocking coefficients in order to apply thresholding methods discussed here.

23.1.5 Choice of basis/transform

An appropriate choice of basis or representing system is often crucial to the success of the associated transform shrinkage method. In particular, it is desirable that the signal coefficients be "sparse" in the transform domain, in the sense that most of the energy in the signal is concentrated in a few components. If it is also the case that the noise is relatively white within blocks, a (block-specific) thresholding strategy can hope to remove the bulk of the noise while leaving intact most of the signal.

For example, smooth signals will often have a sparse representation in the Fourier basis, since most energy is concentrated in low frequencies. Oscillatory signals (such as speech) on the other hand may be better represented in wavelet or cosine packet bases. Wavelet bases are particularly suited to representing functions with point discontinuities or other singularities, while ridglet and curvelet systems are designed for singularities in images that occur across straight (or slowly curving) lines.

23.2 The single sequence problem

In this key section, we focus on the apparently special problem of estimating $\mu = (\mu_i)$ from observations $x_i = \mu_i + z_i$, $i = 1, ..., n$ where the noise variates z_i are assumed to be i.i.d. $N(0, 1)$. This is a natural model for the coefficients within a block of the sort just discussed. It is assumed that the variance is known, and by rescaling, taken as equal to one.

We concentrate on *co-ordinatewise* thresholding strategies, in which the estimate of the i-th co-ordinate of the signal $\hat{\mu}_i(x) = \eta(x_i, \hat{t})$ depends only on x_i, the i-th component of the data, and a threshold $\hat{t}$, where the hat indicates that the choice of t may depend on the full data (x_i). The classical (and extreme) examples of thresholding are:

$$\begin{array}{ll} \textit{Hard}: & \eta(x_i, t) = x_i I\{|x_i| > t\} \\ \textit{Soft}: & \eta(x_i, t) = \text{sign}(x)(|x_i| - t)_+ \end{array}$$

These may be regarded as special cases of a more general class of *threshold shrinkage rules*, which are defined by the properties

$$\begin{array}{ll} \text{odd:} & \eta(-x, t) = -\eta(x, t), \\ \text{shrinks:} & \eta(x, t) \le x \ \text{ if } \ x \ge 0, \\ \text{bounded:} & x - \eta(x, t) \le t + b \ \text{ if } \ x \ge 0, \text{ (some } b < \infty), \\ \text{threshold:} & \eta(x, t) = 0 \text{ iff } |x| \le t. \end{array}$$

Two examples (among many) of threshold shrinkage rules are provided by a) $\eta(x, t) = (1 - t^2/x^2)_+ x$ which arises in the study of the non-negative garrote [Bre95], and b) the posterior median, to be discussed further below.

The choice of the threshold shrinkage rule η and the selection of threshold t are somewhat separate issues. The choice of η is problem dependent. For example, hard thresholding exactly preserves the data values above the threshold, and as such can be good for preserving peak heights (say in spectrum estimation), whereas soft thresholding forces a substantial shrinkage. The latter leads to smoother visual appearance of reconstructions, but this property is often at odds with that of good fidelity – as measured for example by average squared error between estimate and truth. In the remainder of this paper, we will focus mainly on the question of threshold selection once the non-linearity class η has been chosen.

Remark: In the statistics literature, there has been considerable study of James-Stein shrinkage and its variants. In simplest form, this is given by

$$\hat{\mu}_i(x) = (1 - \hat{s})_+ x_i \qquad \hat{s} = \frac{(n-2)\sigma^2}{\sum x_i^2}.$$

While this estimator does threshold the entire signal to zero if the total energy is small enough, $\sum x_i^2 < (n-2)\sigma^2$, it otherwise applies a common, data-determined *linear* shrinkage to all co-ordinates. When the true signal is sparse, this is less effective than thresholding, because either the shrinkage factor either causes substantial error in the large components, or fails to shrink the noise elements - it cannot avoid both problems simultaneously.

23.2.1 Threshold choice

Often, one may know from previous experience or subjective belief that a particular choice of threshold (say 3σ or 5σ) is appropriate. On the other

hand, one may seek an *automatic* method for setting a threshold, and this will be the focus of subsequent discussion.

The simplest automatic methods set a fixed threshold in advance of observing data. One may use a fixed number of standard deviations $k\sigma$, or a more conservative limit, such as the *universal* threshold $t = \sigma\sqrt{2\log n}$. This choice is motivated by the observation that in pure noise $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} N(0,1)$

$$P\{\max_{1\le i\le n} |Z_i| > \sqrt{2\log n}\} \to 0.$$

When combined with the soft thresholding non-linearity, the universal threshold leads to visually smooth reconstructions, but at the cost of considerable bias and relatively high mean squared error (cf. [DJKP95]).

As seen already in Figure 23.2, it is often desirable to choose thresholds from the data, in such a way that blocks, or levels, with sparse signal lead to high thresholds, and blocks with relatively "dense" signal lead to lower choices.

This can be made more explicit with an example. Consider two signals: $\mu^{(1)}$ is relatively "dense": 18% of its components equal 3σ with the remainder zero, while $\mu^{(2)}$ is "sparse": only 0.2% of its components equal 3σ. For a fixed configuration μ, the mean squared error of hard thresholding $MSE(t,\mu) = n^{-1}\sum_1^n E_\mu[\eta(x_i,t) - \mu_i]^2$ depends on t, and we can evaluate the optimizing threshold t^*. For the dense signal $\mu^{(1)}$ the optimal threshold is $t_1^* = 2\sigma$, while for the sparse signal it is higher, namely $t_2^* = 4\sigma$. Furthermore, there is a significant penalty to using the wrong threshold: using the 4σ threshold on the dense signal increases the MSE by a factor of 2.46 over using the optimal 2σ. For the sparse signal, using 2σ instead of 4σ increases MSE by a factor of 15!

Finding a numerically simple and stable method satisfying these desiderata has proven to be elusive. A plethora of methods for choosing thresholds has been proposed (see for example [Vid99, Chapter 6]). The empirical Bayes approach sketched below appears promising, having both empirical and theoretical support. As background, we present two other methods which have been accompanied by some theoretical analysis of their properties.

a) *SURE* In principle, it would seem attractive to choose t to minimize the mean squared error of reconstruction. Since this depends on the signal and so is unknown, one can try instead to use instead Stein's Unbiased Risk Estimate (SURE) for the mean squared error of soft thresholding. Thus, we choose $\hat{t}_{SURE}$ as the minimizer (within the range $[0, \sqrt{2\log n}]$) of

$$\hat{U}(t) = n + \sum_1^n x_k^2 \wedge t^2 - 2\sum_1^n I\{x_k^2 \le t^2\}.$$

This does indeed have some good theoretical properties [DJ95], but the same theoretical analysis, combined with simulation and practical experi-

ence, shows that the method can be unstable (see [DJ95, JS]) and that it does not choose thresholds well in sparse cases.

b) *FDR* This method is derived from the principle of controlling the False Discovery Rate in simultaneous hypothesis testing [BH95] and has been studied in detail in the estimation setting [ABDJ99]. Order the data by decreasing magnitudes: $|x|_{(1)} \geq |x|_{(2)} \geq \ldots \geq |x|_{(n)}$., and compare to a *quantile boundary*: $t_k = \sigma z(q/2 \cdot k/n)$, where the false discovery rate parameter $q \in (0, 1/2]$. Define a crossing index $\hat{k}_F = \max\{k : |x|_{(k)} \geq t_k\}$, and use this to set the threshold $\hat{t}_F = t_{\hat{k}_F}$. Although FDR threshold selection adapts very well to sparse signals [ABDJ99], it does less well on dense signals of moderate size.

23.2.2 Empirical Bayes thresholding

We now describe, in a little more detail, an approach which has some of the good properties of both SURE and FDR thresholding and transitions between the two in a stable manner.

We adopt a Bayesian formulation, in which the components (μ_i) are drawn i.i.d. from a prior distribution. The notion that the signal might be sparse is captured by requiring that the prior distribution have a mixture form

$$f_{\text{prior}}(\mu) = (1 - w)\delta_0(\mu) + wa\gamma(a\mu). \tag{23.5}$$

Thus, it is assumed that with probability $1 - w$, there is no signal: $\mu_i = 0$, while with probability w, the value of μ_i is obtained by a draw from the density $\gamma(\mu)$, with scale parameter a. In principle, the density γ could be quite general, but for purposes of implementation, we have found advantages in using a heavy-tailed density for γ, for example the Laplace density

$$\gamma(u) = \tfrac{1}{2}\exp(-|u|) \tag{23.6}$$

or the mixture density given by

$$\mu|\Theta = \theta \sim N(0, \theta^{-1} - 1) \text{ with } \Theta \sim \text{Beta}(\alpha, 1). \tag{23.7}$$

The latter density for μ has tails that decay as $\mu^{-2\alpha-1}$, so that, in particular, if $\alpha = \frac{1}{2}$ then the tails will have the same weight as those of the Cauchy distribution.

To get a point estimate $\hat{\mu}_i$ we might use the posterior mean (which minimizes posterior expected squared or L_2 error) or the posterior *median* (which minimizes the posterior expected absolute or L_1 error). We prefer the use of the posterior median, since it leads to a genuine threshold shrinkage rule, with threshold zone $[-t(w), t(w)]$, – the posterior mean is close to, but not exactly zero throughout this range. It turns out that the threshold $t(w)$ varies inversely with w: for w small, the threshold is large and vice versa.

Integrating out the prior distribution, the marginal density of the data $\{X_i\}$ is given by the density $(1-w)\phi(x) + wg(x)$, where $g = \gamma \star \phi$ is the convolution of γ with the Gaussian density ϕ. To estimate the threshold from data, we treat w as a parameter which can be estimated by marginal maximum likelihood (MML). Thus the MML estimate $\hat{w}$ is obtained by maximizing

$$\ell(w) = \sum_{i=1}^{n} \log\{(1-w)\phi(X_i) + wg(X_i)\} \qquad w \in [w_n, 1], \tag{23.8}$$

where $t(w_n) = \sqrt{2\log n}$. Finding the zero of the univariate function $w \to \partial l/\partial w$ is easily accomplished numerically.

In both cases (23.6) and (23.7) the posterior distribution of μ given an observed X, and the marginal distribution of X are tractable, so that the choice of w by marginal maximum likelihood, and the estimation of μ by posterior mean or median, can be performed in practice. (see [JS] for software information).

The method automatically adjusts to sparsity: if a considerable number of 'large' X_i are present then $\hat{w}$ will be large and vice versa. It is also possible to extend the method to estimate other parameters in γ such as the scale a. In practice, with prior (23.6), the fixed choice $a = 0.2$ works well.

Table 23.1 shows one summary of a simulation designed to test the performance of threshold selection methods over a range of models for sparse behavior. Further details may be found in [JS]. Twelve configurations were created giving the first K components $\mu_1 = \cdots = \mu_K = \mu_0$ a signal strength μ_0, and the remaining components $\mu_{K+1} = \ldots = \mu_n = 0$ are designated as "noise". The parameter $K \in \{5, 50, 500\}$ controls sparsity, while $\mu_0 \in \{3, 4, 5, 7\}$ controls signal strength. The average squared error of method $\hat{\mu}_{meth}$ on configuration μ_c is measured (over 100 replications of the experiment)by

$$ASE(\hat{\mu}_{meth}, \mu_c) = N^{-1} \sum_i [\hat{\mu}_{m,i} - \mu_{c,i}]^2.$$

We may compare such an ASE to the best possible ASE observed among all methods of Table 23.1 for that configuration:

$$\text{Inefficiency}(\hat{\mu}_{meth}, \mu_c) = 100 \times \Big[\frac{ASE(\hat{\mu}_{meth}, \mu_c)}{\min_{meth} ASE(\hat{\mu}_{meth}, \mu_c)} - 1\Big]$$

Small inefficiencies mean that $\hat{\mu}_{meth}$ is not much worse than the best among the methods tried for that configuration. Table 23.1 summarizes the inefficiences of various methods over the 12 configurations (3 sparsities $\times$ 4 signal strengths).

TABLE 23.1. Inefficiencies of methods showing the median, mean, maximum and tenth largest inefficiency over the 12 cases considered. Exphard refers to the EB choice of thresholds with exponential prior and use of the hard thresholding rule. More detail on the other methods is in [JS].

	median	mean	10th	max
exponential($a = 0.2$)	19	19	30	48
cauchy	20	25	42	48
postmean	25	28	40	96
exphard	37	45	62	95
SURE	35	121	151	676
adapt	104	224	303	1283
FDR q=0.01	44	56	91	210
FDR q=0.1	20	35	39	140
FDR q=0.4	74	170	214	848
universal soft	529	643	1283	1367
universal hard	50	101	159	359

It is striking how the empirical Bayes method outperforms the various SURE and FDR variants, to say nothing of the fixed "universal" threshold choices.

While the simulations lend support to the idea that empirical Bayes threshold choice adapts well to both sparse and dense signals, they are necessarily selective, and so it is reassuring that theoretical analysis points to similar conclusions. The goals for a theoretical analysis are to capture two properties. Firstly *flexible adaptation* to sparsity: specifically if the ℓ_p norm $n^{-1}\sum|\mu_i|^p$ of the signal is small, then we hope for correspondingly small estimation error. Secondly, a *robustness* property that the error of $\hat{\mu}_{EB}$ be bounded, no matter what the configuration μ.

Such properties are not *a priori* obvious for the MML estimated threshold $t(\hat{w})$, since for general configurations μ the mixture prior (23.5) is in general wrong!

To describe briefly the results established in [JS], introduce the mean ℓ_q error $R_q(\hat{\mu},\mu) = n^{-1}\sum_1^n E|\hat{\mu}_i - \mu_i|^q$ for $q \le 2$. In the sparse case, where for $p < 2$ and η small we have $n^{-1}\sum|\mu_i|^p < \eta^p$, it is shown that with high probability $\hat{t}$ is large (and of rough order $\sqrt{2\log\eta^{-p}}$), and that the mean error

$$R_q(\hat{\mu}^{EB},\mu) \le C_1\eta^{p\wedge q}(\log\eta^{-p})^{(q-p)_+/2} + C_2 n^{-1}\log^3 n. \tag{23.9}$$

The first term on the right side is (up to constants) the best possible estimation error attainable by any method subject only to the given sparsity information. The point of the inequality is that without having to know either p, or more importantly, the degree of sparsity η, the empirical Bayes method chooses a threshold with the best possible order of estimation error. By comparison, such a result is not known at this level of generality for either the FDR or SURE methods of threshold estimation.

For the "robustness" property, assume that the tails of the prior γ are exponential or heavier, as in the two examples (23.6) and (23.7). It is shown that if the number of large components: $n^{-1}\#\{|\mu_i| \geq \tau\}$ is large, then the estimated threshold $t(\hat{w})$ is 'small', and then for *all* μ and n large that

$$R_q(\hat{\mu}^{EB}, \mu) \leq C_3. \tag{23.10}$$

It should be noted that both results (23.9) and (23.10) hold true for *any* bounded shrinkage threshold rule so long as the Empirical Bayes threshold $t(\hat{w})$ is used. They also hold for the posterior mean (not a threshold rule) so long as $q > 1$ (but are false if $q < 1$!).

23.3 Consequences for wavelet thresholding

In the basic wavelet shrinkage settings of (23.2) and Figures 23.1 and 23.2, EB thresholding is applied level by level. We illustrate on the ion channel data in Figure 3. The thresholds chosen by SURE (dashed line) are reasonable at the coarse scales 6, 7 and 8, but are too small at the fine scales 9 - 11 where the signal is sparse. [The reconstruction of Figure 23.2 was obtained by manually replacing SURE by universal thresholds at these fine scales]. By contrast, the empirical Bayes threshold choices increase monotonically with scale in a reasonable manner. In particular, the universal thresholds at levels 9-11 are found automatically. Two reconstructions using the same EB thresholds are shown in panel (b): one using the posterior median shrinkage rule, and the other using hard thresholding rule. The hard threshold choice tracks the true signal better, echoing the earlier remark that choice of threshold shrinkage rule is problem dependent, and somewhat separate from the issue of setting threshold values.

Theoretical results and simulations support the conclusions seen in this example. We summarize results for model (23.1) described in detail in [JS]. Suppose, first, that mean squared error is used to assess a reconstruction $\hat{f}$: $R_N(\hat{f}, f) = N^{-1} \sum E[\hat{f}(t_i) - f(t_i)]^2$. Then, if the empirical Bayes threshold choices are used (in conjunction with *any* threshold shrinkage rule), then we might say that $\hat{f}$ attains the right (minimax) rate of convergence over *all* the right function classes. Specifically, if $f \in B^{\alpha}_{p,\infty}(C)$, a Besov function

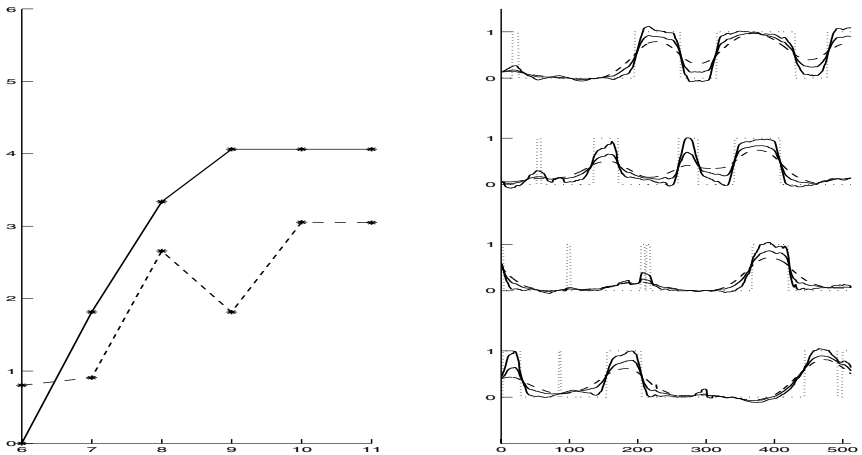


FIGURE 23.3. (a) dashed line: SURE thresholds, solid line: EB thresholds. (b) Both solid lines use EB-thresholds, but one uses a hard thresholding rule and tracks the true signal better, while the other uses posterior median shrinkage.

class of smoothness index α, with $\alpha > 1/p - 1/2$ and $0 < p < 2$, then

$$\sup_{f \in B} R_N(\hat{f}, f) \le cC^{2\alpha/(2\alpha+1)} N^{-2\alpha/(2\alpha+1)} + c \log^4 N/N. \tag{23.11}$$

In other words, the EB threshold based estimator automatically (i.e. without knowledge of the function class) adapts to obtain the best possible rate of convergence for that function class. This rate adaptivity holds over a wider range of function classes than has been established for any other *implemented* estimator. The fact that corresponding generality has not been shown for SURE based thresholds is a theoretical reflectoin of its practical shortcomings. Finally, we note that the bounded shrinkage property of the threshold rule is essential for the validity (23.11) – for example, a linear shrinkage rule (such as yielded by a Gaussian prior) would have excessive bias on some signals.

A simulation, reported in [JS], compared TI versions of E-Bayes, SURE and $\sqrt{2 \log n}$ thresholding on samples of size $N = 1024$ from model (23.1) using the four test functions of [DJ94] and two noise levels. Also included were two default non-wavelet based smoothing methods from the *SPlus* package: spline smoothing with GCV choice of regularization parameter, and Tukey's 4(3RSR)2H. In summary, empirical Bayes threshold choice leads to better MSE than the other methods (and the posterior median with the double exponential prior usually beats other variants of the EB method included in the test). Empirical Bayes also wins out in a comparison using the orthogonal (non-redudant) transform.

Turning briefly to images, Figure 23.4 shows the effect of applying empirical Bayes thresholds – computed separately in each channel within level

– to a standard image with Gaussian noise added. Nine realizations were generated, and the signal to noise ratio ($SNR = 20\log_{10}(\|\hat{f} - f\|_2/\|f\|_2)$) calculated for both thresholding at 3σ and for the EB-thresholds. The actual images shown correspond to the median of the nine examples (ordered by increase in signal to noise ratio SNR).

As shown in the table below, the EB thresholds increase monotonically as the scale becomes finer (SNR = 33.83). They are somewhat smaller in the vertical channel, as the signal is stronger there in the peppers image. Fixing the threshold at 3σ in all channels leads to small noise artifacts at fine scales (SNR = 33.74), while fixing the threshold at $\sigma\sqrt{2\log n}$ (not shown) leads to a marked increase in squared error (i.e. reduced signal-to-noise ratio). Of course, the quantitative SNR measure does not necessarily correspond to visual perception of relative quality.

Channel/Level	3	4	5	6	7
Horizontal	0	1.1	2.3	3.2	4.4
Vertical	0	0	2.0	3.0	4.4
Diagonal	0	1.7	2.7	4.1	4.4

23.4 Concluding remarks

We have focused on a class of problems which after transformation take the form already indicated in (23.4), namely

$$y_{Jk} = \alpha_J \theta_{Jk} + z_{Jk}, \qquad k \in B_J, J \in \mathcal{J}.$$

On the assumption that the data within individual blocks, $\{y_{Jk}, k \in B_J\}$ are approximately exchangeable, and *possibly* sparse, we have described various approaches to thresholding. In particular, the mixture prior (23.5) is a reasonably simple codification of this assumption. An empirical Bayes approach leads to threshold choices (and bounded shrinkage rules) that are easy to compute from data, and show a reasonable response to varying sparsity of signal across levels. Furthermore, these rules have good performance both in theory and simulations.

23.5 Acknowledgments

The work on Empirical Bayes choice of thresholds described here is joint work with Bernard Silverman. Earlier work (on SURE and FDR thresholds) was joint with Felix Abramovich, Yoav Benjamini and David Donoho. Preparation of this paper supported in part by NSF DMS 0072661 and NIH CA72028.

FIGURE 23.4. Translation invariant hard thresholding applied to a noisy version of the "peppers" image. For original image and noisy version see, e.g. [Mal99, Figure 10.6]. Panel (a) uses fixed threshold at 3σ, Panel (b): Level and channel dependent EB thresholds as shown in table.

23.6 References

[ABDJ99] Felix Abramovich, Yoav Benjamini, David Donoho, and Iain Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. Technical Report, Submitted to *Annals of Statistics* , 1999.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B.*, 57:289–300, 1995.

[Bre95] Leo Breiman. Better subset selection using the non-negative garotte. *Technometrics*, 37:373–384, 1995.

[CD99a] E. Candès and D. L. Donoho. Curvelets: A surprisingly effective nonadaptive representation of objects with edges. Technical report, Stanford University, 1999.

[CD99b] E. Candès and D. L. Donoho. Ridgelets: the key to high dimensional intermittency? *Philosophical Transactions of the Royal Society of London, Series A*, 357:2495–2509, 1999.

[CNB98] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902, 1998.

[CS99] T. Cai and B. W. Silverman. Incorporating information on neighboring coefficients into wavelet estimation. Technical report, University of Bristol, 1999.

[CT01] L. Cavalier and A. B. Tsybakov. Penalized blockwise stein's method, monotone oracles and sharp adaptive estimation. Technical report, Université de Paris VI, 2001.

[Dau92] I. Daubechies. *Ten Lectures on Wavelets.* Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia, 1992.

[DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[DJ95] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.

[DJKP95] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, 57:301–369, 1995. With Discussion.

[Don95] D.L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied Computational and Harmonic Analysis*, 2:101–126, 1995.

[EP84] S.Yu. Efroimovich and M.S. Pinsker. A learning algorithm for nonparametric filtering. *Automat. i Telemeh.*, 11:58–65, 1984. (in Russian), translated in *Automation and Remote Control*, 1985, p 1434-1440.

[HKP99] P. G. Hall, G. Kerkyacharian, and D. Picard. On block thresholding rules for curve estimation using kernel and wavelet methods. *Annals of Statistics*, 26:922–942, 1999.

[JS] I. M. Johnstone and B. W. Silverman. Empirical bayes estimates of sparse sequences, with applications to transform shrinkage. manuscript in preparation.

[JS97] I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B.*, 59:319–351, 1997.

[KM99] J. Kalifa and S. Mallat. Minimax deconvolution in mirror wavelet bases. Technical report, Ecole Polytechnique, Palaiseau, 1999.

[Mal99] Stéphane Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1999. 2nd, expanded, edition.

[MC97] Francois G. Meyer and Ronald R. Coifman. Brushlets: A tool for directional image analysis and image compression. *Applied Computational and Harmonic Analysis*, 4:147–187, 1997.

[PS00] Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–71, 2000.

[Vid99] Brani Vidakovic. *Statistical Modelling by Wavelets.* John Wiley and Sons, 1999.

Commentary by Jean-Luc Starck[2]

Iain Johnstone has provided, from a statistician point of view, a very interesting talk about data filtering using wavelets, including some ideas

[2]Centre d'Études Atomique, Paris

from the recent statistical literature. In the following, I will quickly describe a few points which I believe to be important to astronomers and I will provide some examples which illustrate different restoration strategies on astronomical data.

Wavelet Filtering of Astronomical Images

In order to remove the noise contained in the data using wavelets, we need to answer the three following questions:

1. Which wavelet function is the best for astronomical data ?

2. Is it really necessary to use an undecimated wavelet transform instead of a decimated one which is faster and needs less memory ?

3. Which thresholding method should be used ?

And the goal of the filtering is to detect the faintest objects with the minimum of false detection, and to estimate accurately the photometry (i.e., integrated intensity) of the detected objects. The residual image (i.e. noisy image minus filtered one) gives us a good idea how well the photometry is preserved. Indeed, if the sources can still be distinguished by eye in the residual image, it means that a part of their flux has been lost during the filtering process.

Fig. 23.5 shows a simulation. Bottom, the simulated objects (15 sources) and the noisy data. Middle left and right shows the haar filtered image and its residual. Bottom left and right shows the undecimated haar filtered image and its residual. We see in this example the importance of the wavelet function. In the Haar filtered image, all features look like square, which is the shape of the Haar wavelet function. Using an undecimated Haar transform, these artifact have partially disappeared. However, in both cases, the residual is not very clean, and the faintest source is not detected.

Fig. 23.6 top shows the undecimated WT filtering using 7/9 filters (Antonini et al, 1992). The quality of the filtered image is much better than using the Haar filters. The residual is still not perfect. Fig. 23.6 bottom shows the restoration using the à trous algorithm and an iterative method (Starck et al, 1998). We can see that the faintest source has been detected, and the residual is much better. This is due to the fact that the à trous WT is an isotropic transform, while the undecimated WT has three privileged directions. Therefore the à trous is better adapted to detect gaussians. Iterating allows us also to better clean the residual.

To answer the two first questions, we could say that a non decimated transform should always be preferred to a decimated one, except in cases where we have strong computation time constraints. The à trous wavelet wavelet transform is very well suited to most astronomical images, which

contains more or less gaussians of different sizes, or diffuse structures without edges. For planetary images, or images with very anisotropic features, an undecimated wavelet transform with the 7/9 filters should be better.

Thresholding methods

Many thresholding methods have been presented in this paper. For astronomers, the important point is to know what is the probability that a feature in the restored data is true. Therefore, for a given wavelet coefficient $w_{j,k}$, we need to know the probability that the noise produces a coefficient of the same amplitude:

$$\begin{cases} \text{Prob}(W > w_{j,k}) & \text{if} w_{j,k} > 0 \\ \text{Prob}(W < w_{j,k}) & \text{if} w_{j,k} < 0 \end{cases} \tag{23.12}$$

Depending on the noise modeling, a detection threshold T can be derived, corresponding to a given confidence interval ϵ. For example, in case of Gaussian noise, a 3σ detection level corresponds to a probability of false detection of 0.27%. All coefficients with an absolute value lower than T are set to zero (hard thresholding).

Our experiments have shown that for astronomical data, this simple approach is better than the universal thresholding, the soft thresholding, the SURE method, the Wiener method and the hidden Markov field. Furthermore, it can easily be generalized to other kind of noise, which is not the case of the others. The noise in astronomical data is often not Gaussian. For CCD images, it is a mixture of Gaussian and Poisson noise, for X-ray image, it is a Poisson noise, and very often we have the error (or the noise standard deviation) for each pixel. This error map must be taken into account for a correct restoration.

Two other thresholding methods, FDR and Adaptive thresholding, have also been presented in this paper. They seem very attractive, but have not yet been tested for astronomical data. This should be investigated in the future.

References

- J.L. Starck, F. Murtagh, and A. Bijaoui. *Image Processing and Data Analysis: The Multiscale Approach*. Cambridge University Press, Cambridge (GB), 1998.
- M. Antonini, M. Barlaud,P. Mathieu and I. Daubechies, Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1(2), 1992.

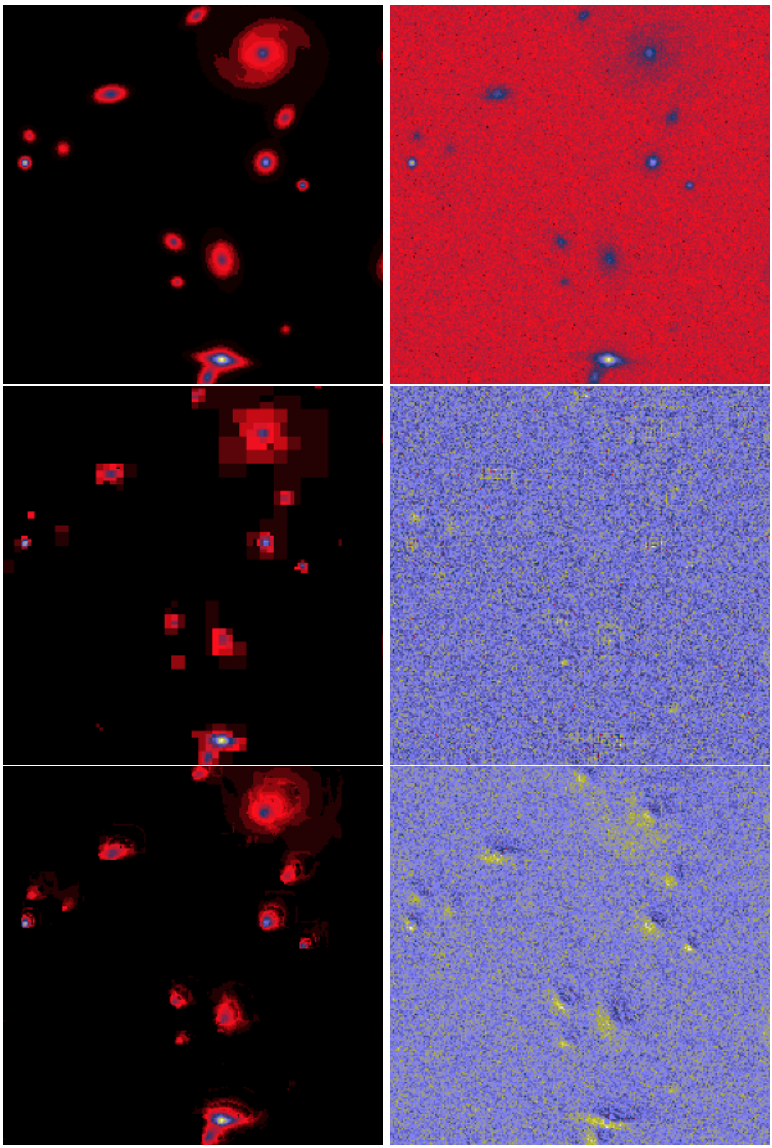


FIGURE 23.5. Top, simulated sources and simulated noisy image. Middle, haar filtered image and residual image. Bottom, undecimated Haar filtered image and residual image.

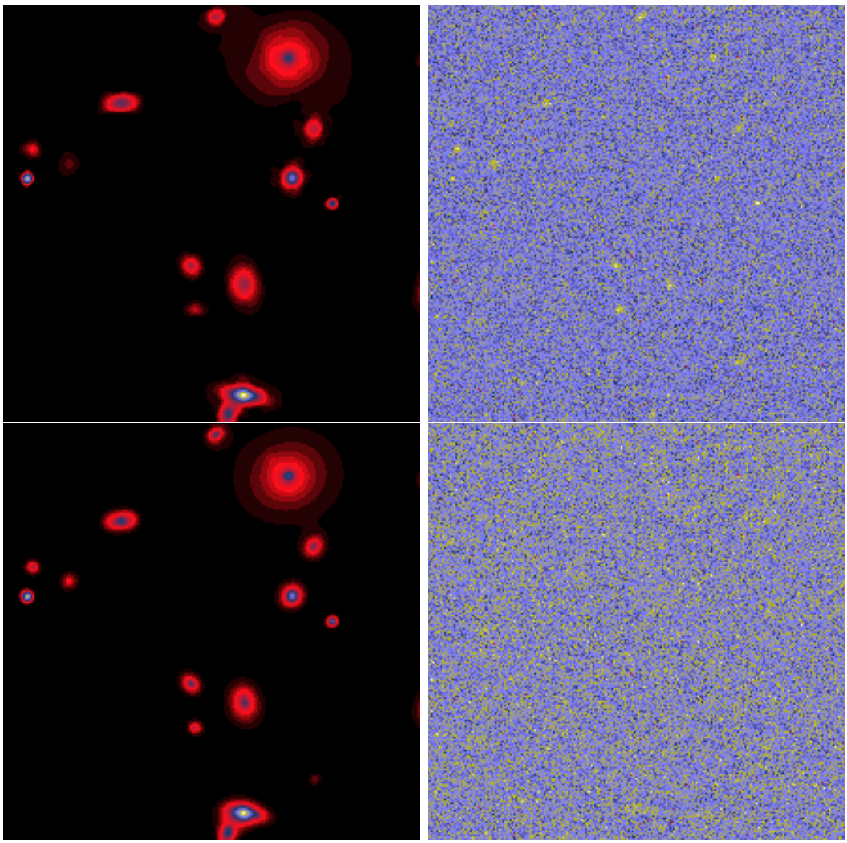


FIGURE 23.6. Top, undecimated wavelet filtering (7/9 filters) and residual. Bottom, filtering using the à trous algorithm and an iterative method, and residual.

24

The Statistical Challenges of Wavelet-Based Source Detection

Peter E. Freeman[1], **V. Kashyap, R. Rosner and D. Q. Lamb**

ABSTRACT Wavelet functions are proving extremely useful for detecting sources in binned, two-dimensional photon counts images. In this chapter, we describe the mission-independent source detection algorithm `WAVDETECT`, part of the *Chandra Interactive Analysis of Observations (CIAO)* software package, and discuss the statistical challenges we have faced in its development, such as: what is the best way to estimate the local background in each pixel, if it is *a priori* unknown? What is the best way to eliminate false detections caused by instrumental variations? And what is the significance of a detected source?

24.1 Introduction

Wavelets are scaleable, oscillatory functions with finite support (i.e. they are non-zero within a limited spatial regime) and an overall normalization zero.[2] They provide a superior means by which to analyze data in binned, two-dimensional photon count images, as their properties allow the simultaneous characterization of the locations, strengths, and dominant length-scales of astronomical sources.

Aside from source characterization, wavelet-based algorithms are being shown to outperform the standard "sliding cell" algorithm [2], which is rapidly being supplanted as the algorithm of choice in the field of source detection. Damiani et al. [3] were the first to present a general wavelet-based source detection algorithm, one appropriate for analyzing data observed by telescopes with nearly Gaussian point-spread functions (PSFs) in the high-background-count limit ($B \gtrsim \frac{0.1}{\sigma^2}$ ct pix^{-1}, where σ is the scale size of the analyzing wavelet). This algorithm is also the first to use exposure maps to mitigate the effect that exposure variations (caused by, e.g., support-

[1]Harvard-Smithsonian Center for Astrophysics

[2]For an introduction to the theory of wavelet functions, see, e.g., Mallat [1].

rib shadows and the edge of the FOV) have upon the rate of false source detections, although their treatment of features must be altered to suit differect detectors (see, e.g., Micela et al. [4]).

In Freeman et al. [5], we describe a more general source-detection algorithm that has been implemented as the *Chandra Interactive Analysis of Observations (CIAO)* application `WAVDETECT`.[3] Our algorithm can: (1) operate effectively in the low background counts regime, which is crucial because of the low particle and cosmic background count rates for the *Chandra* detectors (the overall rate being $\sim 10^{-6}$ and 10^{-7} ct sec^{-1} pix^{-1} for the *Chandra* ACIS and HRC detectors, respectively); and (2) operate effectively regardless of the PSF *shape*, also crucial because of the (non-Gaussian) nature of the off-axis *Chandra* PSFs. It also (3) treats exposure variations in a general, non-detector-specific manner. Thus our algorithm may be immediately adapted for the analysis of data from virtually any other photon-counting detector.

In this chapter, we provide a minimalist introduction to the `WAVDETECT` source detection and characterization algorithm (§24.2), then discuss the statistical challenges that we have faced (and continue to face) in its development. We ask the reader to consider the following questions:

Are our solutions to statistical problems, when indeed we have them, optimal solutions? And are there better methods, approximations, etc., which we should use that are completely (or nearly) independent of detector details, and that do not excessively increase computation time or the use of computational resources?

24.2 Algorithm

24.2.1 Source Detection

A typical analysis of a counts image involves correlating it with a sequence of wavelet functions. In `WAVDETECT`, the Marr, or "Mexican Hat" (MH) wavelet function is used:[4]

$$W(\frac{x}{\sigma},\frac{y}{\sigma}) = \left[2 - \frac{x^2}{\sigma^2} - \frac{y^2}{\sigma^2}\right] e^{-\frac{x^2}{2\sigma^2}-\frac{y^2}{2\sigma^2}} \tag{24.1}$$

This function, which has a positive kernel (PW) surrounded by a negative annulus (NW) and which differs significantly from zero only within a radius of $\approx 5\sigma$ (see Figure 24.1), has several advantages which motivate its use for source detection: (1) the Gaussian-like PW has a shape similar

[3] `WAVDETECT` is composed of `WTRANSFORM`, a source detector, and `WRECON`, a source list generator. The *CIAO* package is available at `http://cxc.harvard.edu`.

[4] Our algorithm allows the use of asymmetric MH wavelets ($\sigma_x \neq \sigma_y$) but for simplicity we assume rotational symmetry in this work.

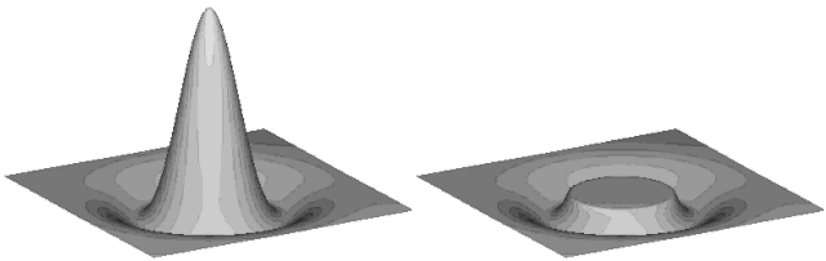


FIGURE 24.1. *Left:* The two-dimensional Marr, or "Mexican Hat," wavelet function (eq. 24.1). *Right:* The negative annulus of the Mexican Hat function, used in background estimation.

to a canonical PSF; (2) it is insensitive to both flat and constant gradient components of any underlying function in the image; and (3) its Fourier transform has limited extent in Fourier space, so that limited, discrete sampling of values of σ (e.g., at values separated by factors of two) is sufficient to sample the entire frequency domain.

In the simplest case, where the background amplitude at each pixel, $B_{i,j}$, is known *a priori* and where the exposure map[5] is constant (or varies linearly), source detection at a given scale σ proceeds in two steps. First, there is the computation of the correlation map

$$\begin{aligned} C_{i,j} &= \sum_{i'} \sum_{j'} W_{i-i',j-j'} D_{i',j'} \\ &\equiv < W \star D >_{i,j} \ . \end{aligned} \tag{24.2}$$

where i and j are pixel indices and $i-i'$ and $j-j'$ are the discrete equivalents of x and y in eq. (24.1).[6] Second, $C_{i,j}$ is compared with a source detection threshold $C_{o,i,j}(S_o, B_{i,j})$ (the computation of which is described in §24.3.1), where S_o is a user-defined significance value (e.g. the inverse of the number of analyzed pixels, for one false source detection in the image); if $C_{i,j} > C_{o,i,j}$, we associate the pixel (i, j) with a source.

[5]If an exposure map is not provided, a flat one is assumed in order to account for the edge of the FOV.

[6]Our notation deviates from that of Mallat, in which $< W \star D >_{i,j}$ would be written $W \star D[i, j]$; however, we feel our notation makes complicated expressions involving transforms more easily interpretable.

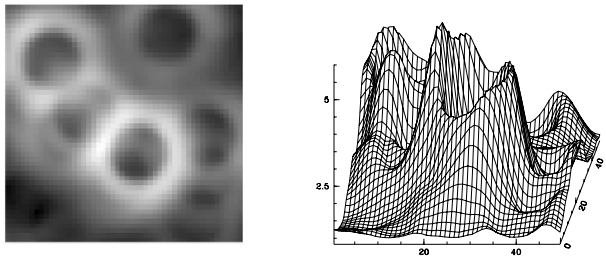


FIGURE 24.2. Illustration of how sources located within the negative annulus of the wavelet cause "rings" in an initial background estimate.

24.2.2 Background Estimation

If a background map B is not provided, then background maps are computed at each scale σ by averaging the raw data around each pixel, using the exposure map E and wavelet negative annulus NW as weighting functions:

$$B_{i,j} \;=\; E_{i,j} B_{\mathrm{norm},i,j} \;=\; E_{i,j} \frac{< NW \star D >_{i,j}}{< NW \star E >_{i,j}} .$$

(See Figure 24.1.) B_{norm} is the normalized (i.e. flat-fielded) number of expected background counts.[7] Our method of background estimation is independent of the details of the PSF and allows `WAVDETECT` to detect sources of arbitrary size.

In Freeman et al., we discuss the ways in which sources in the FOV can bias the local background estimate. In particular, the local background amplitude will be overestimated in rings of radius $\approx 2\sigma$ around sources (this being the radius at which NW achieves its minimum value), adversely affecting the detection of weak sources, and the estimation of source properties (see Figure 24.2). To mitigate their effect, we employ an "iterative cleansing" algorithm, in which `WAVDETECT`: (1) identifies pixels to be cleansed using the probability sampling distribution (PSD) $p(C|B_1)$, where B_1 is the initial background map; (2) replaces the data D (or D_1) in these pixels with B_1, creating a new image, D_2; (3) estimates $B_2(D_2)$; (4) computes $p(C_2|B_2)$ and identifies pixels to be cleansed; etc. The resulting background is then used to make the final determination of source pixels.

24.2.3 Treating the Effect of Exposure Variations

To be effective, a source detection algorithm must distinguish between astronomical and instrumental sources, the latter of which are detected at or

[7]Note that the distinction between vignetted and non-vignetted components of the background (e.g., the particle background) is ignored in our estimate.

near regions where the exposure varies greatly (e.g. support rib shadows or chip boundaries). One way to exclude instrument sources is to generate the PSD $p(C|B_{i,j}, E)$ for each and every observation; however, this is computationally impractical. Instead, `WAVDETECT` attempts to remove the effect of the variation from the observed correlation coefficient:

$$C_{i,j} \;=\; < W{*}B >_{i,j} + < W{*}S >_{i,j} + \Delta C_{i,j}\,.$$

B is the estimated (i.e. noise-free) background amplitude, S is the source counts amplitude, and ΔC is the noise contribution. While $< W{*}B >_{i,j} = < W{*}EB_{\rm norm} >_{i,j}$ is affected by variations in exposure, $E_{i,j} < W{\star}B_{\rm norm} >_{i,j}$ is not, thus we correct the correlation coefficient as follows:

$$C_{{\rm cor},i,j} \;=\; C_{i,j} - < W{\star}(EB_{\rm norm}) >_{i,j} \; + E_{i,j} < W{\star}B_{\rm norm} >_{i,j}\;.$$

$C_{{\rm cor},i,j}$ should only contain information of astrophysical value, thus we use it to compute source detection thresholds. We ignore the effect of exposure variations upon $< W{*}S >_{i,j}$, while noting that it is not affected by exposure variations if the source counts are from a point source, since point source count rates depend only upon the exposure at the center of the PSF. We also ignore the effect of variations upon the quantity $\Delta C_{i,j}$; we return to this point below, in §24.3.4.

24.2.4 Source Characterization

Final Background Map. To determine source properties (e.g. net counts), we need to construct a final background map $B'_{\rm norm}$ from those that were generated during the detection process:

$$B'_{{\rm norm},i,j} \;=\; \frac{\sum_{k=1}^{N} \epsilon_{i,j,k}\sigma_k^2 B_{{\rm norm},i,j,k}}{\sum_{k=1}^{N} \epsilon_{i,j,k}\sigma_k^2}\,. \tag{24.3}$$

N is the number of scales at which the data were analyzed, and ϵ is 1 if $\sigma_k \gtrsim r_{{\rm PSF},i,j}$ and 0 otherwise. This factor helps eliminate bumps caused by source counts in background maps: at scales $\sigma \lesssim r_{\rm PSF}$ (the PSF "size"), source counts will unavoidably overlap the NW, causing "bumps" in the background map which peak at the source centroid.

Source Cells. Source properties are estimated within a source cell, a collection of pixels which we will associate with the source. A source cell is created using a source counts image:

$$SC_{i,j} \;=\; \max\left(\frac{< PW{\star}D >_{i,j}}{< PW{\star}E >_{i,j}} \;-\; B'_{{\rm norm},i,j}, 0\right)\,.$$

The PSF size $r_{{\rm PSF},i,j}$ is used to determine the appropriate smoothing scale for a given source. For an isolated source, the extent of the source cell is

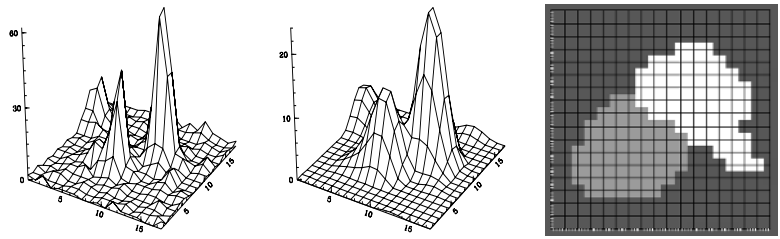


FIGURE 24.3. Illustration of how source cells are created for two nearly overlapping sources. *Left:* the raw counts data. *Middle:* the source counts image, created smoothing the counts data with a PW function of size $\sigma_x = \sigma_y = 2$ pixels, then subtracting the estimated background. *Right:* the source cells defined using the source counts image data. The saddle point seen in the middle image defines the boundary between the cells.

determined by the smallest zero-amplitude contour surrounding its location; tests indicate this contour will encompass nearly 100% of counts. In a crowded field, saddle points in the source counts image are also used to determine cell boundaries (*e.g.* Figure 24.3).

Once a source cell is created, the source location, etc., are computed using the raw data (and estimated background) in the cell. Most computations involve straightforward summations. Of particular interest is the computation of location; for instance, the x-coordinate of a source is estimated using the equation $\sum_{SC} D_{i,j} i / \sum_{SC} D_{i,j}$, where the summation occurs over all pixels in the source cell. This estimate is not optimal in some situations, a point which we will return to below.

24.3 Statistical Challenges

24.3.1 Calculation of Source Detection Thresholds

The fundamental question in source detection is: should an image pixel be associated with a source, or with the background? To answer this, we compare the value of the statistic $C_{i,j}$ with a PSD that is a function of the local background $B_{i,j}$ in each pixel: $p(C|q_{i,j} = 2\pi\sigma^2 B_{i,j})$. The test significance (or Type I error) $S_{i,j}$ would then be calculated by computing the tail integral of this PSD from $C_{i,j}$ to infinity:

$$S_{i,j} = \int_{C_{i,j}}^{\infty} dC p(C|q_{i,j}) \,. \tag{24.4}$$

However, `WAVDETECT` does not compute $S_{i,j}$ directly, because while this distribution tends asymptotically to a zero-mean Gaussian with width $\sigma = \sqrt{q_{i,j}}$ as $q_{i,j} \to \infty$, the PSDs are no longer smooth in the low-counts limit

and cannot be approximated by analytic functions. Instead, we performed a sufficient number of simulations (involving $\approx 5\times10^{10}$ pixels) so that source detection thresholds $C_{o,i,j}$ could be determined directly for user-defined threshold significances $S_o \gtrsim 10^{-7}$, for -7 $\lesssim \log q_o \lesssim$ 3 (where the data in each 1024×1024 simulation are sampled from a Poisson distribution with expectation value $q_o/2\pi\sigma^2$). Note that each simulated image was analyzed with a MH wavelet function of scale $\sigma = 4$ pixels, since, for instance, the threshold for $\sigma = 4$ pixels and $B_{i,j} = 1$ count is the same as the threshold for $\sigma = 2$ pixels and $B_{i,j} = 4$ counts.

We learned a number of lessons when carrying out these simulations. (1) One must record the distribution $p(C_{i,j}|q_{i,j})$, and *not* $p(C_{i,j}|q_o)$. The relationship between the two distributions is

$$p(C_{i,j}|q_{i,j}) = \int dq_o p(C|q_o)p(q_o|q_{i,j})$$

and they only become asymptotically equivalent in the high-count limit. Threshold values estimated from the distributions $p(C_{i,j}|q_{i,j})$ are more conservative than those derived from $p(C_{i,j}|q_o)$, markedly so in the low-count limit. (2) We found that we could determine thresholds for small significance values (i.e. beyond $S = N^{-1}$, where N is the total number of simulated pixels for a particular value of q) by fitting two-dimensional functions to the observed data $C_o(S_o, q)$ (see Freeman et al. for more details on the functional forms). We took 25 observed values of $C_o(S_o, q)$ from 25 sets of simulations and used the mode and central 68% of the values to determine each data point and one-sigma error. The fit functions can then be arbitrarily extrapolated (and tested for validity). (3) While Damiani et al. determine that analytic representations of the PSDs work well if $\log q_{i,j} \gtrsim 3$, we find our threshold functions to be more conservative and thus we use them even in the high-counts limit.

We also note two threshold computation issues that we have yet to fully explore. (1) The computation $p(C|2\pi\sigma^2 B_{i,j})$ does not take into account variations in the background amplitude that may occur *within* the PW, say at the edge of a region in which X-ray shadowing is evident. (2) The threshold correlation values do not take into account uncertainties in the threshold function parameters, i.e. the error in a given threshold estimate is currently not computed.

24.3.2 What is the Significance of a Source?

The question posed above is not answered by `WAVDETECT`,[8] since while we can in principle determine significances on a pixel-by-pixel basis in cor-

[8]Current program output includes the `src_significance` for each source, but this value should be ignored: it assumes that the estimated total background counts B in the source cell are sampled from a PSD with "Gehrels variance" $[1 + \sqrt{B + 0.75}]^2$ (Gehrels

relation space, we cannot easily determine the significance of a group of pixels. This is because correlation values in adjacent pixels are not independently sampled (see eq. 24.2). While covariance terms can be estimated, the computational cost is prohibitive (see §24.3.7). Is there a computationally inexpensive way to determine the significance using correlation maps? Or is using correlation maps to determine source significances the wrong way to proceed?

24.3.3 Background Estimation

What is the best *PSF-independent* approach to determining the local background amplitude in each image pixel, if it is unknown *a priori*? Note that our motivation for a multi-scale PSF-independent approach is not necessarily that we may more easily apply it to data from different detectors, but rather to avoid introducing a bias against the detection of extended objects (e.g. clusters) into our algorithm (cf. Damiani et al., who estimate background once, and not once per scale, using a sliding box whose width is dependent upon the PSF size). However, an even more fundamental question which we must raise is, is our approach of determining the background on a scale-by-scale basis *during* the detection process the wrong approach? Should we first estimate a scale-independent background map, then use that map during the source detection phase? If so, what is the best way to compute that map? While we do create a "final" background map (eq. 24.3), it is used for only for source characterization and it may not be sufficiently accurate for source detection. In particular, the weighting used to create the "final" map, which is meant to minimize the effect of systematic overestimates in the background at small scales, is certainly not optimal for determining the *local* background at a given pixel.

A second fundamental question is: even if our scale-by-scale approach is theoretically sound, is our iterative approach to determining the background at each scale the best approach to use? Should the same significance criterion be used for cleansing data as for detecting sources? To minimize the effect of weak undetectable sources on the background estimate, we advocate an aggressive approach to iterative cleansing: S_o should be set high during cleansing, e.g. to $S_o = 10^{-2}$. Should the data in a pixel to be cleansed be replaced with the estimated background amplitude, as it is now, or some other quantity? (We cannot simply mask out the affected pixel, since then we cannot use FFTs in our algorithm.) Last, we note that there are no rigorous quantitative rules governing how one should specify the number of iterations, as that can depend on the crowdedness of the

[6]). This variance is generally (much) larger than the background variance derived by `WAVDETECT`, thus the `src_significance` can be (very much) an underestimate of the true significance.

field, the source distribution, the source strengths, and the wavelet scale size, etc. Thus, we leave the stipulation of stopping rules to the user.

24.3.4 Treating the Effect of Exposure Variations

Our exposure correction algorithm successfully reduces the number of detected instrument "sources."[9] However, there are still a number of issues with this algorithm which have yet to be fully explored.

First, $C_{\mathrm{cor},i,j}$ cannot be directly compared with the PSD $p(C|q_{i,j})$ because the noise term $\Delta C_{i,j}$ is itself uncorrected. Concentrating on the issue of false positives, the important question here is: is the asymptotic width of the distribution from which the unknown quantity $\Delta C_{\mathrm{cor},i,j}$ is sampled *smaller* than the width of the distribution from which $\Delta C_{i,j}$ is sampled? If so, then the rate of false detections will still be greater than expected. Given that this width is, at least in the high-counts limit, proportional to the observed number of background counts (a number affected by exposure), this should be a problem if and only if $E_{i,j}$ is smaller than the average exposure over the spatial extent of the wavelet centered at (i, j), i.e. this is only a problem within troughs or beyond the edge of the FOV. Thus we suggest that one should carefully scrutinize all sources detected in lightly exposed regions.

Second, by using B_{norm} to adjust correlation values, we "contaminate" $C_{\mathrm{cor},i,j}$ with low-frequency information. As a passband filter, NW is most sensitive to constant components of the data,[10] while W is, as advertised, most sensitive to variations in the data at length-scales similar to the scale size of W itself. The user must keep this contamination in mind if the analysis goal is to characterize detected sources by examining their properties in correlation space.

24.3.5 Source Property Estimation

When computing source properties, we use the raw data $D_{i,j}$ as the weighting function, instead of the source fluence $D_{i,j} - E_{i,j}B'_{\mathrm{norm},i,j}$. We do this because using the latter quantity can greatly complicate the estimation of variances, since values of $B'_{\mathrm{norm},i,j}$ in adjacent bins are correlated. However, using $D_{i,j}$ does not always lead to optimal results; for instance, when a source cell is large and/or asymmetric, and the number of background counts is large compared with the number of source counts, the background

[9] This is not an easily quantified statement, since the reduction depends upon the accuracy of the background map and the specifics of the exposure map, in particular the energy spectrum that is assumed when it is created.

[10] As it should be: we are not seeking a scale-by-scale decomposition of the background, but rather simply to determine its local (presumably constant-component-dominated) amplitude in each pixel.

counts can unduly bias the source properties. This issue has been observed in analyses of *Chandra* data, where estimated source positions for weak sources are not well-determined far off-axis (N. Brandt, F. Bauer, private communications).

Is there a better, more robust and PSF-independent method to determine source properties? Should source locations, e.g., be determined using the fluence, even if the errors cannot be computed rigorously? Should they be determined using some other function of D, or using the mode of the data's distribution?

24.3.6 Extended Source Identification

The ability to detect extended sources in part drives `WAVDETECT`'s design, but the actual algorithm for identifying them is not robust. A source "size" s is computed using the number of pixels in the source cell n (listed in the column `npixsou` of the `WAVDETECT` output file): $s = \sqrt{n}/2\pi$. The ratio of s to $r_{\rm PSF}$ (`psf_size`) is then listed in the `WAVDETECT` output file as `psfratio`. While a large `psfratio` indicates that a source is extended, the PSD for it is unknown (and may in general be unknowable considering the number of factors that can influence it), so we do not currently calculate a significance.

Is there a robust method for identifying extended sources using the raw data and minimal (or no) information about the PSF, and which does not involve actual fits of the PSF to the data? Or is extended source identification best done outside of `WAVDETECT` entirely?

24.3.7 Computation of Variances

We conclude our discussion of statistical challenges by touching upon the issue of error estimation. In principle, we would like to estimate variances using the formula (Eadie et al. [7], p. 23)

$$\begin{aligned} V[Y] &= V[\sum_i \sum_j a_{i,j} X_{i,j}] \\ &= \sum_i \sum_j a_{i,j}^2 V[X_{i,j}] + 2 \sum_i \sum_{i'>i} \sum_j \sum_{j'>j} a_{i,j} a_{i',j'} \mathrm{cov}[X_{i,j}, X_{i'j'}] \end{aligned}$$

where Y is the quantity of interest and $X_{i,j}$ are the random variables (i.e. either $D_{i,j}$ or functions of $D_{i,j}$, with each datum $D_{i,j}$ assumed to be independently sampled from a Poisson distribution with variance $D_{i,j}$). However, an exact computation has a staggeringly high computational cost, as we will demonstrate by deriving the variance for the two-iteration background map B_2:

$$V[B_{2,i,j}] = \sum_{i'} \sum_{j'} (N_{i,j} N W_{i-i',j-j'})^2 V[D_{2,i',j'}]$$

$$+ \quad 2\sum_{i'}\sum_{i''>i'}\sum_{j'}\sum_{j''>j'}(N_{i,j}NW_{i-i',j-j'})(N_{i,j}NW_{i-i'',j-j''})\times \\ \mathrm{cov}[D_{2,i',j'}, D_{2,i'',j''}]\,,$$

where $N_{i,j} = E_{i,j}/ < NW{*}E >_{i,j}$.

First, $V[D_{2,i',j'}]$ is

$$V[D_{2,i,j}] \quad = \quad \begin{cases} D_{i,j} & \text{uncleansed pixel} \\ \sum_{i'}\sum_{j'}(N_{i,j}NW_{i-i',j-j'})^2 D_{i',j'} & \text{cleansed pixel} \end{cases}$$

(Note that in the actual computation of $V[B_{2,i,j}]$ carried out by `WAVDETECT`, however, $V[D_{2,i,j}]$ is approximated as $D_{2,i,j}$.)

The estimation of $\mathrm{cov}[D_{2,i',j'}, D'_{2,i'',j''}]$ is, not surprisingly, more complicated. For the two-iteration background case, there are three possibilities: $D_{2,i',j'} = D_{i',j'}$ and $D_{2,i'',j''} = D_{i'',j''}$; $D_{2,i',j'} = D_{i',j'}$ and $D_{2,i'',j''} = B_{1,i'',j''}$ (or vice-versa); and $D_{2,i',j'} = B_{1,i',j'}$ and $D_{2,i'',j''} = B_{1,i'',j''}$. In the first case, the covariance is zero. We can estimate the covariance in the second case using the approximation (Eadie et al. p. 27):

$$\begin{aligned} & \mathrm{cov}[D_{i',j'}, B_{1,i'',j''}] \\ \approx \quad & \sum_k\sum_l\sum_{k'}\sum_{l'}\left(\frac{\partial D_{i',j'}}{\partial \mu_{k,l}}\right)\left(\frac{\partial B_{1,i'',j''}}{\partial \mu_{k',l'}}\right)\mathrm{cov}[D_{k,l}, D_{k',l'}] \\ = \quad & D_{i',j'}N_{i'',j''}NW_{i''-i',j''-j'}\,, \end{aligned}$$

where μ represents the expectation value of the sampling distribution for D. (We assume $\mu = D$.) Making a similar calculation, we find in the third case that

$$\begin{aligned} & \mathrm{cov}[B_{1,i',j'}, B_{1,i'',j''}] \\ = \quad & \sum_k\sum_l N_{i',j'}NW_{i'-k,j'-l}N_{i'',j''}NW_{i''-k,j''-l}D_{k,l}\,. \end{aligned}$$

A rigorous estimate of $V[B_2]$ takes a factor of $\sim \mathcal{O}(d_x d_y \sigma_x^2 \sigma_y^2)$ longer to compute than `WAVDETECT`'s current approximate estimate (where the image size is $d_x \times d_y$). We find that including covariance terms increases the variance by $\lesssim 10\%$, with the maximum increase adjacent to (but not at) the location of strong sources. We stress that this is a source-strength- and source-geometry-dependent result that obviously cannot be blindly applied to all fields.

24.4 A Concluding Question: Can WAVDETECT Be Completely PSF-Independent?

The current algorithm is not completely independent of detector details, as the source list generator relies on knowledge of the PSF size at a given

pixel in order to: (1) compute the "final" background map, (2) determine the appropriate source-counts image to use to generate source cells, and (3) identify extended sources. The necessity of coding this knowledge into the algorithm prevents, e.g., `WAVDETECT` in CIAO 2.1 from being using used effectively with *XMM-Newton* data: the algorithm will run to completion, but the PSF size will be assumed to be the smallest input scale, severely affecting the determination of source properties. While an experienced user of `WAVDETECT` could work around this issue by running it multiple times with different input values to compute one source list, or while `WAVDETECT` could be altered to allow the user to read in a map of PSF sizes, it is worth asking if the PSF is needed at all. Could the three calculations mentioned above, or even just the first two, be done effectively using just the raw data? We leave that as an exercise for the reader.

Acknowledgements The authors acknowledge the support of the CXC Beta Test Site grant at the University of Chicago and NASA grants NAG5-3173, NAG5-3189, NAG5-3195, NAG5-3196, NAG5-3831, NAG5-6755, and NAG5-7226.

24.5 References

[1] Mallat, S. 1998, A Wavelet Tour of Signal Processing (London: Academic Press)

[2] Harnden, F. R., Jr., Fabricant, D. G., Harris, D. E., & Schwarz, J. 1984, SAO Special Report 393

[3] Damiani, F., Maggio, A., Micela, G., & Sciortino, S. 1997, ApJ, 483, 350

[4] Micela, G., et al. 1999, A&A, 341, 751

[5] Freeman, P. E., Kashyap, V., Rosner, R., & Lamb, D. Q 2002, ApJS, 138, 185

[6] Gehrels, N. 1986, ApJ, 303, 336

[7] Eadie, W. T., Drijard, D., James, F. E., Roos, M., & Sadoulet, B. 1971, Statistical Methods in Experimental Physics (Amsterdam: North-Holland)

25

Reflections on SCMA III

John Rice[1]

25.1 Introduction

It has been a great privilege to participate in this fascinating meeting and a great challenge to be asked to comment on the wide variety of issues that have arisen. I will try to place the papers we have heard here in some perspective and to outline some current and future challenges and opportunities lying in the intersection of statistics and astronomy. I ask the reader to bear in mind that the "seeing conditions" are poor.

25.2 A spectrum of statistical methodology

In considering the wide variety of statistical methodology relevant to astronomy, it may be helpful to view it on a spectrum ranging from procedurally based methods to methods based on highly specified stochastic models. Different regions on this spectrum are relevant to different types of problems and individual statisticians often have "personal equations" that influence where their contributions fall.

The following example is illustrative: Smoothing splines were first proposed as a procedure for passing a smooth curve through a noisy scatter plot of observations (y_i, x_i), $i = 1, \ldots, n$. It was desired that the curve not be of any simple parametric form, such as a low degree polynomial, but merely be "smooth." No explicit stochastic structure was assumed for the data. In Reinsch (1967) proposed choosing the curve, $g()$ as the minimizer of

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 \tag{25.1}$$

subject to the constraint $\int[g''(x)]^2 dx \leq \Omega$. (This basic idea had been around for some time—see Wahba 1990 for more complete references). Using a Lagrange multiplier, the problem can be written as that of choosing

[1]Department of Statistics, University of California, Berkeley

$g()$ to minimize

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int [g''(x)]^2 dx. \tag{25.2}$$

Although this optimization problem has some heuristic appeal, the proposal would not have had much impact were it not the case that the minimizing $g()$ is a cubic spline and that there are fast and stable numerical algorithms for its computation. The solution depends upon the choice of the smoothing parameter, λ: small values of λ give rise to highly oscillatory functions and large values to very smooth ones. It was left to the user to interactively determine a satisfactory choice.

The next stage in the study of this method was an examination of its properties by statisticians with frequentist personal equations. Thus, an explicit stochastic element was added to the structure to produce a statistical model: it was assumed that the data were of the form $y_i = g_0(x_i) + \epsilon_i$ where the ϵ_i are independent random variables with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. Note that this itself is quite an idealization. Frequentist properties of the spline estimate were the subject of intensive theoretical and numerical research and are now quite well understood.

If the random errors are modeled as Gaussian, the procedure can be viewed as an example of penalized likelihood in which the log likelihood is the first term in (25.2) and the second term penalizes rough solutions. It is a canonical example of nonparametric regression, in which there is a tradeoff between bias and variance that does not typically occur in parametric models. (It could of course be argued that any parametric model is an approximation and hence may well give rise to bias, so that the distinction between parametric and nonparametric models is illusory). One of the widely used ways of selecting λ in a data-driven way to achieve this balance is cross-validation (Wahba & Wold 1975).

Finally, a statistician with a Bayesian personal equation examining (25.2) will see the sum of a log likelihood and the log of a prior. Wahba (1993) identified the prior as a doubly integrated Brownian motion where λ is the variance parameter of this stochastic process. Now another layer of idealization has been introduced and can be formally used to construct posterior credible regions, for example.

The bottom line: what really matters is how well a method works. Are there efficient and stable computational algorithms? How well does it work on a suite of simulated data? On a variety of real data sets? How is it affected by outliers? How is it affected by spacings in the x_i? How does it compare to alternative methods for doing nonparametric regression? Such assessments are made in a variety of ways, and not only with respect to a single figure of merit, such as integrated squared error. Whether procedurally or model generated, a method must be assessed by its effectiveness.

I find it helpful to think about statistical contributions to astronomy as being arranged along this spectrum as well. At the risk of oversimplifica-

tion, this can be exemplified by many of the presentations at this meeting. At the procedural end there was the presentation of Cook, showing us some wonderful tools for exploring multivariate data. The presentations of Breiman, Freeman, Murtagh, and Starck showed us some widely applicable procedures that are based on rather minimal modeling structures. Similarly Djogorvski and Nichols were primarily interested in procedures rather than models. The papers of Shafer, Szalay, Wasserman, and Martinez were primarily at frequentist wavelengths, while the Bayesian frequency band was occupied by the presentations of Berger, Bretthorst, Connors, van Dyk, Jaffe, Kolaczyk, Loredo, and Scargle.

The presentations of Raftery and Johnstone contained a mix of Bayesian and frequentist perspectives. Johnstone's exemplified how these two perspectives can enrich each other. In the smoothing spline example above, the Bayesian formulation was rather an afterthought, but in Johnstone's use of priors on different resolution levels we see the Bayesian formalism being adopted for the purpose of generating smoothing procedures which can be explored from a frequentist perspective or merely viewed as empirical procedures. Clearly, no one would take these priors seriously as quantification of personal belief in a game with a bookie; rather, they are devices that hopefully generate useful methods of averaging (what statistics is all about). The procedures presented by Kolaczyk stemmed from a Bayesian formalism and can be viewed in a similar way—his parameters α and β are quite analogous to the λ for a smoothing spline and I can imagine turning a (α, β) knob to explore differing degrees of smoothing without introspection as to the state of my "belief" about (α, β).

Statistical models may be viewed as filters through which data are analyzed, and, as William James wrote, "We must be careful not to confuse data with the abstractions we use to analyze them." But we need these filters/abstractions: as George Box wrote, "All models are wrong, but some are useful." Models are useful and effective to the degree that they provide a mechanism for accurately extracting information of scientific interest from the data. As one moves towards the Bayesian end of the spectrum, models become more detailed and highly specified, as can be seen in the contributions to this meeting.

We really need to be careful in using models, especially in situations in which there is such a large quantity of data that model accuracy cannot be readily checked visually. More elaborate models tend to be more fragile. Despite extensive effort, "de-glitching" may be incomplete. Even beyond glitches, there may be sources of noise not properly accounted for in the model—do the Gaussian or Poisson variables in the model really reflect all the sources of noise, such as cosmic rays, image motion, and crowding, for example?

For these reasons, robustness has a long and honorable tradition in statistics and is increasingly relevant in this age of data floods. Figure 25.1 shows light curves in the form of fourth order trigonometric polynomials fit to

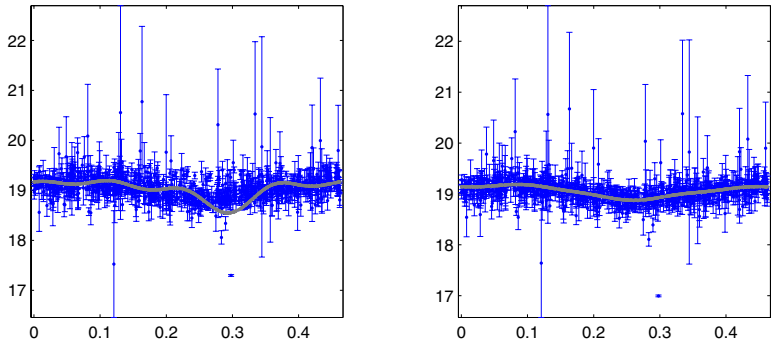


FIGURE 25.1. Weighted least squares fit (left panel) and robust fit (right panel) to phased data from an RR Lyrae with period 0.46 days.

phased observations at the fundamental frequency (0.46 days) of an RR Lyrae of type "d." The individual observations of magnitude were accompanied by error bars and, for one particular point, the error bars were far too small. The weighted least squares fits of the light curves are formed in accord with the model, but are very sensitive to deviations from it. The outlying point (just one observation out of 845!) pulls the curve locally and causes global rippling. The robust fitting procedure is designed to do reasonably well if the model holds and to be resistant to outliers.

Even without such phenomena, one needs to be concerned about the biases that are incurred by analyzing the data only through the filters of the model. One needs to ask how crucial the assumptions are. Are the important conclusions sensitive to distributional assumptions and assumptions of independence in a frequentist model? In the case of a Bayesian analysis, one should seriously examine the consequences of the choice of a prior. This is not easy for complex hierarchical models and often receives only cursory attention. In his contribution to this meeting, Jaffe makes some references to the difficulty of choosing a prior and the influence the prior has on inferences about key cosmological constants. The smoothed lightcurves that Berger displays resulting from priors on wavelet coefficients produce rather suspicious structure precisely in the regions where there is no data (near zero).

25.3 Challenges and opportunities

The meeting has been very exciting in illustrating many opportunities for application of existing statistical methodology and challenges for the development of new approaches. Let me highlight a few:

25.3.1 Large scale structure:

I suspect that there are real opportunities for going beyond two-point and higher order correlation functions for both characterizing structure and discriminating amongst theories. Might not other functionals offer sharper characterization and discrimination? The size of the data, the complexity of coverage patterns, and the presence of selection biases makes this endeavor even more challenging.

25.3.2 Separation of source and background

This problem is omnipresent in astronomical data and we heard about some very interesting developments in the presentations of Freeman and Starck. The problem of removing foreground in studies of the CMB was alluded to by Jaffe. Perhaps because of its high dimensionality and spatial aspects, this problem does not seem to me to fit very well into our standard paradigms of statistical inference and decision theory. I suspect that some gains can be made by taking more advantage of the fact that the same kind of problem is often faced repeatedly (see the section on empirical Bayes below).

25.3.3 Parameter estimation from massive data sets

Jaffe's paper gave us a hint of the kinds of problems of this type that will be faced in the near future. How will we meet the corresponding computational challenges? As mentioned by Djorgovski, one possibility is to forgo computing estimates with high precision and/or to forgo notions of statistical optimality (the best may be the enemy of the good). Algorithms derived from the literature on stochastic approximation and on-line gradient methods may turn out to be important. There are close relationships between parameter estimation, coding, information theory and data compression (Rissanen & Yu 2002). A sufficient statistic provides marvelous data compression, but these rarely exist. We may need a notion of an "almost sufficient" statistic. Nichol's use of KD trees is in this spirit. For some current developments on using compression of astronomical data also see Bond et al. (2000) and Tegmark et al. (1997). There has always been a strong coupling of between inference and computation and the prevalance of massive data sets coupled with the computational power of "the grid" will have profound effects on the nature of the discipline of statistics.

25.3.4 Massive data sets and multivariate analysis

Here we have heard of data sets which would seem to correspond to a multivariate statistician's dream: enormous n and bounded p. But are we really ready to live out our dreams? The challenges were exhilaratingly described in Djorgorvski's and Strauss' presentations. The staggering size

of the data sets begs for multiscale procedures, for adaptive stratification, for adaptive sequential procedures, and for new methodology.

Although we have heard of some very promising developments from Murtagh and Raftery, I think that there remains a great deal to be done in finding clusters of widely varying morphology and other structures in massive data sets. The complexity and heterogeneity of astronomical data offers further challenges. The structures are likely to be quite different from those encountered in generic market-basket data mining: the strong physical constraints operative in astronomical data and good precision measurements should result in concentrations along low dimensional (nonlinear) manifolds. Local linear embedding (Roweis & Saul 2000) and ISOMAP (Tenenbaum et al. 2000) are two interesting recent developments along these lines that may be relevant. Both of these exploit the fact that although nearest neighbors are generically quite distant in high dimensions, they are not if the points lie on relatively low dimensional manifolds. Thus other methods based on nearest neighbors may turn out to be important, too.

How can rare objects be spotted? Can serendipity, so important in the history of astronomy, be automated? This is not just a matter of identifying outliers, although that's important, too.

25.3.5 Time series analysis

The fascinating irregularly spaced time series found in astronomy have have been a stimulus for time series analysis for a long time and challenges remain. The large statistical literature on non-linear time series is rather thin in compelling examples and scientifically plausible analyses and could be enriched and stimulated by confronting such series as those of Miras archived by the AAVSO—see the poster sessions of Foster, Hawkins, and Mattei. Although not discussed in this meeting, I think that the large collections of irregularly spaced time series, such as those of variable stars gathered by micro-lensing surveys (Ferlet et al. 1997) pose methodological challenges for time series. The challenges we have heard in this meeting go beyond one dimensional time series to random fields. How to diagnose or test for non-Gaussianity in the CMB is one example.

25.3.6 Empirical Bayes

Astronomers often do the same type of analysis repeatedly: sources are separated from backgrounds, periods and light curves of variable stars are fit, spectra of similar objects are measured. A basic intent of empirical Bayes procedures (Carlin & Louis 2002) is to "borrow strength" across objects rather than treating each object *de novo*. Large ensembles of similar objects are being measured and astronomers are often more interested in properties of the ensemble than in the individuals. When interested in better estimating individuals, strength can be borrowed from the ensemble.

The "empirical" in "empirical Bayes" refers to the fact that these procedures attempt to estimate the prior distribution of the ensemble. Estimates of individuals are then constructed using this prior. Suppose one has a noisy measurement of an object of interest: $Y = O + N$, that a collection of templates, T_i, have been empirically constructed for such objects, and that the templates have *a priori* probabilities $P(T_i)$. Then an estimate of the object of interest would be

$$E(O|Y) = \frac{\sum_i T_i P(Y|T_i) P(T_i)}{\sum_i P(Y|T_i) P(T_i)}. \tag{25.3}$$

To make these notions more concrete, consider an idealized version of the problem of estimating a periodic function from noisy data, where the period is effectively known. The function might be the light curve of a Cepheid, as in the presentation of Berger *et al.* Suppose that there is a whole collection of Cepheid light curves of interest. For simplicity of notation, suppose that there are n time points equally spaced over phase and corresponding observations $Y = (Y_1, \ldots, Y_n)$. (For a more general setup see Rice & Wu 2001). Consider fitting the function as a Fourier series

$$f(x) = \sum_k [A_k \cos(2\pi k x) + B_k \sin(2\pi k x)] \tag{25.4}$$

where the series is truncated at some point (for simplicity, the mean is taken to be 0). If the measurement errors are modeled as independent with means zero and variances σ_e^2, the ordinary least squares estimate of a Fourier coefficient is

$$\hat{A}_k = \frac{1}{n} \sum_{j=1}^{n} Y_j \cos(2\pi k j / n). \tag{25.5}$$

Taking the point of view that the light curve at hand is drawn from the ensemble, one could estimate A_k by $E(A_k|Y)$, the computation of which would involve the distribution of A_k over the ensemble. Alternatively, one might consider the best linear approximation to this quantity. In a linear empirical Bayes analysis, the variance parameters σ_e^2 and σ_k^2 are estimated from the entire collection of light curves. The linear empirical Bayes estimate of A_k is then

$$E(A_k|Y) = \hat{A}_k \frac{n\sigma_k^2}{n\sigma_k^2 + \sigma_e^2}. \tag{25.6}$$

The ordinary least squares estimate is thus damped by the ratio of variances, so that high frequency terms with small variances (i.e. those that are typically small) will not contribute much the the estimate of $f()$, especially if n is small The amount of damping, or tapering, of the Fourier series is determined empirically by the collection of curves at hand.

25.3.7 Contemporary nonparametrics

There is a large literature on nonparametric function estimation: the point of view of this area is that the parameter to be estimated is infinite dimensional, typically a function. Wasserman *et al.* gave some examples in their presentation. There has been extensive work on how to choose smoothing parameters automatically.

There has been an explosion of research in high dimensional nonparametric function estimation, discrimination, and clustering, often referred to as "machine learning" in the computer science literature. See the contribution of Breiman to this meeting. Astronomers are generally aware of neural nets and decision trees, but there have been other interesting recent developments, such as support vector machines, bagging, boosting, and graphical models.

I would also like to note developments in semi-parametric estimation (Bickel et al. 1993) in which one is interested in estimating both infinite dimensional and scalar parameters. For example, the problem of estimating the period light curve of a variable star can be viewed in this way—the light curve is infinite dimensional and the period is a scalar parameter. See Hall et al. (2000) for a detailed analysis.

25.3.8 Model selection

The recent statistical literature is marked by an increasingly explicit recognition that models are approximate and not given *a priori*. Rather, there is typically a subtle interplay between data analysis and a set of potential models. Model selection and model averaging (George 2002) are active areas of research in statistics that are likely to have an impact in astronomy. It is interesting that in this meeting we heard about the use of model averaging from two quite different perspectives, those of Berger and Breiman. (Of course, the fundamental activity across the spectrum of statistical methodology is figuring out how to average effectively.)

25.3.9 Statistical computing

There have been recent interesting developments in statistical computing which may well be useful for astronomers. We heard about ggobi from Diane Cook. I recommend that astronomers also check out two open-source projects: R (http://www.r-project.org/) and Omega (http://www.omega-hat.org/). The computational demands posed by modern astronomy will also hopefully act as an impetus for further developments in statistical computing. Are we ready to compute on "the grid?"

This list is hardly exhaustive. For wider coverage see the recent collection of very readable vignettes (Raftery et al. 2002) which covers a number of areas of current research in statistics.

In summary, the data revolution in modern astronomy offers a rich feast for statisticians, of whom there are relatively few working in the area. It is essential for statisticians to be open-minded, flexible, and creative since, in the words of Leo Breiman, "To a man who only has a hammer, every problem looks like a nail." There are lots of problems out there, spread over the entire statistical spectrum, and not many of them are best viewed as nails.

25.4 Conclusion

Contributions of statistics to the analysis of astronomical data are not likely to be limited to straight technology transfer. The discipline of statistics thrives on being confronted with new problems and there are fundamental perspectives underlying statistical methodology that can hopefully be brought to bear to address the exciting challenges of modern astronomy.

The most important and enduring contributions of statistics to astronomy, and of astronomy to statistics, are likely to flow from long, close collaborations. It takes a long time for a statistician to appreciate the underlying theory, the measurement process, the vocabulary, the contexts of particular problems, what is really important to the science of astronomy, and thus what statistical approaches will be most fruitful. Similarly it takes a long time for an astronomer to understand the language of statistics and the intuition and heuristics underlying contemporary statistical methods. Establishing such collaboration is not easy, however: it takes a great deal of time and patience, and time is scarce in our too-busy professional lives.

There is a strong basis for collaboration. The two disciplines have been linked for centuries during which probabilistic ideas have been central to astronomy. They continue to evolve in parallel: now both are confronting the world of massive data sets. The institution of the National Virtual Observatory (Brunner et al. 2001) will hopefully bring us closer together. I hope that we also have more meetings like this one!

References

1. P. J. Bickel, C. A. J. Klaasen, Y. Ritov and J. A. Wellner, 1993, *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: JHU Press

2. J. R. Bond, A. H. Jaffe and L. Knox, 2000, 'Radical Compression of Cosmic Microwave Background Data', *Astrophys. J.*, 533, 19-37

3. R. J. Brunner, S. G. Djorgovski and A. S. Szalay, 2001, *Virtual Observatories of the Future*, San Francisco:Astron. Soc. Pacific. Conf. Ser. 225

4. B. P. Carlin and T.A. Louis, 2002, in *Empirical Bayes: past, present, and future*, London:Chapman & Hall/CRC, 312-318

5. Roger Ferlet, Jean Pierre Maillard and Brigitte Raban, 1997, *Variable Stars and the Astrophysical Returns of Microlensing Surveys*, Editions Frontieres

6. E. I. George, 2002, in *The variable selection problem*, London:Chapman & Hall/CRC, 350-358

7. P. Hall, J. Reimann and J. Rice, 2000, 'Nonparametric estimation of a periodic function', *Biometrika*, 87, 545-557

8. A. E. Raftery, M. A. Tanner and M. T. Wells, 2002, *Statistics in the 21st Century*, London:Chapman & Hall/CRC

9. C. Reinsch, 1967, 'Smoothing by spline functions', *Numerische Mathematik*, 10, 177-183

10. J. A. Rice and C. O. Wu, 2001, 'Nonparametric mixed effects models for unequally sampled noisy curves', *Biometrics*, 57, 253-259

11. J. Rissanen and B. Yu, 2002, in *Coding and compression: a happy union of theory and practice*, London:Chapman & Hall/CRC, 229-236

12. S. Roweis and L. Saul, 2000, 'Nonlinear dimensionality reduction by locally linear embedding', *Science*, 290, 2323-2326

13. M. Tegmark, A. Taylor and A. Heavens, 1997, 'Karhunen-Loeve eigenvalue problems in cosmology: how should we tackle large data sets?', *Astrophys. J.*, 480, 22-35

14. J. B. Tenenbaum, V. de Silva and J. C. Langford, 2000, 'A Global Geometric Framework for Nonlinear Dimensionality Reduction', *Science* 290, 2319-2323

15. G. Wahba, 1983, 'Bayesian "confidence intervals" for the ross-validated smoothing spline', *J. Roy. Stat. Soc. B*, 45, 133-150

16. G. Wahba, 1990, *Spline Models for Observational Data*, Soc. Indust. Appl. Math. vol. 59

17. G. Wahba and S. Wold, 1975, 'A completely automatic French curve: fitting spline functions by cross-validation', *Comm. in Statistics*, 4, 1-17

26

An Astronomer's Perspective on SCMA III

Joseph Silk[1]

26.1 Introduction

This has been a remarkable conference. Statisticians and astronomers have addressed each other and explored some of the current issues at the frontiers of their respective fields. It is clear that the two communities have much to learn from each other, and that now is an especially opportune time to explore more extensive collaborations.

I am somewhat of a novice in statistics, and it has been a revelation to hear the continuing vigorous and occasionally acrimonious debate between schools of statisticians that centers on rival methodologies. The only comparable battle in physics is that between the rival interpreters of quantum theory. Here the debate becomes especially shrill when it focuses on Schrodinger's cat which has a 50 percent probability of being alive or dead, when exposed to a radioactive decay-induced trigger of poison gas. The macroscopic world is unhappy with a zombie-like cat that is neither alive nor dead, and this dilemma has led to a still-unresolved ontological and metaphysical crisis in the interpretation of the quantum theory. I spent restless nights at this meeting grappling with the discords between the frequentists and the Bayesians. When I did succeed in sleeping, I had the following dream.

A controlled experiment was being performed to elucidate the reactions of a carefully selected representative group of statisticians and astronomers. An image of a swan was flashed on a giant screen. The swan was white but covered with large black spots. The audience was then interviewed on their reactions.

The astronomers present divided into two groups. The theorists said: let us adopt a model of a spherical swan. A gaussian-distributed field of expanding black dots is added to the swan. The swan is now evolved forwards in time. We conclude that swans are borne white and die black. The observers took a different tack. They were quite conservative. At least one side of one swan has black spots, they said. We need a much larger sample. We

[1]Astrophysics, Oxford University

will write a proposal to our funding agency to provide us with the resources to conduct a full-multidimensional survey of all the swans in Pennsylvania.

The statisticians also divided into two groups. The frequentists concluded that based on the one swan sample, there was at least a fifty percent probability that all swans had black spots. The Bayesians, on the other hand, began with the prior that swans are white, since this confirmed to their previous experience, and concluded that either this is a very sick swan, or someone has been playing a joke by painting black spots on the swan. Curiously, the one ornithologist in the audience was in complete agreement with the Bayesians.

In fact there is a sociological analogy between the communities of statisticians and astronomers. Astronomers divide into three types: observers, theorists and data analysts. The observers also occasionally but too rarely build instruments. The theorists further subdivide into fundamentalists and phenomenologists. The data analysts are a relatively recent breed of theorist who are having a difficult time in being accepted into the traditional community of physics and astronomy departments, being neither fish nor foul, not completely acceptable as either observer or theorist.

Similarly, statisticians divide into frequentists, for whom theory is relatively unimportant or even abhorrent, and Bayesians, who begin by laying down subjective priors that are essentially of theoretical or empirical origin. There are also the data analysts, who despair with the rival philosophies, and adopt a hybrid approach. I cannot tell whether the latter class of statisticians is meeting similar resistance in their own community as are their astronomer counterparts, but I can assure them that the astronomical community would welcome them with open arms.

There is another philosophical difference between the fields of astronomy and statistics. To an astronomer, a statistician seeks the most efficient method of joining up the dots in multidimensional parameter space. To a statistician, the astronomer is diverted by his obsessions with the urge to answer fundamental questions such as how and why the observed structures originated and evolved. There is clearly fertile ground to be ploughed in the terrain that separates the two communities.

26.2 Sociology

I classified the abstracts of papers presented at this meeting into the following categories: galaxy surveys (23) predominated, followed by x-ray sources (11) and stars (7). The cosmic microwave background radiation had 4, and there were a half dozen that I classified as generic. The moral is clear: statisticians are about 5 years behind the "hot" areas that the astronomers are currently developing. There is progress of course: if we had had this meeting a decade ago, stars would have been the dominant category. One

decade from now, the cosmic microwave background will be the dominant field, but the more astute statisticians may wish to move more quickly. They should jump now!

26.3 Highlights

I can best summarize some key points that emerged from the meeting by quoting the proponents. Bayesians have to be careful about their choice of priors. It is by no means obvious that, for example, a smooth prior is suitable for spiky data, in which only rare spikes may actually be real. The use of improper priors was summarised thus by Eric Kolaczyk: "Garbage in, garbage out."

Astronomers historically have been obsessed by classification, from Charles Messier onwards. But astronomers must not follow too closely the example of the botanists, as Jeffery Scargle stressed in his talk: "Classification is not an end in itself."

Galaxy clusters are a modern example of intensively studied astronomical objects that are confusingly rich in morphology. The more wavelengths that are studied, from the optical (galaxy counts and gravitational shear maps), to x-ray and microwave (intracluster gas) and radio (high energy electrons and intracluster magnetic fields), the more complex a picture emerges. Rien van de Weygaert reminded us: "don't take any one cluster seriously." And when it came to structural analysis of clusters, substructure is prominent. This means, as emphasised by Adrian Raftery, that counting the space density of clusters is a non-trivial problem for the Bayesians: one has to incorporate "the peas versus the pod: priors on shape affect the cluster density."

26.4 Cosmic Microwave Background Radiation

Vast data sets are envisioned from future surveys of the cosmic microwave sky. The science is quickly summarized. We are studying temperature fluctuations on the sky. These are the fossil seeds of large-scale structure. The universe was opaque for the first 100,000 years, during which the cosmic blackbody radiation thermalized and inflation generated quantum fluctuations on the macroscopic scales that characterize galaxies and large-scale structures. Linear gravity amplified the fluctuations once the universe was matter-dominated, and gravitational instability generated the observed structures during the ensuing transparent phase of the expansion. Imprinted on the "surface" of last scattering of the cosmic microwave photons are the fossil fluctuations which later seeded structure formation. Detection of these fluctuations is an immense challenge: the amplitude is

only 0.001 percent of the 2.736 Kelvin cosmic blackbody radiation that one observes as a weak all-sky glow in a terrestrial environment at 300 Kelvin, in competition with dominant atmospheric, solar system, galactic and extragalactic backgrounds.

Satellites have provided all-sky maps that are essential for adequate separation of the foregrounds, beginning with the discovery of the fluctuations by the COBE satellite in 1992 in an experiment that had some 6000 pixels. Current data sets such as that of the BOOMERANG balloon experiment have some 10^5 pixels with coverage of at most a few percent of the sky. The near future will see release of all-sky data by the MAP satellite in Jan 2003 on some 10^7 pixels on the sky, to culminate with the Planck satellite, to be launched in 2007 and with 10^8 pixels on the sky.

Such massive data sets require an unprecedented analysis effort. Map analysis requires inversion of a correlation matrix, since the pixels are correlated on the sky, and the computing time scales as N^3, where N is the number of pixels. One has to work in at least a 10-dimensional data space for extraction of the cosmological parameters that are the principal goal of these experiments. Even here, the input prior is highly simplified. For example a gaussian random field is adopted for the initial density fluctuations as is a smooth power spectrum. The cosmological model imprints gaussian ripples due to the effect of sound waves in the primordial baryonic plasma, and these manifest themselves as a series of peaks and troughs. The prior must necessarily be structured if one wants to optimize one's input from basic physics.

Non-gaussianity gets us further away from known models. Very few specific models are available that incorporate nongaussian initial conditions. This is non-trivial: it is all very well to search for a lost dog, but how do we locate a non-dog? Four of the posters were devoted to cosmic microwave background radiation issues, and nongaussianity was a key theme. I draw your attention to discussions of a neural net approach to the search for patterns on the sky, and to detection algorithms based on wavelet decomposition on the sphere.

To search for non-gaussianity, one needs to utilize information both on the power spectrum and on the phases of the Fourier decomposition of the sky map. In effect, one is looking for patterns on the sky. Consider the following example of a non-gaussian pattern on the sky that is motivated by the possible topology of the universe. Cosmology theory says nothing about topology. Nor does quantum gravity address global issues such as topology. Space is known to be approximately Euclidean, from determination of the angular scale of the first peak in the cosmic microwave background at $\ell = 210 \pm 15$. The total density is inferred to be close to critical: $\Omega = 1.00 \pm 0.04$, and the universe is flat to within a few percent uncertainty.

The inferences for topology are remarkable. Naively, we approximate the geometry of a flat universe by the 2-d analogy of an infinite sheet. In two dimensions, there are five topologies for flat space: the sheet, cylinder, torus,

Mobius strip and Klein bottle. In fact, in a flat 3-d universe there are 18 possible global topologies. Some are compact in only one direction, such as the surface of a cylinder. Others are fully compact, such as the hypertorus. In fact, only 10 of these allowed flat 3-d topologies are compact. Of these 10, only 6 are orientable.

Now orientability is a necessary condition for physical cosmology, as it guarantees conservation of parity under time translation. The universe has also been argued to be compact from considerations of quantum cosmology, as might be required by invoking the probability of universe creation from a principle of least action or of 3-space generation via higher dimensional compactification. With a relatively limited number of motivated options, one can begin to test the hypothesis of compactness via its signature in the sky.

In a topologically small universe, light rays circumnavigate the universe, so that ghost images are generated. More generally, light rays can take different paths in different directions between any two points. This leads to anisotropy patterns in the microwave sky as viewed on the surface of last scattering of the photons. Consider the surface of a torus, the two-dimensional analogue of one of the flat topologies. This is a compact surface that is globally anisotropic, unlike an infinite sheet. A pattern on the sky is inevitable even if the radius of the torus is large compared with the horizon. There are a limited number of flat, compact topologies to explore. Each induces a characteristic and distinct pattern on the microwave sky. Pattern searches will require the high resolution of future experiments. Sophisticated statistical techniques will be needed to pick out patterns from the noise, both foreground and cosmological.

26.5 Large-scale structure

Analysis of galaxy surveys was the dominant theme of the posters at this meeting. The aim is to discern the pattern of the evolution of the universe. Structure develops inexorably with time, as correlations grow stronger betwen neighbouring fluctuations. As one looks back into the universe, one should be able to distinguish, via the correlations, between the rival cosmological models. For example, in a low density universe, gravity is less important at recent times than in a universe at critical density. Hence, for example, the correlation length is expected to change less, as is the number of rich clusters, with epoch as one surveys the universe back to a redshift of 1 or 2.

The first question one might ask is whether gravity is necessary. Can purely geometrical models account for large-scale structure? Fractal models have been tried, and they fail. We see homogeneity on a scale of 100 Mpc and beyond, in complete contradiction to a fractal universe of mean

density zero, as explained at the meeting by Martinez. Another promising direction that merits exploration is that of Voronoi foam models for the galaxy distribution. Van de Weygaert produced some impressive maps of large-scale clustering of galaxies. The jury remains out however until predictions of higher order correlations are forthcoming.

If the 3-point correlation of Voronoi foam matches what is observed, I will be amazed. All indications are that the universe is assembled under the action of gravity. Simulations demonstrate that gravity is capable of reproducing the observed structure, and have had excellent success at accounting for the data. All indications are that gravity is responsible for the observed structure. To paraphrase one eminent Princeton cosmologist, if it looks like a duck and quacks like a duck, it surely is a duck.

The results that have most stirred up cosmologists in recent years centre on the observations of high redshift supernovae. At redshift unity, SNIa are about 20 percent dimmer than expected if they have the same intrinsic luminosities as their nearby counterparts. The only interpretation to date has been that the universe is accelerating due to the current domination of dark energy, usually interpreted as the cosmological constant.

In practice, the SNIa measure acceleration via the difference between the mean density of accelerating dark energy Ω_Λ and decelerating dark matter Ω_m. The observations require $\Omega_\Lambda - \Omega_m \approx 0.4$. In practice, one needs independent information to proceed further. Type Ia supernovae, while excellent standard candles, are poorly understood in terms of physics. Hence alternative probes of acceleration are important.

From large-scale structure surveys, one can infer Ω_m, and thereby derive the cosmological constant. One such approach was presented in the poster by Matsubara and Szalay. Their idea is to use the 10^5 bright red galaxies in the Sloan Digital Sky Survey to look at the differential evolution of the clustering of galaxies to $z \sim 1$. The cosmic microwave background provides complementary information, since the curvature of space is the sum of Ω_Λ and Ω_m. This has been measured, as already mentioned. The onus is then on large-scale structure to come up with equally convincing evidence for Ω_m, so that one would have redundancy and presumably greater confidence in the result.

One approaches this goal via the large redshift surveys. These provide a three-dimensional probe of the universe. With precise redshifts, one can map out the peculiar velocity field and thereby determine a combination of Ω_m and the amplitude of the power spectrum, $\Omega_m^{0.65}\sigma_8$, where σ_8 is the ratio of the dark matter to luminous matter fluctuation variance, normalized to a fiducial scale.

The difficulty one immediately encounters is that galaxies are complex systems, whose fundamental properties and correlations are not well understood. One has to decide whether galaxies are indeed mass tracers. If so, then what type of galaxy is most reliable? The clustering length is empirically found to depend on galaxy luminosity, dwarfs being less clus-

tered. Galaxy classification may be influenced by local parameters, including age and gas content as well as morphology, and on non-local parameters such as environment and distance. This means one has to work in a multi-dimensional parameter space.

There are already known trends in the three-dimensional space of luminosity, size and rotation velocity (for spirals) or velocity dispersion (for ellipticals). Optimal projections lead to dispersions as low as 15 percent. Clearly this is merely the tip of the iceberg. The existing correlations utilize catalogues containing tens of thousands of galaxies. Surveys underway are obtaining much larger samples: 250,000 for 2DF and 10^6 for SDSS out to a redshift of 0.2. These surveys will have the spectral quality and depth to probe a higher-dimensional parameter space than hitherto attempted. Future surveys of 100,000 galaxies or more are imminent for the distant universe at $z \sim 1$, so that evolution of galaxy properties will also be studied in detail.

Hitherto, the large galaxy redshift surveys have mostly been restricted to the optical wavelength band. Infrared surveys have typically include 10,000 galaxies to date. The optical SDSS of 10^6 galaxies will have a data volume of 0.2GB. However the VISTA telescope, under construction, will map the entire Southern sky in the near infrared and have an anticipated data volume of 10TB. Data visualization and data mining are areas where astronomers will have much to learn from the statistics community in analyzing the anticipated data flood.

26.6 Stars

While galaxies are the building blocks of the universe, stars are the indispensable building blocks of galaxies. The nature of stars and their statistics are reasonably well understood. However astrometric data, thanks to the Hipparcos satellite, gives adequate coverage only of the solar vicinity, out to 100 pc or so. This situation will change dramatically in 2010 when the GAIA satellite will be launched. Over five years, GAIA will repeatedly observe a billion stars with 5μarc-second precision. GAIA will measure the systemic and random components of star motions throughout the Milky Way galaxy. GAIA will provide an unprecedented data resource. Armed with precise measurements of the locations and motions of stars throughout our galaxy, we will be able to reconstruct our past and predict our future.

An important secondary project for GAIA will be the search for near–Earth asteroids. These could potentially be life-threatening to the Earth. Quick-look data from GAIA will provide an unexcelled means of locating near–Earth asteroids and determining their orbits while they are still far from the Earth. The data archives are immense. The GAIA data volume is

100TB. Statisticians are needed now to bring in novel ways of addressing this data mining challenge via applications to simulated data sets.

26.7 The future

Now is the time to prepare for the new confrontations posed by the astronomical data anticipated over the next decade. The microwave background maps hold clues to our origins, the large redshift surveys will shed light on our evolution. Analysis requires not only immense dedicated computational power, but development of novel algorithms and statistical approaches. Multidimensional parameter space confronts us, and we have to learn how to project this into digestible forms. Multiwavelength analysis is essential for optimising our cosmological parameter extraction procedures. For example, correlating the CMB and galaxy reshift surveys will remove degeneracies that would otherwise plague our analysis.

New ideas are urgently needed. Several were presented at this meeting. Non-parametric Bayesian modelling of data was described by Scargle, as well as by Wasserman. New approaches to multicale methods were discussed by Starck and Kolaczyk. New tools for visualizing multi-dimensional data were presented by Cook. These are just a few of the ideas going the rounds.

What is also clear is that scalability is going to be an issue. For SCMA IV, I suggest that the computer scientists be brought in, as they are best equipped to show us how to surmount this important hurdle, an essential step before we plunge into some of the truly massive data sets that astronomers can simulate now and are soon to appear. Astronomers are proposing ever more grandiose data-taking devices. One under discussion for the VLT will involve the data flow in one or two nights that is equivalent to the entire five year SDSS data volume. The statistical issues that need to be addressed involve visualization, compression, and mining of massive data sets. Within a year of data taking, most data is released into the public domain, so that the opportunity is there for all. The challenges are immense, but so are the potential rewards. This is an opportune moment for statisticians to be exploiting astronomers, and vice versa.

27

Ensembles of Classifiers

D. Bazell[1]

27.1 Introduction

Neural networks and decision trees are the two most commonly used classification methods in astronomy. With both of these methods classification is performed by presenting the algorithm with a training data set consisting of a set of objects that have been previously labelled with a class. The algorithm then tries to produce classifications of the training set objects that agree with the predefined class labels. Once the algorithm classifications and the class labels agree to a certain level of accuracy the learning process is halted and the internal state of the algorithm is saved. We call this a classifier. New objects which have never been seen by the classifier can be labelled using this classifier.

The neural network and decision tree approaches to morphological galaxy classification that have been used to date all rely upon using a single classifier to predict the class of an unknown object. However, ensembles of classifiers can be used to combine the predictions of several individual classifiers to produce a new classifier that often has lower classification error than the individual constituents. In this paper we examine the creation of ensembles using bootstrap aggregation [Breiman1996] of three types of classifiers: neural networks trained with backpropagation, and two decision tree induction algorithms.

27.2 Methods

An ensemble of classifiers can be implemented in a variety of ways. One is to train several individual classifiers whose output decisions can be combined (typically by voting or averaging) to allow classification of new inputs. Bagging is one of the easiest ensemble methods to implement since it only involves resampling of the original data. This algorithm creates the different classifiers by training them on bootstrap replicates of the original training set. Each classifier's training set is created by randomly sampling, with replacement, N examples from the original training set, where N is the

[1]Eureka Scientific, Inc.

number of examples in the original training set. Some examples will appear more than once in the bootstrap replicates while others will not appear at all. When an individual classifier is trained, its overall error may be higher than for a classifier trained on the original training set. However, because the ensemble is created by voting the predictions of each classifier for each test set example, if a plurality of the classifiers make the correct predictions the ensemble will make the correct prediction. In this manner the voting can overcome the increased overall error on the part of individual classifiers.

AdaBoosting [Freund and Schapire 1996] also starts by resampling the data with replacement. However, each input example also has an assigned weight, with all weights initally being uniform. After each training iteration, AdaBoost looks at each example and determines if it was correctly or incorrectly classified. If a given example was incorrectly classified, then its weight is increased for the next iteration. This is effectively the same as increasing the number of times this example is presented to the training algorithm compared to the other examples. This reweighting and retraining takes place of a number of iterations, until the overall training error is reduced below a preset threshold or the maximum number of iterations is reached.

Both bagging and AdaBoosting can be applied to any training algorithm. We have implemented a Perl script for each ensemble method that takes the training algorithm as an input parameter and produces output files in a common format for each algorithm. These scripts are available upon request.

27.3 Results

We ran tests with two data sets created previously by [Naim et al. 1995] and [Storrie-Lombardi et al. 1992] using three classifiers. Our tests using bagging show that we can reduce the classification error by up to 16% for decision tree classifiers but only a few percent for the neural network. The results for AdaBoost were less spectacular, with only a 12% decrease in classification error. Preliminary results were reported in [Bazell and Aha 2001] with a more detailed exposition in preparation.

27.4 References

[Bazell and Aha 2001] Bazell, D. and Aha D.W. 2001, The Astrophysical Journal, 548, 219

[Breiman1996] Breiman, L. 1996, Machine Learning, 24, 123

[Freund and Schapire 1996] Freund, Y., and Schapire, R.E. 1996, Proc. of the Thirteenth Intl. Conf. on Machine Learning

[Naim et al. 1995] Naim, A., Lahav, O., Sodré, L., jr., and Storrie-Lombardi, M.C. 1995, MNRAS, 275, 567

[Storrie-Lombardi et al. 1992] Storrie-Lombardi M.C., Lahav O., Sodré Jr. L., Storrie-Lombardi L.J., MNRAS 259, 8

28

A Model for Brightest Galaxies Using Extreme Value Statistics

S. P. Bhavsar[1] and J. P. Bernstein

ABSTRACT We contend that neither a Normal nor a Gumbel distribution describes the brightest cluster galaxy (BCG) magnitudes. A two-population model fits recent data. This model has a physical basis.

28.1 BCG Magnitude Distributions

BCG magnitudes are highly uniform, with a dispersion of only 0.32 magnitudes [4]. Are BCGs a single population, the statistical tail of ordinary galaxies, or a class of special galaxies; or do they consist of two populations comprising a mix of these [2]? We consider five models (A, B, C, D, and E) that include both one-pop and two-pop hypotheses. One-pop models (like A and B below) may have Gaussian or Gumbel distributions depending on whether the BCGs are special or statistical extremes.

$$A : f_{sp}(M) = f_g = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(M-M_g)^2}{2\sigma^2}}. \tag{28.1}$$

$$B : f_{stat}(M) = f_G = a e^{a(M-M^*)-e^{a(M-M^*)}}, \tag{28.2}$$

where the symbols have their usual meaning. For compactness $M^* = M_G + \frac{0.577}{a}$. For two-pop models [2], we consider models C, D, and E comprising the three possible combinations of f_G and f_g, where d represents the fraction of clusters that contain both types of BCGs.

$$C : f_{Gg}(M) = d \cdot [f_g \cdot I_G + f_G \cdot I_g] + (1-d) f_G \tag{28.3}$$

$$D : f_{gg}(M) = d \cdot [f_{g2} \cdot I_{g1} + f_{g1} \cdot I_{g2}] + (1-d) f_{g1} \tag{28.4}$$

$$E : f_{GG}(M) = d \cdot [f_{G2} \cdot I_{G1} + f_{G1} \cdot I_{G2}] + (1-d) f_{G1}, \tag{28.5}$$

where $I_G = \int_M^\infty f_G(M')dM' = F(M)$; and $I_g = \int_M^\infty f_g(M')dM' = (1 \pm erf|M - M_g|)/2$.

[1]Department of Physics and Astronomy, University of Kentucky

28.2 Results and Conclusion

We used data from Lauer and Postman [5] to fit the models [3]. Figure 1 shows the results. Parameters were determined by maximum-likelihood. The P values determined by the K-S test for rejection of models A, B, C, D, E are 16.2%, 92.6%, 15.8%, 9.8% and 1.4%, respectively. We reject the pure Gumbel (model B), the hypotheses that BCGs are statistical extremes. Two-pop models D and E describe the data adequately, but model E stands out as giving the best overall fit. A second population could evolve from bright ordinary galaxies [1,6]. Model E particularly, has a physical basis.

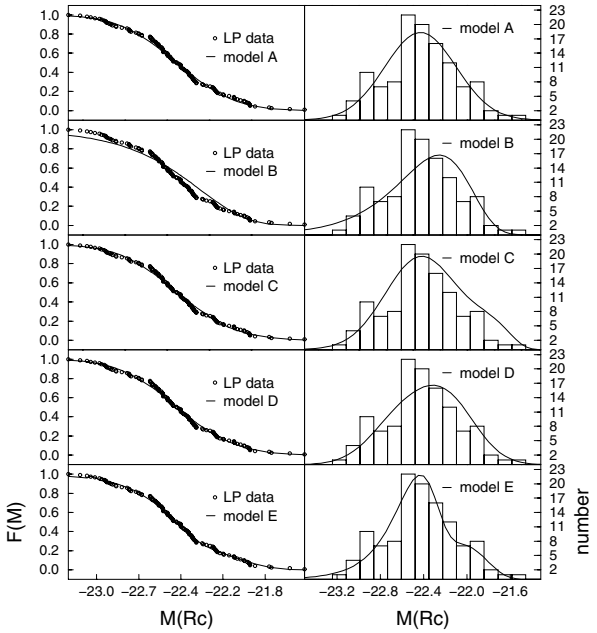


FIGURE 28.1. data and fits for the models, cumulative and histograms

When galaxies with an exponential luminosity function undergo a random boost in magnitude, a Gumbel distribution results [3].

References

Aragon-Salamanca A., Baugh C.M., Kauffmann G., 1998, MNRAS, 297, 427
Bhavsar S.P., 1989, ApJ, 338, 718
Bernstein J.P., Bhavsar S.P., MNRAS, 322, 625
Hoessel J.G., Schneider D.P., 1985, AJ., 90, 1648
Lauer T.R., Postman M., 1994, ApJ, 425, 418
Ostriker J.P., Hausman M.A., 1977, ApJ, 217, L125

29

New Statistical Goodness-of-Fit Techniques in Noisy Inhomogeneous Regression Problems with an Application to the Problem of Recovering of the Luminosity Density of the Milky Way from Surface Brightness Data

Nicolai B. Bissantz[1] and Axel Munk

ABSTRACT Fitting models to regression data is an important part of astronomers everyday work. A common proceeding is based on the assumption, that a parametric class of functions describes the data structure sufficiently well. We present a new method which is applicable in noisy versions of Fredholm integral equations of the first kind, and an associated goodness of fit measure, which works under the assumption that the parametric model in question holds for the data. For this we suggest a bootstrap algorithm which allows an approximation of the distribution of the suggested test statistic.
Then we switch to the assumption that the model under consideration does not hold, and present a method to compare parametric models under this assumption. This second method is based on the same bootstrap algorithm as the first method.
As an example we finally apply our methods to the problem of recovering the luminosity density of the Milky Way from data of the DIRBE experiment on board of the COBE satellite. We present statistical evidence for flaring of the stellar disk inside the solar circle.
Details on our methods can be found in Bissantz & Munk 2001a, 2001b.

29.1 References

[1] Bissantz, N., Axel, A., 2001a, New statistical goodness of fit techniques in noisy inhomogeneous inverse problems. With application to the re-

[1] Astronomisches Institut der Universität Basel

covering of the luminosity distribution of the Milky Way. *A&A* **376**, 735-744

[2] Bissantz, N., Axel, A., 2001b, Comparison of parametric models with the same or a different number of parameters in noisy inhomogeneous regression problems. *In preparation*

30

Measuring the Galaxy Power Spectrum with Multiresolution Decomposition

Yaoquan Chu[1], **XiaoHu Yang, Long-Long Feng, Li-Zhi Fang**

The power spectrum is one of the most important statistical measures to quantify the clustering features of large scale mass density distribution traced by galaxies. Observational data of galaxy redshift survey are rapidly increasing in both quantity and quality, bringing new challenges to data analysis. The standard method for power spectrum estimation is Fourier decomposition of the density field in term of discrete plane wave modes. Here we present an alternative method of measuring galaxy power spectrum based on the multiresolution analysis using the discrete wavelet transformation (DWT). Besides the technical advantages of the computational feasibility for data sets with large volume and complex geometry, the DWT space-scale decomposition provides a physical insight into the clustering behavior in phase space, which are hardly revealed using the Fourier decomposition.

The DWT power spectrum estimator is constructed as following[1][2]. Let $\delta(\mathbf{x})$ be the density fluctuations, the projection onto the multiresolution space spanned by a wavelet basis $\{\psi_{\mathbf{j},\mathbf{l}}(\mathbf{x})\}$ gives the wavelet function coefficients (WFCs) $\tilde{\epsilon}^{\mathbf{l}}_{\mathbf{j}} = \langle\delta(\mathbf{x})|\psi_{\mathbf{j},\mathbf{l}}(\mathbf{x})\rangle$. The estimator of power on the $\mathbf{j}$ scale is obtained by averaging over the $2^{j_1+j_2+j_3}$ measurements in the disjoint volume elements,

$$P_{\mathbf{j}} = \frac{1}{2^{j_1+j_2+j_3}} \sum_{\mathbf{l}} |\tilde{\epsilon}^{\mathbf{l}}_{\mathbf{j}}|^2$$

which is related to the Fourier power spectrum power by

$$P_{\mathbf{j}} = \sum_{\mathbf{n}} W_j(\mathbf{n}) P(\mathbf{n})$$

with the filter

$$W_{\mathbf{j}}(\mathbf{n}) = \prod_{i=1}^{3} \frac{1}{2^{j_i}} |\hat{\psi}(n_i/2^{j_i})|^2$$

[1]Center for Astrophysics, University of Science and Technology of China

The function $\hat{\psi}(n)$ is the Fourier transform of the basic wavelet $\psi(x)$, where $\mathbf{n}$ labels wavenumber by $\mathbf{k} = 2\pi\mathbf{n}/L$. Obviously, P_j represents band averaged power spectrum in the logarithmic spacing of n. For a galaxy distribution, the Poisson sampling effect should be corrected for by subtracting the shot noise contribution.

The DWT estimator can provide two types of power spectra: (1) diagonal power spectrum given by the powers on cubically symmetric modes ($j_1 = j_2 = j_3 = j$); (2) off-diagonal power spectrum given by the powers on other modes, which are more flexible for dealing with complex survey geometry.

We applied the DWT power spectrum estimator to analyze the catalogues of the Las Campanas redshift survey (LCRS)[3]. To assess the accuracy to which the DWT power spectrum is recovered from the LCRS, we performed an analysis for mock LCRS samples extracted from N-body simulation in the CDM family of models. We showed that (1) the slice-like survey geometry in the LCRS does not affect the estimation of the DWT power spectrum in off-diagonal modes. (2) the perturbation powers in the peculiar velocity field which results in redshift distortion are approximately scale-independent. (3) the difference between the diagonal and off-diagonal DWT power spectrum could be employed for measuring the anisotropic velocity fields in galaxy redshift surveys.

Moreover, we measured the DWT power spectrum in the six strips of the LCRS, which is then compared with those from the SCDM, τCDM and ΛCDM models including the effects of non-linear evolution of density perturbations and redshift distortion. We estimated the one-dimensional peculiar velocity dispersion σ_v and redshift distortion parameter β using the least square fitting. It is found that, for instance, in the ΛCDM model, $\beta = 0.46 \pm 0.06$ and $\sigma_v = 250 \pm 72 kms^{-1}$, which are comparable with other estimations using different techniques. The similar results have been found for the IRAS Point Source Catalog Redshift Survey (PSCz)[4]. To account for the redshift distortion effect in DWT representation, we also develop a method of analyzing cosmic velocity fields with a multiresolution decomposition [5].

A full report of this series of works are in Yang et al., references [2]-[5].

30.1 References

[1] Pando, J & Fang, L.Z., 1998, Phys. Rev. E57, 3593.
[2] Fang, L.Z. & Feng, L.L. 2000, ApJ, 539, 5.
[3] Yang, X.H., Feng, L.L., Chu, Y.Q. & Fang, L.Z. 2001, ApJ, 553, 1.
[4] Yang, X.H., Feng, L.L. & Chu, Y.Q., Chin. J. Astron. Astrophys., 2001, 1, 200.
[5] Yang, X.H., Feng, L.L., Chu, Y.Q. & Fang, L.Z. 2002, ApJ, 566, 630.

31

Finding Gamma-Ray Pulsars with Sparse Bayes Blocks

A. Connors[1] and A. Carramiñana

ABSTRACT Beamed radiation from rapidly spinning (periods $\sim 1 - 100$ms), highly magnetized ($\sim 10^12$ gauss fields) neutron stars, or *pulsars*, is notoriously difficult to find in γ-rays. First, one may have to wait > 10 minutes between photons even with a large gamma-ray telescope like CGRO-EGRET viewing a bright source like the Vela pulsar. Second, these γ-ray light-curves (brightness versus time or phase) are very sharply peaked. Current methods (Z_n^2 – [1, 2] and references therein) are carefully studied and well understood but use Fourier components — a bad match to this shape of light curve. Binning a light-curve into increasingly narrow bins then testing for flatness can introduce many free parameters and hence lower detection thresholds. So, why not use a statistic that more directly represents the sharp changes in a pulsar lightcurve? Why not let the data themselves (plus any prior knowledge) set the optimal size of a very few bins? This is what we have done. We test a modified "Bayesian Block" [3] method on simulated light-curves with a variety of signal–to–noise–ratios. Preliminary results are encouraging, showing the "Sparse-BB" method more powerful for detecting very "spiky" light-curves.

31.1 Introduction

Given any (class of) models, Bayesian Inference prescribes how to derive the statistic with the best measure of all the information in the data. Conversely, any likelihood statistic can be thought of as embodying the information in a class of models (e.g. Lomb-Scargle periodograms and Fourier series; or Z_n^2 and n-component exponentiated Fourier series [4, 5, 6, 7]. In this paper, we introduce a new method derived specifically for $\gamma-$ray pulsar detection. The underlying model is extraordinarily simple: one (or a very few) blocks of arbitrary width, phase, and rate ("Bayes Blocks"[3]).

The extremely coherent periodic signals characterizing *pulsars* (rotating neutron stars with over the mass of the sun compressed into $\sim$10 km and magnetic fields compressed to $\sim 10^{12-14}\times$ that of the Earth) have been

[1]Eureka Scientific

detected at all wavelengths, with periods ranging from ms to seconds. The massive, rapidly spinning fields of younger pulsars are thought to power not only the most energetic photons (γ-rays), but also the most energetic of the particles (cosmic rays) bathing the galaxy.

31.2 EGRET sources and γ-ray pulsars

Only a handful of pulsars have been detected in γ-rays, most by the EGRET telescope on board of the *Compton Gamma-Ray Observatory (CGRO)*. It performed the first all-sky survey at photon energies above 100 MeV [8], and the most complete database for γ-ray astronomy for years to come. Of the 271 objects included in the 3rd EGRET catalog, five are identified with radio pulsars, 93 with blazars (about a third with low-confidence), and 163 remain unidentified. The distribution of this unknown source population indicates that most belong to the Milky Way [9]. Although most pulsars are first found at radio wavelengths [8, 10, 11, 12]. the discovery that one of the brightest (Geminga) is a nearby radio-quiet γ-ray loud pulsar [13, 14]. strongly suggests that others may be the same (e.g. [15]).

In all cases timing of the γ-ray data has been performed using Fourier based analysis. It is conceivable that some tentative associations have not been confirmed because the light curve has very narrow components. As Z_n^2 analysis of EGRET data gets close to exhaustion, fresh methods to test for narrow peaks in light curves might give new light to γ-ray pulsars.

31.3 Methods: Sparse Bayes Blocks

Ideally, intuitively, one seeks the simplest method that captures the significance of one (or two) peaks (plus perhaps a bridge) of arbitrary narrowness and height. But this 'intuition' is straightforward to quantify via *change-points*: one (or a very few) 'Bayes Blocks' of arbitrary placement, width, and height [3]. Once this class of models is specified, it is straightforward to derive an optimum statistic (likelihood ratio) via Bayesian probability theory. We restrict it to a very sparse number of 'Bayes Blocks' for speed, simplicity, and greater detection power (see [16] for a discussion of the "Ockham's Razor" penalty built into Bayes Odds ratios). This new method we propose, 'Sparse Bayes Blocks', then includes both the high resolution of finely-binned epcoh–folding [17, 18] and the fewer (implicit) parameters of Z_n^2 [2, 1]), in a fully Poisson way.

We step through the Bayesian procedure (data; null and interesting hypotheses; priors; posteriors; likelihood ratios) below (see [19] for details).

31.3.1 Bayes Applied to Sparse bayes Blocks

Data. The data are intrinsically Poisson: lists of arrival times (plus energies, positions, data quality indicators, etc) measured by the instrument. For pulsar (i.e. period) searches, each arrival time is carefully mapped back to the geometric phase ϕ of the known rotating pulsar [1, 2].

Null hypothesis: Zero changepoints, $\mathcal{H}_0$. For the null hypothesis we assume the rate r_o is constant. Hence the expected total number of counts μ_0 is given by: $\mu_0 = r_{TOT}T_{TOT}$, where T_{OT} designates the total instrumental livetime during the observattion.

Priors, 0. We used an exponential prior on the overall rate r_T, using the inverse of the expected average rate (from previous measurements) as the scale factor β [22, 19].

Simplest Interesting hypothesis, $\mathcal{H}_2$. The model rate r_n is piecewise constant. For a single block (i.e. two changepoints, ϕ_1, ϕ_2) the expected counts $\mu(\phi)$ in each are: $\mu(\phi) = r_1T_1,\ \phi \in (\phi_0, \phi_1]; \ = r_0T_0$ otherwise.

Priors, I. We used individual exponential priors on the model rate for each component (again with scale factor β the inverse of the previously measured average rate). This is a fairly conservative assignment: equivalent to testing for a *new component*. (In the next subsection we compare it with a different prior: assuming beforehand that a source of this flux exists, but that the shape of its (periodic) light-curve is unknown.) For the changepoints, we used a prior $\pi(\phi)$ that is constant in phase (that is, one that is invariant with respect to translations in phase): $\pi(\phi_n|I)\mathrm{d}\phi_n = \mathrm{d}\phi_n$.

Posterior likelihoods, I. Now it is straightforward (if tedious) to 'turn the crank' to obtain the posterior for the null and interesting hypothesis (see [19] for explicit details). Marginalizing over the unknown rates and taking their ratio produces a nice form for the Bayes likelihood ratio as a function of changepoints (ϕ_0, ϕ_1):

$$\Lambda(\phi_0, \phi_1|\mathcal{H}_2, \mathcal{H}_0, I, \{y_i\}) = \frac{\Gamma[Y_1 + 1]\Gamma[Y_{02} + 1]}{\Gamma[Y_{TOT} + 1]} \frac{\beta(\beta + T_{TOT})^{(Y_{TOT}+1)}}{(\beta + T_1)^{(Y_1+1)}(\beta + T_{02})^{(Y_{02}+1)}}.$$

This maps out the likelihood of the changepoints. To find global (or total) odds O, or Bayes factor, of $\mathcal{H}_2$ (one peak, two changepoints) versus $\mathcal{H}_0$ (flat, no changepoints) we marginalize (i.e. numerically integrate) the expression above over all changepoints (ϕ_0, ϕ_1):

$$\mathcal{O}(\{y_i\}) \ | \ \mathcal{H}_2, \mathcal{H}_0, I) \ = \ \int \mathrm{d}\phi_1 \mathrm{d}\phi_2 \Lambda(\phi_0, \phi_1 \ | \ \mathcal{H}_2, \mathcal{H}_0, I, \{y_i\}).$$

Priors, II. This first result has a dependence on the prior parameter β (the inverse of the average rate determined from prior measurements). This is not the case when the problem can be formulated as a question of unknown fractional shapes rather than an unknown extra component.

Rephrasing the interesting hypothesis $\mathcal{H}_2$: Let the total rate be r_T. The fraction of the total counts in the peak is f_1, while the fraction outside

the peak is f_{02}, with constraint $f_1 + f_{02} = 1$. The expected number of counts in each time (or phase) bin δt_i is then:

$$r_i = r_T T_{TOT} \delta t_i \big(f_1/T_1\big), \text{ for } \phi_i \in (\phi_0, \phi_1]; \ = r_T T_{TOT} \delta t_i \big(f_{02}/T_{02}\big) \text{ otherwise.}$$

As before T_1 and T_{02} represent the livetimes accumulated in the peak and background sections, respectively.

Rephrasing the priors. The prior on the total rate has the same form as before: $\pi(r_T \mid I)\mathrm{d}r_T = \mathrm{e}^{(-\beta_T r_T)}\beta_T \mathrm{d}r_T$. However the prior on the fractional rates is new. It is uniform on $[0,1]$ with the constraint that both sum to unity: $p(f_1|I)\mathrm{d}f_1 = \mathrm{d}f_1$, $p(f_1|I)\mathrm{d}f_1 = \mathrm{d}f_1$; with $f_1 + f_{02} = 1$.

Alternate Posterior Likelihoods. Marginalizing and taking a ratio gives our second likelihood ratio:

$$\Lambda(\phi_0, \phi_1 \mid \mathcal{H}_2, I, \{y_i\}) = \frac{{T_{TOT}}^{Y_{TOT}}}{{T_1}^{Y_1}{T_{02}}^{Y_{02}}} \frac{\Gamma[Y_1+1]\Gamma[Y_{02}+1]}{\Gamma[Y_{TOT}+2]}.$$

Notice any dependence on the scale parameter β for the prior on the flux has cancelled out. Notice, too, how similar this is to the form in [18] save that the bins can now have arbitrary width and placement.

One can derive the equivalent marginalized likelihood ratio for three changepoints (and higher):

$$\Lambda(\phi_0, \phi_1 \mid \mathcal{H}_3, I, \{y_i\}) = \frac{{T_{TOT}}^{Y_{TOT}}}{{T_1}^{Y_1}{T_2}^{Y_2}{T_{03}}^{Y_{03}}} \frac{\Gamma[Y_1+1]\Gamma[Y_2+1]\Gamma[Y_{03}+1]}{\Gamma[Y_{TOT}+3]}.$$

31.4 Results on Monte Carlo Data

In tables 1–3, we list the results of our tests on Monte Carlo data. We simulated three kinds of data: 1) flat background; 2) a Vela pulsar–shaped light–curve, with CGRO/EGRET 100 MeV - 10 GeV Obs 00 data used as a template; and 3) a spike in a single 5×10^{-4} wide bin. We approximated the signal to noise ratios one would expect from CGRO/EGRET observations. We analyzed each of these simulated datasets with three methods: 1) the current high energy standard, Z_n^2 with $n = 6$; 2) The Bayesian epoch—folding method of GL92; and 3) our new statistic using 1–3 "Bayes Blocks", with both versions of prior (exponential, and similar to GL92). We note 'GL92' would have performed better had we used a much larger cutoff for m (number of bins), rather than stopping at the default $m = 12$.

Notice that both "One BB' methods outperformed the classical method on the "single spike" pulse–profiles, but not on the double–peaked 100 MeV Vela light-curve. Parametrizing the model with an overall rate and shape parameters improved the $\log Odds$ throughout. Indeed, on Vela the "Three BB" statistic appeared to be roughly as good as Z_6^2.

Monte Carlo Cts	CLASSIC — Z_6^2 n=6	$-\log_{10}$ *Prob*	BAYES — "Sparse BB" $\log_{10}$ $\mathcal{O}_{2,E}$	$\log_{10}$ $\mathcal{O}_{2,GL}$	$\log_{10}$ $\mathcal{O}_{3,GL}$	$\log_{10}$ $\mathcal{O}_{4,GL}$	GL92 $\log_{10}$ *Odds*
134	393.7	76.1	96.6	174.	171.	172.	32.1
74	195.8	34.6	41.2	93.1	91.7	92.3	14.2
32	102.4	15.7	12.9	39.7	38.6	39.0	6.53
13	32.3	2.91	0.44	9.2	8.5	8.5	0.65

TABLE 31.1. Preliminary Monte Carlo Results: Single Spike. "CLASSIC" is *classical probability (frequency of occurence) of the null hypothesis*, rather than a *ratio* of the probabilities of the null and interesting hypotheses, as are the others. 'E' in $\mathcal{O}_{2,E}$ stands for our first choice of parametrization, with an exponential prior on each separate segment. 'GL' stands for the second parametrization, similar to that from [18]; here, 'GL92'. The number tells the number of changepoints used in the model (two, three, or four). GL92 Caclulations provided by P. Freeman, private communication; calculated for up to m = 12 bins. "Vela" means CGRO/EGRET 100 MeV - 10 GeV Obs 00 data used as "template" for source shape.

Monte Carlo Cts	CLASSIC — Z_6^2 n=6	$-\log_{10}$ *Prob*	BAYES — "Sparse BB" $\log_{10}$ $\mathcal{O}_{2,E}$	$\log_{10}$ $\mathcal{O}_{2,GL}$	$\log_{10}$ $\mathcal{O}_{3,GL}$	$\log_{10}$ $\mathcal{O}_{4,GL}$	GL92 $\log_{10}$ *Odds*
561	467.5	91.8	52.4	52.2	70.0	85.2	77.6
277	279.3	52.0	29.1	28.9	41.1	50.0	44.3
138	165.5	28.4	16.2	16.1	26.6	32.1	23.0
72	73.8	10.21	5.65	5.42	7.4	9.9	7.40

TABLE 31.2. Preliminary Monte Carlo Results: CGRO/EGRET Vela. See notes for Table 31.1.

Monte Carlo Cts	CLASSIC — Z_6^2 n=6	$-\log_{10}$ *Prob*	BAYES — "Sparse BB" $\log_{10}$ $\mathcal{O}_{2,E}$	$\log_{10}$ $\mathcal{O}_{2,GL}$	$\log_{10}$ $\mathcal{O}_{3,GL}$	$\log_{10}$ $\mathcal{O}_{4,GL}$	GL92 $\log_{10}$ *Odds*
538	14.9	0.61	-0.431	-0.91	-0.06	0.9	-1.92
258	13.4	0.47	-0.453	-0.91	-0.3	0.5	-1.79
136	9.4	0.17	-0.445	-0.89	-0.5	-0.01	-1.78
71	10.1	0.22	-0.0046	-0.40	0.04	0.4	-0.95

TABLE 31.3. Preliminary Monte Carlo Results: flat background (null). See notes for Table 31.1.

31.5 Prospects and Conclusions

We found this preliminary test of the concept very encouraging. We look forward to applying it to actual data. The possibility of being sensitive to different kinds of light-curves could be very interesting.

Acknowledgments: A. C. acknowledges the hospitality of Wellesley College and UNH; her collaborators in AstroStatistics (esp. J. Scargle, V. Kashyap, T. Loredo, E. Kolaczyk, A. Siemiginowska, and D. van Dyk); and NASA contract NAG5-7984; and her numberless AISRP "AS-DATA" grant.

31.6 References

[1] O.C. De Jager, J.W.H. Swanepoel, and B. C. Rubenheimer, 1989, *A & A*, **221**, 80.

[2] R. Buccheri and B. Sacco, 1985, in: *Data Analysis in Astronomy*, eds: L. Scarsi, V. Di Gesu, P. Crane, and S. Levialdi, Plenum Press, p 15.

[3] Scargle, J, 1998., ApJ, 504, 405 .

[4] T. Loredo, 1992, In *Statistical Challenges in Modern Astronomy*, Springer-Verlag.

[5] Connors, A., 1997, in *Statistical Challenges in Modern Astronomy II*, ed. E. Feigelson and G. Babu, p 39.

[6] Connors, A. 1997, in *Data Analysis in Astronomy V*, ed. Di Gesu, V., Duff, M.J.B., Heck, A, Maccarone, M.C., Scarsi, L. and Zimmeran, H.U., (World Scientific, London), p 251.

[7] Bretthorst, G.L., these proceedings.

[8] Hartman, R., et al, 1999, *ApJS*, 123, 79.

[9] Gehrels, N., et al., 2000, *Nature*, 404, 363.

[10] Ramanamurthy, P.V., et. al., 1995, *ApJL*, 447, L109

[11] Ramanamurthy, P.V., et. al., 1996, *ApJ*,458, 755.

[12] Kaspi, V.M., 2000, *ApJ*, 528, 445.

[13] Halpern, J. & Holt, S., 1992, *Nature*, 357, 222.

[14] Bertsch, D. et al., 1992, *Nature*, 357, 306.

[15] Halpern et al., 2001, *ApJL*, 552, L125.

[16] Jefferys, W. and Berger, J., 1992, *American Scientist*, 80, 64.

[17] Leahy, D.A., Elsner, R.F., Weisskopf, M.C., 1983, *ApJL*, 272, 256.

[18] Gregory, P. C. & Loredo, T. J. 1992, *ApJ*, 398, 146

[19] Connors, A., 2001, these proceedings.

[22] M. West, in: ed. E. Feigelson and G. Babu, Springer-Verlag, New York, 299, 1992.

32

Analysis of the Fractal Structure of the Horsehead Nebula

Srabani Datta[1]

32.1 Introduction

The Horsehead Nebula is 2 x 2 deg in size at a distance of 450 kpc, centered at 5h 40m 59.0s , 5^0 27' 29.99". The Horsehead Nebula was studied in the optical wavelengths using the H_α (6560 A).

The nebula is found to be evolving with a virial mass of 35 $M_\odot$ and radius 0.17 pc with average density 3 x 10^4 / cm ^{3}. Kramer et. al. (1996) have reported that HH objects , IR point sources, condensations in NH_3 and ^{13} CO are found within B 33. ^{13} CO emssion spectra show that a clump exists in the centre of the Horsehead of radius 0.22 pc and mass 95.4 $M_\odot$, average density 2 x 10^3 / cm^3 , column density $N(H_2)$ / Δ v approximately 1.3 x 10^{21} / cm^2 km/s^{-1}. These values imply that B 33 is in virial equilibrium.

32.2 Method

Analysis of B33 consists of estimating the fractal dimensions of the main head and trunk structure (fig.1 of ([1]). To do this, B 33 is considered as a non-empty compact set of a metric space ([3]). Then the Kolmogorov dimension (also known as Minkowski dimension) of B33 is defined it's dimension. However, since such a definition is difficult for practical measurement, an alternative dimension called the grid dimension or box counting dimension is used ([3]).

The grid dimension is equivalent to the Kolmogorov dimension since B33 is a non-empty set. The grid dimension of B 33 was measured using an automated fractal dimension analysis software , Benoit 1.3, procured from Trusoft International Inc., St. Petersburg, USA. Benoit has been reviewed in Science (1999, vol. 285, 1228). For the analysis, a set of ten measurements were taken ([1]). Normality of the sample populations were tested by the Shapiro-Wilkes test ([8]).The value of W is 0.01065 and so the population is normal. Then the Students's t test of significance was applied to the sample population to test for their deviation from the value of the dimension for a Euclidean shape (dimension 1). A further test was made

[1]Department of Applied Mathematics, University of Calcutta

for the fluctuations of the cloud dimension values from that of the Julia set (1.679594).

32.3 Discussion

The structure of molecular clouds have been observed to follow a power-law relation ([7] ;[2]; [9]). Kramer et.al. ([5]) carried out an analysis on the images of the Orion B region, among others, using automated software and found that the clump mass spectra was consistent with a power-law, dN/dM α M $^{-\alpha}$, with α =3D 1.72 ± 0.09, M being the mass of the clump and N is the number of clumps with connection between the power-law index of the above, as also the fact that the region studied is around the Horsehead nebula, it is significant that it's observed dimension(1.6965725) is within the error limits of Kramer et. al. ([5]). As the Horsehead is physically attached to it's parent, it implies that the index applies to it as well and so it can be postulated that the fractal dimension of the parent cloud is also 1.6965725. Results also show that the dimension of the Horsehead is not significantly different from that of the Julia set and so it can be assumed that the structure of the Horsehead is identical to that of the Julia set. This assumption has consequences for the dynamics of cloud formation ([1]).

Acknowledgments The author wishes to thank her guide, Prof. B. Basu, former head of Department, Department of Applied Mathematics, University College of Science and Technology, University of Calcutta; Dr. N.K. Dey, Ramakrishna Mission Residential College, Narendrapur, Calcutta; the staff of the VBO, Kavalur, for expert assistance; Prof. R. Gupta, Y. Wadedkar, Dr.D. Mitra , IUCAA , Pune, India. Thanks also goes to Dr. M. Hart, Dept. of Pure Mathematics,University of Sheffield; Prof. A. Boksenberg, Institute of Astronomy, Cambridge, U.K. The author would like to acknowledge support of the National Science Foundation, USA.

32.4 References

[1] Datta,S. 2001, in Automated Data Analysis in Astronomy, ed. R. Gupta, ; H.P. Singh & C. Bailer-Jones ; Narosa; New Delhi.

[2] Elmegreen, B.G.; 1999 Ap.J. 527, 266

[3] Falconer, K.; 1997,in Fractal Geometry; Wiley & Sons, Chichester

[4] Kapur,J.N. & Saxena, H.C. ;1982; in Mathematical Statistics, S. Chand & Co.; New Delhi

[5] Kramer,C.; et al. ; 1998 A & A, 329, 249

[6] Kramer, C.; et al. 1996 ; A & A ; 307, 915

[7] Larson, R.B. ; 1981 MNRAS, 194, 809

[8] Pearson, E.S. & Hartley, H.O. 1972 in Biometrica Tables for Statisticians vol. II, CUP.

[9] Williams, J.P.; et.al- 1997, in Protostars & Planets IV, 97. astro-ph 9902246

33

On the Statistics of the Gravitational Field

A. Del Popolo[1]

ABSTRACT In this paper we extend Chandrasekhar and von Neumann's analysis of the statistics of the gravitational field to systems in which particles (*e.g.* stars, galaxies) are not homogeneously distributed. We derive a distribution function $W(\mathbf{F}, d\mathbf{F}/dt)$ giving the joint probability that a test particle is subject to a force $\mathbf{F}$ and an associated rate of change of $\mathbf{F}$ given by d$\mathbf{F}$/dt. We calculate the first moment of d$\mathbf{F}$/dt to study the effects of inhomogeneity on dynamical friction.

33.1 Introduction

The study of the statistics of the fluctuating gravitational force in infinite homogeneous systems was pioneered by Chandrasekhar & von Neumann in two classical papers (Chandrasekhar & von Neumann 1942, 1943 hereafter CN43). The analysis of the fluctuating gravitational field, developed by the quoted authors, was formulated by means of a statistical treatment in the case of uniform systems, and with no correlations.

Two distributions are fundamental for the description of the fluctuating gravitational field:

1. $W(\mathbf{F})$ which gives the probability that a test star is subject to a force $\mathbf{F}$ in the range $\mathbf{F}$, $\mathbf{F}+$ d$\mathbf{F}$;

2. $W(\mathbf{F}, \mathbf{f})$ which gives the joint probability that the star experiences a force $\mathbf{F}$ and a rate of change $\mathbf{f}$, where $\mathbf{f} = d\mathbf{F}/dt$.

From a pure theoretical ground we expect that inhomogeneity affects all the aspects of the fluctuating gravitational field (Antonuccio & Colafrancesco 1994; Del Popolo 1996a, b; Del Popolo & Gambera 1998).

[1]Catania Astrophysical Observatory

33.2 $W(\mathbf{F},\mathbf{f})$ and $\overline{f}$ in inhomogeneous systems

Assuming that the system density is described by a power law of index p, the expression of $W(\mathbf{F},\mathbf{f})$ is given following Markoff's method by (CN43):

$$W(\mathbf{F},\mathbf{f}) = \frac{1}{64\pi^6}\int_0^\infty\int_0^\infty A(\mathbf{k},\mathbf{\Sigma})\cdot \{\exp\left[-i(\mathbf{k\Phi}+\mathbf{\Sigma\Psi})\right]\}\, d\mathbf{k}d\mathbf{\Sigma} \tag{33.1}$$

A lengthy calculation leads us (see Del Popolo & Gambera 1998 for a derivation and the meaning of symbols) to find the function $A(\mathbf{k},\mathbf{\Sigma})$:

$$A(\mathbf{k},\mathbf{\Sigma}) = e^{-\tilde{a}k^{\frac{3-p}{2}}}\{1- igp(\mathbf{k},\mathbf{\Sigma}) + \tilde{b}k^{\frac{-(3+p)}{2}}]\cdot[Q(\mathbf{\Sigma})+kR(\mathbf{\Sigma})]\} \tag{33.2}$$

This last equation introduced into Eq. (33.1) solves the problem of finding the distribution $W(\mathbf{F},\mathbf{f})$ and makes it possible to find the moments of $\mathbf{f}$ that give information regarding the dynamical friction. Using this last expression for $A(\mathbf{k},\mathbf{\Sigma})$ and performing a calculation similar to that by CN43 the first moment of $\mathbf{f}$ is given by:

$$\overline{\mathbf{f}} = -\left(\frac{1}{2}\right)^{\frac{3}{3-p}}\cdot A(p)\cdot B(p)^{\frac{p}{3-p}} \cdot\frac{\alpha^{\frac{3}{3-p}}GML(\beta)}{\pi H(\beta)\beta^{\frac{2-p}{2}}}\cdot\left[\mathbf{v}-\frac{3\mathbf{F}\cdot\mathbf{v}}{|\mathbf{F}|^2}\cdot\mathbf{F}\right] \tag{33.3}$$

where

$$L(\beta) = 6\int_0^\infty\left[e^{(x/\beta)^{\frac{(3-p)}{2}}}\right]\left[\frac{\sin x}{x^{(2-p)/2}}-\frac{\cos x}{x^{p/2}}\right]dx - 2\int_0^\infty\left[e^{(x/\beta)^{\frac{(3-p)}{2}}}\right]\cdot\frac{\sin x}{x^{(p-2)/2}}dx \tag{33.4}$$

As shown by Eq. (33.3), in a inhomogeneous system, differently from homogeneous ones, $\mathbf{f}$ is a function of the inhomogeneity parameter p.

At this point we may show how dynamical friction changes due to inhomogeneity. From Eq. (33.3) we see that $\frac{d\mathbf{F}}{dt}$ differs from that obtained in homogeneous system only for the presence of a dependence on the inhomogeneity parameter p. If we divide Eq. (33.3) for the correspondent of CN43 we obtain:

$$\frac{\left(\overline{\frac{d\mathbf{F}}{dt}}\right)_{Inh.}}{\left(\overline{\frac{d\mathbf{F}}{dt}}\right)_{Hom.}} = -\left(\frac{1}{2}\right)^{\frac{6-p}{3-p}}\cdot\frac{3\alpha^{\frac{3}{3-p}}L(\beta)B(p)^{\frac{p}{3-p}}\cdot A(p)}{n\cdot\pi^2 H(\beta)B(\beta)\beta^{\frac{2-p}{2}}} \tag{33.5}$$

This last equation is an increasing function of p. This means that for increasing values of p the star suffers an even greater amount of acceleration in the direction $-\mathbf{v}$ (when $\mathbf{v} \cdot \mathbf{F} \leq 0$) than in the direction $+\mathbf{v}$ (when $\mathbf{v} \cdot \mathbf{F} \geq 0$), with respect to the homogeneous case. This is due to the fact that the difference between the amplitude of the decelerating impulses and the accelerating ones is, as in homogeneous systems, statistically negative, but now larger, being the scale factor greater. This finally means that, for a given value of n, the dynamical friction increases with increasing inhomogeneity in the space distribution of stars. In addition, by increasing n the dynamical friction increases, just like in the homogeneous systems, but the increase is larger than the linear increase observed in homogeneous ones.

33.3 References

[1] Antonuccio-Delogu V., Colafrancesco S., 1994, ApJ 427, 72

[2] Chandrasekhar S., von Neumann J., 1942, ApJ 95, 489

[3] Chandrasekhar S., von Neumann J., 1943, ApJ 97, 1 (CN43)

[4] Del Popolo A., 1996a, A&A 305, 999

[5] Del Popolo A., 1996b, A&A 311, 715

[6] Del Popolo A., Gambera M., 1998, A&A, 342, 34

This page intentionally left blank

34

Cross-identification of Very Large Catalogues

S. Derriere[1], F. Ochsenbein, D. Egret

ABSTRACT Modern astronomy has entered the era of very large catalogues, gathering information for over 10^8 sources. Dedicated methods have been developed at the CDS to handle the huge amounts of data involved: powerful lossless compression, keeping direct access to the data on the basis of celestial position, allow fast queries on those very large catalogues.
We present the use of these very large datasets in the context of a data mining project undertaken by CDS and ESO. The challenge of cross-matching very large catalogues with other data sets is discussed. The question of likelihood of cross-identifications, using a statistical approach for large samples, is also addressed.

34.1 Accessing very large catalogues

Very large catalogues (containing over 10^8 objects) represent huge amounts of data if stored in ASCII tables. In order to handle these large volumes, dedicated tools have been developed at CDS [1].

Reducing I/O (with a lossless binary compression scheme) and keeping direct access to relevant data for positional requests (by indexing compressed data on celestial positions) allow fast queries.

Those very large catalogues are fully integrated in CDS services such as VizieR or Aladin (and are also used by OASIS). Various standardized output (including XML-Astrores [3]) are available.

Available very large catalogues include USNO A2.0, UCAC1, GSC 2.2 and the current DENIS and 2MASS releases. The average query time is a few μs per source (on the CDS server).

34.2 Cross identification

There is a strong interest in performing cross-matching between catalogues of sources observed at various epochs and wavelengths, as well as with

[1]Observatoire Astronomique de Strasbourg

user's own data.

ESO and CDS have been developing data-mining tools to access and combine the data available in those two Centers (*ESO-CDS Data Mining Project*, [4]). For all catalogues in the VizieR catalogue service (nearly 10^5 columns in August 2001), the contents of heterogeneous datasets is precisely described by meta-data, attached to each column of a catalogue, and named UCD's (*Unified Content Descriptor*).

In a first step, a prototype positional cross-correlator was developed for cross-matching VizieR catalogues, or user's data. Using the UCD structure, cross-matching by criteria other than position will be made possible soon.

34.3 Statistical approach

For very large catalogues, the task of cross-identifications must be automated, with statistical validation of the associations, as it is no longer possible to perform identifications "by eye".

Considering the case of positional association between two catalogues, one can build, for the first catalogue, the distribution of distances to the nearest neighbor in the second catalogue. Under simple assumptions (source density locally constant, gaussian errors on position), it is possible to precisely fit this histogram with a statistical distribution law, and to derive, for each source, a likelihood that it has been properly associated [2].

This statistical validation will serve forthcoming cross-identification tools at CDS. With the use of meta-data such as UCD's for multi-criteria cross-matching, this should be an element of the upcoming Virtual Observatory.

34.4 References

[1] Derriere S., Ochsenbein F., and Egret D., 2000, *ASP Conf. Ser.*, **216**, 235

[2] Derriere S. et al., 2001, A&A, in preparation

[3] Ochsenbein, F. et al., 2000, *ASP Conf. Ser.*, **216**, 83

[4] Ortiz, P. F. et al., 1999, *ADASS VIII*, *ASP Conf. Ser.*, **172**, 379

35

Minimal Spanning Tree Technique

A. Doroshkevich[1]

ABSTRACT The application of the Minimal Spanning Three technique to the description of large scale galaxy distribution shows that it can be roughly described as a network of high density 1D filaments and 2D wall-like condensations.

35.1 Minimal Spanning Tree

The MST is an *unique network* associated with a given point sample and connects all points of the sample to a *tree* in a special and unique manner which minimizes the full length of the tree. Cosmological implications of this technique were firstly discussed in [1], [2], and recently in [3].

The probability distribution function of MST edge lengths, (PDF MST), $W_{MST}(l)$, depend on the correlation functions (or cumulants) of all orders. For larger point separations, however, when correlations become small and the cumulants tend to constants, the Poisson-like point distribution can be expected and the PDF MST characterizes the geometry of a point distribution. For the 1D and 2D Poissonian distributions analytical expressions for the PDF MST [4] are:

$$W_{MST}(l) = \frac{1}{\langle l \rangle} e^{-l/\langle l \rangle}, \quad W_{MST}(l) = 2\frac{l}{\langle l^2 \rangle} e^{-(l^2/\langle l^2 \rangle)}. \tag{1}$$

The PDFs MST for 1D, 2D and 3D Poissonian samples are plotted in Figure 35.1 together with fits (1). For 3D Poissonian point distribution the cutoff of the PDF MST at $l \sim 2\langle l \rangle$ decribes the percolation process.

The PDFs MST plotted in Figure 35.2 for the SDSS catalogue is well fitted by the superposition of Rayleigh and exponential functions. This fact indicates that this distribution can be described as a network of 1D filaments and 2D sheets (or walls).

Basically, the MST contains within it all 'friends-of-friends' cluster catalogues for all linking lengths. The set of clusters for a given linking length is extracted by the process of *separating* the MST – i.e., removing any edges

[1]Theoretical Astrophysics Center, Copenhagen

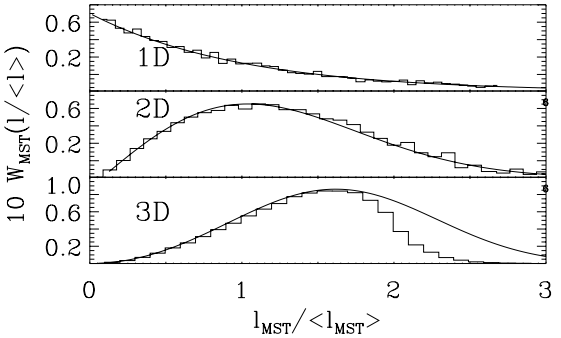


FIGURE 35.1. The PDFs MST for 1D, 2D and 3D Poissonian point distributions. Fits (1) (top and middle panels) and $W \propto l^2 \exp(-l^3/\langle l^3 \rangle)$ (bottom panel) are plotted by solid lines.

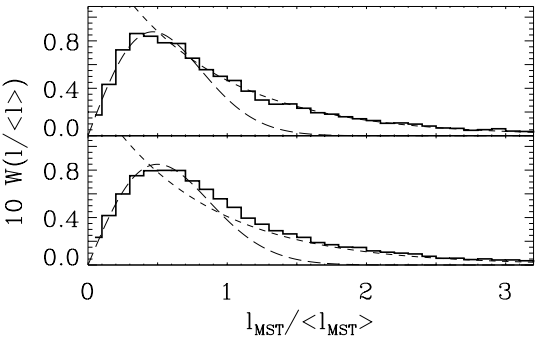


FIGURE 35.2. The PDFs MST for south (top panel) and north (bottom panel) samples of the SDSS. Rayleigh and exponential fits (1) are plotted by long dashed and dashed lines.

from the MST whose length exceeds that linking length. This approach allows to separate wall-like high density and low density regions which are occupied by filaments and poorer clusters.

The morphology of separate clusters can also be characterized by a ratio of the full length of tree, L_{sum}, builded for each cluster with the length of its trunk, L_{tr}, what is the maximal path of the tree. Evidently, for a filamentary-like cluster this ratio $\epsilon = L_{tr}/L_{sum} \sim 1$ can be expected while for walls a value $\epsilon \ll 1$ is more typical.

[1] J. Barrow, S. Bhavsar, D. Sonoda, MNRAS, 216, 17, 1985
[2] R. van de Weygaert, Ph.D. Thesis, University of Leiden, 1991
[3] A. Doroshkevich, D. Tucker, R. Fong, V. Turchaninov, H. Lin, MNRAS, 322, 369, 2001
[4] M. Kendall, P. Moran, Geometrical Probability, (London: Griffin), 1963

36

A Statistical Chromatic Study of Nearby Galaxies

Michel Fioc[1]

The synthesis of the spectral energy distribution (SED) of nearby galaxies puts constraints on their stellar populations, age, metallicity and dust content, but suffers from degeneracies. To break these degeneracies, previous studies in the optical must be extended to the infrared.

In the near-infrared (NIR), most observations are obtained in small apertures and are not comparable to optical data because of the blue-outwards color gradient. Fioc & Rocca-Volmerange (1999) solved this problem by building magnitude vs. aperture growth curves. Using statistical estimators taking into account the intrinsic scatter of colors within one given type, they showed that

1. total optical-to-NIR colors are significantly bluer than the small-aperture colors of Aaronson (1978);

2. they follow a well-defined sequence as a function of type;

3. the dependence of colors on inclination is an efficient tool to determinate the optical depth and the dust content of galaxies;

4. the color-magnitude relation (CMR) of elliptical galaxies – a major constraint on galaxy formation models – is nearly flat, in contradiction with previous studies (Bower et al. 1992) based on aperture colors and thus biased by the small-aperture–color-gradient problem. The CMR is also strongly dependent on the morphological type; so, both the mass and the type characterize the star formation history.

This work is currently extended to the mid- and far-infrared (MFIR) using IRAS data (Fioc & Dwek, in prep.). To avoid the bias of IR-selected samples towards starbursts and active galaxies, "normal" galaxies were selected from the LEDA optical database (Paturel et al. 1997) and their IR counterparts were identified in the Faint Source Survey (Moshir et al. 1992). Because many galaxies are not detected in the IR, survival analysis techniques (Feigelson & Nelson 1985) have been used to compute standard

[1]Institut d'Astrophysique de Paris

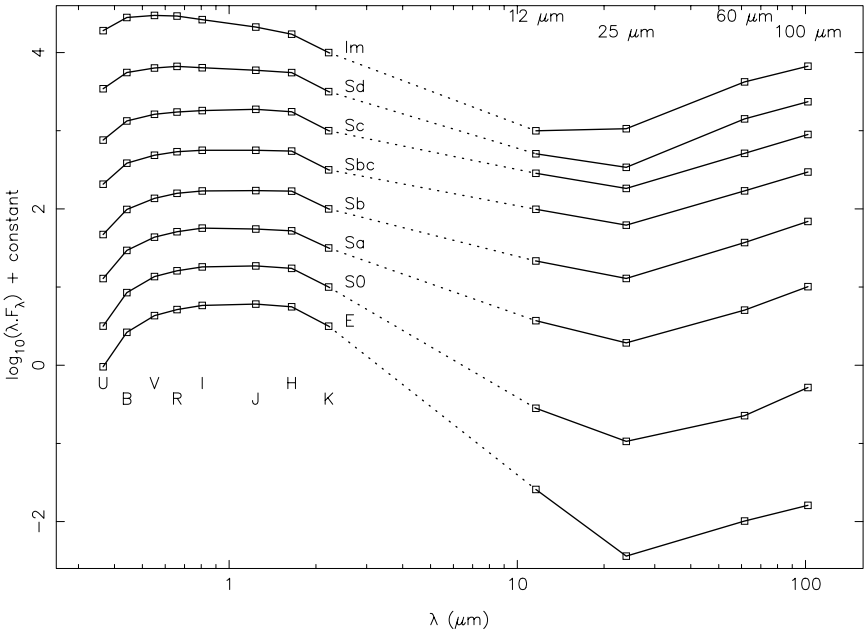


FIGURE 36.1. Template SEDs as a function of type.

MFIR-to-optical flux ratios as a function of morphological type. An important result is that, despite a stronger and harder radiation field in later types, the ratios peak for Sbc (Milky Way-like) galaxies, indicating that early-type spirals contain more dust than late-type ones and irregulars.

Template optical-to-IR SEDs (see Fig. 36.1) derived from this work will be analyzed with the spectral evolution model PÉGASE (Fioc & Rocca-Volmerange 1997) to derive the star formation history of normal galaxies.

Aaronson M. 1978, ApJL 221, L103
Akritas M.G., Bershady M.A. 1996, ApJ 470, 706
Bower R.G., Lucey J.R., Ellis R.S. 1992, MNRAS 254, 601
Feigelson E.D., Nelson P.I. 1985, ApJ 293, 192
Fioc M., Rocca-Volmerange B. 1997, A&A 326, 950
(`http://www.iap.fr/users/fioc/PEGASE.html`)
Fioc M., Rocca-Volmerange B. 1999, A&A 351, 869
Moshir M., Kopman G., Conrow T.A.O. 1992, IRAS Faint Source Survey, Explanatory supplement version 2
Paturel G. et al. 1997, A&AS 124, 109
(`http://cisr.univ-lyon1.fr/~leda/`)

37

Detection of Non-Gaussianity on the Sphere Using Spherical Wavelets

J. Gallegos[1], E. Martínez-González, F. Argüeso, L. Cayón and J. L. Sanz

ABSTRACT We present results showing the efficiency of the spherical Mexican Hat wavelet in detecting non-Gaussian CMB features on the sphere. We compare its performance with that of the spherical Haar wavelet for two families of non-Gaussian fields, both generated using the Edgeworth expansion to introduce skewness and kurtosis respectively. Analyzing the skewness and kurtosis of the wavelet coefficients in contrast to Gaussian simulations, the Mexican Hat is more efficient in detecting non-Gaussianity than the spherical Haar wavelet for all the different levels of non-Gaussianity introduced. The Mexican Hat can detect levels of the skewness and kurtosis of $\approx 1\%$ for $33'$ resolution. These results are relevant to test the Gaussian character of the CMB data and therefore the standard inflationary scenario.

37.1 The Spherical Mexican Hat Wavelet for Non-Gaussianity analysis

Wavelets have demonstrated to be a very useful tool for data analysis due to its space-frequency localization. Recently the Spherical Haar Wavelet, SHW, (Barreiro et al. 2000) and the Spherical Mexican Hat Wavelet, SMHW, (Cayón et al. 2001) have been used to test the non-Gaussianity of the COBE-DMR data. In this work we compare the performance of the two spherical wavelet bases (SHW and SMH) proposed for discriminating between Gaussian (e.g. Standard Inflation) and non-Gaussian models.

The non-Gaussian simulations have been obtained by perturbations of a Gaussian distribution using the Edgeworth expansion. We generate CMB maps from the Gaussian and non-Gaussian distributions, and convolve them with a $33'$ beam. In Figure 37.1, we show the deviations from Gaussianity for two non-Gaussian models for the first five resolution levels of

[1]Instituto de Fisica de Cantabria

TABLE 37.1. Power of the Fisher discriminant at 1% significance level

	Moment $\times 10^{-2}$	SMHW P(%)	SHW P(%)	Temperature P(%)
SKEWNESS	0.9(2.4)	66.8	1.51	2.51
	1.6(2.3)	100	7.09	4.67
	4.6(2.4)	100	36.12	36.85
	6.9(2.4)	100	78.46	73.60
KURTOSIS	0.3(2.6)	15.35	3.00	1.42
	0.8(2.7)	86.89	9.00	3.40
	1.1(2.7)	98.10	16.11	4.90
	1.4(2.6)	99.90	28.43	3.50

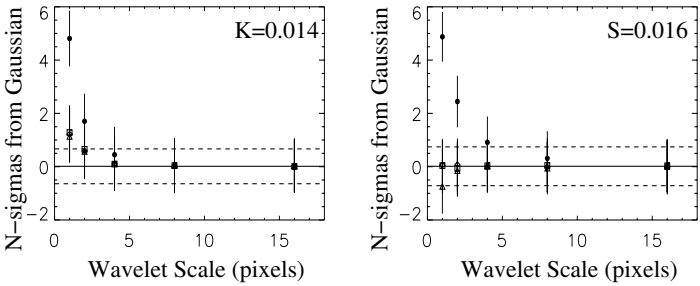


FIGURE 37.1. Comparison of Spherical Mexican Hat wavelet (black circle) and the Spherical Haar Wavelet details. Each point represents the number of sigmas deviated from the Gaussian model.

the wavelets. It is clear that the performance of the SMHW is much better than that of the SHW. The Fisher discriminant t have been applied to distinguish between the non-Gaussian and Gaussian models; its power p is presented in Table 37.1, constructed from the skewness and kurtosis of the SMHW, SHW and temperature. For a more complete description of the method and results see Martínez-González et al. 2001.

Barreiro, R.B., Hobson, M.P., Lasenby, A.N, Banday, A.J., Górski, K.M. & Hinshaw, G. 2000, MNRAS, 318, 475

Cayón, L., Sanz, J.L., Martínez-González, E., Banday, A.J., Argüeso, F., Gallegos, J.E., Gorski, K.M. & Hinshaw, G. 2001, MNRAS, 2001, 326, 1243

Martínez-González, E, Gallegos, J.E., Argüeso, F., Cayón, L. & Sanz, J.L., (submitted to MNRAS), see astro-ph/0111284

38

Characterising Anomalous Transport in Accretion Disks from X-ray Observations

J. Greenhough[1], S. C. Chapman, S. Chaty, R. O. Dendy & G. Rowlands

ABSTRACT We examine the time variation of the total X-ray flux from three sources and find that the signal from the Crab (non-accreting) is uncorrelated, the Cygnus X-1 signal is correlated on timescales up to three years, and in the GRS1915+105 signal correlation may extend to only a few days. The method we use also quantifies the distributions of fluctuations and hence constrains turbulence/instability models of accretion disks.

38.1 Introduction and technique

Non-Gaussianity and non-trivial temporal scaling together are strong indications of correlated processes such as turbulence (Bohr 1998). Applying the differencing and rescaling technique explained below to RXTE data[2], we show how trivial scaling of near-Gaussian fluctuations in the Crab X-ray signal – evidence of diffusive transport – contrasts with non-trivial scaling of non-Gaussian fluctuations in the X-ray signals from Cygnus X-1 and GRS1915+105. The functional forms of these fluctuations can then be used to constrain turbulence/instability models of the accretion disks.

From the raw time-series y we form a set of differenced series Z for a range of time-lags τ, and thence a set of probability density functions (PDFs) $P(Z, \tau)$. If these PDFs belong to a stable distribution, rescaling the axes by a single parameter α collapses them onto one curve whose functional form is characterised by α. For a full discussion of this technique see Greenhough et al. (2001).

38.2 Results and conclusions

Figure 38.1 shows the result of differencing and rescaling the Cygnus X-1 time-series. We find the PDFs of the differenced Crab data are near-

[1]University of Warick

[2]http://xte.mit.edu/XTE/asmlc/ASM.html

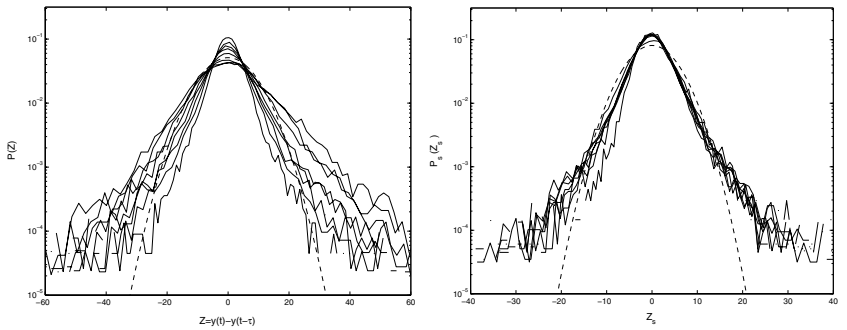


FIGURE 38.1. Unscaled (left) and rescaled (right) PDFs of differenced time-series for Cygnus X-1, 1996 Sep. – 1999 Dec. (mean timestep 77 minutes); dashed line Gaussian. τ in half-integer powers of timesteps to maximum 10^4; $\alpha \approx 8$.

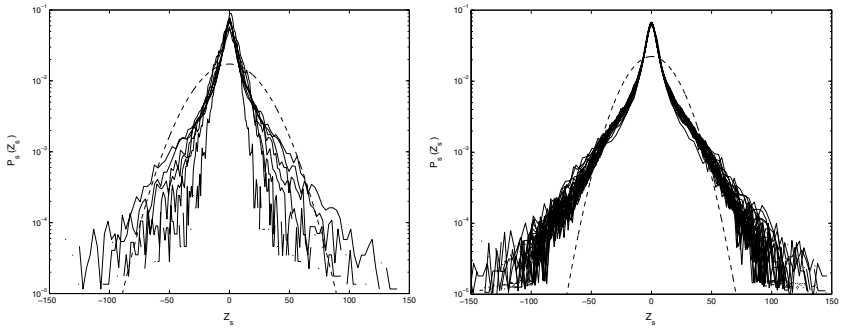


FIGURE 38.2. Rescaled PDFs of differenced time-series for GRS1915+105, 1996 Feb. – 2001 Mar. (mean timestep 96 minutes); dashed line Gaussian. Left: τ in half-integer powers of timesteps to maximum 10^4. Right: τ in single integer timesteps to maximum $10^{1.5}$; $\alpha \approx 6$.

Gaussian and independent of τ, whereas for GRS1915+105 the PDFs rescale approximately for τ up to only a few days as seen in Fig. 38.2. Thus we have evidence that the two accreting objects display a degree of correlation in their X-ray time-series, which is absent from the nonaccreting Crab. This is a quantitative, observational, and model-independent measure of anomalous transport in accretion disks.

References:

Bohr, T. et al. 1998, Dynamical Systems Approach to Turbulence (Cambridge)

Greenhough, J. et al. 2001, astro-ph/0107074

39

A Bayesian Analysis of the Radio Binary LS I +61°303

P. C. Gregory[1]

ABSTRACT Bayesian hypothesis testing and parameter estimation has played a central role in deciphering the complex radio properties of a remarkable radio emitting binary star. We briefly summarize the steps and present a recent confirmation of our earlier conclusions based on more extensive data.

39.1 Introduction

The luminous, massive X-ray binary, LS I +61°303 is remarkable for its periodic radio outbursts every 26.496 days. The optical, infrared, X-ray and γ-ray properties indicate the presence of a neutron star in a highly eccentric orbit, embedded within an equatorial wind from a rapidly rotating massive star of spectral class Be. Orbital variations in wind accretion by the neutron star are thought to be responsible for the periodic radio emission. A 23 year time series of radio measurements exists for this object.

In 1998, armed with 21 years of data, we carried out two sophisticated Bayesian hypothesis testing studies (Gregory 1999, Gregory et al. 1999) of competing models to account for the observed variability (time scale of years) in both the peak flux density and phase of the outbursts. Only the phase behavior is discussed here. The outburst phase is derived from the time of the outburst peak and the assumed orbital period (P), and expressed as a timing residual in days. The 45 outburst timing residuals, available at the time of the 1998 study, are shown in figure 39.1(a), based on an orbital period of 26.496 days.

The four hypotheses considered are indicated in table 39.1. The outburst phase is dependent on the assumed value of P, which is unknown independent of the radio data, so P was itself treated as a parameter in each of the four models. Model H_4 assumed a periodic modulation of the outburst phase and we employed a special version of the Bayesian GL method (Gregory and Loredo 1992) applicable to the case where the noise sampling

[1]Department of Physics and Astronomy, University of British Columbia

distribution is independent Gaussian (Gregory 1999). The GL method addresses the problem of the detection and characterization of a periodic signal in a time series when we have no specific prior knowledge of the existence of such a signal or of its characteristics, including shape.

TABLE 39.1. Hypothesis Space

	Hypotheses	**Odds Ratio**
H_1	Outburst times are consistent with a single period P. The timing residuals are assumed to be independent Gaussian random with an unknown sigma.	1
H_2	Sudden period change from P_A to P_B sometime during the data gap between day $\sim 5500 - 6000$) in figure 1(a), just prior to start of Green Bank monitoring program.	5.2×10^7
H_3	Outburst times are consistent with a single period P and a period derivative $\dot{P}$.	100
H_4	Single period P for all outbursts plus a periodic modulation (P_2) of the timing residuals of unknown shape.	1.4×10^{11}

The probability of each model was obtained by application of Bayes theorem, and involved marginalizing over all the model parameters with suitably chosen priors. The final probability of each hypothesis compared to the probability of H_1 is expressed as an odds ratio in the last column. The odds ratio, $O_{i,1}$, for model H_i compared to model H_1 can be factored according to

$$O_{i,1} = \frac{p(H_i \mid I)}{p(H_1 \mid I)} \frac{p(D \mid H_i)}{p(D \mid H_1)} \equiv \frac{p(H_i \mid I)}{p(H_1 \mid I)} B_{i,1}, \tag{39.1}$$

where D represents the new data, and I, the prior information. The first term, the prior model odds ratio, was assumed = 1. The Bayes factor $B_{i,1}$, is the ratio of the global likelihoods of the models and automatically includes a quantified Occam's razor that penalizes the more complicated model for its extra complexity. The resulting odds ratios strongly support the case for a periodic phase modulation of the radio outbursts. The next step was to estimate the H_4 model parameters. This consisted of computing the marginal probabilities of the orbital period, modulation period and the mean and standard deviation of the modulation shape.

This parameter estimation problem has recently been redone using all the radio outbursts measurements available up to October 2000, when the Green Bank interferometer monitoring program (Ray et al., 1997) ceased operation. Figure 39.1(b) shows the full timing residual data set based on the most probable orbital period of 26.496 days, together with two solid

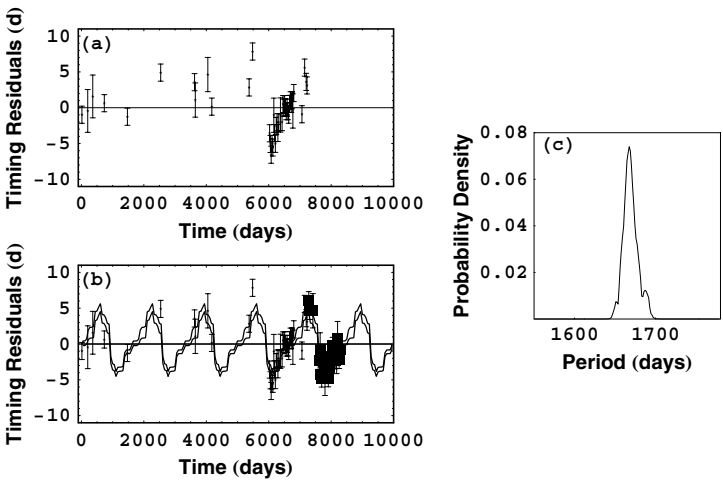


FIGURE 39.1. Panel (a) shows the initial outburst timing residuals. Panel (b) compares the Bayesian estimate of the light curve with the full set of timing residuals. The new data is indicated by a shaded box. The solid curves are the estimated mean light curve, ±1 standard deviation. Panel (c) shows the marginal posterior of the modulation period.

curves which are the Bayesian estimate of the mean light curve ±1 standard deviation. The additional outbursts are indicated by the solid squares. Clearly the timing residuals are well fit by a roughly saw tooth shaped modulation and the new data confirm the periodic pattern discovered on the basis of the smaller data set.

Figure 39.1(c) shows the modulation period marginal probability for only the small portion of the prior period search range (which extended from 800 to 2500 days) that contained significant probability [2]. The new estimate for the modulation period is 1667^{+14}_{-11} days.

This research was supported in part by grants from the Canadian Natural Sciences and Engineering Research Council at the University of British Columbia.

39.2 References

Gregory, P. C., and Loredo, T. J. 1992b, ApJ, 398, 146

Gregory, P. C. 1999, ApJ, 520, 361

[2]This figure is based on the final joint Bayesian analysis of both the phase and peak flux density data.

Gregory, P. C., Peracaula, M. and Taylor, A.R. 1999, ApJ, 520, 376

Ray, P. S., Foster, R.S., Waltman, E. B. et al. 1997, ApJ, 491, 381

40

Accounting for Absorption Lines in High Energy Spectra

Christopher Hans[1] **and David A. van Dyk**

40.1 Overview

The increasing popularity of Markov chain Monte Carlo (MCMC) methods and the limitations of "classical" astrophysical data analysis in the face of a new class of instruments (e.g. the *Chandra X-Ray Observatory*) make Bayesian analysis of high-resolution low-count energy spectra both feasible and attractive. van Dyk et al. (2001) and Sourlas et al. (this volume) describe a Bayesian hierarchical model which directly models counts as a Poisson process, avoiding problems resulting from Gaussian assumptions of standard chi square fitting. We extend this model to account for absorption lines. Parameter estimation is accomplished via an MCMC algorithm whose latent conditional structure allows us to concentrate on the problem of absorption lines outside of other complications such as background contamination, photon pile-up, etc. For specific computational details, see van Dyk et al. (2001).

40.2 Statistical Model and Data Augmentation

To simplify our presentation, we assume there is no instrument response, background contamination, effect of the effective area of the instrument, and that the source model is a simple continuum model with a single absorption line. These assumptions can easily be relaxed within the framework of the Bayesian hierarchical model (see van Dyk et al. 2001). In the absence of the absorption line we model the true counts at energy E_j as independent observations from a Poisson distribution with intensity $f(E_j, \theta)$, where $f(E_j, \theta)$ is the expected counts at energy E_j from a continuum model with parameters θ. If an absorption line is present, some of the photons emitted by the source are absorbed before they reach the detector, meaning that the observation is "incomplete" in the sense that there is a set of photons that

[1]Institute of Statistics and Decision Sciences, Duke University

was emitted but not observed. We can therefore define an "augmented" data set, $Y_j^{\text{aug}} = \{Y_j^{\text{obs}}, Y_j^{\text{mis}}\}$, where Y_j^{obs} is the number of photons detected at energy E_j and Y_j^{mis} is the number of photons absorbed by the line at energy E_j. With this notation in hand, we can explicitly model

$$Y_j^{\text{obs}} | \theta, \phi \overset{\text{indep.}}{\sim} \text{Poisson}\Big(f(E_j, \theta)\pi(E_j, \phi) \Big),$$

where $\pi(E_j, \phi)$ is the probability that a photon is *not* absorbed by a line with parameters ϕ. We allow $f(E_j, \theta)$ to represent any continuum model but restrict $\pi(E_j, \phi)$ to be the double exponential absorption line model used by Freeman et al. (1999),

$$\pi(E_j, \phi) = \exp\left\{ -\tilde{\lambda} \exp\left\{ \frac{-(E_j - \mu)^2}{2\sigma^2} \right\} \right\},$$

where the parameters $\phi = (\mu, \sigma^2, \tilde{\lambda})$ are the location, width and intensity of the line, respectively. Estimation of ϕ is simplified by noting that under a log-log link function, $\pi(E_j, \phi)$ is linear in E_j and E_j^2, reducing the problem to the standard statistical problem of estimation of parameters for a generalized linear model (GLM). To formalize the Bayesian model specification, we can assign flat (non-informative) priors, proper (informative) priors, or a combination of both. This model can be easily expanded to account for multiple lines in the continuum, and algorithms to compute maximum likelihood estimates (MLEs) of the parameters are readily available.

Acknowledgments: The authors gratefully acknowledge funding for this project partially provided by NSF grant DMS-01-04129 and by NASA contract NAS8-39073 (CXC). This work is a result of a joint effort of the members of the Astro-Statistics working group at Harvard University.

40.3 References

[1] reeman, P., Graziani, C., Lamb, D., Loredo, T., Fenimore, E., Murakami, T., and Yashida, A. (1999). Statistical analysis of spectral line candidates in gamma-ray burst GRB 870303. *The Astrophysical Journal* **524**, 753–771.

[2] ourlas, N., van Dyk, D. A., Kashyap, V., Drake, J., and Pease, D. Bayesian Spectral Analysis of "MAD" Stars, in *Statistical Challenges in Modern Astronomy III* (this volume), edited by Eric D. Feigelson and G. Jogesh Babu. Springer-Verlag, New York.

[3] an Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *The Astrophysical Journal* **548**, 224–243.

41

χ^2-method: An Automatic Classification Technique

Evanthia Hatziminaoglou[1] **and the EIS team**

ABSTRACT An automatic classification technique for separating the different astronomical objects in classes based on multi-colour photometry only, is presented. The technique consists of a standard fitting procedure, where the observed spectral energy distributions are compared to a template library. This method is applied on the point-like source catalogue of the Chandra Deep Field South. The spectral library consists of a series of quasar, white dwarf, low-mass star, brown dwarf and main sequence star theoretical templates and/or observed spectra. Over an area of 0.25 square degrees a total of 234 quasar candidates, 48 low-mass star and brown dwarf and 100 white dwarf candidates have been selected.

χ^2-technique, Data and Spectral Library

The χ^2-technique consists of a standard fitting procedure, where the observed spectral energy distributions are compared to a template library. Thus, the traditional multi-dimensional method ($2 \times$ N dimensions, with N the number of the colour-colour diagrams) is reduced to a one-dimensional technique.

The technique is applied on the point-source multi-colour data of the Chandra Deep Field South (CDF-S) provided by the ESO Imaging Survey. The CDF-S (0.25 square degrees) has been covered in U, B, V, R, and I, while its central region of 0.1 square degrees has also been observed in J and $K_{\rm s}$ ([1]; [8]). In the present analysis objects detected in at least three passbands have been considered, and the sub-samples examined comprise 1494 point sources in five passband and 605 in seven passband.

The spectral library currently in use consists of series of model quasar ($0 < z < 6$), white dwarf ($6000{\rm K} < T_{\rm eff} < 10^5{\rm K}$; $\log g = 7 - 9$; [3], [4]), and brown dwarf spectra ($500{\rm K} < T_{\rm eff} < 2800{\rm K}$ corresponding roughly to M < 0.1 M$_\odot$, for $\log g = 4.5$; [2]), three empirical cool white dwarf observed spectra ([5], [6]) and a set of synthetic stellar templates [7].

[1]European Southern Observatory

Selected Targets

Combining the data from the five and seven passband catalogues, one finds a total of **234** quasar candidates with estimated photometric redshifts up to $z \sim 5$, among which 16 have $z > 3.5$. In addition, **48** low-mass star/brown dwarf candidates and **100** white dwarf candidates were identified, including nine with $T_{eff} < 4000\mathbf{K}$.

If the classifications are confirmed, samples comprising **over 100** high-z quasars, $\sim$ **200** low-mass stars/brown dwarves and **over 1000** white dwarves will become available at the end of the survey, expected to cover 3 square degrees. It is worth emphasizing the **contribution of the near-infrared data:** it increases the accuracy of the determination of the photometric redshifts and significantly increases the number of quasar candidates in the redshift interval $2.5 < z < 3.5$. Infrared photometry is also important for tracking very low-mass stars and brown dwarves.

41.1 References

[1] Arnouts, S. et al., 2001, A&A 379, 750

[2] Chabrier G., Baraffe I., Allard F., Hauschild P., 2000, ApJ, 542, 464

[3] Finley, D. S., Köster, D., Basri, G., 1997, ApJ, 488, 375

[4] Homeier D., Köster, D., Hagen, H.-J. et al., 1998, A&A, 338, 563

[5] Ibata, R., Irwin, M., Bienaymé, O. et al., 2000, ApJ, 532, L41

[6] Oppenheimer, B. R., Saumon, D., Hodgkin, S. T. et al., 2001, ApJ, 550, 448

[7] Pickles, A. J., 1998, PASP, 110, 863

[8] Vandame, B. et al., 2001, accepted for publication in A&A, astro-ph/0102300

42

Wavelet Analysis of a Large Period Change in the Mira Variable R Cen

G. Hawkins[1], J. A. Mattei[1], and G. Foster[1]

R Centauri (R Cen) is an oxygen-rich Mira variable with a period of 546 days, range of variation of 5.3 - 11.8 magnitude at V, and spectral type of M4e-M8IIe as listed in the General Catalog of Variable Stars (Kholopov 1985). Visual observations from 1918 to 2001 from the AAVSO International Database show the familiar pattern of double maxima in the light curve of R Cen (Figure 1). The light curve also shows two other unusual properties: 1) the pulsational amplitude has decreased by 3 magnitudes since about 1950, (Figure 1) and 2) the period of the dominant mode has been steadily decreasing from 550 days at JD 2434000 (1951) to its present value of 505-510 days (Figure 2). The decrease in period and pulsational amplitude are probably caused by a He-shell flash in the interior of R Cen, as the period decrease of 1 day/yr is similar to that of other Miras thought to be undergoing a He-shell flash, such as R Hya and R Aql (Wood and Zarro 1981), and T UMi (Mattei and Foster 1995; Gál and Szatmáry 1995).

For our wavelet analysis in Figure 2, we use the Weighted Wavelet Z Transform (WWZ) of Foster (1996), which gives better results than a traditional wavelet transform when the data are unevenly sampled or have seasonal gaps. Further details of our analysis of R Cen are given in Hawkins, Mattei and Foster (2001).

We gratefully acknowledge the dedicated observations of hundreds of variable star observers since 1918 that made this study possible.

42.1 References

[1] Foster, G. 1995 AJ, 109, 1889
[2] Foster, G. 1996 AJ, 112, 1709
[3] Gál, J., and Szatmáry, K. 1995, A & A, 297, 461
[4] Hawkins, G., Mattei, J.A., and Foster, G. 2001, PASP, 113, 501
[5] Kholopov, P.N. 1985, General Catalogue of Variable Stars (4th ed.; Moscow: Nauka)
[6] Mattei, J.A., and Foster, G. 1995 JAAVSO, 23, 106
[7] Wood, P.R., and Zarro, D.M. 1981, ApJ, 247, 247

[1]American Association of Variable Star Observers

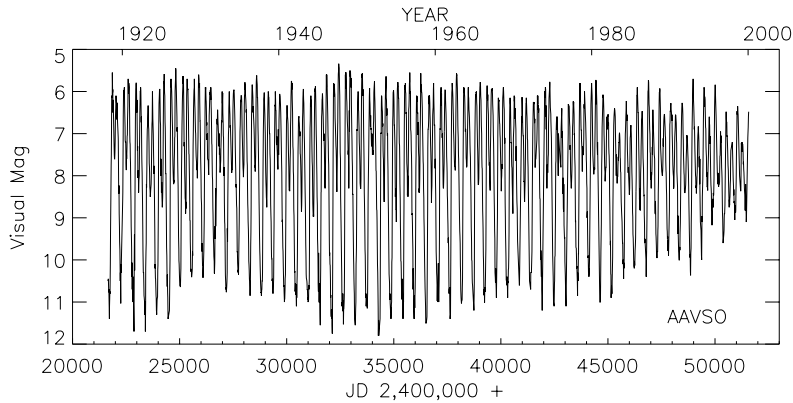


FIGURE 42.1. AAVSO light curve of R Cen from 1918-2000. 10 day averages of the data have been connected by a solid line for visual clarity.

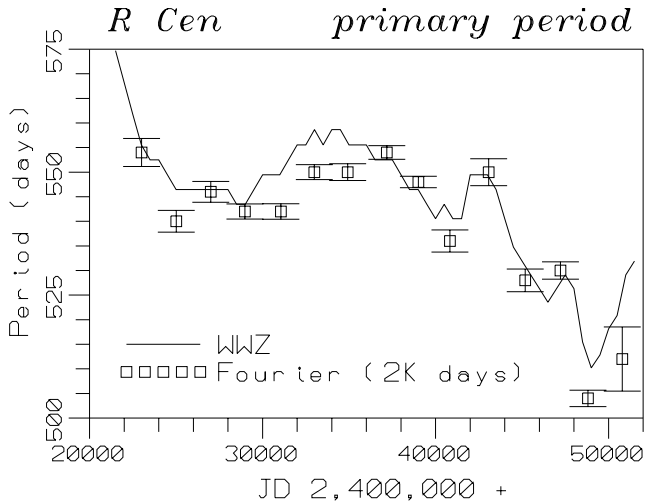


FIGURE 42.2. Wavelet plot (solid line) showing the period change in the primary mode (510-550 days) from 1918 to 2000. The squares show a Fourier analysis of the data in 2000 day segments using the Cleanest algorithm of Foster (1995).

43

Nonparametric Statistical Models of Astronomical Systems

William D. Heacox[1]

Nonparametric statistical modeling is a useful tool for estimation of the statistical distribution of a property characterizing a population of astronomical objects, when that property is observed only in combination with some other properties of *a priori* unknown individual values, but of known or estimated distribution. A trivial and well known example is the integral relation between the distributions of true velocities V and their observed radial components $V_R = V \sin i$, among objects of uniformly distributed but otherwise unknown orientations (*e.g.,* ref. [1]). The technique used to derive this relation can be extended to more complex and interesting such problems in observational astronomy; this is what is meant by the phrase "nonparametric statistical modeling".

As a simple example, the relation of the observed mass function Y to the underlying mass ratio q of a single-lined spectroscopic binary is $Y = q^3 \sin^3 i / (1+q)^2$, where i is the (*a priori* unknown) inclination of the orbital pole to the line of sight. The resulting statistical model, for a presumed uniform distribution of orbital orientations, is (ref. [2]):

$$f(Y) = \int f(q) \left\{ \frac{(1+q)^{4/3}}{3qY^{1/3}\sqrt{q^2 - Y^{2/3}(1+q)^{4/3}}} \right\} dq,$$

where by $f(x)$ is meant the pdf of random variable x. This model is typical in that it takes the form of an integral equation whose kernel $K(Y,q) = \Pr[Y \text{ observed, given q}]$ is the quantity in braces. Models such as these can be unambiguously derived from the identification of the kernel as a conditional probability, and have the essential character of an accounting technique: one adds up all the probabilities of arriving at the observed distribution of the disguised quantity (Y) in order to deduce the only possible distribution of the underlying variable (q) of interest. The resulting integral equation may be inverted to infer the desired underlying distribution from the observed one.

[1]University of Hawaii

This type of statistical modeling is described in more detail, including kernel derivation, in references [4] & [5]; it may readily be applied to a wide variety of problems, including multivariate ones. Its principal limitations are the need to reasonably model the statistical behavior of obscuring variables (*e.g.*, inclination i in the above examples), and the often encountered poor numerical conditioning of the models, leading to inaccuracies in integral equation inversion. To date it has been usefully applied (in non-trivial forms) to the following areas of observational astronomy:

- The inference of the distributions of orbital parameters and primary-secondary mass ratios among the entire population of binary stars with (roughly) solar-like primaries (refs. [2], [4], [7]), including the discovery that all binary orbital dynamical quantities (semi-major axis, angular momentum, binding energy) are distributed approximately as $f(x) \propto x^{-1}$ over wide ranges, a result with probable (if not currently understood) consequences for formation theory. These models have been extended to extrasolar planets (ref. [6]) to demonstrate that, within observable limits, extrasolar planet orbital characteristics are statistically indistinguishable from those of binary star systems with stellar-mass secondaries, with some consequences for the proper interpretation of the nature and formation of these low-mass objects.

- The modeling of observable kinematics within globular clusters and other spherically symmetric, multi-body systems to determine internal mass distributions and the statistical distributions of stellar orbital energies and angular momenta (refs. [3], [5]). Such models require no *a priori* assumptions of total mass or its sources, nor of orbital angular momenta and energies; but employ stellar proper motions (available for some clusters) to unambiguously determine distributions of these quantities, and total cluster mass, from the data themselves. To date this has been used to demonstrate that traditional models employing only radial velocities and central separations cannot constrain the overall mass of globular clusters to within a factor of 3 without inclusion of *a priori* untestable assumptions, such as velocity isotropy or that mass follows light. Application of these models to actual clusters is a computationally intensive task that is currently being undertaken with a parallel-processing supercomputer.

- The modeling of the mass distributions of populations of microlensing objects, in terms of the observed lensing timescale distributions and presumed kinematics of the lensing population. Preliminary application to Galactic halo microlenses observed against the Magellanic Clouds seems to require an asymmetrical mass distribution with significant numbers of lens masses near or below the minimum hydrogen-burning mass limit. This work – and application to the recently dis-

covered low-mass lenses in the globular cluster M22 (ref. [8]) – are ongoing projects.

43.1 References

[1] Chandrasekhar, S. & Münch, G. 1950, Astrophys. J. **111:142**

[2] Heacox, W. D. 1995, Astron. J. **109:2670**

[3] Heacox, W. D. 1997, Astrophys. J. **490:263**

[4] Heacox, W. D. 1998a, Astron. J. **115:325**

[5] Heacox, W. D. 1998b, Astrophys. J. Supp. **114:121**

[6] Heacox, W. D. 1999, Astrophys. J. **526:928**

[7] Heacox, W. D. & Gathright, J. 1994, Astron. J. **108:1101**

[8] Sahu, K. C., Casertano, S., Livio, M., Gilliland, R. L., Panagla, N., Albrow, M. D. & Potter, M. 2001, Nature **411:1022**

This page intentionally left blank

44

Likelihood Estimation of Gamma Ray Bursts Duration Distribution

Istvan Horváth[1]

Two classes of Gamma Ray Bursts have been identified so far, characterized by T_{90} durations shorter and longer than approximately 2 seconds. It was shown that the BATSE 3B data allow a good fit with three Gaussian distributions in $\log T_{90}$ [4]. In the same Volume in the Astrophysics Journal another paper suggested that the third class of GRBs is may exist [11]. Using the full BATSE catalog here we present the maximum likelihood estimation, which gives us 0,5% probability to having only two subclasses. The MC simulation confirms this probability.

In the BATSE current catalog [8] there are 2702 Gamma-Ray Bursts (GRBs), of which 2041 have duration information. [6] have identified two types of GRB based on durations, for which the value of T_{90} (the time during which 90% of the fluence is accumulated) is respectively smaller or larger than 2 s. This bimodal distribution has been further quantified in other papers [7], [5] where a two-Gaussian fit were made. Previously we have published an article [4], where two and three Gaussian fits were made using the χ^2 method, which gave us app. 0,02% significance the third group is needed. This is an agreement with the [11] result, who used a multivariate analysis and find that the probability of existence of two clusters rather than three is less than 10^{-4}. [1] also confirmed this result by statistical clustering analysis, however they suggested the third group was caused by instrumental biases [1], [2]. Recently, remarkable anisotropy was found in the angular distribution of this third group [10]. In this paper we take another attempt at the trimodal distribution, evaluating the probability that the two populations are independent using the maximum likelihood estimation.

For this investigation we have used a smaller set of 1929 burst durations in the current catalog, because these have peak flux information as well. Firstly we take a two Gaussian fit for the duration which gives us a best parameters of the two Gaussian fit, which are very similar than previously was published [4]. Secondly we take a three Gaussian fit. The means are -

[1]Bolyai Military University, Budapest

.25; .63; 1.55 in lg$s$. These fits gives us the best logarithm's of the likelihoods 12320.11 and 12326.25. Twice of the difference of these numbers follows the χ^2 distribution with three degree of freedom because the new fit has three more parameters [12]. The difference is 6.14 which gives us a 0.5% probability. Therefore the third Gaussian fit is much better and there is a 0.005 chance the third Gaussian is caused by statistical fluctuation.

One can check the probability using the Monte-Carlo (MC) simulation. Generate 1929 numbers for T_{90} whose distribution follow the sum of two Gaussian distributions. Find the best likelihood with five free parameters (two means two sigmas and two weights, but the sum of the last two must be 1929). Secondly made a fit with three Gaussian (eight free parameters, three means, sigmas and weights). Take a difference between the two logarithm's of the maximum likelihoods, which gives one number. We do the process 100 times and have a hundred MC simulated numbers. Only one of these numbers is bigger than which the BATSE data has (6.14). Therefore the MC simulation confirm the mathematical low statement and gives us a similar probability if the third group is a statistical fluctuation.

The BATSE on-board software tests for the existence of bursts by comparing the count rates to the threshold levels for three separate time intervals: 64, 256, 1024 ms. The efficiency changes in the region of the middle area because the 1024 ms trigger is becoming less sensitive as burst durations fall below about one second. This means that at the "intermediate" timescale a large systematic deviation is possible. To reduce the effects of trigger systematics in this region we truncated the dataset to include only GRBs that would have triggered BATSE on the 64 ms timescale. Using the Current BATSE catalog CmaxCmin table [9] we choose the GRBs, which numbers larger than one in the second column (64 ms scale maximum counts divided by the threshold count rate). Although this process reduced the bursts numbers very much (only 857 GRBs remain) the significance level still stay below 1%.

It is possible that the three log-normal fit is accidental, and that there are only two types of GRB. However, if the T_{90} distribution of these two types of GRBs is log-normal, then the probability that the third group of GRBs is an accidental fluctuation is less than 0.5-1.0 %.

This research was supported in part through OTKA F029461 and T34549. Useful discussions with M. Briggs, E. Fenimore, J. Hakkila, P. Mészáros, are appreciated.

44.1 References

[1] Hakkila, J., et al. 2000. ApJ, 538, 165

[2] Hakkila, J., et al. 2000. Gamma-Ray Burst Fith Huntsville Symposium. Huntsville, Alabama. AIP 526. Melville. p. 48

[3] Horváth, I., Mészáros, P., & Mészáros, A. 1996, ApJ, 470, 56

[4] Horváth, I., 1998, ApJ, 508, 757

[5] Koshut, T. M., et al. 1996, ApJ, 463, 570

[6] Kouveliotou, C., et al. 1993, ApJ, 413, L101

[7] Kouveliotou, C., et al. 1995, Third Huntsville Symposium on GRB. New York: in AIP Conference Proceedings 384, 84-89

[8] Meegan C. A., et al. 1996, ApJS, 106, 65

[9] Meegan C. A., et al. 2000, Current BATSE Gamma-Ray Burst Catalog, on the Internet http://www.batse.msfc.nasa.gov/data/grb/catalog/

[10] Mészáros, A., Bagoly, Z., Horváth, I., Balázs, & Vavrek., R. 2000, ApJ, 539, 98

[11] Mukherjee, S., Feigelson, E. D., Babu, G. J., Murtagh, F., Fraley, C., & Raftery, A. 1998, ApJ, 508, 314

[12] Press, W. H., Teukolsky, S. A., Vetterling, W.T., & Flannery B. P. 1992, Numerical Recipes in Fortran, Second Edition, Cambridge University Press, Cambridge

This page intentionally left blank

45

Nonparametric Density Estimation and Galaxy Clustering

Woncheol Jang[1]

ABSTRACT We estimate the number density of galaxy clusters as a function of z, redshift. Nonparametric density estimation is used to estimate the galaxy density f given z and then the connected components of the level set $\{f(\cdot|z) > \delta_c\}$ are extracted as clusters. The parameter δ_c is estimated by matching the number density to the Press-Schechter model using a goodness-of-fit criterion. Since δ_c is itself a function of a cosmological parameter, this leads to a confidence interval for the parameter.

45.1 Introduction

Clusters of galaxies provide powerful tools from tracing the large scale structure of the universe to determining the amount of dark matter. Moreover, the mass distribution function of these large scale structures plays a key rule to the nature of primitive density fluctuations. In addition to these observational advantages, clusters can be understood via relatively simple theory, the Press-Schechter model.

Our goals are (1) to estimate the galaxy density f given z, redshift, and extract the connected components of the level set $\{f(\cdot|z) > \delta_c\}$ as clusters. Then, (2) one can find a confidence interval for Ω_m, density parameter for matter, using a goodness-of-fit criterion since δ_c is itself a function of Ω_m and z, therefore it can be estimated by matching the number density to the Press-Schechter model as well.

45.2 Density Estimation and Clustering

Suppose $X_1, \ldots, X_n$ are the locations of n galaxies in a sky survey where $X_i = (X_{i1}, X_{i2}, X_{i3})$=(RA, DEC, redshift). Since we are interested in the evolution of galaxies, we want to estimate the joint distribution of RA

[1]The Pittsburgh Institute for Computational Astrostatistics(PICA)

and DEC given redshift. To do so, we slice the data by redshift and fit a bivariate kernel density estimation. See PICA (2002) for details.

For clustering, a modified version of Cuevas et al. (1998) is proposed. The detail steps are as follows.

Given z and Ω_m, use the Fast Fourier Transform(FFT) to calculate $\hat{f}$ at grid points $t_j = (t_{j1}.t_{j2})$, $j = 1, \ldots, m$. Then, extract contiguous grid points such as $\{t|\hat{f}(t) > \delta_c\}$ as a cluster. Here δ_c is a function of z and Ω_m. See Reichart (1999) for details. After clustering, one assign the data to closest grid points. If the grid point belongs to a cluster, so do the data. Define N_k, $k = 1, \ldots, K$ be the number of galaxies in each cluster and use it as the mass of cluster.

45.3 Goodness-of Fit test

By Press-Schechter theory, the number density is,

$$n(M) = \sqrt{\frac{2}{\pi}} \frac{\delta_c \alpha}{\sigma_M} \frac{\rho}{M^2} \exp\left(-\frac{\delta_c^2}{2\sigma_M^2}\right) \propto M^{\alpha-2} \exp\left(-\frac{\delta_c^2}{2}\left(\frac{M}{M_0}\right)^{2\alpha}\right),$$

which is proportional to generalized gamma distribution (Johnson et al. 1994).

Given z_i and Ω_m, we calculate $\delta_c(z_i, \Omega_m)$, then fit generalized gamma distribution on N_k and get a p-value using a goodness-of-fit test.

Repeat the previous steps for every z_i and Ω_m and use Fisher's meta analysis (Fisher 1932) to combine "p"-values over z:

$$\chi^2 = -2 \sum_{i=1}^{L} \log p_i \sim \chi^2(2L)$$

where p_i is p-value for the goodness-of-fit test given z_i and Ω_m. After, one calculates p-value of χ^2 then converts it into confidence interval for Ω_m.

45.4 Future Work

We proposed a clustering method via density estimation and estimating confidence interval for Ω_m with goodness-of -fit test. Once the Sloan Digital Sky Survey is available, it will be addressed. A long version of this document (Jang 2001) will be available at PICA's website (www.picagroup.org) Double truncation and measurement error need to be considered.

45.5 References

Cuevas, A., Febrero-Bande, M. and Fraiman, R. (2000). Estimating the number of clusters. *The Canadian Journal of Statistics*, 28, 367-382.

Fisher, R.A. (1932). *Statistical Methods for Research Workers.* 4th Ed. Oliver and Boyd.

Jang, W. (2001). Nonparametric Density Estimation and Clustering. In preparation.

Johnson, N.L., Kotz, S. and Balakrishnan. (1994). *Continuous Univariate Distributions. Vol. 1.* 2nd Ed. John Wiley, New York.

The Pittsburgh Institute for Computational Astrostatistics (PICA). (2002). Nonparametric Inference in Astrophysics. *To appear in Statistical Challenges in Modern Astronomy III*, . Springer, New York.

Reichart et al. (1999) *Astrophysical Journal*, 518, 521

This page intentionally left blank

46

Teaching Bayesian Statistics Through Simulation

William H. Jefferys[1]

ABSTRACT I describe an introductory graduate course on Bayesian statistics taught in the Spring of 2001 at the University of Texas. The course made extensive use of simulation through Markov Chain Monte Carlo, with students completing a number of projects to introduce them to the basic ideas of MCMC simmulation and Bayesian reasoning.

46.1 Course Description

The course was designed primarily for physical scientists with no statistical background who wished to learn practical Bayesian inference and techniques. It actually attracted a wider audience: students from astronomy, mathematics, statistics, aerospace engineering, biology, management, and public affairs. The main idea of the course design was to concentrate on the ideas behind Bayesian inference, to get the students "thinking like Bayesians." I decided to deemphasize exact results and special situations such as conjugate priors and normal distributions, in favor of Markov Chain Monte Carlo (MCMC) as a generalized tool for practical solution of complex problems not amenable to specialized techniques.

Markov Chain Monte Carlo (MCMC) simulation techniques have been developed over the past 10–15 years into a powerful tool for producing a draw from the full posterior distribution, which can then be used to provide marginal distributions, medians and quantiles, and averages of various sorts to summarize the results of a statistical investigation. Essentially anything of interest can be calculated from the sample. Therefore, in keeping with the philosophy of the course, MCMC simulation techniques were introduced early; students were assigned a sequence of problems of increasing sophistication which, though simple, illustrated the application of MCMC in various useful contexts, and which could be generalized to more complex problems in obvious ways. Students were encouraged to work in teams and to program their solutions in a language of their choice.

[1]Department of Astronomy, University of Texas

Many MCMC examples were run in class using a computer attached to an LCD projector. This allowed us to experiment as we varied parameters. For in-class examples, I used the free statistical language R. I introduced it early in the course in order to give students a practical tool for solving problems. Most students also used R for their assignments.

46.2 Evaluation of the Course

Although I had used them before, I was not completely satisfied with my texts (Sivia 1996, Schmitt 1969). They are are good books, but I had a very different audience from what I had expected. Also, the books don't discuss MCMC (I had to present this *de novo*). I am looking seriously at Gelman *et. al.* (1996) when I teach the course again.

I would have liked to have assigned even more simulations but ran out of time. The next time I teach the course I will introduce R and the ideas of simulation even earlier and in parallel with the initial topics on probability theory, and reduce the discussion of some theoretical issues, to allow more time for such assignments.

I felt that the emphasis on simulation as a tool improved the students' connection with and understanding of Bayesian inference. I believe that they came out of the course with confidence that they would be able to attack even complex problems in their own field of interest effectively. The strategy to de-emphasize special situations like normal errors in favor of early examples using Cauchy and Poisson data worked well. I wanted to make it clear to the students that they need to examine the fundamentals of their problems rather than automatically to assume normality, and also to show them that they had the tools to attack such problems.

The students were enthusiastic and class participation was excellent. Students asked challenging questions, and several are already applying what they learned to their own research. Course reviews were excellent. Detailed information about the course, including assignments and presentation materials, may be found at `http://bayesrules.net/ast383.2001.html`

46.3 References

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995), *Bayesian Data Analysis*. London: Chapman and Hall.

Schmitt, Samuel (1969). *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics*. Reading, MA: Addison–Wesley. This book is out of print but was published as a course packet with the permission of the copyright holder.

Sivia, Devender (1996). *Data Analysis: A Bayesian Tutorial*. New York: Oxford University Press.

47

New MCMC Methods to Address Pile-up in the Chandra X-ray Observatory

Hosung Kang [1], David A. van Dyk, Yaming Yu, Aneta Siemiginowska, Alanna Connors, and Vinay L. Kashyap

47.1 Pile Up

Pile-up occurs in X-ray detectors when two or more photons arrive in an event detection island during the same frame. Such coincident events are counted as a single higher energy event or lost altogether if the total energy goes above the on-board discriminatory. Thus, for bright sources pile-up can seriously distort both the count rate and the energy spectrum. Accounting for pile-up is perhaps the most important outstanding data-analytic challenge for Chandra. Here, we outline how Bayesian hierarchical models can be designed to account for pile-up in X-ray detectors and how they can be fit via Markov chain Monte Carlo. Model fitting is accomplished using the *Gibbs Sampler*. Roughly speaking, this method sequentially fits one model component at a time conditionally on all the others. The power of this approach is that it allows us to ignore other model components when we account for pile-up. Specifically, we stochastically separate a subset of the observed counts into multiple counts of lower energy based on the current iteration of the particular spectral/spatial model being fit. The spectrum is then updated given the 'unpiled' counts. Because of the complexity of the pile-up process this is a challenging statistical task requiring simulation of highly structured multi-modal distributions. Nonetheless, the Bayesian framework is promising because it allows the inclusion of other sources of information. For example, event grades (i.e, a description of the likelihood of the degree of pile-up based on the spatial distribution of the charge) can be used to improve the fit.

[1]Department of Statistics, Harvard University

Table 1: Summaries of fitted models.

Data	Pile Up in model?	Model Fit[a] $\Gamma < 2$keV	$\Gamma > 2$keV	% piled
ACIS-S/HETG	no	1.70 ± 0.06	1.05 ± 0.05	n/a
ACIS-S/HETG	yes	1.70 ± 0.05	1.07 ± 0.05	00.6%
ACIS-S	no	1.53 ± 0.03	1.12 ± 0.04	n/a
ACIS-S	yes	1.69 ± 0.03	1.29 ± 0.05	14.3%
ACIS-S (to 15 keV)	yes	1.74 ± 0.03	1.07 ± 0.04	15.1%

[a]Error bars are one posterior standard deviation; Γ is the powerlaw parameter.

47.2 Unpiling 3C273 ACIS-S spectrum

We have applied our method[2] to the Chandra ACIS-S/HETG and ACIS-S observation of 3C273, a strong X-ray point source. We exclude the core from the analysis of ACIS-S data, because piled events exceed the threshold and have been removed from the telemetry signal (there are no counts in the core region). We do not consider any corrections to the readout streaks in this analysis. CIAO 2.0 software and the recent calibration data are used to construct the RMF and ARF files for this observation.

A broken power law model (break at 2keV) is assumed to fit the data within energies 0.5-8 keV. We fit the model to both data sets as is summarized in Table 1. We expect little pile-up in the ACIS-S/HETG grating data. Thus, accounting for pile-up has little effect on the fit; see Table 1.

The ACIS-S Spectrum is fit three ways: (1) without accounting for pile up, using data with energies 0.5-8.0 keV; (2) accounting for pile up, using data with energies 0.5-8.0 keV; and (3) accounting for pile up, using data with energies 0.5-15.0 keV. The three fits are compared with the ACIS-S/HETG fit in Table 1. In a heavily piled observation, photons may be recorded as events with significantly higher energy. For a proper fit, we must include these high energy events in our analysis. Thus, all of the pile up corrected analyses do well below 2 keV but for higher energies we need to include events of higher energy, as in fit (3), which does a remarkable job of reproducing the ACIS-S/HETG fit[3].

Acknowledgments: This project is funded in part by NSF grant DMS-01-04129 and by NASA contract NAS8-39073 (CXC) and is a joint effort of the members of the Astro-Statistics group at Harvard University.

[2]For simplicty, we further assume each event corresponds to either one or two photons and that the PSF is flat. i.e., we observe a point source whose photons are spread evenly across a region of the detector.

[3]The fit to ACIS-S data up to 15 keV included a large count in the highest energy channel (32 counts at 15 keV). Although some of these counts are undoubtable mis-recorded higher energy events they were included in the analysis. This strategy seems likely to induce less bias than removing these counts from the data.

48

Modeling Stellar Microflares

Vinay Kashyap[1], Jeremy J. Drake, Manuel Güdel, and Marc Audard

48.1 Overview

An open question in the field of Solar and stellar astrophysics is the source of heating that causes stellar coronae to reach temperatures of millions of degrees. One possibility is that the coronae are heated by a large number of small flares (see Audard et al. 2000 and Drake et al. 2000). On the Sun, microflares are distributed with energy as a power-law of the form $\frac{dN}{dE} = k \cdot E^{\alpha}$, with $\alpha = 1.8$, and α appears to increase to values 2.2-2.9 for flares of lower energy (cf. Asch et al. 2000). If the slope exceeds the critical value of 2, then in principle the entire coronal energy input may be ascribed to flares that are increasingly less energetic, but are more numerous. We have developed a new method to model these weak flares in photon arrival-time data.

48.2 Method

Model: Because flare onset is stochastic in nature, the light curves cannot be modeled directly. Instead we compare the distribution of **arrival-time differences** between the data and the model. We consider a 3-parameter model $\mathbf{M} = \{\alpha, r_F, r_B\}$ where α is the index of the power-law, r_F is the average count rate due to flares, and r_B is a constant "background" component. The Poisson-distributed model counts in an interval dt, $c(t)\,dt \;\sim\; \phi(t)\; Poisson[\, r_B(t)\,dt \;+\; f(t)\,dt\,]$, where $\phi(t)$ is a correction factor that takes into account Primbsch, dead-time, and GTIs, and the flare component, $f(t) = \sum_{j=1}^{N_f} \Theta(t - t_j) F_j e^{-(t-t_j)/\tau}$. Here τ is a fixed flare decay timescale, F_j are flare peak intensities sampled from the power-law distribution, and $\Theta(x)$ is a step function to represent instantaneous flare onset. Note that not only will the placement t_j and intensity F_j of the flares vary for each simulation, but so will the total number of flares N_f. Within the bounds of Poisson statistics, we expect that for any given

[1]Harvard-Smithsonian Center for Astrophysics

simulation, $\sum_{j=1}^{N_f} F_j\,\tau \approx r_F \cdot \Delta T$ where ΔT is the total duration of the observation. The model parameter r_F fixes the normalization k by equating the total counts due to the flare component with the counts expected from the power-law distribution.

Algorithm: We follow a Bayesian formalism and derive the joint posterior probablity of the model parameters, $p(\mathbf{M}|D,I) \propto p(\alpha|I)p(r_F|I)p(r_B|I) \cdot p(D|\mathbf{M},I)$ where D represents the data. The prior distributions are taken to be non-informative and flat over the limited parameter ranges considered. The likelihood is computed as the probability density of obtaining the observed χ^2 value for N degrees of freedom (see Eadie et al. 1971, Equation 4.22), $p(D|\mathbf{M}) = \frac{\frac{1}{2}\left(\frac{\chi}{2}\right)^{\frac{N}{2}-1}e^{-\frac{\chi}{2}}}{\Gamma(N/2)}$. The basic steps of the algorithm are: First, derive the distribution of photon arrival-time differences $f_D(\delta\,t) \propto \sum_i \rho_i\lambda_i e^{-\lambda_i\delta\,t}$, where ρ_i is the fraction of the time that a source spends at the intensity λ_i; then obtain realizations of the photon event list over a grid of parameter values; and compare the simulated $f_M(\delta\,t)$ with $f_D(\delta\,t)$.

Limitations: Unlike existing methods that rely on detection of flares, this method is best-suited to investigate the effects of very weak flares on stellar coronal emission. However, in the process of deriving $f(\delta\,t)$, the sequential information inherent in the light-curve is lost. For example, we cannot take advantage of the known fact that flares decay in intensity. Further, because the model is stochastic, a large number of simulations are necessary to obtain a stable result, leading to very lengthy computations. Finally, the method loses sensitivity for $\alpha \sim 3$ as the model approaches the limiting case of constant emission.

Results: We find strong evidence in favor of the slope of the flare distribution to be greater than 2 for active stars such as FK Aqr, Wolf 630, AD Leo, β Per. We find that $\alpha_{FK\,Aqr} = 2.68\pm0.25$, $\alpha_{V1054\,Oph} = 2.62\pm0.21$, $\alpha_{AD\,Leo} = 2.17 - 2.3$, and $\alpha_{\beta\ Per} = 2.84(> 2.41)$ The flare component contributes to 70%, 85%, 80%, and 75% respectively.

Acknowledgments: We would like to thank David van Dyk, Alanna Connors, and Eric Kolaczyk for useful discussions. VK was supported by NASA AISRP grants during the course of this research. JJD was supported by the Chandra X-Ray Center NASA contract NAS8-39073.

Aschwanden, M.J., *et al.* 2000, ApJ, 535, 1047

Audard, M., Güdel, M., Drake, J.J., & Kashyap, V.L. 2000, ApJ, 541, 396

Drake, J. J., Peres, G., Orlando, S., Laming, J. M., & Maggio, A. 2000, ApJ, 545, 1074

49

Canaries in the Data Mine: Improving Trained Classifiers

V. G. Laidler[1] **and R. L. White**

ABSTRACT Supervised classification uses a training set to construct a classifier such as a decision tree. Normally, the training set is discarded once the training process is complete. By imprinting information about the training population onto the classifier, we can make use of the extrema at each node as "canaries", warning us that we have left the well explored area of parameter space and have crossed into a domain where the classifier is unreliable. This technique can identify training set deficiencies; provide reliability estimates for decision tree classifiers; improve the results of multi-tree voting; and provide helpful visualization tools. See `http://www-gsss.stsci.edu/PublishedPapers/Canaries_SCMA.htm` for the poster version of this paper.

Motivation

All supervised classification techniques begin with the construction of a training set that is to be representative of the test population. A good training set must be of sufficient size and extent in parameter space to probe the entire domain occupied by the test population. The construction of the training set is of critical importance in supervised classification, for a bad training set will mislead even the best classification algorithm.

Defining the technique

This work was done with the GSC2 classification problem, which uses an oblique decision tree, OC1 (Murthy et al. 1994), to classify astronomical images into stars, nonstars, and plate artifacts based on ranked values of 30 image features (Laidler et al. 1996, White 1997). Recall that a decision tree operates by determining a set of decision surfaces that best separate classes of objects in the training set.

The imprinting technique records the minimum and maximum values of each feature for all the training set objects at each node, along with the actual coefficients defining the decision surface. Thus, it "imprints" a

[1]Computer Sciences Corporation at Space Telescope Science Institute

bounding box (hypercube) of the training set at each node.

When the decision tree is applied to a new object, the object's (normalized) distance outside this bounding box is computed, producing a Training Set Domain Distance (TSDD). Objects that resemble the training set will lie within the bounds, and have a TSDD of zero. Objects that lie in parts of parameter space that were not probed by the training set will have nonzero values for the TSDD. The distance increases as the object moves further away from known (training set) space.

Results

Imprinting successfully identified a known deficiency of bright objects in the original GSC2 training set: most bright objects in several test sets had high values of the TSDD. When tested against another subpopulation of deblended, or "child" objects, it confirmed our previous suspicion that these objects were not well represented in the training set, and that deblended objects occupied a different part of parameter space than clean single objects. External software can select training set candidate objects based on the TSDD, resulting in directed improvement of an existing training set.

The TSDD can also be used as a reliability measure. Since the decision surface is defined over the domain of the training set, any part of the surface that falls outside this domain is extrapolated, and the classifications derived therefrom are similarly unreliable. The TSDD indicates how far the test object is from the well defined (interpolated) zone.

Classification can be improved by voting (Heath et al. 1996). The GSC2 classifier votes a committee of 5 trees independently grown from the same training set. When the TSDD is used as a suitably scaled weighting function for the voting, the outcome changes for 2-3% of the objects on a plate. Of these changed objects, 70-80% of the changes result in a correct classification of objects that were previously misclassified.

References

Heath, D., et al., 1996, *Cognitive Technology: In Search of a Human Interface*, eds. Gorayska & Mey.

Laidler, V. G., et al., 1996, Bull.Am.Astron.Soc., 188, 5421

Murthy, S. K., Kasif, S., & Salzberg, S. 1994, J. Artificial Intelligence Research, 2, 1

White, R. L. 1997, *Statistical Challenges in Modern Astronomy II*, eds. G. J. Babu & E. D. Feigelson (New York: Springer), p. 135.

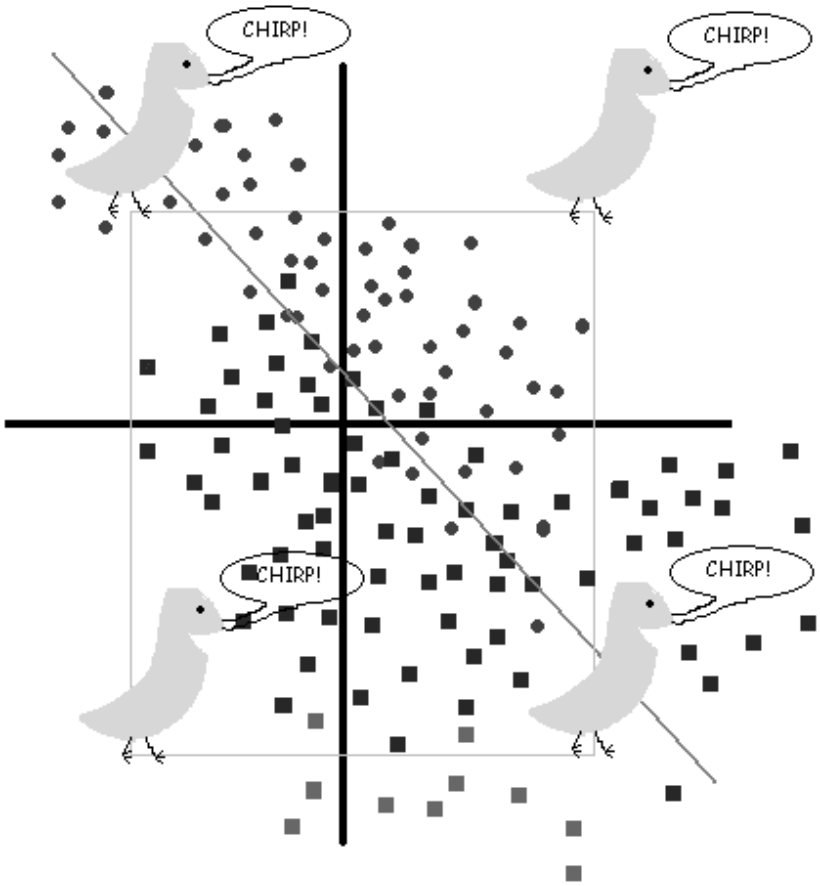


FIGURE 49.1. The cartoon illustrates the imprinting technique in a two dimensional, two class problem. The diagonal line is the decision surface that separates squares from circles. Inside the rectangle marking the training set domain, this line results in a good separation of classes. In the extrapolated domain outside the box, the separation breaks down in some areas. The training set domain distance (TSDD) is the normalized distance of an object from the bounding box. It is defined to be zero inside the box.

This page intentionally left blank

50

Wavelet Analysis of Heteroscedastic, Unevenly Spaced Data: The Case of OJ 287 Revisited

Harry Lehto[1]

ABSTRACT We present a method for calculating the statistical significance in a wavelet transform and apply it to the optical light curve of OJ 287. The original data is heteroscedastic and unevenly spaced and it appears to show transient oscillations. The statistical significance of the detection of these kind of variabilities have remained elusive in previous studies.

OJ 287 is a blazar showing rapid and quite extreme variations in intensity. The data of such intensity measurements are characterized by following properties:

1) Sampling in uneven
2) Quality of data varies from point to point
3) Variance of the measurement is known at each point.

We have applied a Morelet wavelet transform to the data. The transform for evenly spaced data can be defined as

$$W(f,t) = f\left(\left(S_k\right)^2 + \left(C_k\right)^2\right), \text{ where}$$

$$C_k + iS_k(f,t) = \sum_k m_k \exp i(2\pi f(t_k - t)) \exp(-\frac{1}{2} f^2 (t_k - t)),$$

where m_k and t_k refer to the intensity and the time of the kth measurement.

Let us assume that each observed datapoints can be expressed as a sum of a noiseless signal term and a noise term and that the noise is uncorrelated from one point to another, i.e. white noise. Note that there is no assumption of gaussianity is made at this point. If we further assume that the expected value of the data point is the unknown true value of the noiseless signal (with a delta function distribution) and that there is no correlation between the signal term and the noise term, then we can write

[1]Tuorla Observatory, University of Turku

$$W_{obs}(f,t) = W_{signal}(f,t) + W_{error}(f,t)$$

Since the distribution of the error in a single data point is known, one can in principle calculate the second term in the above equation and obtain an estimate of the true wavelet transform of the signal.

To determine the distribution function of $W_{error}(f,t)$, consider

$$S_k = (\sum_k n_k \sin(\phi_k) w_k).$$

Here $\phi_k = 2\pi f(t_k - t)$ and $w_k = \exp(-\frac{1}{2} f^2 (t_i - t))$ depend only on the sampling and not on the values of the noise. Let's combine the two terms and write our equation as

$$S_k = (\sum_k n_k a_k).$$

If the noise in individual measurements n_k have a Gaussian $N(0, \sigma_k^2)$ distribution then S_k will have a distribution equal to $N(0, \sum_k \sigma_k^2)$. This means that S_k^2 will have a χ_1^2-like distribution suitably scaled in variance. Similarly C_k^2 will also have a χ_1^2-like distribution. The probability distribution is then the convolution of these two distributions. If these two distributions happen to have similar variances then the resulting distribution mimics closely a χ_2^2-like distribution.

Even if the noise in individual points is non-gaussian, we may proceed as above, expect that the χ_1^2 distribution has to be replaced with a suitable distribution before convolution of the two distributions.

This scheme provides us with a quantitative estimate for the full distribution function of the noise terms in a wavelet transform enabling the calculation of the significance of actual signal's transform. The point that may need further investigation in this approach is the independence of S_k^2 and C_k^2 when calculating the combined distribution.

Nearly all the peaks in the transform of OJ 287 turned out to be highly significant ($p > 0.001$). The details of the analysis will be published in a forthcoming paper.

Acknowledgements The data for this experiment was provided by an international collaboration OJ-94. The work was funded by grants number 71355 and 44011 from the Finnish Academy.

51

Estimating Large-Scale Structure From QSO Absorbers: Using Across-Line Information

J. M. Loh[1], J. M. Quashnock and M. L. Stein

The clustering of QSO absorption-line systems, or absorbers, is on the same comoving scale as that traced by the voids and walls of galaxy redshift surveys of the local universe (see e.g. [2]). Thus it appears that the absorbers are effective probes of very large scale structure of the universe ([1]).

Previous investigations of clustering using QSO absorbers have, on large scales, been confined to considering absorbers occurring on the same lines of sight (see e.g. [4] and [6]). Absorber pairs lying on different lines of sight contain information about the clustering of absorbers. The use of such across-line-of-sight absorber pairs may improve the efficiency of estimates of clustering. Furthermore, lines of sight are generally about 400 h^{-1} Mpc long, limiting the distance at which clustering can be investigated using only absorber pairs lying on the same lines. With estimates that also use pairs of absorbers on different lines, there is no such limitation.

We have developed an estimator $\hat{K}(r)$ of the reduced second moment function $K(r)$ for QSO absorbers observed on a set of lines of sight using all possible absorber pairs. The main assumptions are that the absorbers are spheres of small constant radius and that the process of absorber centers is stationary and isotropic. The full details of this procedure can be found in [3]. Here, we describe the main results of using this estimation procedure.

We performed a simulation study to compare the new estimator $\hat{K}(r)$ with $\hat{K}_{\parallel}(r)$, an estimator that uses only absorber pairs on the same lines. We defined a conic section with half-angle of 45^0 and Earth at its tip, bounded by comoving distance 2000 to 3300 h^{-1} Mpc from the Earth. This region is similar to the region in which the Sloan Digital Sky Survey (SDSS) will find QSO lines of sight. We placed m lines uniformly and randomly in this region, with $m = 100$, 1000, 10,000 and 100,000. With each simulation of a Poisson process on these lines, we calculated $\hat{K}(r)$ and $\hat{K}_{\parallel}(r)$. The ratio of the standard errors of the two estimators is shown in the figure.

Our simulations show that, with 100,000 lines of sight, using $\hat{K}(r)$ instead

[1]Department of Statistics, Columbia University

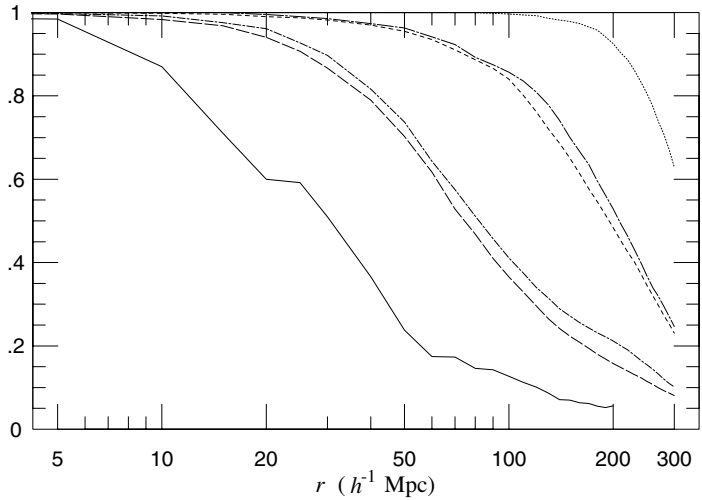


FIGURE 51.1. Ratio of standard errors of $\hat{K}(r)$ to $\hat{K}_{\parallel}(r)$ for $m = 100$ (dotted line), 1000 (short-dashed line), 10,000 (long-dashed line) and 100,000 (solid line). Also shown is the same ratio, for $m = 100$, but with 10 times (long-dashed and dotted line) and 100 times (short-dashed and dotted line) higher angular line density.

of $\hat{K}_{\parallel}(r)$ results in a reduction in standard error by a factor of 2 to 20 on scales of 30 to 200 h^{-1} Mpc. This is effectively an increase in sample size by an extra factor of 4 to 400 on large distances.

Vanden Berk and Quashnock (private communication) provide an extensive absorber catalog consisting of 276 lines of sight and 345 Carbon IV absorbers drawn from the literature. Using the new estimator with this catalog, we find strong evidence for clustering on scales up to 100 h^{-1} Mpc, and possibly up to 150 h^{-1} Mpc, similar to that found by [5]. We also calculated $\hat{K}(r)$ for r up to 1000 h^{-1} Mpc and do not find any evidence of clustering for $r > 150$ h^{-1} Mpc.

51.1 References

[1] Crotts, A. P. S. 1985, ApJ, 298, 732.

[2] Landy, S. D.,, et al. 1996, ApJ, 456, L1.

[3] Loh, J. M., Quashnock, J. M. & Stein, M. L. 2001, ApJ (to appear).

[4] Quashnock, J. M. and Vanden Berk, D. E. 1998, ApJ, 500, 28.

[5] Quashnock, J. M. and Stein, M. L. 1999, ApJ, 515, 506.

[6] Vanden Berk, D. E., et al. 1996, ApJ, 469, 78.

52

Point Source Detection on the Sphere Using Wavelets and Optimal Filters

E. Martínez-González[1], P. Vielva, D. Herranz, J. Gallegos and J. L. Sanz

ABSTRACT We present an analysis of simulated microwave data to detect point sources using wavelets and optimal filters. We search for point sources in the Time Order Data (TOD), using optimal adaptive filters, and in the map using the Spherical Mexican Hat Wavelets (SMHW). The SMHW provides a whole sky point source catalogue at 30 GHz of ≈ 900 sources with a flux limit detection of 0.53 Jy and a mean error of 16%. The optimal filter is able to detect ≈ 250 sources from the TOD reaching a flux limit of 0.97 Jy and a mean error of 20%.

52.1 Introduction

A critical issue in the analysis of microwave data is the component separation process. Microwave data consist in a mixture of emissions coming from different sources: Cosmic Microwave Background (CMB), the Galaxy and extragalactic sources. Several methods have been proposed to disentangle those emissions, based on Maximum Entropy, Bayes Theory, neural netwoks, ... Generically they are relatively good in dealing with the separation of the diffuse emissions (CMB, Galactic). On the other hand, it has been shown that compact sources are better identified with adaptive filters/wavelets more optimal for localised objects (Tegmark and Oliveira-Costa 1998, Cayón et al. 2000, Sanz et al. 2001). Following this last approach we search for point sources in microwave data both in the TOD and in the whole sky map. The data has been simulated with the Planck Pipeline Simulator and represents 6 months run of the 30 GHz LFI28 channel of the Planck mission. The simulations include all relevant Galactic and extragalactic emissions from both diffuse and compact sources. Also white and $1/f$ noises are present in the data and the antenna response has a FWHM of $33'$ and slightly differs from a circular Gaussian one. The sim-

[1]Instituto de Fisica de Cantabria

TABLE 52.1. Results of point source detections in the TOD and in the map.

Data	Number	Spurious	Mean error	Bias	Min. flux(Jy)
TOD	257	5%	20.4%	-8.4%	0.97
Map	926	5%	19.2%	-4.1%	0.53

ulated data cover the whole sky except for two circles of 1°.82 around the ecliptic poles.

52.2 Results

Details about the detection of point sources in the TOD using optimal filters have been given in Herranz et al. (2001). As it is shown in that paper optimal adaptive filters are efficient in detecting and extracting sources with a given profile embedded in a background of known statistical properties. In particular they could be used to obtain a real-time preliminary catalogue of extragalactic sources which would have a great scientific interest, e.g. for follow-up observations. The method based on the Mexican Hat wavelet has been shown to perform very well detecting point sources on maps representing small patches of the sky and also complementing the Maximum Entropy Method for the separation of all components (Vielva et al. 2001a,b). Here we demonstrate the performance of the method using the SMHW on all sky maps and for more realistic simulations which takes into account deviations from the ideal case of a perfect circular Gaussian antenna response and pure white noise. The results of detecting point sources in the TOD, using optimal filters, and in the whole sky map, using the SMHW, are given in the table.

52.3 References

Cayón, L. et al. 2000, MNRAS, 315, 757
Herranz, D., Gallegos, J., Sanz, J.L. and Martínez-González, E. 2001, submitted to MNRAS.
Sanz J.L., Herranz D. and Martínez-González E. 2001, ApJ, 512, 484
Tegmark, M. and Oliveira-Costa, A. 1998, ApJ, 500, 83
Vielva, P., Martínez-González, E., Cayón, L., Diego, J.M., Sanz, J.L. and Toffolatti, L. 2001a, MNRAS, 326, 181
Vielva, P., Barreiro, R.B., Hobson, M.P., Martínez-González, E., Lasenby, A.N., Sanz, J.L. and Toffolatti, L. 2001b, MNRAS, 328, 1

53

Constraining the Cosmological Constant from Large-Scale Redshift-Space Clustering

Takahiko Matsubara[1] **and Alexander S. Szalay**

ABSTRACT We show how the cosmological constant can be estimated from cosmological redshift distortions, using maximum-likelihood techniques. Using a simple idealized survey geometry, we compute the Fisher matrix for $\Omega_{\rm M}$ and Ω_Λ. We also estimate confidence contours for real survey geometries, using the SDSS LRG as an specific examples.

53.1 From Correlations to Fisher Matrix

To generically investigate how a given redshift survey can constrain the cosmological constant, we construct a rectangular box, in which the Gaussian smoothed cells are placed on lattice sites i in the box so as to have the smoothed density fluctuation vector $d_i = \rho_i/\langle\rho_i\rangle - 1$. A correlation matrix, $C_{ij} = \langle d_i d_j\rangle$, theoretically specifies all the statistical information for a given data set. First, the theoretical form of the correlation matrix is calculated from Matsubara & Suto (1996) with smoothing effect taken into account.

Once the correlation matrix is theoretically calculated in any cosmological model, the Fisher information matrix is used to estimate how well the model parameters can be measured:

$$F_{\alpha\beta} = -\langle\partial^2 \ln L/\partial\theta_\alpha\partial\theta_\beta\rangle = {\rm Tr}(\boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta})/2,$$

where $L(\boldsymbol{d};\boldsymbol{\theta})$ is a probability distribution for the data vector $\boldsymbol{d}$, which depends on a vector of model parameters $\boldsymbol{\theta}$. The Cramér-Rao bound states that the maximal likelihood estimate constrains the model parameters with a minimum variance $\langle\theta_\alpha\theta_\beta\rangle \geq (F^{-1})_{\alpha\beta}$.

[1]Department of Physics and Astrophysics, Nagoya University

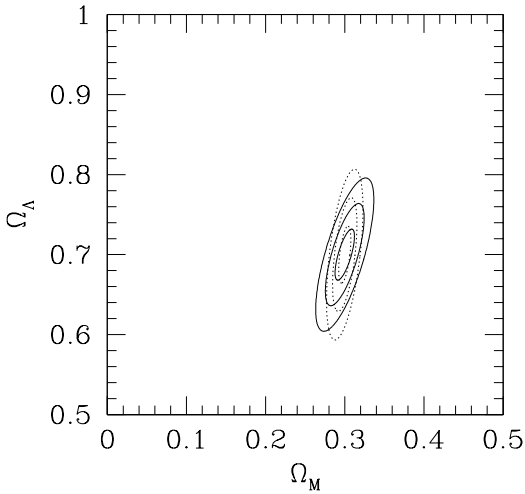


FIGURE 53.1. Concentration ellipses corresponding to 1σ, 2σ, 3σ confidence levels for approximate geometries of the 100,000 galaxies in the SDSS LRG (Luminous Red Galaxy) sample. Dotted lines assume a bias factor of $b = 1.5$, solid lines has $b = 2$.

53.2 Results

In this work, the power spectrum $P(k)$ and the bias parameter b are fixed throughout. The power spectrum is a CDM-type with $\Gamma = 0.2$, $\sigma_8 = 1$. We consider the cosmological constant parameter Ω_Λ and the density parameter $\Omega_{\rm M}$ to be estimated.

We have considered several different survey layouts for both galaxies and quasars. The best survey to perform these tests seems to be the Luminous Red Galaxy (LRG) sample of the Sloan Digital Sky Survey (SDSS). We simulate the geometry of LRG sample as a composite of the generic $200\,h^{-1}{\rm Mpc}_z$ boxes at the mean redshift $z = 0.3$. The shot noise is approximately given by $(20\,h^{-1}{\rm Mpc}_z)^3\bar{n} = 0.5$.

The resulting concentration ellipses are shown in Figure 53.1. This shows that the shot noise level and the depth of the survey volume are suitably balanced to constrain the geometry of the universe in the SDSS LRG survey.

The Cramér-Rao bound for Ω_Λ is only $[(F^{-1})_{\Lambda\Lambda}]^{1/2} = 0.04$ for $b = 1.5$, and $[(F^{-1})_{\Lambda\Lambda}]^{1/2} = 0.03$ for $b = 2$. This shows that the shot noise level and the depth of the survey volume are suitably balanced to constrain the geometry of the universe in the SDSS LRG survey. Unfortunately, the currently ongoing QSO redshift surveys, like Sloan Digital Sky Survey and 2dF QSO redshift survey, have too low sampling rates for QSOs, $\bar{n} \sim 10^{-3}/(40\,h^{-1}{\rm Mpc}_z)^3$, to obtain comparable constraints.

Matsubara, T. & Suto, Y. 1996, ApJ, 470, L1
Matsubara, T. & Szalay, A. S. 2001, ApJ, 556, L67

54

Multivariate Monte Carlo Methods with Clusters of Galaxies

J. R. Peterson[1], J. G. Jernigan, S. M. Kahn, F. B. S. Paerels, J. S. Kaastra, A. Miller, J. Carlstrom

ABSTRACT We describe a novel Monte Carlo approach to both spectral fitting and spatial/spectral inversion of X-ray astronomy data, and illustrate its application in the analysis of observations of clusters of galaxies. The X-ray events are directly compared with simulations using multivariate generalizations of the Kolmogorov-Smirnov and the Cramér-von Mises statistic. We demonstrate this method in studying the soft X-ray spectra of cooling-flow clusters with the Reflection Grating Spectrometers (RGS) on the XMM-Newton observatory. We also show preliminary results on simultaneously inverting X-ray and interferometric microwave Sunyaev-Zeldovich cluster data using a Monte Carlo technique. Various techniques are applied to simulate radiative transfer effects, model spatially-resolved sources, and simulate instrument response. We then apply statistical tests in the multi-dimensional data space.

Clusters of galaxies contain large amounts of X-ray emitting plasma. It can be used to study important physical processes and answer many cosmological questions concerning the chemical and thermodynamical evolution of dense regions of the universe. The analysis of X-ray data from clusters poses interesting data analysis problems. X-ray photons are detected with three measurements related to two spatial positions and the intrinsic photon energy. This makes the data multi-dimensional. Additionally, only 10^4 to 10^5 photons sparsely fill the multi-dimensional data space.

We have employed a number of Monte Carlo techniques to study X-ray clusters of galaxies to attempt to reproduce the detected data (Peterson, Jernigan, and Kahn, in preparation). A spectral model that varied spatially was used along with an instrument Monte Carlo of the Reflection Grating Spectrometers on the XMM-Newton observatory to study the soft X-ray spectrum of the galaxy cluster Abell 1835 (Peterson et al. 2001, A&A 365).

[1]Department of Astronomy, Columbia University

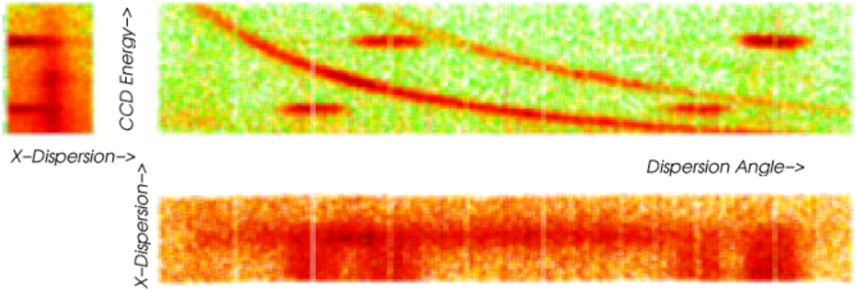


FIGURE 54.1. Three projections of the data from the galaxy cluster, A S 1101. The two curved lines in the upper right plot are the first and second order dispersed spectrum. The three detected values are the dispersion angle, the CCD energy, and the sky angle perpendicular to the spectrometer (x-dispersion).

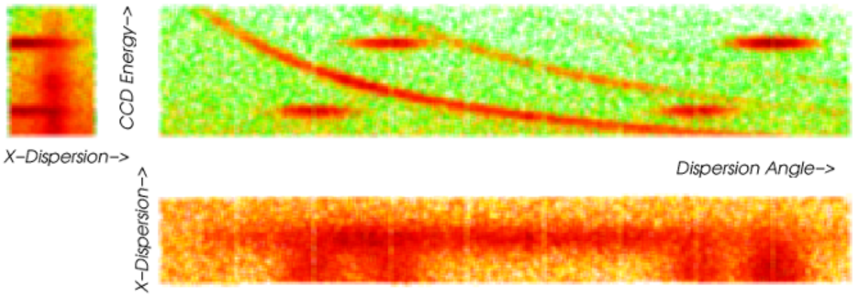


FIGURE 54.2. Simulation of the cluster in Figure 1.1 using a spectral-spatial model for the X-ray emission. It is compared globally using multivariate statistics and then details of the astrophysical model can be compared to find inconsistencies in the model. Spatial variations of emission lines can also be studied.

Figures 1.1 and 1.2 shows the detected photons and simulated photons of a galaxy cluster. A Monte Carlo approach also naturally handles difficult radiative transfer problems (Xu et al., ApJ submitted, 2001).

X-ray data and measurements of the distortion of the cosmic microwave background through the Sunyaev-Zeldovich (SZ) effect can give joint constraints on the density, temperature, and clumping of the intracluster medium at each projected spatial position. Figure 2 shows the inversion of interferometric SZ data through Monte Carlo techniques. Future analysis may allow us to place further constraints on the complex thermal and spatial structures in the cluster plasma.

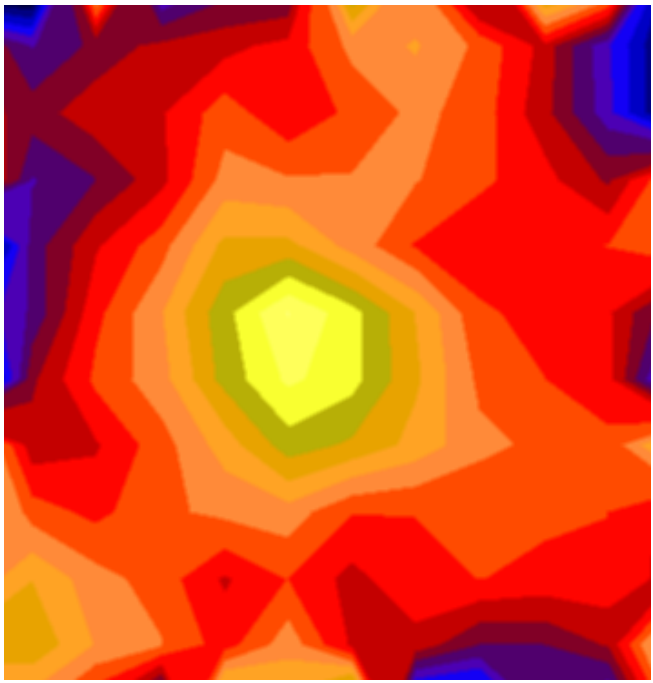


FIGURE 54.3. Inverted SZ image of the galaxy cluster Abell 1835 using Monte Carlo techniques.

This page intentionally left blank

55

A New Tool for Automated Classification of Astronomical Images

Ninan Sajeeth Philip[1], **Yogesh Wadadekar, Ajit Kembhavi and K. Babu Joseph**

Difference Boosting Neural Network (DBNN) is a variant of the Naive Bayesian Neural Network that assume parameter independence for computing the Bayesian Probability. Parameter independence is generally uncommon in practice and the performance of Naive Bayesian Networks degrades when the condition is not satisfied. DBNN, however, does not strictly require the parameters to be independent.

The underlying principle used by DBNN is that even when the parameters are correlated, there always exist some range of allowed values for each parameter given the range of the other parameters. DBNN uses a lookup table and a window function to make fast and robust guesses about the parameters while predicting the class of an example. The lookup table is generated during the training of the network. The table holds the lowest and the highest values allowed for each parameter given the class of the object. While computing the Bayesian Probability for membership to a class, the computed value is reduced to one fourth its value if the range of the parameter happens to be outside the value specified in the lookup table. This is the job of the window function. There could exist situations in which such a window function alone is not able to make adequate classification of objects. DBNN handles this situation by assigning a weight function to each of the parameters given the class of the object. The weight function is updated during the training cycle in such a way that the differences in the parameters are highlighted to make the classification.

We constructed our training set from the R band image of the publicly available NOAO Deep Wide Field Survey (NDWFS) images. We chose to use this data because it has a high dynamic range, large area coverage and high sensitivity that allowed us to maintain uniformity between the moderately large training set and numerous test sets. The training set was carefully constructed from a randomly selected subimage of 2001x2001 pixel region in the R-band image. This image has the best seeing conditions

[1]Cochin University of Science and Technology

among the data currently released. The objects were largely in the Kron-Cousins magnitude range 20-26.

For classification we used three derived parameters from the parameters extracted by the SExtractor package.

- **Elongation measure:** This is the logarithm of the ratio of second order moments along the major and minor axis of the lowest isophote of the object. For a star, the ratio should be near unity. For our training set, this ratio is different from unity because of the slightly elliptical point spread function.
- **Standardized FWHM measure:** This is the logarithm of the ratio of the full width half maximum (FWHM) of the object (obtained from a Gaussian fit to the intensity profile) to the FWHM of the point spread function for the image.
- **Gradient Parameter:** This is the logarithm of the ratio of the central peak count to the standardized FWHM measure of the object.

Our training procedure on the $\sim$ 400 objects in the training dataset took 0.23 seconds on an Intel Pentium III processor running at a clock speed of 700 MHz. Such short training times are invaluable when one has to optimally deal with large datasets that are collected and processed over a significantly wide span of time, demanding repeated retraining of the classifier to account for variations in observing conditions and the parameters chosen for classification. Data from large surveys fall into this category.

The performance of the network was tested on two sub regions of the NDWFSJ1426p3456 field. The object catalogs for the test sets were constructed using the same SExtractor configuration as for the training set. The results are shown in Table 1. The classification accuracy is seen to be marginally better than that of SExtractor.

TABLE 55.1. Comparison of classification accuracy of the DBNN and SExtractor on the NDWFS data.

Label	Stars	Galaxies	Total	Accuracy SExtractor	Accuracy DBNN
Training	87	321	408	97.55 %	
Test 1	72	233	305	97.05 %	97.38 %
Test 2	99	289	388	97.94 %	98.45 %

The source code and the full documentation for the DBNN software described here may be downloaded from the URL:

http://www.iucaa.ernet.in/$\sim$ nspp/dbnn.html

This work made use of images and data products provided by the NOAO Deep Wide-Field Survey (Jannuzi and Dey 1999).

56

Parameter Estimation via Neural Networks

Nicholas G. Phillips[1] **and A. Kogut**

ABSTRACT We use neural networks for astrophysical parameter estimation in the context of models of the cosmic microwave background (CMB). Our method allows for a Bayesian analysis and recovers results comparable to standard maximum likelihood methods when tested on simulated CMB anisotropy maps. We find the computational cost for this method scales with the map size as $N_{\rm CPU} \sim N_{\rm pix}^{1.5}$.

Neural nets can estimate parameters even from stochastic models where the input patterns are intrinsically random. We use Multi-Layer Perceptron neural networks with back propagation training [1]. There is one input neuron per input pixel, a single hidden layer and one output unit. The details of this work can be found in [2].

Focusing on estimating a single parameter, we start by choosing a pair of parameter values that bracket the range to test. We generate realizations of the model at each value and train networks to differentiate between the two sets. The networks are then presented with sets of realizations drawn on a grid of parameter values spanning values between the training values above. By using a committee of 50 networks, we determine the parameter probability distributions for any given committee consensus on the network output. Once this process is completed, we have all the priors necessary to conduct a Bayesian analysis of an unknown input pattern.

To test our method, we simulate COBE-DMR full-sky maps of the CMB anisotropy, parameterized by the spectral index n [3]. Instrumental effects are accounted for by including noise [4] and excluding the pixels dominated by foreground galaxy emission [5]. galaxy cut. patterns, the networks the pixels spectral indices for 1000 patterns for $n = 1.40$. The mean recovered value is $n = 1.30$ and from a Bayesian analysis, we find we have a 68% confidence interval of $[0.94, 1.66]$, an uncertainty similar to an maximum likelihood analysis [6].

We determine how the computational needs scale with the problem size by presenting our networks with different size patterns. Our patterns are circular patches of our CMB maps, with the range of patch size covering 1.5

[1]RITSS, NASA Goddard Space Flight center

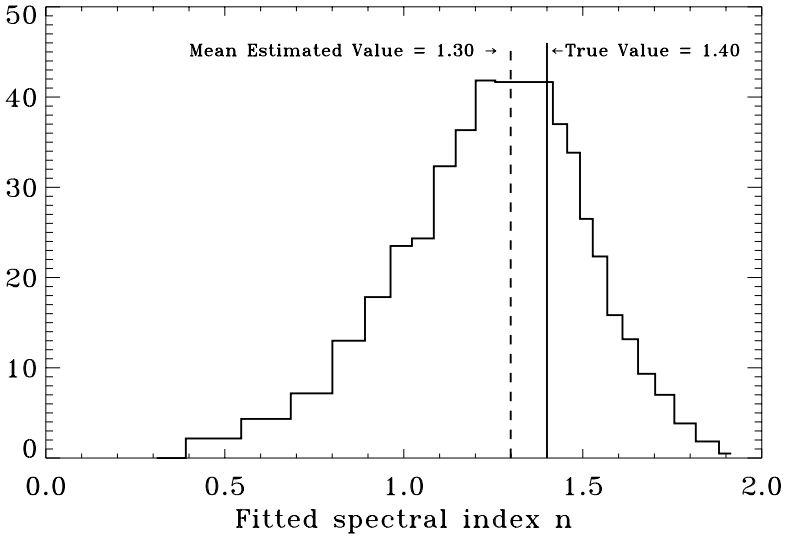


FIGURE 56.1. Sample outputs for given input maps

orders of magnitude of pattern size. For each patch size, we determine how many hidden units and number of training passes are needed to achieve a preset level of discrimination. The results are shown in Figure 2, from which we find the CPU cost scales with the patch size as $N_{\rm CPU} \sim N_{\rm pixel}^{1.5}$. We expect our method to readily scale to anticipated mega-pixel datasets.

56.1 References

[1] Rumelhart, D.E., Hinton, G.E., and McClellend, J.L. 1986, in *Parallel Distributed Processing*, Eds. D.E. Rumelhart, J.L. McClellend and the PDP Research Group (MIT Press: Cambridge)

[2] Phillips, N. G. and Kogut A., submitted to *ApJ*; preprint astro-ph/0108234

[3] Bond, J. R., and Efstathiou, G. 1987, MNRAS, 226, 655

[4] Bennett, C. L., *et al.* 1996, ApJ, 464, L1

[5] Banday, A. J., *et. al* 1997, ApJ, 475, 393

[6] Górski, K. M., *et. al* 1994, ApJL, 430, L89

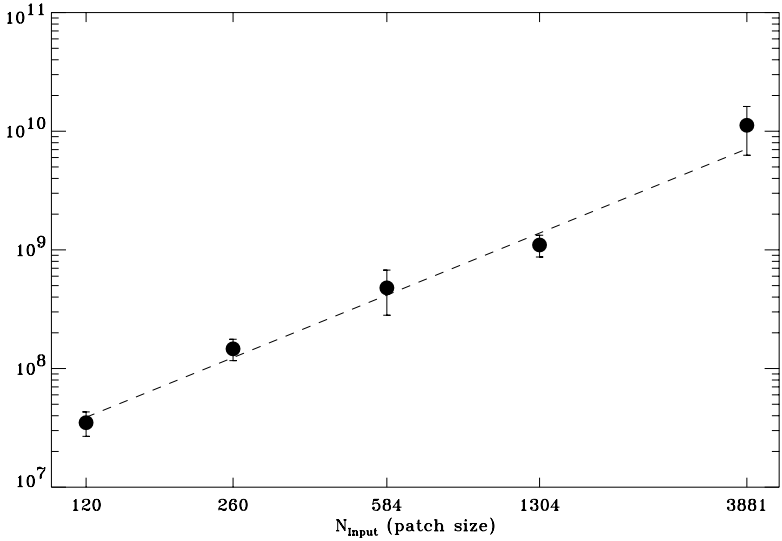


FIGURE 56.2. Scaling of computational cost for CMB data

This page intentionally left blank

57

Correlations at Large Scale

M. J. Pons–Bordería[1], V. J. Martínez, B. López–Martí and S. Paredes

ABSTRACT We show point processes generated in different ways and having different structure, presenting very similar power-law two–point correlation functions at small scales and quite different shapes at large scales.

The two-point correlation function $\xi(r)$ measures the excess probability —with respect to a Poisson distribution— of, given a point of a process, finding another point at a distance r of the first one ([2]). It is well-known that $\xi(r)$, for the galaxy distribution, fits well a power law at small scales ($r < 10h^{-1}$Mpc). Here we analyze several point processes having similar power-law shapes at small scales, but different visual aspect. The differences are encapsulated in the behavior of the correlation function at large scales as well as in other statistical measures ([3]). The analyzed point processes are the following:

1. **COX** A segment Cox process has been produced by randomly scattering segments of length $l = 10h^{-1}$Mpc with a density $\lambda_s = 0.0013$ within a cube of side $100h^{-1}$ Mpc, and then randomly distributing points on the segments with density $\lambda_l = 0.76923$ per unit length. An analytical expression for $\xi(r)$ depending on these parameters is known ([4]).

2. **VORONOI** We have considered the vertices of a Voronoi tessellation ([5]) constructed from a binomial field with 1500 nuclei. There are 10085 vertices (events of the point process) within a cube of sidelength $100\sqrt{2}h^{-1}$Mpc.

3. **VIRGO** From a Λ-CDM N-body simulation of the Virgo Consortium, a sample of simulated galaxies has been constructed by the GIF project ([1]). The sample contains $N = 15445$ galaxies within a cube of sidelength $141.3h^{-1}$Mpc.

At small scale the behavior of $\xi(r)$ is very similar for the three clustering models — power-law functions with comparable exponents. The differences

[1]Department Matemática Aplicada y Estadística, Univ. Politécnica de Cartagena (Spain)

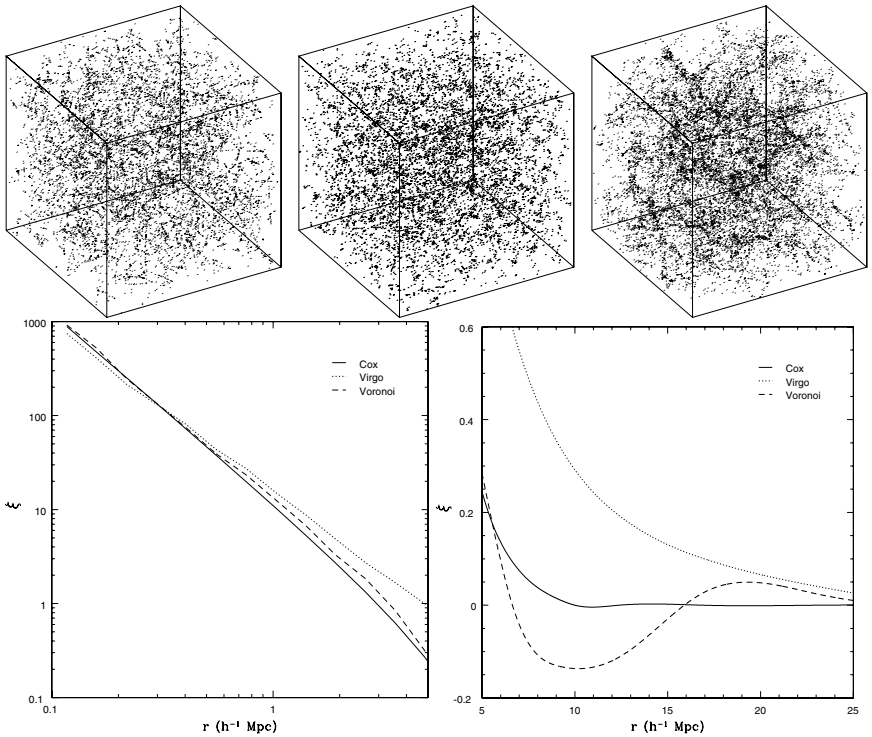


FIGURE 57.1. Top: from left to right, the COX, VORONOI and VIRGO point processes described in the text. Bottom: ξ for the three processes at small scales (left panel) and at large scales (right panel).

of the clustering properties of the three point processes are better appreciated at large scale. For the Cox process $\xi(r) = 0$ for $r \geq l$ whereas for the N-body simulation $\xi(r)$ approaches zero more gradually, taking place the first zero crossing at $\sim 30h^{-1}$Mpc. For the Voronoi vertices model, $\xi(r)$ behaves with damping oscillations around the zero value.

We conclude by stressing that the behavior of ξ at large scales provides us with crucial information about the clustering properties of point processes presenting similar power-law shapes at small scales. Appropriate estimators had to be used to obtain this information, that can be complemented with other statistical measures ([3]).

Acknowledgments This work was supported by the Spanish MCyT project AYA2000-2045.

[1] Kauffmann, G. et al., 1999, MNRAS, 303, 188
[2] Martínez, V. J. and Saar, E., 2002, this volume.
[3] Martínez, V. J. et al. 2002, in preparation.
[4] Pons-Bordería, M.J. et al., 1999, ApJ, 523, 480
[5] van de Weygaert, R. and Icke, V., 1989, A&A, 213, 1

58

Constraining Cosmological Models by the Cluster Mass Functions

Nurur Rahman[1] and Sergei F. Shandarin

Cluster abundance test puts strong constraints on the cosmological parameters such as matter density ($\Omega_0 = \Omega_b + \Omega_{cdm} + \Omega_{hdm}$) in the Universe and the amplitude of the mass density fluctuations (σ_8). We present a comparison between two observational and three theoretical mass functions for eight cosmological best-fit models suggested by the data from recently completed BOOMERANG-98, MAXIMA-1 Cosmic Microwave Background anisotropy experiments as well as Peculiar Velocities and type Ia Supernovae observations. Further details of this work can be found in the Astrophysical Journal Letters 550:L121 (2001).

Analytical mass functions are obtained from three sources: Press & Schechter (ApJ, 187, 425 1974), Lee & Shandarin (ApJ 500, 14 1998) and Sheth & Torman (MNRAS 308, 119 1999).

Our results are shown in the accompanying table and figures. The findings may be summarized as follows:

1. We find that no model is in agreement with the X-ray clusters abundance at $\sim 10^{14.7} h^{-1} M_\odot$.
2. The BOOM+MAX+*COBE*:I, Refined Concordance and ΛMDM models are in good agreement with the optical clusters abundance.
3. The P11 and Concordance models predict slightly lower cluster abundances than observed at $\sim 10^{14.6} h^{-1} M_\odot$.
4. The BOOM+MAX+*COBE*:II and PV+CMB+SN models predict slightly higher cluster abundances than observed at $\sim 10^{14.9} h^{-1} M_\odot$.
5. The non-flat MAXIMA-1 model is inconsistent with the observation at the entire mass range.

Our analysis shows that: 1) the Universe has low matter density ($0.3 < \Omega_0 < 0.4$) and high dark energy ($\Omega_{dark} > 0.6$) density; 2) a small amount of neutrino density ($\Omega_\nu \sim 0.03$) can be reconciled with the observation; 3)

[1]Department of Physics and Astronomy, University of Kansas

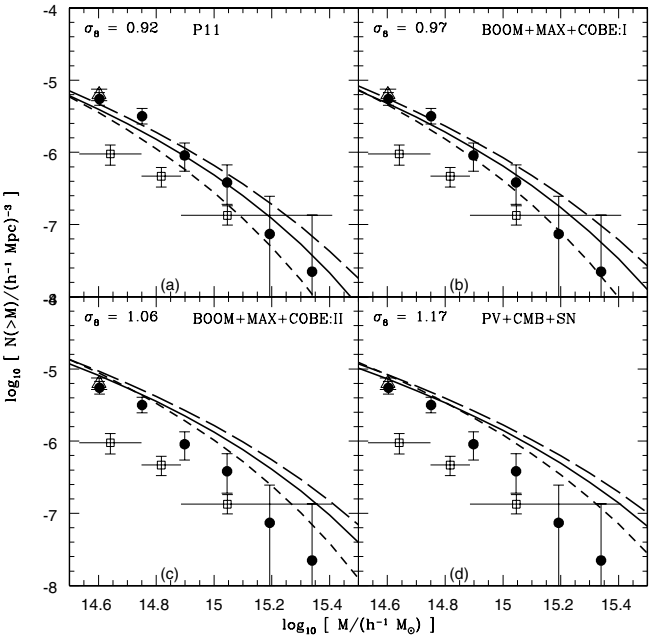
a relatively low normalization, $0.8 < \sigma_8 < 1.0$, suggesting a slight galaxy formation bias value ($b = \frac{1}{\sigma_8}$). The analysis justifies the present notion of the low matter density ($\Omega_0 \sim 0.40$) Universe dominated by some unknown dark energy density ($\Omega_{dark} \sim 0.60$).

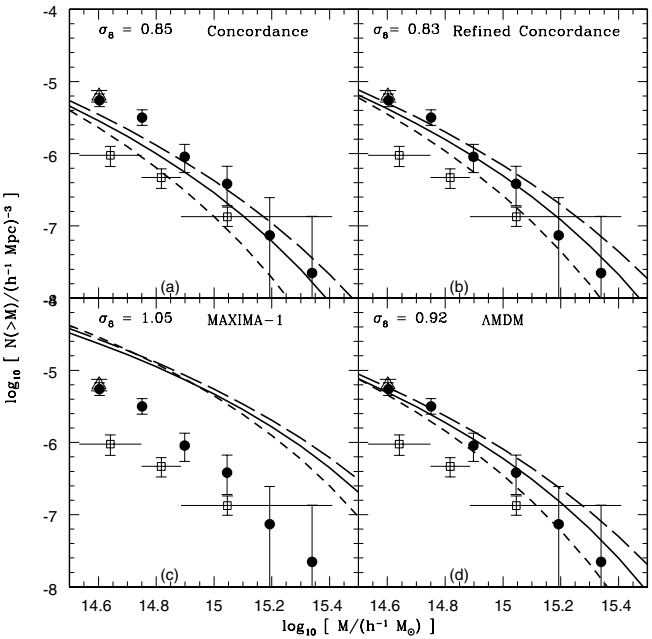
Table of the Cosmological Models

Models	Parameters					
	Ω_b	Ω_{cdm}	Ω_Λ	n_s	h	σ_8
1	0.045	0.255	0.7	0.95	0.82	0.92
2	0.045	0.255	0.7	0.975	0.82	0.97
3	0.036	0.314	0.65	0.95	0.80	1.06
4	0.035	0.245	0.72	1.0	0.74	1.17
5	0.03	0.27	0.7	1.0	0.68	0.85
6	0.05	0.33	0.62	0.91	0.63	0.83
7	0.07	0.61	0.23	1.0	0.60	1.05
8	0.037	0.303	0.69	1.02	0.71	0.92

1) P11: Lange et al. 2001, Phys. Rev. D., 63, 042001;
2) BOOM+MAX+*COBE* I: Jaffe et al. 2001, Phys. Rev. Lett. 86, 3475;
3) BOOM+MAX+*COBE* II : Hu et al. 2001, ApJ, 549, 669;
4) PV+CMB+SN: Briddle et al. 2001,MNRAS, 321, 333;
5) Concordance: Ostriker & Steinhardt 1995, Nature, 377, 600;
6) Refined Concordance: Tegmark et al. 2001, Phys. Rev. D, 63, 04007;
7) MAXIMA-1($\Omega_{tot} = 0.91$):Balbi et al. 2000, ApJ, 545, L1;
8) ΛMDM ($\Omega_{tot} = 1.06$): Durrer & Novosyadlyj 2001, MNRAS, 324, 560.

FIGURE 58.1. (following page) Observational cmfs (cumulative mass functions) measured for virial mass are compared with different theoretical predictions. Top panel: (a) P11, (b) BOOM+MAX+*COBE*: I, (c) BOOM+MAX+*COBE*: II and (d) PV+CMB+SN. Bottom panel: (a) Concordance, (b) Refined Concordance, (c) MAXIMA-1 and (d) ΛMDM. The short dash line is n_{PS}, long dash line is n_{λ_3} and solid line is n_{ST}. The filled circles are the observational data points corresponding to virial masses determined by Girardi et al. (ApJ, 506, 45, 1998). The open squares are those determined by Reiprich et al. (X-Ray Astronomy 2000, R. Giacconi et al. eds, ASP Conf. 234, 405, 2001) The error bars are in 1σ limit along the vertical direction. Horizontal bars indicate the bin size. The open triangle is the value of the cmf for masses estimated within the $1.5h^{-1}$ Mpc radius by Girardi et al. (1998).


σ8 = 0.92 P11
σ8 = 0.97 BOOM+MAX+COBE:I
σ8 = 1.06 BOOM+MAX+COBE:II
σ8 = 1.17 PV+CMB+SN
(a)
(b)
(c)
(d)
log10 [N(>M)/(h⁻¹ Mpc)⁻³]
log10 [M/(h⁻¹ M⊙)]


σ8 = 0.85 Concordance
σ8= 0.83 Refined Concordance
σ8 = 1.05 MAXIMA−1
σ8 = 0.92 ΛMDM
(a)
(b)
(c)
(d)
log10 [N(>M)/(h⁻¹ Mpc)⁻³]
log10 [M/(h⁻¹ M⊙)]

This page intentionally left blank

59

Analysing Cosmic Large Scale Structure using Surrogate Data

C. Räth[1], W. Bunk, P. Schuecker, J. Retzlaff, M. Huber, G. Morfill

ABSTRACT Methods derived from nonlinear time series analyses are applied to three-dimensional point distributions as they are typical in the analysis of the cosmic large scale structure. Using the technique of constrained randomisation we generate for a given data set surrogate data sets which have the same linear properties (power spectrum) as well as the same density amplitude distribution but different morphological features. It is shown that the original data set can be discriminated from the surrogates by analysing the local scaling properties of the point sets as measured by weighted scaling indices.

With the method of constrained randomisation (Theiler et al. 1992) an ensemble of surrogate data sets, which share properties of a given point distribution, is generated. The analysis of the original and surrogate data sets with measures, which are sensitive to nonlinearities, yields valuable information about the existence of nonlinear correlations in the data. On the other hand one can test whether given statistical measures are able to account for higher order and/or nonlinear correlations by applying them to original and surrogate data sets. In this work we want the surrogate data sets to have the same power spectrum and the same amplitude distribution as a given data set. A refined approach which fulfills these requirements quiet well is called iteratively refined surrogates (Schreiber & Schmitz 2000). It consists of alternating fourier transformation and rescaling steps. By construction, the data sets have the same two-point correlation function whereas their topological features are very different. Nonlinear structural measures (e.g.Halsey et al. 1986) can account for these morphological differences in point sets. In this study weighted scaling indices are calculated for characterising the *local* scaling properties of a point set. Consider a set of N points $P = \{\vec{r}_i\}, i = 1, \ldots, N$. For each point the local weighted cumulative point distribution ρ is calculated. With a class of exponentials as weighting functions it can be written as $\rho(\vec{r}_i, R) = \sum_{j=1}^{N} e^{-(d_{ij}/R)^n}$, $d_{ij} = \|\vec{r}_i - \vec{r}_j\|$,

[1]Centre for Interdisciplinary Plasma Sciences, Max-Planck-Institut für extraterrestrische Physik

where we use $n = 2$. The weighted scaling indices $\alpha(\vec{r}_i, R)$ are defined as the logarithmic derivation of $\rho(\vec{r}_i, R)$ with respect to the length scale R, $\alpha(\vec{r}_i, R) = \frac{\partial \log \rho(\vec{r}_i, R)}{\partial \log R}$. Structural components of a point distribution are characterised by the values of α of each point belonging to a certain kind of structure (e.g. $\alpha \approx 0$: cluster, $\alpha \approx 1$: filaments, $\alpha \approx 2$: sheets, etc.). The scaling indices for the whole point set under study are comprised in the probability distribution $P(\alpha)$.

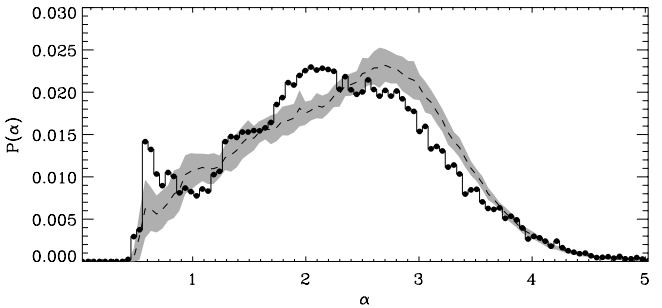


FIGURE 59.1. Spectrum of scaling indices for $R = 4$ Mpc/h (black line) of the original data set (open cold dark matter model) and the 1σ-error region (gray area) as derived from 20 surrogate data sets with the same power spectrum and amplitude distribution. The original data set can clearly be discriminated against the surrogates.

The results of our work yield further evidence that the linear global measures like the two-point correlation function and power spectrum are only of limited use characterising the morphology of a given structure. This is due to the fact that these second order statistical measures are insensitive to the distributions of Fourier *phases* which are responsible for the fine details of cosmic structures. The development of more sophisticated *nonlinear local* descriptors which are based on the analysis of the scaling behaviour of the point distribution can offer new possibilities to refine statistical methods so that previously ignored subtle but important features can now be both detected and quantitatively characterised. This may allow for a better discrimination between models with very similar power spectrum. In this context the method of surrogate data is a vital tool with which the quality of newly developed measures can be assessed in terms of sensitivity to different topological features and in terms of discriminative power.

Theiler J., Eubank S., Longtin A. et al., 1992, *Testing for Nonlinearity in Time Series: The Method of Surrogate Data*, Physica D **58**, 77.

Schreiber T., Schmitz A., 2000, *Surrogate Time Series*, Phys D **142**, 346.

Halsey T.C., Jensen M.H., Kadanoff L.P. et al., 1986, *Fractal Measures and their Singularities: The Characterization of Strange Sets*, Phys. Rev. A **33**, 1141.

60

Delaunay Recovery of Cosmic Density and Velocity Probes

W. E. Schaap[1] **and R. van de Weygaert**

ABSTRACT Optimally resolved one-dimensional density and velocity profiles through cosmological N-body simulation are constructed by means of the Voronoi-Delaunay tessellation reconstruction technique. In a fully self-adaptive fashion a strikingly detailed view of the density features and the corresponding cosmic motions is recovered.

In essence, N-body simulations of cosmic structure formation are supposed to represent a discrete sampling of underlying continuous density and dynamical fields. The recovery of the corresponding continuous fields is a less than trivial exercise. They are often distorted by manipulated, user-dependent and therefore biased reconstruction schemes. This makes it in particular cumbersome to deal self-consistently with the characteristically multi-scale hierarchical nature of cosmological density fields. As significant is the failure to recover crucial structural aspects of the salient and frequently sharply defined anisotropic – filamentary and wall-like – patterns in the cosmic matter distribution.

Recently, Schaap & van de Weygaert [3] have developed a fully self-adaptive and unbiased method to reconstruct density and related dynamic fields from a discrete and in general nonuniformly sampled set of point locations. It is based on the stochastic geometric concept of Voronoi/Delaunay tessellations and forms an elaboration on the formalism first proposed by Bernardeau & van de Weygaert [1] for the case of assessing the statistical properties of cosmic velocity fields.

The application of the method to a large 256^3 GIF N-body simulation (LCDM, $141.3h^{-1}$Mpc, courtesy: S. White) [2, 3] provides a beautiful illustration of its sizeable promise. The top panel of fig. 1 presents the particle distribution in a slice through this simulation. The corresponding density field determined through the Delaunay technique is shown in the adjacent panel. Notice how much better than the saturated particle plot this density field manages to elucidate the wealthy and detailed structural features present in this cosmic volume, superbly rendering its high density contrasts.

While the image of the density field already provides evidence of its oper-

[1]Kapteyn Institute, University of Groningen

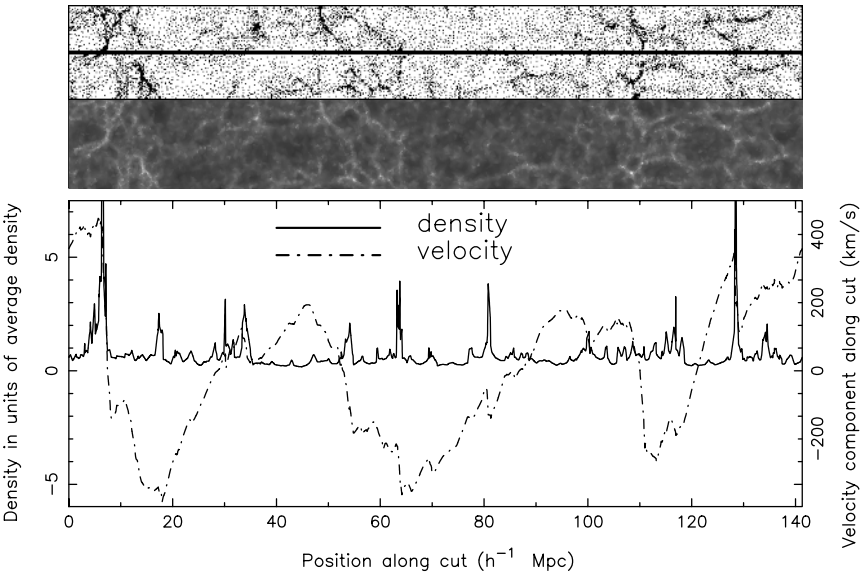


FIGURE 60.1. Slice through a GIF N-body simulation (top), the Delaunay recovered density (center) and density and velocity profiles along the central line (bottom). (We are grateful to S. White for initiating these calculations.)

ation, it is through the objective assessment of the density profile along the central axis of the slice that the success of the method is fully manifested (solid line, lower panel). Evidently, the Delaunay technique yields a faithful representation of the density field over an impressive dynamic range, encompassing gently varying and extended low-density regions as well as the high density contrasts found in compact objects, be it either condensed clumps or the flattened dimension(s) of filaments and walls. Even more compelling is the correlation with the corresponding velocity field along the same line (dashed line). Largely superseding the poor velocity resolution in the conventionally shotnoise-dominated void regions it succeeds in reproducing the matter depleting *super-Hubble* like peculiar velocity flows (e.g. void at $\approx 123h^{-1}$Mpc). Even more striking are the sharp velocity transitions encountered at the location of high density peaks, indicating the large induced infall motions along various directions towards these features (e.g. the peak at $\approx 7h^{-1}$Mpc).

60.1 References

[1] Bernardeau, F., van de Weygaert, R., 1996, MNRAS 279, 693

[2] Kauffmann, G., et al., 1999, MNRAS 303, 188

[3] Schaap, W.E., van de Weygaert, R., 2000, A&A 363, L29; 2001, in preparation

61

A Large Proper Motion Survey of the Pleiades Cluster

J. Souchay[1] and E. Aleshkina

61.1 Introduction

The accurate determination of proper motions of the stars can be achieved through the analysis of Schmidt photographic plates, at the condition that these plates have been obtained at epochs covering a relatively large time span. Applied to the Pleiades cluster, this kind of analysis can drastically improve the knowledge of the cluster, noticeably with respect to two points : the first one is that it can lead to a very trustable evaluation of the membership probability of a given star of the field to the cluster. At second, when this probability has been evaluated, photometric information obtained through various filters, can lead to fundamental astrophysical information. For this purpose, we have gathered at all 51 plates: 25 of them have been taken from the Tautenburg Schmidt telescope (134/203/401) in Germany, 13 of them with the CERGA Schmidt telescope (90/152/316) in France , and 13 with the Kizo Schmidt telescope in Japan.

61.2 Procedures for astrometric measurements

For each plate the first step of our analysis consists in identifying, as far as it is possible, any star of the field with its corresponding counterpart in a master plate.The recognition of the common stars is made possible by the intermediary of PPM astrometric standards (Roser and Bastian,1988). Then one of the plates, selected among the set of 51 ones according to appropriate criteria (colour, deepness, center, intermediate time), is selected as a master plate.

Each of the plates has been scanned with the Machine Automatique a Mesurer pour l'Astronomie (MAMA) located at Paris observatory. The pixel size of each of the 1024 photodiodes representing the linear scanning bar of the MAMA is 10μm. According to technical tests, the precision

[1]Observatoire de Paris/DANOF

of the measurements is of the order of 0.1 μas, which corresponds to a few milliarcseconds as given the scale of the plates (around 60"/mm). For further details concerning the specific characteristics and performance of the MAMA, we can refer to Berger et al. (1991).

Once the recognition of common stars with those of the master plate has been done, the reduction consists in converting the bi-dimensional coordinates x_i, y_i of a given star of the i-th plate into fictitious equivalent coordinates on the master plate, symbolized by X_i^{pm}, Y_i^{pm}. The corresponding algorithm has the following form:

$$X_i^{mp} = \sum_{1<j,k<n} a_i^{jk} x_i^j y_i^k \quad Y_i^{mp} = \sum_{1<j,k<n} b_i^{jk} x_i^j y_i^k \tag{1}$$

The coefficients a_i^{jk} and b_i^{jk} are chosen by a least-square algorithm in such a way that the coordinates X_i^{mp} and Y_i^{mp} calculated through (1) become as closest as possible to the corresponding respective positions X_0^{mp} and Y_0^{mp} in the master plate. In other words $\sum(X_i^{mp} - X_0^{mp})^2$ and $\sum(Y_i^{mp} - Y_0^{mp})^2$ are set to their minimum values by the judicious choice of the coefficients a_i^{jk} and b_i^{jk}. Practically we choose: n=4 as the maximal degree of our polynomials.

In order to optimize the quality of the plate-to-plate transformations, we adopt an iterative procedure consisting in eliminating step by step any star whose the absolute difference $\left[X_i^{mp} - X_0^{mp}\right]$ or $\left[Y_i^{mp} - Y_0^{mp}\right]$ between the converted coordinate and the reference coordinate in the master plate, exceeds a threshold value, which has been set to $2.0\sigma_X$ and $2.0\sigma_Y$, where σ_X and σ_Y represent respectively the following r.m.s.:

$$\sigma_X = \sqrt{\frac{\sum_{i=1,N}\left(X_i^{mp} - X_0^{mp}\right)^2}{N}} \quad \sigma_Y = \sqrt{\frac{\sum_{i=1,N}\left(Y_i^{mp} - Y_i^{mp}\right)^2}{N}} \tag{2}$$

N being the total number of stars. The iteration stops when no rejection occurs. Then, we can notice the evolution w.r.t. the time and a unique reference frame, of the positions $X_{mp}^{t_i}$ and $Y_{mp}^{t_i}$ of a given star at a time t_i (t_i being the date when the ith. plate has been taken), that is to say its proper motion, which is not absolute but evaluated with respect to a zero value which characterizes the average proper motions of all the stars of the field, each of them being dragged by galactic rotation.

61.3 Proper motion determinations

To accept a given star in our proper motions survey this star has to be identified in 3 different plates separated by a large time span. When this constraint is satisfied, we gather all the positions (X_{mp}^l, Y_{mp}^l) of the given star, l ranging between 1 and m, m having a value between $m = 3$ and

$m = 50$ according to the star, t_l being the corresponding epoch. Then each of the time series $X^l_{mp}(t_l)$ and $Y^l_{mp}(t_l)$ is fitted independently by a straight line whose the slope gives directly the proper motion according to the Y axis ($\mu_\alpha \cos\delta$) and the X axis (μ_δ). Here also the fit is carried out by the intermediary of least-square analysis, the two parameters to be determined for each axis being the original value ((X^0_{mp} or Y^0_{mp}) at $t_0 = 1961.05$, and the value of the proper motion $\mu_\alpha \cos\delta$ or μ_δ. As it was the case for the plate-to-plate correspondence (last section) the procedure is iterative. Two tests have been carried out, the first one with a rejection threshold at 2σ and the second one with a rejection threshold at 1.5σ.

61.4 Membership probability

We have plotted the VPD (Vector Point Diagram) for the proper motions of the Pleiades field stars respectively with a 2.0 $\times\sigma$ rejection threshold and a 1.5 $\times\sigma$ rejection threshold, after fitting the locations on the master plate with a straight line, by the means of least-square analysis. We notice exactly the same kind of pattern that was already shown and analyzed by Schilbach et al.(1995) and Meusinger et al.(1996), which consists in a large clustering of the proper motions around the origin, which represents the motions of the field stars, and a compact secondary clustering, centered on the point with coordinates $[\mu_\alpha \cos\delta]_m = +17.0 mas/y$ and $[\mu_\delta]_m = -39.3 mas/y$. The statistical way to evaluate the membership probability P_{pm} of a given star to the cluster of a star inside the clustering VPD zone is described in detail by Meusinger et al. (1996)

$$P_{pm} = exp\Big[-\frac{1}{2}\Big(\frac{\Delta\mu}{\sigma_\mu(V)}\Big)^2\Big] \text{with}$$

$$(\Delta\mu)^2 = (\mu_\alpha \cos\delta - [\mu_\alpha \cos\delta]_m)^2 + (\mu_\delta - [\mu_\delta]_m)^2 \qquad (3)$$

assuming a non-correlated bi-dimensional Gaussian distribution, with dispersions: $\sigma^2_\mu = \sigma^2_{\mu,meas} + \sigma^2_{\mu,intr}$ where $\sigma_{\mu,intr}$ is the intrinsic velocity dispersion, and $\sigma_{\mu,meas}$ represents the accuracy of the proper motion measurement. According to the law above, a star is considered as a cluster member if the distance of its locus on the VPD from the center of the Pleiades is smaller than a prescribed distance limit, $\Delta\mu_{lim} = k\sigma_\mu(V)$ so that the selection of the Pleiades stars set is given by the condition: $P^{Pleiades}_{pm} > exp(-k^2/2)$ the value of k being empirically determined. In a previous study, Souchay and Schilbach (1995) have found that 332 probable Pleiades stars are located inside a circle with a radius of 4 mas/year around the center of the clustering whose the coordinates have been determined above.

References

Berger J., Cordoni J.P., Fringant A.M., Guibert J., Moreau O., Reboul H., Vanderriest C., 1991, *Astron. Astrophys.* **87**, p. 389

E. Schilbach, N. Robichon, J. Souchay, J. Guibert, 1995, *Astron. Astrophys.* **299**, p. 696-709

H. Meusinger, E. Schilbach, J. Souchay, 1996, *Astron. Astrophys.* **312**, p. 833-844

Roser, S., Bastian, U., 1988 , *Astron. Astrophys.* **XXX**, p. XXX

Souchay,J., Schilbach, E., *The Future Utilisation of Schmidt Telescopes*, ASP Conferences Series, Vol. 84, 1995.

62

Bayesian Spectral Analysis of "MAD" Stars

Nondas Sourlas[1], David A. van Dyk, Vinay Kashyap, Jeremy Drake and Deron Pease

62.1 Overview

Computing reliable estimates of coronal metallicity (Z) from X-ray spectra obtained with instruments such as ASCA/SIS and Chandra/ACIS is very difficult, because the sole determinant of Z is the ratio of line to continuum fluxes, which is not well-determined for low-resolution spectra. Here we propose new Bayesian methods which directly model the Poisson nature of the data. Our model also accounts for the Poisson nature of background contamination, blurring due to instrument response, and the absorption of photons in space. The resulting highly structured hierarchical model is fit using the Gibbs sampler, data augmentation, and Metropolis-Hasting. We demonstrate our methods with the X-ray spectral analysis of several apparently coronal metal abundance deficient ("MAD") stars.

62.2 A Poisson Spectral Model

The model is designed to summarize the relative frequency of the energy of photons arriving at a detector. We model the photon counts in each bin as independent Poisson random variables. Specifically, we model the high energy tail of the ASCA spectrum (2.5-7.5 keV) as a combination of a Bremsstrahlung continuum and ten narrow emission lines, included at positions of known strong lines. This source model is combined with instrument response, the effective area of the instrument, and background contamination to model actual observed counts [12].

Statistical analysis is based on two observations of the same source and two of the same background. We use sequential Bayesian analysis for the two source observations; the posterior distribution from the first analysis is used to construct an informative prior for the second. Non-informative

[1]Department of Statistics, Harvard University

priors were used in the first stage. Model fitting proceeds by using the EM algorithm to check for multimodality in the posterior distribution. A Markov chain Monte Carlo algorithm is constructed to sample from that posterior. Three chains are used at each step of the analysis in order to assess convergence. Posterior inference is based on the second half of the draws of all the chains. The sensitivity of our results to the choice of prior was investigated by altering the prior. We used residual plots for both source and background to diagnose the fit of our models.

62.3 Results

We have measured the coronal metallicity for 4 stars, α Aur (Capella), σ Gem, YY Gem, and II Peg. Our result for Capella ($Z = 0.73^{<2.4}_{>0.24}$) are consistent with the independently determined value of $Z = 0.57 - 0.78$ (Brickhouse et al. 2000). Based on our analysis of σ Gem ($Z = 0.81^{<1.1}_{>0.59}$), we find that contrary to classical analyses ($Z = 0.25$) it is *not* a MAD star (photospheric $Z = 0.6$ (Randich et al. 1994). YY Gem is found to have sub-Solar metallicity ($Z = 0.46^{<0.71}_{>0.23}$) but higher than EUVE/SW measurements ($Z \sim 0.1$ (Kashyap et al. 1998). For II Peg, we find $Z = 1.1^{<1.3}_{>0.88}$, higher by a factor ~ 10 than preliminary results from Chandra/HETG; this discrepancy may be due to anomalies in the abundances of other elements, or to a high-temperature component to the emission measure.

Acknowledgments: The author gratefully acknowledges funding for this project partially provided by NSF grant DMS-01-04129 and by NASA contract NAS8-39073 (CXC). This chapter is a result of a joint effort of the members of the Astro-Statistics working group at Harvard University.

62.4 References

[1] van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2001) *The Astrophysical Journal* **548**, 224.

[2] Brickhouse, N. S., Dupree, A. K., Edgar, R. J., Liedahl, D. A., Drake, S. A., White, N. E., and Singh, K. P. (2000) *The Astrophysical Journal* **530**, 387.

[3] Randich, S., Giampapa, M. S., and Pallavicini, R. (1994) *Astronomy and Astrophysics* **283**, 893.

[4] Kashyap, V., Drake, J., Pease, D., and Schmitt (1998) *American Astronomical Society* **192**, 82.01.

63

Stellar Membership in Open Clusters Using Mixture Densities

Antonio Uribe[1], Ruth Barrera, Mario A. Higuera G. and Alvaro Montenegro

ABSTRACT
A view of differrent parametric methods to solve the stellar membership problem in open clusters is given and a new approach is obtained using the EM Algorithm.

Different multivariate normal mixtures density models are found in the literature to discriminate between the open cluster stars and field stars. These models overlap bivariate normal components as parametric statistical models of proper motions data. A first one overlaps a circular and an elliptic normal component (Sanders, 1971; Zhao et al, 1982; Cabrera C., Alfaro E., 1985, Uribe A., Brieva E., 1994, Lattanzi et al.,1991) ; a second one overlaps two elliptic normal components (Sabogal, M. et al., 2001); in a third one the errors of proper motions are taking into account (Zhao et al, 1990; Brieva, E., Uribe A., 1996). The maximum likelihood method is followed in all these cases to find estimates of the parameters of the mixture density, and membership probabilities are found by Bayes theorem and the Bayes rule of minimum error rate of missclassification. Certainly another, nice and precise approach to solve this problem, is found in Dinescu (Dinescu et al., 1996).

We now solve the stellar membership in open clusters using the EM algorithm as explained in Dempster, Laird and Rubin 1977, Wolfe 1970, and mainly using the McLachlan EMMIX software to solve the maximum likelihood equations of an incomplete data problem (McLachlan G. et al.1997, 2000, and http://www.maths.uq.edu.au/ gjm/emmix/emmix.html). This last approach had led to new membership solutions for several galactic clusters, in good agreement with other results found in the literature, as in in the case of the Pleyades cluster, and in NGC654, NGC6530, NGC2244

[1]Observatorio Astronómico Nacional and Departamento de Matemáticas y Estadística, Universidad Nacional de Colombia

and NGC6823.

We start, as it is usual, from a mixture of two weighted normal bivariate heteroscedastic components as a density model of the proper motions data for the field and the cluster stars in a considerd region of the sky. The fact that we do not know to which component each star belongs, leads to an incomplete data problem, and allows the use of the EM algorithm approach to find estimates of the parameters of the mixture density. This procedure requires to find the maximum likelihood functions L and L_c, respectively, for the incomplete and the complete data problem. Then, considering $\log(L_c)$, and following the expectation and the maximization steps that define the EM algorithm, a maximum value of L is found. In the general case of a mixture of g components this is a local maximum of an unbounded likelihood (Hand, 1981). In the case of the mixture of two bivariate normal heteroscedastic components, the no singularity of the covariances matrices Σ_1 and Σ_2 implies the continuity of the Likelihood L, and this local maximum is unique if we restrict the parameters to vary in compact sub intervals of the real line. The spread of the cloud of proper motions data allows to hold these restrictions.

The so called Wolfe equations (Cabrera C., Alfaro E., 1985) are generated by the EM algorithm when we work with a mixture of two normal components. Its solutions lead to estimates of the parameters of the mixture density, and then to a solution of the membership problem.

It also seems appropiate to say that EMMIX is a powerful software to find the parameters of a general finite mixture of g normal or t components following the EM algorithm.

Acknowledgments: We thank for the given support the SCMA III organizers and the the Observatorio Astronómico Nacional, Universidad Nacional de Colombia.

Bibliography

Brieva E. and Uribe A., 1996, Rev. Acad. Col. de Ciencias, 20, 7

Cabrera J. and Alfaro E., 1985, Astron. Astrophys, 150, 298

Dempster A. P., Laird N.M., Rubin D.B., 1977, J.Royal Astr. Soc. B, 39, 1

Dinescu D., Girard T., Van Altena W., 1996, Astron. J.,111, 1205

Hand D. J., 1981, Discrimination and Classification, John Wiley

Lattanzi M., Massone G., Munari U., 1991, Astron. J, 102,177

McLachlan G. and Peel D., 2000, Finite Mixture Models, John Wiley

McLachlan G. and Krishnan Thriyambakam, 1997, The EM Algorithm, John Wiley

Sabogal-Mártínez B.E., García-Varela, J.A., Higuera G. M.A., Uribe A., Brieva E., 2001, Revista Mexicana de Astronomía y Astrofísica, 37, 105
Sanders W., 1971, Astron. Astrophys, 14, 226

Wolfe, J. H. 1970. Multivariate Behavioral Research, 329

Uribe A. and Brieva E., 1994, Astrophysics and Space Science, 214, 171

Zhao J.L., Tian K., Su Z., Yin M., 1982, Chin. Astron. and Astrophys., 6, 293

Zhao J.L., He Y. P., 1990, Astron. Astrophys, 237, 54

This page intentionally left blank

64

Comparison of Object Detection Procedures for XMM-Newton Images

Ivan Valtchanov[1]

Procedures based on current methods to detect sources in X-ray images are applied to simulated XMM images. All significant instrumental effects are taken into account, and two kinds of sources are considered – unresolved sources represented by the telescope PSF and extended ones represented by a β-profile model. Different sets of test cases with controlled and realistic input configurations are constructed in order to analyze the influence of confusion on the source analysis and also to choose the best methods and strategies to resolve the difficulties.

In the general case of point-like and extended objects the mixed approach of multiresolution (wavelet) filtering and subsequent detection by SExtractor gives the best results. In ideal cases of isolated sources, flux errors are within 15-20%. The maximum likelihood technique outperforms the others for point-like sources when the PSF model used in the fit is the same as in the images.

The classification using the half-light radius and SExtractor stellarity index is successful in more than 98% of the cases. This suggests that average luminosity cluster of galaxies ($L_{[2-10]keV} \sim 3{\times}10^{44}$ erg s^{-1}) can be detected at redshifts greater than 1.5 for moderate exposure times in the energy band below 5 keV, provided that there is no confusion or blending by nearby sources.

We find also that with the best current available packages, confusion and completeness problems start to appear at fluxes around 6×10^{-16} erg s^{-1} cm^{-2} in [0.5-2] keV band for XMM deep surveys.

Comprehensive analysis of the detection procedures for simulated XMM images can be found in Valtchanov, Pierre & Gastaud (2001, VPG).

Here we briefly display the application of the procedures over real XMM data – the XMM deep survey in the Lockman Hole (Hasinger et al. 2001). Only the best performing procedures were used: cell detection + maximum likelihood (EMLDETECT in XMM-SAS), wavelet detection (WAVDETECT in CIAO, see also P. Freeman in these proceedings) and mixed approach

[1]Service d'Astrophysique, Centre d'Études de Saclay

consisting of wavelet filtering using Poisson noise model and subsequent SExtractor detection (MRSE).

The corresponding number of detections and cross-identification with MRSE catalogue inside $12'$ of the FOV are presented in the following table:

Method	N_{det}	N_{cross}
EMLDETECT	189	141
WAVDETECT	161	146
MRSE	175	175

There is a systematic difference for the bright sources counts with respect to the MRSE inferred counts: $\sim 15 - 20\%$ more flux with EMLDETECT, while the difference is less than 10% with WAVDETECT. There is no such difference when using simulated images which can be explained as arising from the different real PSF shape from the one used in the maximum likelihood fits – the image is a composite from three XMM-EPIC instruments and five different orbits.

Both MRSE and WAVDETECT classification for the two well known clusters is successful with some more extended sources candidates using the MRSE criteria.

As a conclusion, for unresolved sources EMLDETECT gives the best results when the model PSF used in the maximum likelihood fit is the same as the true one. In realistic situations, if one combine different instruments and observations from different orbits, the performance of the maximum likelihood technique could be worse. Wavelet based techniques are more robust and do not rely on detailed PSF shape information and give similar results in terms of positional and photometric accuracy. The mixed approach of wavelet filtering with Poisson noise model with SExtractor detection is the best procedure when detection, classification and characterization of extended sources are concerned.

Acknowledgments: We thank G. Hasinger and T. Miyaji for providing us with calibrated event lists for the XMM observations in the Lockman hole.

64.1 References

[1] Hasinger, G., Altieri, B., Arnaud, M., 2001, A&A, 365, L45

[2] Valtchanov, I., Pierre, M., Gastaud, R., 2001, A&A, 370, 689

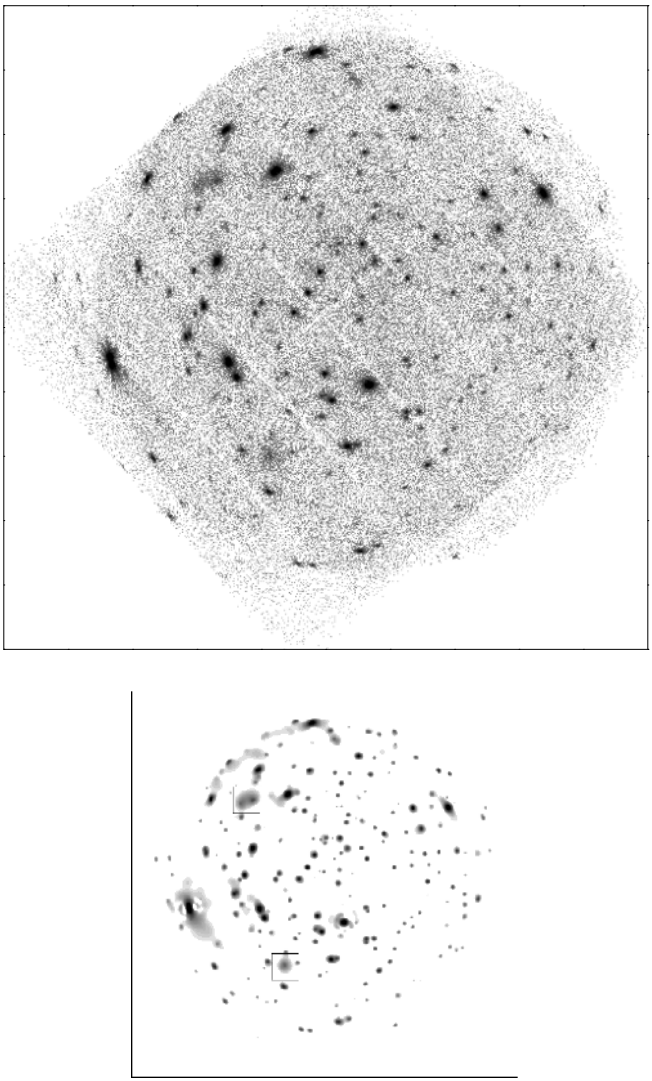


FIGURE 64.1. XMM observations in the Lockman Hole. Exposure is about 100 ks in [0.5-2] keV energy band, the three EPIC instruments are taken together. Raw image (left) and wavelet filtered (right) with Poisson noise model with significance $\sim 10^{-4}$ ($\sim 4\sigma$). The two well-known clusters of galaxies are indicated by squares.

This page intentionally left blank

65

Astronomical Aspects of Multifractal Point-Pattern Analysis: Application to the DENIS/2MASS Near-Infrared and BATSE Gamma-Ray Data

Roland Vavrek[1], Lajos G. Balázs, Attila Mészáros, István Horváth and Zsolt Bagoly

ABSTRACT Two applications of the multifractal (MFR) point pattern analysis are presented. First, we study the angular distribution of subclasses of gamma-ray bursts (GRBs), then we analyse the structure of extinction maps of dark molecular clouds obtained by near-infrared (NIR) star counts.

Multifractality is a generalization of the fractal description of self-similar objects or point fields. A monofractal set is characterized by a measure, which is globally self-similar, however, it is possible that self-similarity is only local and different scaling behaviors are observed at different scales and locations. Therefore, the MFR description of point sets provides a powerful tool to characterize them on wide range of scales. A MFR on a point process can be defined as unification of subsets of different (fractal) dimensions [2, 6]. The contribution of this subsets to the whole pattern is not necessarily equally weighted, practically it depends on the relative abundances of subsets. The functional relationship between the subsets and the corresponding fractal dimension is called the *MFR or Hausdorff spectrum*, $f(\alpha)$. In the vicinity of point i one can measure from the neighborhood structure a local dimension α_i or *pointwise dimension* giving a possibility to construct the MFR spectrum which characterizes the whole (finite) pattern. If the pattern *is not a fractal*, the MFR spectrum remains sensitive to the inhomogeneities and anisotropies of the point set.

Besides MFR analysis, we carried out several statistical tests [5] to verify the null-hypothesis of the intrinsic isotropy of the angular distribution of subclasses of the gamma-ray bursts (GRBs) at BATSE Catalog [3]. In order to determine the confidence levels pseudo-random samples were also generated by Monte Carlo simulations taking into account BATSE's non-

[1]Konkoly Observatory, Hungarian Academy of Sciences

uniform exposure function (see Table 65.1) .

The long GRBs are distributed isotropically - the positive result from two-point angular correlation function is probably an unknown instrumental effect. There are indications for the anisotropy of short GRBs both excluding two tests. Note that the shortest "tail" $T_{90} < 0.1$ s is doubtlessly anisotropic (T_{90} is the time interval, during which the 90% of gamma rays from a burst are detected). The intermediate subclass *is anisotropic*; the isotropy is rejected on a satisfactorily high confidence level which "fluctuates" between $92.0 - 99.9$ %. The character of anisotropy of intermediate subclass is peculiar, because the "dimmer" half of this subclass is more anisotropic [4]. In addition, there is no concentration toward the Galactic or Supergalactic planes.

TABLE 65.1. Survey of results of the isotropy tests. High confidence level indicates that the subsample significantly differs from the MC simulated patterns, the null hypothesis of intrinsic randomness in the angular distribution of the subsample was rejected.

	Long $T_{90} > 10s$	Short $T_{90} \leq 2s$	Inter. $2s < T_{90} \leq 10s$
Multifractal analysis	No	$> 99.9\%$	$> 99.9\%$
Minimal spanning tree	No	$> 96.0\%$	$> 92.0\%$
Voronoi tesselation	No	$> 99.9\%$	$> 99.2\%$
Spherical harmonics	No	No	$> 97.0\%$
Counts in cells	No	No	$> 96.4\%$
Two-point correlation	$> 98.8\%$	$> 99.2\%$	$> 99.8\%$

The most dense regions of the Chamaeleon I and ρ Ophiuchi molecular clouds were analysed in order to quantify the scaling properties of dust extinction using multicolour star counts on data provided by DENIS/2MASS in the I, J, H, K_S NIR bands. We draw out the maps of local monofractal dimensions which refer to lower (~ 1) projected dimensions in the cloud cores. This result may related to the assumed initial mass segregation of young stellar objects [1, 7].

This research was supported by OTKA grants T024027 (L.G.B), F29461 (I.H.) and by Research Grant J13/98: 113200004 (A.M.).

[1] D. Chappell, J. Scalo:ApJ **551** 712 (2001)
[2] V.J.Martínez, S. Paredes, E. Saar: MNRAS **260** 365 (1993)
[3] C.A. Meegan, et al.: Current BATSE Gamma-Ray Bursts Catalog, http://gammaray.msfc.nasa.gov/batse/grb/catalog/current/ (2000)
[4] A. Mészáros, Z. Bagoly, I. Horváth, L.G. Balázs, R. Vavrek: ApJ **539** 98 (2000)
[5] A. Mészáros, Z. Bagoly, L.G. Balázs, I. Horváth, R. Vavrek: In Proc. of*"Gamma-Ray Bursts in the Afterglow Era - Second Workshop", Rome, Italy, ed. N. Masetti, in press* (2001)
[6] G. Paladin, A. Vulpiani: Physics Reports, **156** N4 (1987)
[7] R.Vavrek: PhD Thesis (2001)

66

Higher-order Correlations of Cosmological Fluctuation Fields

Licia Verde[1]

ABSTRACT Traditionally, the standard way to describe the statistics of cosmological fluctuation fields has been the power spectrum. However, higher-order correlations (HOC) contain a wealth of information e.g., on the initial conditions, evolution and clustering properties of cosmic structures. With the recent observational progress and the advent of new large galaxy surveys it will be possible to perform high-precision study of HOC.
We developed a generating functional approach that allow one to compute the expected HOC of cosmological fields, in real and Fourier space. In this approach it is straightforward to include the effects of discreteness, selection function and redshift space distortions. This has many applications, in particular I illustrate possible applications of the bispectrum and the trispectrum.

In the current cosmological model of the Universe, structures we see today (e.g., the galaxy distribution) grew by gravitational instability from initial fluctuations. Models of these fluctuations can be divided in two classes: Gaussian and non-Gaussian. Even if initial conditions were Gaussian, gravitational instability introduces deviations from Gaussianity and leaves a characteristic signature on higher-order correlations (HOC) and in particular on the three-point function (or its Fourier space counterpart: the bispectrum). Moreover, theory predicts the clustering properties of the mass, while what we can primarily observe is the distribution of luminous material (e.g. galaxies). The process of galaxy formation is highly complicated and thus galaxies might be biased tracers of the mass. Biasing has its own effect on HOC, specifically on the bispectrum. Furthermore, convincing evidence against or for Gaussian initial conditions would point us towards a physical theory for the origin of structures. The 4 point function (or its Fourier counterpart, the trispectrum) is a particularly suitable tool.

The advent of large three-dimensional galaxy surveys (e.g. SDSS) is now for the first time making possible to accurately measure HOC. The observed

[1]Department of Astrophysical Sciences, Princeton University and Department of Physics & Astronomy, Rutgers University

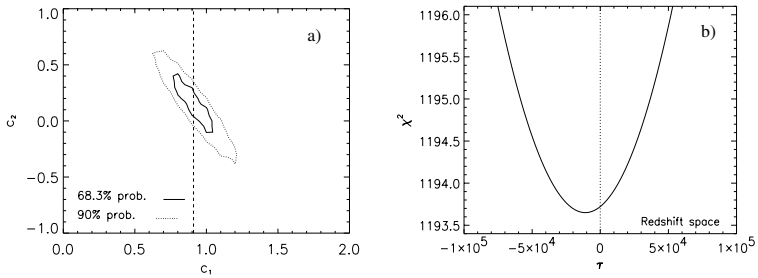


FIGURE 66.1. a): Forecast of the likelihood analysis of the bispectrum for the bias parameters from a survey of about 200^3 Mpc^3 volume. b): χ^2 analysis of the trispectrum from an N-body simulation with Gaussian initial conditions ($\tau_{true} = 0$).

galaxy distribution, however, presents additional complications due to discreteness (shot noise), selection function, window function, and redshift space distortions. We have developed a method that allows us to evaluate the HOC in real and Fourier space for continuous and discrete fields, and models the presence of these complicating effects [1,2].

The Bispectrum: Gravitational instability, even from Gaussian initial conditions, and biasing, both generate a non-zero bispectrum B. This effect is described by: $B(\vec{k}_1, \vec{k}_2, \vec{k}_3) \propto [c_1 J(\vec{k}_1, \vec{k}_2) + c_2) P(k_1) P(k_2) + cyc.$; where P denotes the galaxy Power spectrum, c_1 and c_2 are bias parameters, J is a function which expression can be found in [1] and is modified for redshift space as in [2], and cyc. stands for two cyclic terms $\{k_2, k_3\}$ and $\{k_3, k_1\}$. Since the galaxy bispectrum and power spectrum are observable quantities it is possible to measure the bias parameters via a likelihood analysis of the bispectrum [1]. Fig. (66.1a) shows the forecast of the likelihood contours for these parameters for a survey of about 200^3 Mpc^3 volume [2].

The Trispectrum: The trispectrum, being zero for a Gaussian field, is a particularly useful discriminant between Gaussian and non-Gaussian initial conditions because is only weakly modified by gravitational instability [3]. In Fig. (66.1b) the deviation from Gaussianity is parameterized by τ. The figure shows the result of a χ^2 analysis on the trispectrum of a N-body simulation with Gaussian initial conditions ($\tau_{true} = 0$).

I would like to thank my collaborators in this work A. F. Heavens and S. Matarrese

66.1 References

[1] Matarrese S., Verde L., Heavens A. F., 1997, MNRAS, 290, 651 *Large-scale bias in the Universe: bispectrum method*

[2] Verde L., Heavens A. F., Matarrese S., Moscardini L., 1998, MNRAS, 300, 747 *Large-scale bias in the Universe II Redshift space distortions*

[3] Verde L., Heavens A. F., 2001, ApJ, 533, 14 *On the trispectrum as a gaussian test for cosmology*

67

Bayesian Multiscale Deconvolution Applied to Gamma-ray Spectroscopy

C. A. Young[1], A. Connors, E. Kolaczyk, M. McConnell, G. Rank, J. M. Ryan, and V. Schoenfelder

ABSTRACT A common task in gamma-ray astronomy is to extract spectral information, such as model constraints and incident photon spectrum estimates, given the measured energy deposited in a detector and the detector response. This is the classic problem of spectral "deconvolution" or spectral inversion [2]. The methods of forward folding (i.e. parameter fitting) and maximum entropy "deconvolution" (i.e. estimating independent input photon rates for each individual energy bin) have been used successfully for gamma-ray solar flares (e.g. [5]). Nowak and Kolaczyk [4] have developed a fast, robust, technique using a Bayesian multiscale framework that addresses many problems with added algorithmic advantages. We briefly mention this new approach and demonstrate its use with time resolved solar flare gamma-ray spectroscopy.

Recent treatments of Poisson inverse problems have augmented the likelihood equations with a regularization or penalization term [4]. This regularization term stabilizes the otherwise ill posed ML problem. The regularization term can take the form of a Bayesian prior so that the MLE is replaced with the Maximum a posteriori (MAP) estimator. Nowak and Kolaczyk [4] developed a deconvolution technique that uses a Bayesian multiscale framework. The technique uses an Estimator Maximization (EM) algorithm that has a closed-form step. Under reasonable choice of the multiscale priors, the EM algorithm converges to a unique, global MAP estimate [4] and is computationally simple. Unfortunately, errors or confidence intervals in the traditional sense do not follow [3]. The most straightforward method for this is to use a parametric bootstrap [1].

Figure 67.1 shows the light curve for a gamma-ray solar flare divided into 3 time intervals and the deconvolved spectra for each of the time intervals. We have shown this technique to be very useful for the analysis of solar flare gamma-ray spectra.

[1]NASA Goddard Space Flight Center

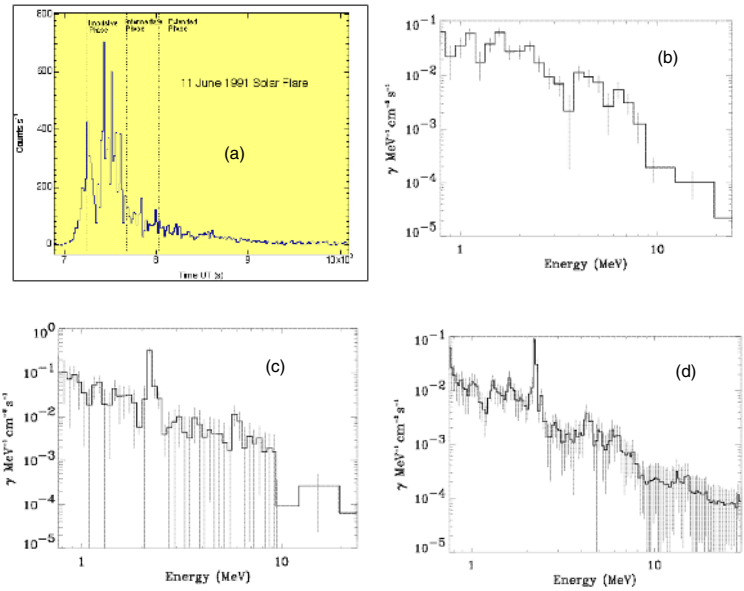


FIGURE 67.1. The light curve for a gamma-ray solar flare (a). Also shown is the deconvolved spectra for the time intervals defined in the light curve (b-d). Error bars extending to the bottom of the plots are 1 σ upper limits.

Acknowledgments: This work was supported through NASA contract NAS5-26645 and NASA's Supporting Research and Technology program.

[1] Connors, A., Private Comm. (2000).
[2] Craig, I. J. D., and Brown, J. C., *Inverse Problems in Astronomy*, Boston: Adam Hilger, Ltd., (1986).
[3] Kolaczyk, E. D., in Bayesian Inference in Wavelet-based Models, edited by M. A. Vidkovic, New York City: Springer-Verlag, (1999).
[4] Nowak, R. D., and Kolaczyk, E. D. in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (1999).
[5] G. Rank *Ph.D. Thesis* University of Munich (1997).

Index