Computational Finance

and its Applications

II

EDITORS

M. Costantino & C.A. Brebbia

WITPRESS

Computational Finance
and its Applications II

WIT*PRESS*

WIT Press publishes leading books in Science and Technology.
Visit our website for the current list of titles.
www.witpress.com

WIT*eLibrary*

SECOND INTERNATIONAL CONFERENCE ON
COMPUTATIONAL FINANCE

COMPUTATIONAL FINANCE II

CONFERENCE CHAIRMEN

M. Costantino
Royal Bank of Scotland Financial Markets, UK

C. A. Brebbia
Wessex Institute of Technology, UK

INTERNATIONAL SCIENTIFIC ADVISORY COMMITTEE

D. Anderson
D. Bloch
H. Chi
O. Criner
J. P. Lawler
M. Mascagni
D. Tavella
H. Tutek
M. Wahde

Organised by
Wessex Institute of Technology, UK

Sponsored by
WIT Transactions on Modelling and Simulation

D F Cutler
Royal Botanic Gardens
UK

G De Mey
Ghent State University
Belgium

Q H Du
Tsinghua University
China

A El-Zafrany
Cranfield University
UK

S Finger
Carnegie Mellon University
USA

M J Fritzler
University of Calgary
Canada

G S Gipson
Oklahoma State University
USA

K Hayami
National Institute of Informatics
Japan

D B Ingham
The University of Leeds
UK

D L Karabalis
University of Patras
Greece

E R de Arantes e Oliveira
Insituto Superior Tecnico
Portugal

J Dominguez
University of Seville
Spain

S Elghobashi
University of California
Irvine
USA

P Fedelinski
Silesian Technical
University
Poland

J I Frankel
University of Tennessee
USA

L Gaul
Universitat Stuttgart
Germany

S Grilli
University of Rhode Island
USA

J A C Humphrey
Bucknell University
USA

N Kamiya
Nagoya University
Japan

J T Katsikadelis
National Technical
University of Athens
Greece

H Lui
State Seismological Bureau Harbin
China

R A Meric
Research Institute for Basic Sciences
Turkey

K Onishi
Ibaraki University
Japan

M Predeleanu
University Paris VI
France

S Rinaldi
Politecnico di Milano
Italy

G Schmid
Ruhr-Universitat Bochum
Germany

X Shixiong
Fudan University
China

V Sladek
Slovak Academy of Sciences
Slovakia

J Stasiek
Technical University of Gdansk
Poland

M Tanaka
Shinshu University
Japan

T Tran-Cong
University of Southern Queensland
Australia

W J Mansur
COPPE/UFRJ
Brazil

J Mikielewicz
Polish Academy of Sciences
Poland

E L Ortiz
Imperial College London
UK

D Qinghua
Tsinghua University
China

T J Rudolphi
Iowa State University
USA

A P S Selvadurai
McGill University
Canada

P Skerget
University of Maribor
Slovenia

T Speck
Albert-Ludwigs-Universitaet
Freiburg
Germany

S Syngellakis
University of Southampton
UK

N Tosaka
Nihon University
Japan

W S Venturini
University of Sao Paulo
Brazil

J F V Vincent
The University of Bath
UK

J R Whiteman
Brunel University
UK

Z-Y Yan
Peking University
China

K Yoshizato
Hiroshima University
Japan

G Zharkova
Institute of Theoretical and Applied
Mechanics
Russia

Computational Finance
and its Applications II

Editors

M. Costantino
Royal Bank of Scotland Financial Markets, UK

C. A. Brebbia
Wessex Institute of Technology, UK

WIT*PRESS* **Southampton, Boston**

M. Costantino
Royal Bank of Scotland Financial Markets, UK

C. A. Brebbia
Wessex Institute of Technology, UK

*The texts of the papers in this volume were set
individually by the authors or under their supervision.
Only minor corrections to the text may have been carried
out by the publisher.*

Preface

This book contains the edited version of the papers presented at the conference Computational Finance 2006, held in London in June 2006. This conference follows the success of the First International Conference of Computational Finance and its Applications which was held in Bologna, Italy, in 2004.

In the last two years, several major events have characterised the international financial markets. The most significant one was certainly the explosion of the price of commodities, in particular of oil, which has recently reached a level of 74 dollars.

This prompted several investment banks and traditional commodity players to strongly increase their presence in the commodities trading area. Several new trading groups have been established, which marks a strong expansion in this area after the collapse of the previous generation of players, such as Enron.

Surprisingly, the impact on the economy of such high oil prices has been so far rather limited.

On the contrary, share prices have shown a prolonged growth, fuelled by a relative strong economic growth, low unemployment, strong profits from oil-related companies and a new wave of mergers and acquisitions. Technology shares, such as Google, have also grown strongly, with some analytics comparing this performance with the years of the .COM boom.

Analysts are now divided on where the markets will go next. Some argue that growth will continue, while others are warning that we could be already in the middle of a new stock market bubble.

Because of the interests at stake in the financial markets and in particular because of the uncertainty of the direction of the economy and financial markets, investment in research in the field of finance has remained extremely strong.

Finance has continued to be one of the main fields of research where the collaboration between the industry, such as investment banks, and the wider research community is strongest.

Within this context, the purpose of this conference has been to bring together leading experts from both the industry and academia to present and share the results of their research, with the aim to become one of the main forums where such collaboration takes place.

This book contains many high quality contributions reporting advances in the field and focussed on the following areas: Financial service technologies in the 21st century; Advanced computing and simulation; Derivatives pricing; Forecasting, advanced computing and simulation; Market analysis, dynamics and simulation; Portfolio management and asset allocation; Risk management; Time series analysis and forecasting.

This volume would not have been possible without the help of the members of the International Scientific Advisory Committee, whose help is gratefully acknowledged. Their help in reviewing the papers has been essential in ensuring the high quality of this volume.

The Editors
London, 2006

Contents

Section 3: Derivatives pricing

Section 4: Forecasting, advanced computing and simulation

Section 5: Market analysis, dynamics and simulation

Section 6: Portfolio management and asset allocation

Section 7: Risk management

Section 8: Time series analysis and forecasting

Section 1
Financial service technologies in the 21st century
(Special section edited by J. Lawler and D. Anderson)

This page intentionally left blank

Community e-kiosk portal technology
on Wall Street

J. Lawler & D. Anderson
Pace University, USA

Abstract

The community of downtown Wall Street in New York City continues to cope
with economic disruption, due to the World Trade Center disaster of September
11. This case study explores design factors of engagement in the implementation
of a Web-based e-kiosk portal, which is furnishing residents of the community
with critical cultural, financial and social information on the re-building of the
downtown economy. The e-kiosk portal was an emergency project implemented
by computer science and information systems students at a major metropolitan
university. The preliminary findings and implications of the study indicate the
importance of social and technical cachet in the design of a Web portal
community. The study introduces a framework for research into civic Web
communities that empower its member residents.
*Keywords: community, e-government, government-to-citizen (G2C), Internet,
kiosk, portal, touch-screen technology, World Wide Web.*

1   Background

Community is considered to be a critical characteristic of the Internet Armstrong
and Hagel III [1]. Community is concretized as a "feeling of membership in a
group along with a strong sense of involvement and shared common interests …
[that] creates strong, lasting relationships." Rayport and Jaworski [2].
Definitions of community consist of "a social grouping which exhibits … shared
spatial relations, social conventions … and an on-going rhythm of social
interaction Mynatt et al. [3]. Features of community are empowered by
connection and communication functionality of the World Wide Web. This
functionality helps consumers and citizens in continuing to engage in dialogue

on the Web with business and governmental entities. The culture of community is enabled not only in an off-line but an additional on-line context.

Communities in an on-line context are characterized as those of fantasy, interest, relationship and transaction [1]. Communities of fantasy are illustrated in chat and discussion games on ESPNet. Communities of interest are indicated in financial Motley Fool Web-blogs and forums, and communities of relationship are indicated in interpersonal Cancer discussion Forums of help. Communities of transaction are indicated in Land's End forums of product inquiry friends and shoppers.   Communities can have elements of each of the forums on the Web [1].

The design of an on-line community on the Web is considered a constant challenge Ginsburg and Weisband [4] for technologists. The first intent is to enable social capital, defined as a "network" Cohill and Kavanaugh [5], Nahapie and Sumantra [6] and Schuler [7] or "web of social relationships that influences individual behavior and … [impacts] economic growth" Lesser [8] and Pennar [9]. These networks of relationships furnish empowering information to citizen and consumer members of a "trusted" Putnam [10] community. Interaction in customized forums of citizens and governmental agencies is further indicated in an "empowered deliberative democracy" Fung and Wright [11], which may help disadvantaged members. Empowerment is enabled in the implementation of a community design that is considerate of diverse concerns of community members and residents. Community design is facilitated in the introduction of a government-to-citizen (G2C) portal that is currently transforming the home page of a traditional Web site.

2   Introduction

An on-line portal is defined as a dynamic or static site on the Web that collects content for a group of members that have common interests Heflin [12]. A portal can be considered horizontal and public, as in a G2C or business-to-consumer (B2C) portal, or vertical and private, as in a business-to-business (B2B) extranet, personalized business-to-customer (B2C), or business-to-employee (B2E) intranet portal Donegan [13]. Portal in this study is defined as horizontal and public to members and residents of a distinct community. Members can contribute and get information on the horizontal portal from other members and from other sources of interest for the members. The immediate benefit of the Web portal is the integration and interoperability of diverse information sources.

Designers of a community G2C portal are challenged by the heterogeneous nature of information sources, in order to have a common standard for information display and exchange and a highly functioning and intelligent site Gant and Gant [14]. Though a portal is the framework for the federal government to develop its electronic (e-Government) strategies through the Web Fletcher [15], internal issues in the agencies of the government are frequent in the development of e-Government portals Liu and Hwang [16]. State governments are not even distinguishable in the efficiency, functionality and innovation of

their information portals Watkins [17]. Figure 1 below indicates the slowness in
the implementation of e-government portals in the United States, in phases of
transformation: information publishing "portal", interactive and transactional
"portal", multi-functional portal, personalization of portal, clustering of common
services on portal, and full collaboration and transformation of portal Wong [18].
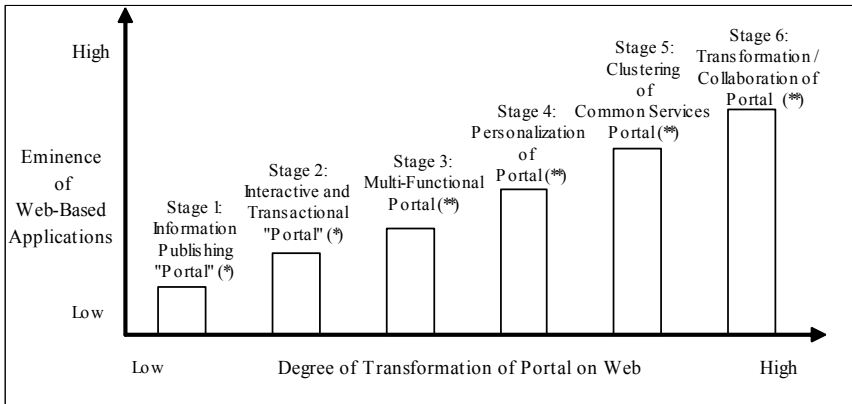
(*) Individual departments of government; (**) Multiple departments of government.
Source: Wong [18] (Adapted).

Figure 1:   E-government portal transformation in United States.

Design of a community portal is concurrently impacted by the perception of
the portal by members and residents in the community. Studies in the literature
frequently indicate the importance of trust, usefulness and ease of use in e-
Government services on the Web Warkentin [19]. Openness of services is often
indicated to be important on the portal site Demchak et al. [20]. Perception of
ease of use may be facilitated by increased innovation in electronic (e-kiosk)
information and self-service touch-screen Web-based systems Boudioni [21].
Such systems may be failures though Dragoon [22], if friendly and simple
graphical user interfaces and screen layouts and intuitive navigational tools
Cranston et al. [23] are not evident for distinct Mendelsohn [24], limited literate
Ekberg [25], and health impaired members. Residents may be disadvantaged in
the community due to unanticipated catastrophe. Few studies in the literature
have analyzed further factors specific in the design of an on-line community
portal that may be helpful to potentially disadvantaged or challenged members
and residents in solving immediate issues arising from a catastrophe.

3   Case study

This study analyzes a design of an emergency Web-based e-kiosk portal, for a
community of citizens in the Wall Street district of New York City. The citizens
consist largely of local disadvantaged residents and small businesspersons that

continue to cope with the dislocation of apartments and offices and the disruption of the downtown economy and life, due to the World Trade Center disaster of September 11 Rosenberg [26]. The function of the e-kiosk portal is to be a catalyst for economic development, in an initial facility for furnishing employment information and financial and governmental information on loan procedures, local rebuilding programs and social and cultural projects that are enabling the recovery of the economy. Its function further includes instillation of confidence in the recovery of the city and the World Financial District on Wall Street. Funded by grants from the Center for Downtown New York of Pace University, a member of the community, the e-kiosk portal is an extracurricular outreach implementation by graduate and undergraduate students of the Ivan G. Seidenberg School of Computer Science and Information of the university. These students responded enthusiastically to the post September 11 impact.

The e-kiosk consists of the following features: *Who Are We*, *What's New Downtown*, *What's New with the Rebuilding*; *Want to Learn More about Downtown*, *Want to Have Lunch and Shop*, *Want to Volunteer*, and *Want to Talk to Us*. These features are enabled in a pleasant and simple graphical Windows interface and intuitive and navigational touch-screen system, illustrated in Figure 2.

To enable community, the e-kiosk is not only an off-line physical facility of information, in installable downtown locations, but also an on-line virtual Web portal of interactivity that links small businesspersons and residents, and also tourists, to cultural, economic, employment, financial and governmental agencies. This portal is beginning to enable a bona fide citizen community that includes institutions and members beyond downtown and in New York State and in the Northeast Corridor of the United States. Students of the university, along with the citizens, are already members of the community.

4   Focus of analysis

The focus of the analysis is centered on factors contributing to citizen engagement in the e-kiosk community. Rayport and Jaworski define factors in a design method that introduces *cohesion*, *effectiveness*, *help*, *language*, *relationship* and *self-regulation* [2] in the functionality of a Web community. The factors are defined below:

- *cohesion*, element of design from which members have a feeling of belonging in the community;
- *effectiveness*, element from which members have a feeling of personal impact from the community;
- *help*, element from which members have personal help from the community;
- *language*, element from which members have a forum for specialized languages in the community;
- *relationship*, element from which members have interaction and friendship in the community; and
- *self-regulation*, element from which members regulate their interactions in the community [2].

These factors are imputed to facilitate fulfillment, inclusion, influence and emotional experience sharing [2] in a Web-based community.   Though the students applied the factors in their implementation of the e-kiosk portal, in iterative prototyping and usability review, its extension as a model to other civic Web communities is not substantiated empirically by theorists.   This study analyzes these design factors of engagement in the e-kiosk Web portal community, and its preliminary findings are demonstrating the importance of the factors in a functioning economic and social Web community in the Wall Street neighborhood.

Figure 2:   E-kiosk portal on Wall Street (sample screen).

5   Methodology

The methodology of the case study is analyzing the e-kiosk community portal, in the downtown New York Wall Street neighborhood, in three stages.

In stage 1 a controlled off-line sample of students, of the School of Computer Science and Information Systems at Pace University, not members of the e-kiosk implementation team was surveyed by questionnaire by the authors.   The questionnaire surveyed the students on perceptions of the importance of the *cohesion*, *effectiveness*, *help*, *language*, *relationship* and *self-regulation* factors in the e-kiosk Web community, on a simple *high*, *intermediate*, or *low* scale.

These students were mature subjects and were surveyed as though they were downtown residents and small businesspersons.

In stage 2 the survey is being currently expanded to include an on-line sample of non-student downtown residents, small businesspersons, and tourists. In stage 3 the findings of stages 2 and 1 will be analyzed through descriptive and statistical interpretation, with the final study to be finished in early 2007.

6 Preliminary analysis

From stage 1, and a limited stage 2, of the preliminary study, a summary of the analysis disclosed that most of the sampled subjects indicated *help*, *effectiveness* and *cohesion* factors as *high*, in importance ranking in e-kiosk engagement functionality. The subjects indicated *relationship* as *intermediate* in importance. They indicated *self-regulation* and *language* as *low*, in importance ranking in the functionality. They indicated What's New with the Rebuilding and What's New Downtown as *high* in feature importance on the portal. Want to Talk to Us was indicated as *intermediate* in importance, while Want to Volunteer, Who Are We and Want to Have Lunch and Shop were indicated as *low* in importance on the portal site.

The e-kiosk on Wall Street was indicated in the analysis to be at lower stages of e-Government information publishing and multi-functional "portals" in 2004–2005. It will be at higher stages of interactive and transactional, personalized, serviced and transformational portals in 2006–2008, if fully integrated with New York City and New York State portal systems. The stages of transformation are indicated in Figure 3.
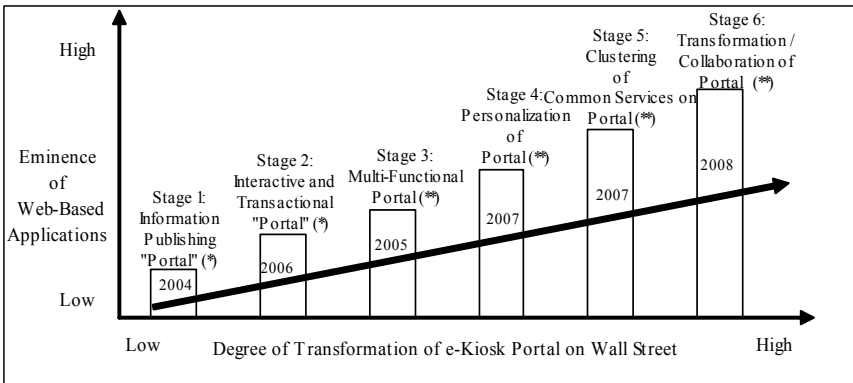
(*) Individual departments of government; (**) Multiple departments of government.
Source: Anderson and Lawler, 2005 and Wong [18] (Adapted)

Figure 3:   E-kiosk portal system on Wall Street (2004–2008).

The study needs further analysis and interpretation in stages 2 and 3 of the methodology, in order to evaluate the creditability of the initial methodology. Stage 2 will be finished in fall 2006, and stage 3 will be finished in winter 2007.

Though the findings of the study will not be final until 2007, the preliminary findings are helpful in analyzing a civic portal Web community.

(Further information on statistical findings in stage 1 will be furnished upon request of the authors.)

7   Implications

The preliminary findings from stage 1 of this study imply the design importance of the *cohesion*, *effectiveness* and *help* factors in the downtown New York community. The factors of *help*, *effectiveness* and *cohesion* are indicated to be high in importance in the e-kiosk portal, in expediting financial aid and employment for disadvantaged residents and small businesspersons in downtown New York and Wall Street. The e-kiosk is important in helping the Small Business Development Center of the university, in informing the small businesspersons and residents of over $10 million in governmental and economic injury loans and job services. This e-kiosk is further instrumental in informing residents, businesspersons and tourists of neighborhood recovery and social programs. Factors of *help* and *effectiveness*, furnished in the e-kiosk portal system, give the disadvantaged residents and the small businesspersons, if not the tourists, the feelings of increased confidence and pride in the recovery of downtown New York.

Factors of *relationship* and also *self-regulation* and *language* are indicated to be respectively *intermediate* and *low* in importance in the functionality of the e-kiosk Web portal community. Friendships of the residents and the small businesspersons, as members of the community in interaction on the network, are not currently forming social capital, as the e-kiosk is not community-driven Zhdanova and Fensel [27] and functioning as an information portal. However, the World Wide Web is helpful inherently in integrating members in a community Preece [28], fostering social capital. Further capital may be formed in integration of the downtown community with other constituencies in New York and on the Northeast Corridor of the United States. Though the benefit of an on-line virtual community *to the community* is its social capital, the residents and small businesspersons have a good foundation and process Fernback [29] in the existing e-kiosk portal system to enable a later social structure.

Findings indicated the design importance of interface on an e-kiosk community portal system. On-line kiosks are indicated in the literature to enable inclusion of senior citizens that might otherwise be excluded from an information society Ashford et al. [30]. Students, in a limited stage 2 of the study, learned that senior residents in the downtown Wall Street community were not excluded socially or technologically as members of the system. Touch-screens on off-line physical portals in the neighborhood facilitated interface to What's New with the Rebuilding and What's New Downtown, for senior residents frequently hesitant in keyboard and Web technology Coleman et al. [31] and Cranston et al. [32]. Usability of the touch-screens facilitated social inclusion [21].

Further findings indicated the importance of external e-Government projects in initiating kiosk community Web portals. Students in the School of Computer Science and Information Systems at Pace initiated the e-kiosk information publishing "portal" on Wall Street in less than three months in 2004, and the multi-functional "portal" in less than one month in 2005, as indicated in Figure 3. Internal state and city governments may often be slow in initiating service solutions through Web portal sites Douglas [33], as indicated in Figure 1. Governments may be limited by internal legacy systems. Full integration of the e-kiosk portal system on Wall Street with New York City and New York State systems is however a next step in the university.

Other findings of the preliminary study confirm the benefits of including self-motivated and mature students in a Web community portal project Alavi et al. [34]. The students that implemented the portal system indicated increased learning in the technological context of community Web design. They also learned design in the social context of the implemented e-kiosk portal Web community for downtown members and residents. The students were sensitive to socio-technical systems design Eason [35]. Residents and small businesspersons are as a result inquiring of further empowerment in a functionally enhanced informational e-kiosk portal system, to be implemented with requested student volunteers of the university. In short, the community of downtown New York on Wall Street and Pace University continue to benefit from a fruitful partnership.

8 Limitations and opportunities for research

The study needs empirical evaluation of the exploratory findings from the survey of students and of the forthcoming results from the survey of non-student residents in the Wall Street neighborhood, in order to extend generalizability. Further research will be initiated in future integration of audio podcasting, digital interactive television, and hand-held mobile tools with the e-kiosk portal system. Integration of the system with the New York City and New York State portal systems, and possibly with the portal system and its technologies in Washington, D.C., is intended in the near future and will be a new opportunity for research.

9 Conclusion

The study identified design factors of importance in engagement in an e-kiosk portal Web community. Further empirical research is needed in an expanded study, in order to analyze the factors of importance in the implementation of civic Web communities. This study of the downtown New York City Wall Street community is facilitating an evolving and new framework.

Acknowledgement

References

[1]   Armstrong A. & Hagel III, J., The real value of on-line communities. *Harvard Business Review*, May –June, p. 135-138, 1996.

[2]   Rayport, J.F. & Jaworski, B.J., *Introduction to e-Commerce*, McGraw-Hill: New York, pp. 204-206, 2002.

[3]   Mynatt, E.D., Adler, A., Ito, M. & Oday, V.L., Design for network communities. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computer Systems*, 1997.

[4]   Ginsburg, M. and Weisband, S., Social capital and volunteerism in virtual communities. *Proceedings of the IEEE Conference–Virtual Communities*, January, p. 1, 2002.

[5]   Cohill, A.M. & Kavanaugh, A.L., *Community Networks: Lessons from Blacksberg, Virginia*, Artech House: Norwood, MA, 1997.

[6]   Nahapie, J. & Sumantra, G., Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review*, 23(2), pp. 242-266, 1998.

[7]   Schuler, D., *New Community Networks: Wired for Change*, ACM Press - Addison-Wesley: Reading, MA, 1996.

[8]   Lesser, E., *Knowledge and Social Capital*, Butterworth-Heinemann: Boston, 2000.

[9]   Pennar, K., Ties that lead to prosperity. *Business Week*, 15 December, 1997.

[10]   Putnam, R., *Bowling Alone: The Collapse and Revival of American Community*, Simon and Schuster: New York, pp. 65-78, 1995.

[11]   Fung, A. & Wright, E.O., Deepening democracy: innovations in empowered participatory governance. *Politics and Society*, 29(1), p. 127, 2002.

[12]   Heflin, J., Web ontology language (owl) use case and requirements. *W3C Working Draft*, 31 March, 2003.

[13]   Donegan, M., Contemplating portal strategies. *Telecommunications* (International Edition), February, 2000.

[14]   Gant, J. P. & Gant, D. B., Web portals and their role in e-government. *Proceedings of the Seventh Americas Conference on Information Systems*, p. 1617, 2001.

[15]   Fletcher, P. D., *Digital Government: Principles and Best Practices*, Idea Group Publishing: Hershey, PA, pp. 52-62, 2004.

[16]   Liu, S. & Hwang, J.D., Challenge to transforming information technology in the United States government. *IT PRO*, May-June, 2003.

[17]   Watkins, S., Potent portals: in the 2005 best of the web contest, Delaware's web portal came out on top, *Government Technology*, 18(10), October, pp. 40-46, 2005.

[18]   Wong, W.Y, The dawn of e-government, *Deloitte & Touche Report*, 2000.

[19]   Warkentin, M., Encouraging citizen adoption of e-government by building trust. *Electronic Markets*, 12(3), 2002.

[20] Demchak, C. C., Friis, C. & LaPorte, T.M., Webbing governance: national differences in constructing the face of public organizations. *Handbook of Public Information Systems*, ed. G. David Garson, Marcel Dekker Publishers: New York, 2000.

[21] Boudioni, M., Availability and use of information touch-screen kiosks (to facilitate social inclusion). *Proceedings of the ASLIB*, 55 (5/6), pp. 320-329,331, 2003.

[22] Dragoon, A., Six simple rules for successful self-service. *CIO*, 15 October, p. 59, 2005.

[23] Cranston, M., Clayton, D.J. & Farrands, P.J., Design and implementation considerations for an interactive multimedia kiosk: where to start. *Proceedings of the Adelaide Conference*, 1996.

[24] Mendelsohn, F., KISS, *Kiosk Business*, September/October, 2001.

[25] Ekberg, J., Public terminals. *Include*, 3 January, p. 3, 2003.

[26] Rosenberg, J.M., Small businesses in varying state of recovery after September 11. *The Times*, 9 January, p. F4, 2004.

[27] Zhdanova, A. V. & Fensel, D., Limitations of community web portals: a classmate's case study. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 19-22 September, Compiegne, France, p. 1, 2005.

[28] Preece, J., *On-Line Communities: Designing Usability, Supporting Sociability*, John Wiley & Sons, Ltd.: New York, p. 12, 2000.

[29] Fernback, J., There is a there there (Notes towards a Definition of Cyber-Community). *Doing Internet Research: Critical Issues and Methods for Examining the Net*, ed. S. Jones, Sage Publications: Thousand Oaks, CA., pp. 203-220, 1999.

[30] Ashford, R., Rowley, J. & Slack, F., Electronic public service delivery through on-line kiosks: the user's perspective, *Proceedings of the EGOV Conference*, Springer-Verlag: Berlin, Germany, p. 1, 2002.

[31] Coleman, N., Jeawody, F. & Wapshot, J., Electronic government at the department for work and pensions–attitudes to electronic methods of conducting benefit business, *DWP Research Report*, 176(CDS), 2002.

[32] Cranston, M., Clayton, D.J. & Farrands, P.J., Design and implementation considerations for an interactive multimedia kiosk: where to start. *Proceedings of the Adelaide Conference*, p. 96, 99, 1996.

[33] Douglas, M., Virtual village square: what does it take to transform a lackluster municipal web site into a vibrant community meeting place, *Government Technology*, 18(10), October, pp. 66-68, 2005.

[34] Alavi, M., Wheeler, B. & Valacich, J., Using information technology to reengineer business education: an exploratory investigation of collaborative tele-learning. *MIS Quarterly*, 19(3), pp. 293-313, 1995.

[35] Eason, K.D., *Information Technology and Organizational Change*, Taylor Francis: London, UK, 1988.

Management of the productivity of information and communications technology (ICT) in the financial services industry

J. W. Gabberty

Ivan G. Seidenberg School of Computer Science and Information Systems, Pace University, USA

Abstract

Financial service firms were among the earliest users of information and communications technology (ICT). As introduced in this study, investment in this technology in the banking sector of the industry, initiated in 1970, enabled automation of numerous functions, including loan payment scheduling and automated teller systems. Besides hastening the pace at which functions are performed in the sector, these time-saving improvements reduced the cost of labor, as banking tellers by the thousands were replaced by automated systems. These investments later resulted in fee revenue from customers of the teller systems. The replacement of traditional interest calculation tables, together with spreadsheet programs, resulted in the customization of interest-paying consumer loans. Transaction processing is indicated in this study to have satisfied increasingly larger databases that facilitated the explosion of consumer credit cards and further revenue for the banking sector. The frequent perception that investments in information and communications technology would continue to lower the cost of business while concomitantly and perpetually increasing revenue was the maxim in the sector in 1970–1990. Massive investment by the banking sector in 1990–2000 failed however to support this phenomenon. The failure of the industry to match increasing labor productivity rates was manifest in the sector, as the sector immediately curtailed spending on information and communications technology in 2000–2005. This study evaluates the new relationship of labor productivity and technology, and introduces steps for firms to mitigate the risks of overdependency on the technology. This study will benefit management practitioners and users researching information and communications technology in financial service firms.
Keywords: ICT productivity, productivity paradox, United States banking productivity.

1   Background

The tangential issues accompanying information and communications technology (ICT) driven productivity have their origins in a diverse body of research that includes economics, accounting, marketing, management, finance, and information assurance and security; in totality, these diverse topics sources form the basis of current thinking on the topic.

Rarely does an economic indicator garner more attention than the term '*productivity*' - especially when used in the context of information and communications technology. Far more meaningful than the terms 'current account', 'unemployment' and 'inflation', productivity is probably the most often used and most misunderstood term used by technology pundits (and the general public) to provide some proximal measure of the impact of technology in a project, process, or enterprise. In simplest terms, a firm is either productive or not - nothing can be simpler - and that is precisely the reason why nearly everyone can understand the broader meaning of the term. But when it comes to measuring both the *tangible* and *intangible* aspects of ICT productivity, the *risk* components associated with ICT, or even the concept of *value* as applied within the universe of ICT deployment, most managers are hard pressed to fully comprehend the real impact that ICT bears on any firm. Yet implicitly, the terms 'ICT' and 'productivity' seem to go hand in hand; indeed the proliferation of the computer in all aspects of our society has virtually cemented the notion that spending on information and communications technology *always* results in heightened productivity, though nothing could be farther from reality. The generally accepted perspective upheld universally is that ICT has produced a fundamental change, in particular within the economy of the United States, and has lead to a *permanent* improvement in growth prospects, as studied by Greenspan [1] and Jorgenson [2]. The final resolution of this perspective however, i.e., the conclusory evidence linking ICT to productivity, has yet to be found.

Economists and academic scholars search in vain for the "killer application", thinking that some elusive program (or suite of programs) will form the core of a new framework for ICT productivity measures to complement those already found in Paul Schreyer's (2001) OECD Manual, *Measuring Productivity*. But while that search continues, the objective of this paper is to attempt to bring into focus the obfuscated issues surrounding ICT and productivity and their place in the banking sector of the United States.

2   Relationship of ICT to productivity

Businesses, especially in the U.S., continue to pump billions of dollars into information and communications technologies. Apparently, these firms perceive ICT as having a value in excess of the aggregate sum of the total monies spent on hardware, software, licensing, programmers, analysts, middleware, training, and all other tangential costs that go into building a firm's ICT arsenal. But how is the total return on these investments measured? Clearly, top executives at

these companies believe that their investments must be made in lieu of other important assets that the firm could acquire as alternatives to ICT investment(s). These substitute investments might include more staff personnel, additional office space or office locations, higher research and development investments, more pay incentives for key personnel, additional money spent on marketing and sales initiatives, and so on. The use of computers in business, though not entirely new, is still a field shrouded in a sea of complexity, disillusionment, and misunderstanding.

As for calculating the return on ICT investments, at least from the national level, improvement in raising the Gross Domestic Product (GDP) per capita is widely regarded as the best single measure, not only of economic well-being, but also of the aggregate impact of ICT. That measure is simply labor productivity (how many goods and services a given number of employees can produce) multiplied by the proportion of the population that works. Figure 1 illustrates the GDP of the United States on a per capita basis. Also listed is the share of the GDP figure that stems from information technology related industries. Logically, information technology leads to higher productivity levels but the improvements in output capabilities are not reflected by the statistics.
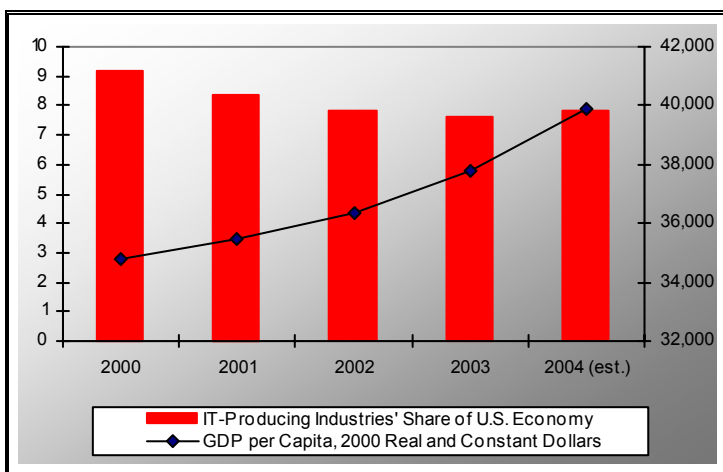
Figure 1:   Source: Statistical Abstract of the United States 2006, Table 1113: *Gross Domestic Income in Information Technologies (IT) Industries* and Table 657: *Selected Per Capita Income and Product Measures in Current and Real (2000) Dollars*.

Productivity, however, varies enormously among industries and explains many of the differences in GDP on a per capita basis. Thus, to understand what makes countries rich or poor, one must understand what causes productivity to be higher or lower. This understanding is best achieved by evaluating the performance (i.e., output) of employees in individual industries as well as the degree of automation and computerization used in production processes - both manufacturing- and service-related, since a country's productivity is the

aggregate of total factor productivity for each industry. Such a micro-level approach is extremely costly and time-consuming to perform, yet, if accomplished, would reveal an important fact about productivity: not only does it vary from firm to firm but also varies between industries and also varies widely from country to country.

2.1 Ascendancy of ICT

Just over twenty years ago, a study by Warner [3] of Future Computing Inc. posited that 1.4 million personal computers were in use by 10% to 12% of office employees in Fortune 2000 firms. The next year, Dell Inc. was founded by Michael Dell. By 2000, that firm would be selling personal computers and peripherals via the Internet in excess of $50 million per day, and twenty years after Future Computing's study, Dell's 2004 sales exceeded $49 billion with an employee base of 55,000. Clearly, ICT had taken hold both in the public and private sector; however the level of complexity associated with calculating both the impact on productivity and costs accompanying ICT also grew at a phenomenal rate.

It is important to note that the lateral costs associated with ICT are not easily calculable and, from this context, relates to costs not easily accounted for, such as administration and upkeep of technology. For example, while a $1,000 purchase for a personal computer can be accounted for in terms of costs throughout its useful lifetime (adding peripherals, memory, internet access costs, etc.), assessing the total 'true' cost to create a database on that machine is extremely difficult and varies from computer to computer and from industry to industry. Besides the costs of the database software and licenses, additional (latent) costs can be found in the costs of the administrator's time to have questions answered by existing and planned users, posing a very difficult task for the assessor. Off-line questions asked by the database programmer touch numerous employees as the database expands in complexity and completion. In calculating the costs associated with peripheral employees, whose input is frequently sought throughout construction of a simple database, only then do these true costs become identified and accounted for. Hence, a $200 database package that has been customized by a database programmer earning $50,000 salary (with an additional $15,000 in fringe costs), installed on a personal computer with an initial cost of $1,000, may bring the total cost of this single installation to a cost in excess of $250,000 during its useful lifetime, excluding fixed costs such as rent, electric, HVAC, etc. Obviously, the higher the associated costs of building and maintaining such a straightforward database system drives downward the level of productivity as inputs get consumed to create outputs.

3 ICT in the financial services industry

Corporate America and its fascination with and dependence on computers are known internationally. In 2004, for example, the World Economic Forum ranked

the United States #1 in its annual *Overall Networked Readiness Index*, used to provide insight into overall preparedness of a country to participate in and benefit from the networked world. Similar rankings by the *World Economic Forum Competitiveness Index* also place the United States in the top 3 positions for each of the past 5 years. Much of this competitive performance has been the result of using information and communications technology throughout firms in practically every industry, including financial service firms, such as banks.

Throughout the 1970s, 1980s and early 1990s, banks were among the avid consumers of ICT, enabling them to push decision-making downward in the organization, as discussed by Drucker [4]. Concomitantly this brought about new sources of revenue streams for banks in the form of credit card processing and consumer loans. The rapid transaction processing capabilities of mainframe computers were the backbone of corporate strategic plans for numerous large banks which offered their customers access to cash dispensing machines throughout large metropolitan areas made available 24 hours per day, seven days per week. Large mainframes would also eventually lead to the creation of even newer sources of revenue, in the form of transaction fees for these dispersed automated banking systems.

By the mid-1990s, mergers among American banks increased at faster rates than exhibited previously, and it seemed like technology would continue to be a source of competitive advantage both operationally and strategically ad infinitum. But immediately after the consequent passage of the 1996 Telecommunications Act, making possible the use of the Internet for commercial usage, the banking industry's voracious appetite for information and communications technology had begun to surpass the high rate of return that senior executives had become accustomed to then. In fact, during the late 1990s, productivity trends in retail banking continued to disappoint and began to slide underneath the productivity trajectories exhibited in other industries, as illustrated in Figure 2. While the banking industry's information technology investments accelerated substantially, the sector consistently yielded disappointing labor productivity growth rates, even though these rates were higher than the economy-wide average, declining from 5.5% during the period 1987 - 1996 to 4.1% during the period after 1995, as identified by Olazabal [5]. Research into this paradox reveals that the relationship between information technology and labor productivity is more complicated than merely adding the former to lift the latter.

3.1 Interoperability problems in banking

Throughout the build-up that continued through the mid-1990s, the ICT investments made by banks were primarily done without taking into consideration the enterprise-level view of the firm, and specifically, how these systems would eventually interoperate. Instead, most of the investments were made in consumer services departments and marketing tools for customer information and support; still other investments were made in back-end applications that automated various corporate functions. This approach was a
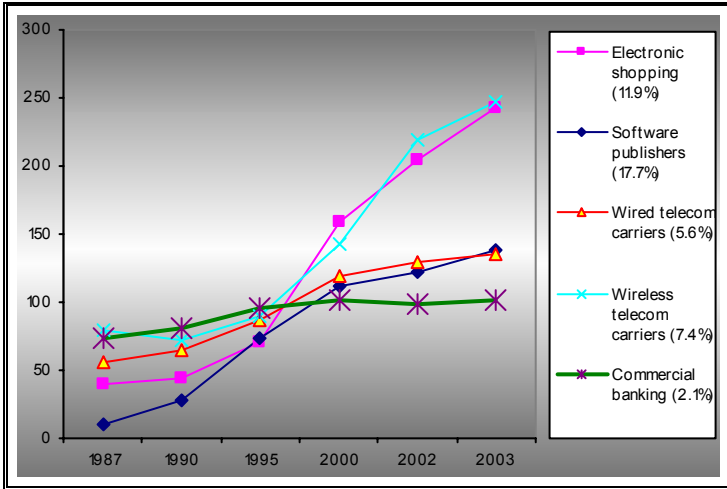
Figure 2:   Source: Statistical Abstract of the United States 2006, Table 622:
             Annual Indexes of Output Per Hour for Selected NAICS Industries.

departure from the traditional, more functionally organized method of implementing change around product lines, such as deposit accounts, loans, and credit cards. As a result, coordination among departments was loose and uncoupled, leading to an erosion of customer information flow throughout the organization.

To mitigate this problem, banks attempted to create a single customer interface, forcing them to integrate databases and downstream ICT systems. Once accomplished, banks adopted newer applications or, more succinctly, customer-relationship-management tools designed to improve customer retention and to facilitate marketing. This massive attempt to interlink the banks' databases required significant investments in personal computers for branch employees and call-center representatives, as well as the integration of complex systems. Also, upgrades in operating systems in the late 1990s caused banks to be burdened with keeping pace with technological change while simultaneously servicing customer needs.

To make matters worse, further effort was put into attracting new customers, primarily with credit card schemes based on elaborate pricing options. At the same time, bank mergers were getting larger. Although the industry consolidated at a steady pace before and after 1995, the size of the banks engaged in mergers grew, largely because of a 1997 regulatory change that lifted the prohibition against interstate bank mergers, which tend to involve larger banks. The average assets of bank merger participants increased from $700 million (1994–96) to $1.4 billion (1997–99). Naturally, the integration of larger systems involved greater complexity.

Lastly, banks were among the horde of firms rushing headstrong into the Internet frenzy of the late 1990s. New, unproven technologies, coupled with third party startup firms, helped banks gain a toehold in the Internet space prior

to 2000, and this, in essence, not only cost the firms additional dollars but also diverted attention away from the consolidated practices taking place in data centers internationally. As a result, market share among large banks began to slip as smaller, more nimble banks were able to withstand radical changes to their business operations as a result of the technology shifts evidenced in the late 1990s.

3.2 Net result of poor productivity

While some of the first attempts brought about by the Internet revolution produced tangible results, such as consumer convenience of conducting transactions on-line and the improved availability of account information available through call centers, automated teller machines, and World Wide Web sites (all of which went under the radar screen of 'captured' productivity improvements, since productivity measures quantity, not quality, of transactions), the costs to integrate disparate systems was enormous. Further, since on-line transactions account for only a small percentage of the banking's revenue stream, these qualitative improvements were not enough to reverse the losses of productivity growth manifest throughout the industry.

It is noteworthy to mention that technology managers are not inherently skilled to the degree that they would include measurement of the myriad intangible aspects for ICT improvements; thus, it become inordinately more difficult to appraise the true value imparted to productivity levels for implementing a costly investment in customer-relationship-management sales tools, for example.

Overinvestment in ICT also resulted from the manner in which banks make their technology purchases. To simplify maintenance of a personal computer for instance, firms often buy either a single or a small number of computer models, meant to satisfy the most demanding users, giving unnecessarily powerful computers to the majority of users. Further, since the more sophisticated end-users also demand newer computers more frequently than do average users, costly department- or even enterprise-wide upgrades become commonplace. Managers at the line level have little knowledge of the larger consequence of their actions and therefore no incentive to oppose this purchasing pattern, causing perpetual and unnecessary ICT investments to be made by firms.

4   Conclusion

Despite the generally disappointing results, banks have made enormous investments in information and communications technology. Some of these investments have led to increasing the flow of new customers, lured by the availability to maintain their account information on-line. Other investments were transparent to the end user, such as integrating disparate operating platforms and sophisticated databases. Further, the attention paid to maintaining a secure operating environment has driven upward the cost to the firm of keeping up with competitors. As a result, the level of measured productivity has dropped

in recent years, as only some of these operational improvements have been captured by the methods employed to measure productivity.

However, the availability of more sophisticated technology obtainable throughout banking translates into a future view that may be characterized as one of significantly improved performance levels, as the investments made over the past ten years reach their full payoff level, and as ICT spending slows. For early adopters of ICT, this is good news; for laggards, the picture is not so rosy. The pressure on banks to offer similar customer service levels, such as the capability for customers to view cashed checks, configuring call centers to automating customer calls using information technology, implementing bill consolidations program for demanding users, etc. places significant burden on a financial services industry searching for additional sources of revenue. As back-office reengineering continues, the dawn of a new era of significantly higher levels of productivity beckons.

References

[1] Greenspan, A., *Challenges for Monetary Policy-Makers*, Board of Governors, Federal Reserve System, Washington, October 19, 2000.
[2] Jorgenson, D., Information technology and the G7 economies. *World Economics*, 4(4), October - December, pp. 139-169, 2003.
[3] Warner, E., Universities promoting micro use in MBA curriculum. *Computerworld*, 24 September, pp. 40-41, 1984.
[4] Drucker, P., The coming of the new organization. *Harvard Business Review,* January – February, reprint, 1988.
[5] Olazabal, N. G., Banking: the IT paradox. *McKinsey Quarterly*, pp. 1,47-51, 2002.

Collaborative support for on-line banking solutions in the financial services industry

H. Krassnigg & U. Paier
Evolaris Research Lab, Graz, Austria

Abstract

Building and enhancing consumer trust in on-line banking on the World Wide Web is a critical factor in the success of e-Commerce systems. Though the number of on-line banking customers is increasing constantly in firms, there is a definite opportunity in convincing consumers to become customers. This study contributes insight into the development and growth of co-browsing and collaboration, as functionality in enabling improved on-line banking customer service and trust on the Web. Defined in the study are the benefits of building components of e-Services, consisting of collaborative guidance tools, pre-emptive support tools, and responsive service tools. The focus of the study is on benchmarking a sample of financial service firms and of tools of trust and on introducing an *interactive advisor* as a collaborative on-line banking service and tool of trust. The paper evaluates as an in-depth case study the functionality of interactive advisor tools and the benefits of the tools in enabling trust for on-line banking customers on the Web. This study will benefit business management practitioners and researchers in the financial services industry that are exploring continued opportunity and risk in on-line banking solutions of trust on the Web.
Keywords: customer retention and recovery, e-Services, interactive help desk, tools, trust, trust building and trust building components.

1 Introduction

1.1 Lack of customer trust

An important success factor for on-line banking is to be able to create and increase the customer's trust in e-Commerce service [1]. Although the number

of on-line banking users is constantly increasing [2], there is still far more potential to reach and convince new customers.

   With the cumulative acceptance and spread of internet based technologies, potential business in the field of electronic commerce grows. Despite the boom in on-line banking [3], there are uncertainties which are strongly anchored in the consumers [4]. The uncertainties, as perceived by the consumer, stem especially from the lack of trust in the use but also through the virtuality and the spatial separation and the aggravated assessment of the trust of the supplier [5]. Furthermore, there are insecurities due to the security of Internet communication. Due to the numerous amounts of suppliers furnishing the customer a service, the customer is often uncertain of the service [6].

   The cultivation of trust, however, can be positively influenced by the supplier with appropriate measures. This can also occur when many exogenous factors take effect on the cultivation of trust, which lie outside of the realm of influence of the supplier. Trust management will then find access to, for example, the areas of customer retention management, complaint management, service management, etc. The goal of trust management is to overcome (with on-line banking) risks and to build up a long term and continual trust relationship between the supplier and the customer [7].

1.2 Customer retention and recovery through raising trust

1.2.1   Process oriented approach

In e-Commerce, the sales process, according to Riemer and Totz [8], is classified in four phases: information, initiation, development and after sales.   The contention of the paper is that this must be complemented with a fifth phase, namely the area of exception handling, which should be observed in each of the four phases in Figure 1.
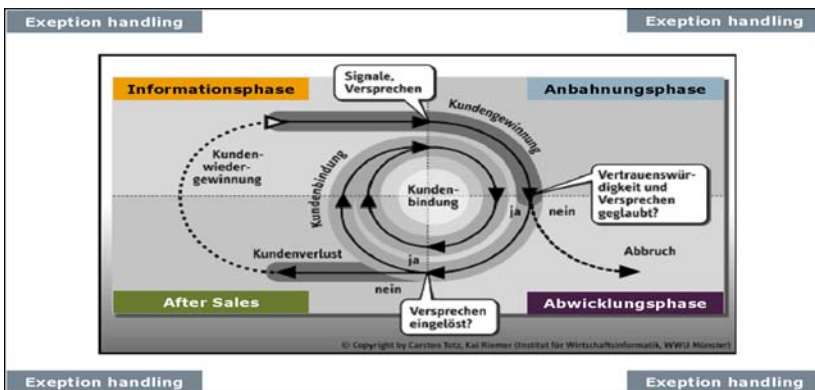
Figure 1:   Customer retention cycle from Riemer and Totz expanded through exception handling.

With customer retention, all phases will be run through. For customer recovery, it is necessary to run through all four phases, and retention requires more run throughs.

The pre-conditions for a successful business transaction are the creation of trust, which should be signalled by the supplier to the customer during the information phase. If the expectations of the customers are fulfilled or the supplier exceeds them, then the customer will step into a new business relation with the supplier, which can then lead to the retention of the customer [9]. If the customer trusts the supplier, then there is a possibility for customer retention. This will then be consolidated through the rising trust and maintained measures and will be kept intact [10]. Also, in the field of financial service, all four phases of the e-Commerce problems could arise, which could lead the customer not to close a transaction and to leave the cycle. If, however, in the scope of exception handling, he/she is optimally handled, then he/she can be brought back into the cycle. A customer dissatisfied and the consequential loss of trust can then be turned around, and this then considerably strengthens the customer retention.

1.2.2 Problem oriented approach

According to the problem oriented approach, the contentedness and the trust of the customer can be increased through the support of the simple tools and appplication, in case a problem with utilization occurs with the tools [11]. These can increase the trust, which is why some of them can be viewed upon as trust building components [12, 13], which can be divided into three different types of on-line customer services below and in Figure 2 [14].

- Responsive service tools: These enable the customer service inquiry to begin and offer an automatic support without being steered by a person. In this manner, customers have the possibility to initiate a service when they have a problem, which is usually without personal support of an employee in helping to solve the problem. Possible examples for these types of applications are: search, virtual agents, frequently asked questions, automated e-mail, self-service transactions, interactive knowledge base, checking images, online statements, and Avatare.
- Collaborative guidance tools: These accomplish a personal connection between customer and agent, in which the customer is able to request human help during a sales or service interaction. In order to enable the customer to have personal support during a transaction, a connection with an employee will be generated by the tools. Examples of applications are chat, collaboration, co-browsing, joint form filling, and instant messaging.
- Pre-emptive support tools: Through pro-active service, specific circumstances of the customer can be solved already before a problem occurs and, therefore, exceed customer expectation. Examples are news, account based alerts, status alerts, and actionable two-way alerts.

2 Methodology

The evolaris research lab of the authors of the study analyzed the largest 20 banks in their qualitative benchmarking study, and the findings detected 17 confidence-raising tools and applications. The benchmarking study reached the goal of identifying applications and technologies which would contribute to the confidence rising and, therefore, assist in customer retention in on-line banking. This study incorporated financial service businesses [15] and businesses from other trust sensitive branches [16], in order to identify e-Services and tools, which could support a rising in trust for digital transactions.

Differences in the applications were found mainly in the area of customer support in case of a problem. According to the available results, there is a clear trend in favor of responsive services. The most common examples in this connection are frequently asked questions, different demonstration applications, and calculators. These services are geared so that the customer can solve the problem without the interaction of a bank employee, such as frequently asked questions, downloads, and search engines.

Only a few banks are able to offer their customers applications, which furnish an interaction between the customer and the bank employee (interactive consultancy and chat applications). In the area of collaborative guidance, applications yield to new possibilities in customer service through co-browsing (cross screen comparison) and a direct approach with the customer via an available telephone connection (call back).

The least available are applications from the area of pre-emptive support, in which the customer is pre-supplied with information before the problem occurs or is even prepared with alternatives in order to avoid problems. In this group, are mainly those with a definable automatic alert via different channels, such as e-mail. Information is mainly transmitted for account coverage, and subscription respites from shares [17].

Financial service firms must check in which phase they need to catch up in. Once this is determined by firms, a systemic application of the trust building applications and technologies in this phase can follow [18].

An increase in the customer satisfaction and a decrease in the rate of abort during a transaction can be obtained if during all occuring problems support is offered to the consumer on the basis of simple applications [19]. This support through value added services raises the effectiveness and usage, lowers the costs and leads therefore to a customer recovery and retention. Successful firms use a combination of the three types on on-line services.

In order to enable customers to have personal support in the course of a transaction in on-line banking, the bank has to set up a connection between the customer and the employee, who can assist during the problem solving [20]. Examples of applications in this area are collaborative guidance, which is technically based solutions, such as chat, collaboration, co-browsing, joint form filling, and distant messaging.
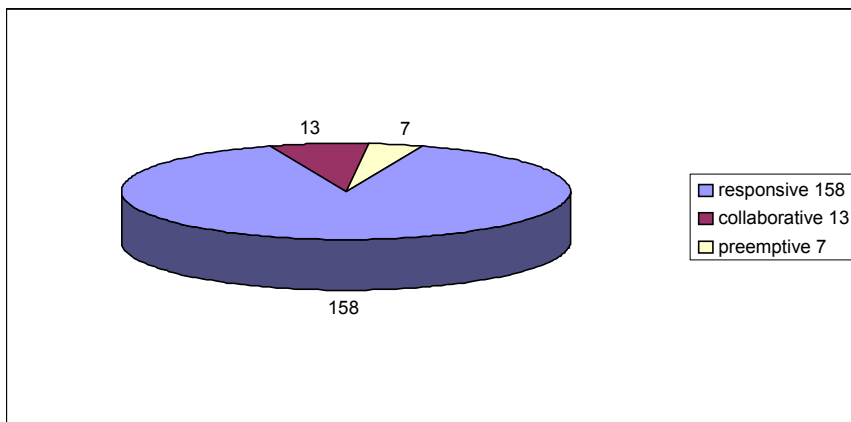
Figure 2:   Results of division of e-Services from evolaris benchmarking study.

This paper is focused on the collaboration and co-browsing service, known as *interactive advisor*. Following will take a look at the function principle, the utilization for the customer and the bank, as well as potentials from this application. Also, technical aspects of this collaborative service will be covered in the paper.

3   Analysis

3.1 Functional description of *interactive advisor*

The *interactive advisor* in Figure 3 allows customer support when problems occur in connection with an on-line banking transaction.

During the navigation (filling out a form or using the Web based calculation module), the customer clicks on the integrated button, to *interactive advisor* on the Web form. In the next step, a pop-up opens, in which the customer has the possibility to type in his telephone number or e-mail address. Depending on which configuration, the customer now sees either a new browser window for the text-chat, or the screen sharing would be initiated by the tools. In case of screen sharing, the customer has to activate a signed Java applet through a download. Screen sharing can also be initiated during a telephone call, in which the employee gives the customer an ID, which he would then use in order to have a connection.

In this manner, one can fill out an application form together or one can explain the problem at hand to the customer. This is established by screen comparison. The consultation through the Web, also known as e-Consulting, relates to the technical questions as well as content problems with the content

matter, which would be in connection with the products and services offered by the bank [21].

Figure 3:   *Interactive Advisor* of Hypovereinsbank (calculation form).

3.2 Economical aspects

3.2.1 Supplier
There are various software suppliers in this area [22]. The following descriptions arise from the product, Aspect Web Interaction, from the firm, Aspect Communications Corporation [23]. This is used for example by the Hypovereinsbank [24] .

3.2.2 Utilization
The customer can use, free of charge, this tool in on-line help and resolve problems. Through the help of the employee, who guides the customer through different processes, the financial service firms can decrease the rate of abort during on-line orders and calculations and can increase customer satisfaction through the service advantage, and this is where the customer retention is initiated by the tools. Customer trust can be increased through the advice performance, which is a crucial success factor in on-line banking. Technologically experienced customers are approachable through the innovative e-Service, which in turns enhances customer recovery [25].

    The sessions, including the text chats, can be recorded, which gives hints as to how to improve the service and the information content [26].

3.2.3 Implementation

In order to have such an extensive system, the bank has to first of all have a customer service center or a call center as the case may be in the firm. The implementation of this service would offer the financial service firm the following features [27]:

- Call-Back-Service on any of the clients' desired Web site;
- Automatic call back to the customer from the best suited agent in the customer care center;
- Synchronized browsing in a joint session (escorted browsing);
- Simultaneous telephone calling and browsing during simultaneous utilization from two canals through ISDN or cellular telephone);
- Interactive marking and flagging of the content through the agent on the Web site;
- File-Transfer during the telephone call; and
- Web-Conferencing (meet me).

Prerequisites for a successful implementation furnish the following:

- Provision of necessary hard and software (e.g. Fujitsu-Siemens Primergy H250 and W2K Advanced Server);
- Provision of necessary technical infrastructure in the customer care center (e.g. Siemens Hirocm 300E telephone system and ProCenter Software for the agents' work stations); and
- Documentation of the dialogue and communication process between the customer and the consultant on the Web site.

3.2.4 Risks

The visualized system requires encroachments in the technology and influences the function of the customer consultant in the branches. There is the risk that the qualified consultant does not accept the tool, because he fears that he will be replaced by call center employees.

Because of this, employee training is necessary. There is also a risk of customer acceptance. For one thing, the customer is not familiar with the procedures, and for another, there are technological utilization barriers with the customers, although this deals with on-line banking clients [28]. In any case, it will be difficult to charge for something in the long run which is being offered for free in the short run, or to charge for the service in the beginning. Also, the accessibility of potentially available customer relationship management (CRM) solutions, or customer data bank associated with a certain amount of expenditure, is a risk.

4   Technology

4.1.1 System structure and interfaces

The Web interaction server includes the Web interaction application server in Figure 4, which runs under the IBM WebSphere and serves primarily the provision of the Web based client for the customer interface. It is not necessary to have a connection to the corporate Web server, in order to enable an

integration of the application on the customer's Web site without a break. The automatic forward of questions from the logged-in customer is directly handled on the desktop of a consultant from Aspect Enterprise Contact Server, which delivers all relevant information for the follow-up chat with the consultant in the bank or in the call center of the firm.
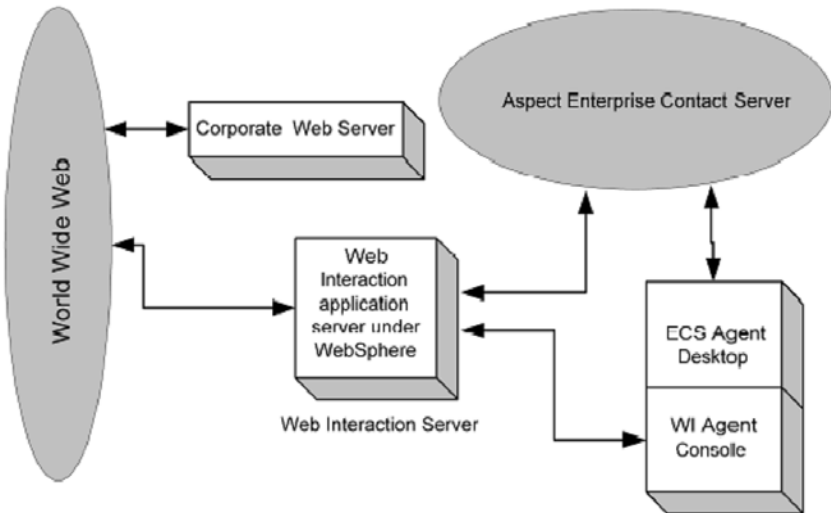
Figure 4:   Comprehensive depiction of application server [30].

4.1.2 Availability

The system can apply the load-balancing possibilities of the underlying server. In this manner, the extensive calculations and large amounts of questions can be channelled through numerous systems, all functioning concurrently.

The manner of distributing the processes to the processors has a huge influence on the whole performance of the system, because, for example, the Cache-content is local for every processor.

In responding to the http requests, systems are already switched on (front-end server), which then distribute the individual questions to the back-end server, according to assigned criterion.

The load balancing is a mechanism to safeguard against failure. Through the building of a cluster and the distribution of requests to single systems, one reaches a high level of safeguarding against failure, as long as the failure of a system is recognized by the tools, and the requests are automatically passed on to another system [29]. This function is only limited through the possibilities of the customer adviser employee and only reaches in a cost effective variation the business hours.

4.1.3 Security

The software allows itself to be conducted not only in internal networks (possible with IIS 4.0, IIS 5.0 and iPlanet 4.x Enterprise Edition on Solaris), but also in a demilitarized zone (DMZ). With this variation the server is in the DMZ. In this case, the MS IIS 5.0 webserver should be used by the system.

Following ports are employed by the system:

- Port for Web Interaction Application Server on Websphere: 80 (http) and 443 (https) respectively. Two way TCP/IP and http or as the case may be https with SSL coded variation;
- Aspect DataMart (responsible for reporting): Port 5001 two way TCP/IP; and
- Enterprise Contact Server ECS: Port 9001 2-Weg TCP/IP (blocks at external firewall).

4.1.4 Anti-virus Software

This software is to be used in combination with the available Web server, wherefore the same anti-virus protection measures can be met by the system. The customer has no possibility to distribute a feasible code from his/her computer to the workstation from the employee. Therefore, the risk to get infected is credible.

The customer has to activate an applet (for Internet Explorer), or as the case may be a signed script (Netscape), through a download. These are, however, declared safe because these are signed and therefore are rated in Internet Explorer and Netscape as harmless.

In order to further increase security, one could install "agent desktop" software behind the firewall. These could be obtained from the terminal services.

4.1.5 Back-up

Just like the topic concerning anti-virus protection, the same system that was used for the Web server can be used for the back-up. In due time, a three way distribution for the back-up will be necessary because of the backing up of individual systems. Concerning the back-ups of the individual systems, soon it will be necessary to use a tri-section back-up because of the difficulties experienced in retrieval.

In this manner, Aspect enterprise contact server can be securely separated from both of them in the DMZ situated systems (Web interaction server and corporate Web server). The advantage of the backed up data is that when a failure of the entire system occurs, then the corporate Web server can be recovered independently of the other systems.

5   Other technology

Due to the high costs a call center creates for a financial service firm, the firm has the possibility to use a virtual customer consultant agent, instead of a personal contact like with the *interactive advisor* in Figure 5. In order to find appropriate examples, paper refers readers to Sparda-bank Nürnberg eG [30] and the Bausparkasse Schwäbisch Hall AG [31] in Germany. Both firms place

emphasis on technology from Kiwilogic.com AG [32], which has already implemented approximately 100 virtual agents internationally [33].
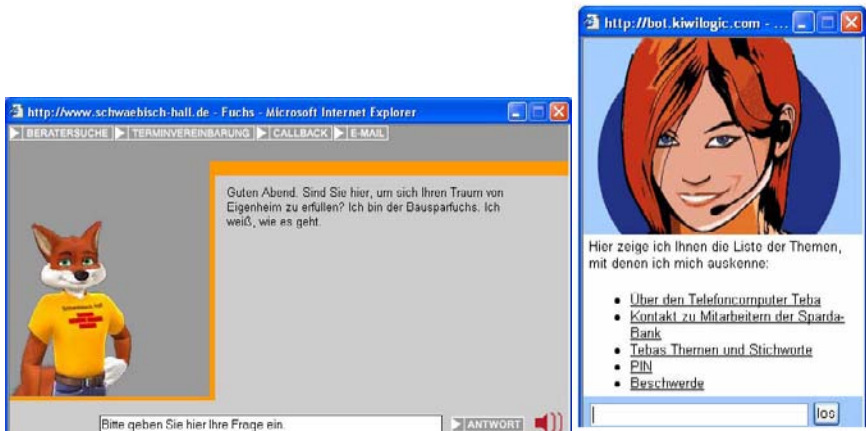
Figure 5:   *Interactive advisor* (entry form).

The above *illustrated adviser* [34] receives the re-entries from the customer through a simple HTML-form located in the Web browser. After dispatching the re-entries, the entries from the Web engine are further prepared through a common gateway interface (CGI). The answer is then searched within the knowledge base. The transferred keywords of the customer will then be compared with the questions already deposited in the data bank. Resulting consistencies will then be sent back to the customer. The answer text, as well as graphics depicting the mood, will be sent to the customer. Finally, those Web sites containing information that the customer has requested will be automatically recalled by the system.

References

[1]   Kundisch, D., Building trust – the most important crm strategy?. *Proceedings of the 3rd World Congress on the Management of Electronic Commerce*: Hamilton, pp. 1-14, 2002.
[2]   Jung, C., Internet und on-line banking: warum offliner offliner sind. *Die Bank*, 4, pp. 282-283, 2004.
[3]   Forrester Research, Inc., Efficient multi-channel banking. February, 2002.
[4]   Forrester Research, Inc., Experience - the key to on-line security issues. February, 2002.
[5]   Petrovic, O., Fallenböck, M., & Kittl, C., *Der paradigmenwechsel in der vernetzten wirtschaft*: von der sicherheit zum vertrauen, in Petrovic, O., Ksela, M., Fallenböck, M., & Kittl, C., (eds.). *Trust in the Network Economy*, Springer-Verlag: Vienna and New York, pp. 3-28, 2003.

[6] Demunter, Internetnutzung in Europa: Sicherheit und Vertrauen, http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-NP-05-025/DE/KS-NP-05-025-DE.PDF, On-line banking security: give customers more control and reassurance. January, 2006.

[7] Riemer, K. & Totz, C., *Nachhaltige Kundenbindung durch Vertrauensmanagement*, in Klietmann, M., (ed). *Kunden im E-Commerce. Verbraucherprofile - Vertriebstechniken – Vertrauensmanagement*, Symposium Verlag, pp.175-199, 2000, and Riemer, K. & Totz, C., Vertrauensmanagement — Loyalität als Schlüsselgröße, in Internetshopping
Report 2001: Käufer, Produkte, Zukunftsaussichten, p. 339, 2001.

[8] Riemer, K &, Totz, C., in Klietmann, M., (ed). Kunden im E-Commerce, p.183, 2001.

[9] Riemer, K &, Totz, C., in Klietmann, M., (ed.). Kunden im E-Commerce, pp.183-185, 2001.

[10] Petrovic O. & Kittl, C., Trust in digital transactions and its role as a source of competitive advantage in the network economy, *Proceedings of the IADIS International Conference*, Carvoeiro, Portugal, 2003.

[11] For a model of trust, refer to Petrovic, O., Fallenböck, M., Kittl, C., Wolkinger, & T., Vertrauen in digitale Transaktionen. *Wirtschaftsinformatik*, 45(1), pp. 53-66, 2003.

[12] Ba, S. & Pavlou P., Evidence of the effect of trust building: technology in electronic markets: price premiums and buyer behaviour. *MIS Quarterly,* 26(3), pp. 243-268, 2003.

[13] For trust building components and strategies, refer to Urban, G., Sultan, F. &, Qualls, W., Placingt trust at the center of your internet strategy. *Sloan Management Review,* Fall, pp. 39-48, 2000.

[14] Forrester Research, Inc., On-line service: the next generation, September, 2002.

[15] Citigroup, Bank of America, Egg, UBS, Advance Bank, Credit Suisse, Abbey National, Deutsche Bank, ING Postbank, National Australia Bank, Commerzbank, ICBC, HypoVereinsbank & Lloyds TSB Bank, 2002.

[16] Trust sensitive branches, such as insurance and notary health care, 2002.

[17] Evolaris Benchmarking StudieVertrauenssteigerung durch neue e-Services als Trust building components, 2002.

[18] Gomilschak, M. & Kittl, C., The role of trust in internet banking. *Proceedings of the MIPRO 2004 Conference*, May, Opatija, Kroatien, pp. 24-28, 2004.

[19] For rate of cancellation and reasons for not using the on-line banking, refer to Forrester Research, Inc., Why on-line banking users give up, May, 2002.

[20] Dührkoop, Strategische Erfolgsfaktoren für das Online- Angebot von Privatbanken, October, 2001.

[21]   The description of the *interactive advisor* was conducted during proceedings with the Hypovereinbank, 2001.
[22]   Chordiant Software, Inc., http://www.chordiant.com/home.html.
[23]   Aspect Communications Corporation, http://www.aspect.com/index.cfm.
[24]   Bayerische   Hypo-   und   Vereinsbank   AG, http://www.hypovereinsbank.de/pub/home/home.jsp.
[25]   Naef, A., Maintaining customer relationships across all channel *Proceedings of Financial Services Europe*, 13 October, London, UK, 2005.
[26]   Holzhauser, A., E-CRM - E-Service, http://www.factline.com/154848.0/.
[27]   Aspect Communications Corporation.
[28]   Jung, C., Internet und On-line banking: warum offliner offliner sind.   *Die Bank*, 4, pp. 282-283, 2004.
[29]   For load balancing, refer to Article *Lastverteilun,* in: wikipedia, die freie Enzyklopädie,   Bearbeitungsstand,   16   December,   2005, http://de.wikipedia.org/w/index.php?title=Lastverteilung&oldid=1170701 4.
[30]   http://www.sparda-telefonbank.de/wer.html.
[31]   http://www.schwaebisch-hall.de/.
[32]   http://www.kiwilogic.de/.
[33]   http://www.kiwilogic.de/.
[34]   Kiwilogic Lingubot Software, Kiwilogic.com AG.

Time value of the Internet banking adoption and customer trust

Y. T. Chang
ESRC Centre for Competition Policy, University of East Anglia, UK

Abstract

Studies on adoption of new technologies have focused mainly on the behaviour of adoption and on efficiency gains from advancement in the state of technology. The contention of this study is that it is more appropriate to regard the adoption of technology in the banking industry in dual aspects by banks and by customers, given the intermediary role of banks. Despite growing interest in e-Commerce and financial activities, consumer choice decisions as to whether to adopt banking on the Internet has not been fully investigated in the literature. Applying data from Korea on the adoption of on-line banking, the study evaluates consumer characteristics that affect the adoption decision. The study focuses insight on whether the time value perceived by consumers affects their adoption decision to banking on the Internet, introducing decision criteria. The study furnishes helpful information for managers in the banking industry, regarding customer characteristics of trust and risk factors that determine adoption of banking on the Internet.
Keywords: consumer adoption, Internet banking, perceived time value, risks and trust.

1   Background

The banking industry has been significantly influenced by evolution of technology. The growing applications of computerized networks to banking reduced the cost of transaction and increased the speed of service substantially. In particular, the nature of financial intermediaries made banks improve their production technology, focusing especially on distribution of products. In other words, the evolution of banking technology has been mainly driven by changes in distribution channels, such as the development of over-the-counter (OTC),

automated-teller-machine (ATM), phone banking, tele-banking, personal computer (pc) banking, and, most recently, Internet banking (IB).

Applications of new technologies, including the Internet, have created new methods of doing business. For instance, e-Commerce and e-Finance have clearly changed the business environment. However, there are only a few studies on consumer behaviour, relative to the vast amount of literature on the behaviour of firms regarding technology adoption, especially in the field of banking and finance. The paper posits that customer trust and risks associated with Internet banking are useful areas of investigation, and that perceived time value of Internet banking adoption is one of the important customer characteristics for the Internet banking adoption.

This paper uses on-line survey data from Korea on Internet banking to analyze the Internet banking adoption pattern across customers. The determinants of IB adoption by customers are identified in a dynamic framework, in order to explain why new banking technologies are not always taken up by the mass market. Differences in the characteristics of early adopters and late delayed adopters are presented in the paper, while customer trust and risks are further discussed for those who have not yet adopted Internet banking.

In the context of trust and risks, the evolution of new technologies in banking and finance has raised additional concerns. As indicated in the survey by the Bank of International Settlements BIS [1], most governments believe that new supervisory or regulatory measures are necessary for Internet banking, although it will take time for them to prepare prudential regulatory guidelines. On the basis of results in this paper, the study shows that the relevant banking regulation has an important implication for adoption of a new banking technology.

The next section describes the new banking technology of Internet banking and factors likely to affect its diffusion, followed by an investigation of the theoretical and empirical literature. The study then presents a duration model for Internet banking adoption and the results, with further discussion on factors preventing Internet banking adoption. The last section concludes with discussions of policy.

2 Introduction

One could notice that the evolution of banking technology from CD and ATM to Internet makes banking transactions more mobile (or less location restricted) at a lower fee at the terminal. In addition, the Internet added a new feature of information search in banking, when it retains the advantage of various information types, e.g. in text and audio-visual, which are furnished by CD and ATM. However, despite the benefits of Internet banking, this medium has not yet replaced traditional banking channels, and the banking industry seems to maintain the multi-channel distribution approach.

Since banking technology has been deployed in pursuit of reduction of distribution costs, Internet banking can be considered as a process innovation, with which both banks and customers save time and money. It also allows new customers to visit virtual banks through the public Web network, while phone

banking and personal computer banking provide only a closed network limited to the existing clients.

Increasing competition among the leading banks also promotes product and service differentiation. Despite the nationwide Internet banking system developed in 1999 by the consortium led by Korea Telecom and several banks, most leading Internet banking firms now use their own system to differentiate from rivals. Currently, all 17 commercial banks in Korea are offering Internet banking. Although they may vary, four main areas of Internet banking services are information search engine, balance check, fund transfer, and loans, in addition to the basic services, such as opening an account and financial product sales. Internet banking does not have the same capacity as CDs and ATMs in delivering cash; however, there are numerous informational features which enable customers to search for appropriate products and services, make a decision, and act on it over the Internet. One important observation to make is that customers need to become more proactive in their information search, in the absence of bank tellers or financial advisors on the telephone.

3   Focus

Davies [2] indicates that society fully benefits from a process or product innovation only when the innovation is diffused enough to enhance the productivity of firms or the utility of consumers. However, most of the earlier literature on technological progress focused on the behaviour of firms, analyzing how process innovation would improve its productivity, while the consumer behaviour in relation to innovation has been less frequently discussed in the literature. Gourlay and Pentecost [3] indicate that the inter-firm diffusion of new technology has been relatively less researched for the financial industry, compared to other industries. In particular, study on customer behaviour of financial technology adoption is almost next to none.

Mansfield [4] indicates that commonly used epidemic models of diffusion can draw an analogy between the contact among firms or consumers and the spread of disease in an epidemic sense. For example, some consumers adopt a new technology before others, because they happen to become infected first. Similarly, some technologies diffuse faster than others, as they are more contagious, due to its profitability and risk factors. On the other hand, Karshenas and Stoneman [5] indicate diffusion into three different mechanisms of rank effects, stock effects, and order effects, which explain the cases where firms with sufficiently high ranking adopt an innovation first, early adopters obtain higher returns on the new technology with diminishing returns in time, and adoption is profitable for only early adopters who secure access to the critical input.

Hannan and McDowell [6] indicate strong evidence for rank effects in the diffusion of ATMs, while rejecting the existence of epidemic effects. However, their approach has to be further tested as they excluded the aspects of consumer adoption. More recently, Akhavein et al. [7] indicates few quantitative studies on the diffusion of new financial technologies and the weakness where the technology is limited to ATMs.

However, more recent developments in the literature focus on trust and risks associated with the Internet, in general, and unidentified risks in Internet banking and finance, in particular. In the general context, Mansell and Collins [8] furnish a comprehensive collection of recent literature on on-line trust and crime. On the other hand, Kim and Prabhakar [9] indicate that a possible reason for the delayed adoption of the Internet as a retail distribution channel is the lack of trust consumers have in the electronic channel and in Web merchants. Similarly, Bauer and Hein [10] indicate that some of the hesitation to adopt Internet banking is due to perceived risks.

4 Methodology

The on-line survey data from Korea were collected by sending out 3200 e-mails to predetermined addresses, based on a systematic and stratified sampling. The explanatory variables included in the analysis were drawn from the data for the following categories: demographics, exposure to Internet banking, awareness, banking behaviour, and customer time value and risks.

A duration model is used in order to investigate the dynamics of the Internet banking adoption process. The determinants of early adopters versus delayed adopters are identified as the data contain the sequential information of adoption time. Given the interest in the length of time that elapsed before customers adopt a new banking technology (Internet banking), a hazard rate is estimated for IB adoption in each month, conditioning on the fact that the customer has not adopted Internet banking by that time, as indicated in eqn. (1).

$$\lambda(t) = \lim_{\Delta \to 0} \frac{\Pr(t \leq T \leq t + \Delta \mid T \geq t)}{\Delta} = \lim_{\Delta \to 0} \frac{F(t + \Delta) - F(t)}{\Delta S(t)}$$
$$= \frac{f(t)}{S(t)} = \lambda p (\lambda t)^{p-1} \tag{1}$$

Then, the probability density function and the associated survivor and failure functions are written as follows:

$$f(t) = \lambda p (\lambda t)^{p-1} \cdot S(t) = \lambda p (\lambda t)^{p-1} \cdot e^{-(\lambda t)^p} \tag{2}$$

$$S(t) = \Pr(T \succ t) = 1 - F(t) = e^{-(\lambda t)^p} \tag{3}$$

$$F(t) = \Pr(T \leq t) = 1 - S(t), \quad where \ \lambda \equiv \exp(\beta'X) \tag{4}$$

The hazard rate $\lambda(t)$ appears to be the conditional probability of having an exact spell length of t, i.e. adopting Internet banking in interval $[t, t + \Delta t]$, conditional on survival up to time t in equation (1), but one should note that the hazard rate is not a probability in a pure sense, since it can be greater than 1 for

positive duration dependence $(p \succ 1)$. Now the hazard function is derived by conditioning on survival up to time t, and the survival function is written as in equation (3). Subsequently, the failure function takes the form, $1 - S(t)$, as in equation (4).

5   Analysis

Figure 1 illustrates the initial reasons why customers adopt Internet banking. Not surprisingly, more than 50% of the respondents indicated 'time saving' as their initial reason for using Internet banking, followed by 'easy payments' (28%). This draws research attention to time value of customers and justifies the inclusion of survey response time (a proxy for time value of customers) to the duration analysis.
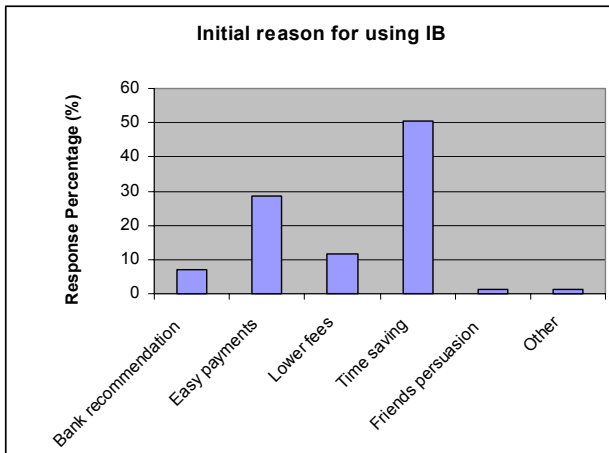
Figure 1:   Initial reason for using Internet banking.

On the other hand, those who have not yet adopted Internet banking seem to be most concerned about the on-line security risks (48%), and many of them do not feel the urge to adopt Internet banking, since they find their banking convenient enough without Internet banking (37%), as indicated in Figure 2. This obviously brings forward policy discussions on how to regulate and/or manage security risks that arise from Internet banking and how to educate the customers about the benefits of Internet banking. If one believes the arguments indicated by Davies [2], current society is not fully benefiting from the new technology (Internet banking), and there is opportunity for enhancement of social welfare if appropriate policy measures are in place.
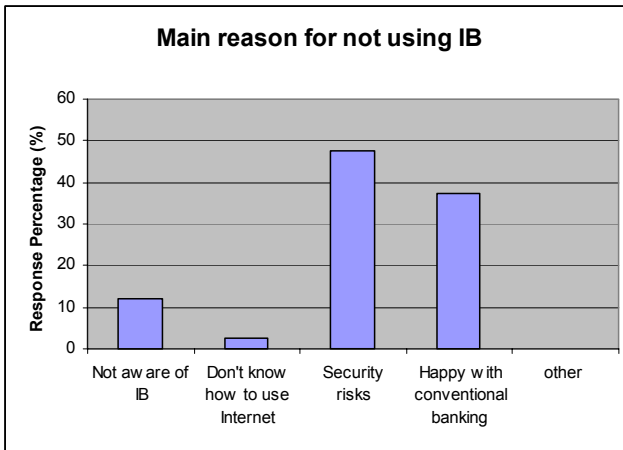
Figure 2:   Main reason for not using Internet banking.

The results from the duration analysis of Internet banking adoption are presented in Table 1, of which the second column presents the hazard ratio of each explanatory variable, and the last column shows the predicted marginal effects on adoption time, measured in months. Although the number of variables with statistically significant hazard ratio is limited, most of the variables furnish useful insight.

Gender, marital status, and residential area seem to matter more significantly than other variables. According to the predicted marginal effects on adoption time, males would adopt Internet banking 3.55 months earlier than females at the median, which is not surprising as the IDC [11] report on adoption of wireless communication shows that young male groups are more likely to adopt earlier. Education is not significant in determining customer adoption time of Internet banking, but the duration dependence is negative, which means that further education delays the adoption, perhaps due to risk-aversion. Age does not seem to have much impact, nor does personal income, although the effects seem to be non-linear. Singles are less likely to be early adopters, perhaps given their lower time value or lack of complex banking activities. Another important finding is that residents in the Seoul metropolitan area seem to delay their Internet banking adoption than residents in the provincial areas. This coincides with the time saving reason, as provincial residents may need to travel further to bank branches, and hence they save more incentives to adopt Internet banking than those who have many bank branches or ATM machines nearby in the metropolitan area.

In terms of banking behaviour, Internet banking recommendation does not have much impact on the adoption time or else seems to have rather averse effects, by making customers suspicious and delay the adoption. On the other hand, those who are well aware of interest information tend to adopt Internet banking earlier, given the benefits of fast on-line information services.

Table 1:   Duration estimation results.

Independent variables	Parametric Weibull Hazard ratio (Z)	Marginal effects on adoption time (predicted median t=22.51) dt/dx
Sex (1=male; 0=female)	1.3287 (1.78)*	-3.5493*
Education (1=Univ/College or above)	.8738 (-.54)	1.5780
Age	1.0101 (.19)	-.1206
Age squared	.9999 (-.03)	.0002
Personal income	1.0021 (1.13)	-.0247
Personal income squared	.9999 (-.96)	.0000
Single	.5166 (-1.79)*	8.2873*
Married	.5756 (-1.49)	6.5485
Outright owned house	.9200 (-.58)	.9987
Seoul metropolitan residence	.7695 (-1.81)*	3.0973*
IB recommended	.8143 (-1.05)	2.3791
Interest rate awareness	1.2208 (1.42)	-2.3753
First mover bank dummy	1.1424 (.74)	-1.5661
Market leader bank dummy	1.2086 (1.35)	-2.2598
Concerned about bank's reputation	.9750 (-.19)	.3049
Survey response time	1.0108 (.81)	-.1289
Survey response time squared	.9999 (-.59)	.0014
Ln(p)	.626 (11.15) ***	
Parameter P	1.877	
χ^2	27.28	
Log likelihood	-264.94	
p-value	.0541	
No. of adoptions	246	
Time at risk	6260	
Unobserved heterogeneity	Not significant	

Note:
1. Standard errors are in the parentheses.
2. *,**,*** Z-values significant at the 5%, 2.5%, and 1% levels respectively.
3. *,**,*** χ^2 -values significant at the 5%, 1%, and 0.1% levels respectively.
4. Hazard ratio greater than 1 indicates a positive duration effect on adoption, i.e. more likely to be an early adopter.

However, the hazard ratio does not seem to vary much whether customers are banking with the first Internet banking introducer (order effects or first mover advantage) or the large market leader bank (rank effects), although these two bank dummies have positive duration dependence, i.e. early adoption.

It is disappointing not to see any significant results for reputation criteria of banks and the survey response time, but the signs of the duration dependence support the earlier discussion on customer trust, risks and time value in this section. Customers who care about reputation of banks can be risk averse and hence delay the adoption of Internet banking. By contrast, those who took longer in responding to the survey response are more likely to adopt Internet banking earlier, given their high time value, but at a diminishing rate.
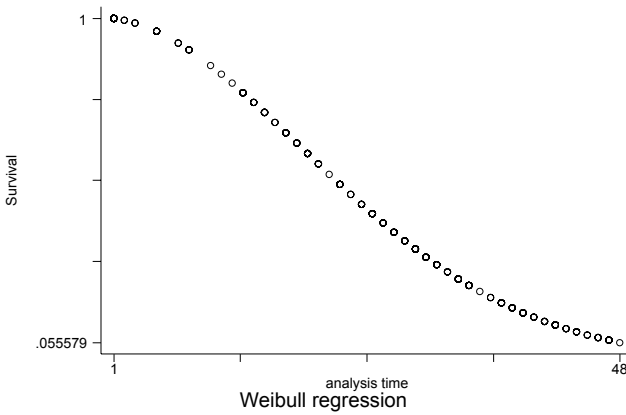
Figure 3:   Cumulative survival function for IB non-users using Weibull
              distribution.

The aggregate estimated pattern of Internet banking adoption is shown in Figure 3, in terms of a survival function, which indicates the number of Internet banking non-users in an S-shaped decline over time. This adoption pattern is also significant as indicated in the duration parameter, $Ln(p)$. Whether the society can reach the optimal level of Internet banking adoption is up to when and where public policy intervenes in the adoption path of Internet banking. When customers face unidentifiable levels of risk associated with Internet banking, such as human errors in inputting data on the Web or security breakdown on the protection of personal information, public policy has a role in reducing the potential welfare loss associated with the event. We are living in a society increasingly reliant on the Internet, but unfortunately the Internet is mainly unregulated, and the current regulation makes it hard oversee the global network, due to the openness of the Internet.

6   Conclusion

The results presented in this study furnish strong evidence that the adoption of Internet banking and its timing are affected by individual characteristics, in particular, gender, marital status and residential area. The analysis also included other individual characteristics in terms of demographics, exposure to the Internet banking, information seeking behaviour, general banking behaviour, and the customer trust and time value, which were not statistically significant, but reassured the time value. However, the duration dependence is significantly positive showing that the earlier literature on epidemic effects of technology diffusion is rightly put forward. More importantly, the descriptive illustration of the initial reasons for Internet banking adoption and the reasons for not adopting it, furnishes us an important field where policy makers and managers could intervene in industry. If security and trust issues are the main concerns for both

adopters and non-adopters, appropriate public policy and regulation are required to mitigate the potential loss of welfare arising from financial accidents on the Internet as well as to optimise the speed of adoption. The analysis and the discussion in this study only focused on the adoption of Internet banking, but the lessons from Internet banking adoption in Korea shed light on investigation of new industries based on the Internet.

Acknowledgements

References

[1]   BIS, Electronic finance: a new perspective and challenges. *Bank for International Settlements*, BIS Papers (7), 2001.
[2]   Davies, S., *The Diffusion of Process Innovations*, Cambridge University Press: Cambridge, UK, 1979.
[3]   Gourlay, A. & Pentecost, E., The determinants of technology diffusion: evidence from the UK financial sector. *The Manchester School*, 70(2), pp. 815-203, 2002.
[4]   Mansfield, E., *The Economics of Technical Change*, Norton: New York, NY, USA, 1968.
[5]   Karshenas, M. & Stoneman, P., Rank, stock, order and epidemic effects in the diffusion of new process technologies: an empirical model. *The RAND Journal of Economics*, 24(4), pp. 503-528, 1993.
[6]   Hannan, T.H. & McDowell, J.M., The impact of technology adoption on market structure. *The Review of Economics and Statistics*, 72(1), pp. 164-168, 1990.
[7]   Akhavein, J., Frame, W.S. & White, L.J., The diffusion of financial innovations: an examination of the adoption of small business credit scoring by large banking organizations. *Federal Reserve Bank of Atlanta*, Working Paper Series 2001-9, 2001.
[8]   Mansell, R. & Collins, B.S., (eds). *Trust and Crime in Information Societies*, Edward Elgar: Cheltenham, UK and Northampton, MA, USA, 2005.
[9]   Kim, K. & Prabhakar, B., Initial trust and the adoption of B2C e-commerce: the case of internet banking. *The Database for Advances in Information Systems*, 35(2), Spring, 2004.

[10] Bauer, K. & Hein, S.E., The effect of heterogeneous risk on the early adoption of internet banking technologies. *Journal of Banking and Finance*, forthcoming 2006.

[11] IDC, Unwiring the internet: end-user perspectives. *International Data Corporation Asia / Pacific Report*, (AP181102J), 2002.

Financial assurance program for incidents induced by Internet-based attacks in the financial services industry

B. G. Raggad
Pace University, USA

Abstract

This paper furnishes an analytical model for the generation of a risk-driven financial assurance program capable of preventing, detecting, and responding to financial incidents (FAPG) for a general support system. Risk is defined in the paper as a basic belief assignment. The study reviews a single general support system with a known basic risk, integrating ids evidence and meta-evidence obtained from security management, in order to estimate the current system security risk position. The study shows the functioning of the FAPG, by generating a risk-driven financial assurance program, for a relatively small general support system in a firm in the financial services industry. This study is focused on financial incidents induced by Internet-based attacks but introduces a framework for further research.
Keywords: financial assurance, Internet, risk, security, World Wide Web.

1   Background

The story of financial fraud that affects consumers and firms is abundant in the literature. Forensic audits in general continue to indicate earnings overstated by millions if not billions of dollars in the United States. There is no doubt that corporate fraud in the United States has affected market values of firms, public pension funds, and consumer savings plans. Firms globally however continue to engage in a diversity of illegal and non-ethical accounting schemes. Effectiveness and timeliness of auditors in identifying fraud are of concern to industry internationally. It is important to discern what a firm can do if auditors fail to detect fraud. Is a computer information system capable of examining financial statements and detecting financial fraud? Efforts from investors and

auditors help in furnishing information critical in designing such a system and in clarifying the context of the financial statements and the content that may lead to early warning signs of earnings mismanagement.

In order to enable the feasibility of a fraud detection information system, the paper of this study posits a basic financial taxonomy as a framework for the design of this system. The organization of financial fraud generates an actual taxonomy based on the discrimination parameters of method of delivery, imposter, victim, and attack. The method of delivery has distinct values of phone, mail, media, and the e-Banking Internet. The imposter parameter has distinct values of user and business, and the victim parameter has similar values. The method of attack parameter has values of impersonation, decoy, information corruption, information leakage, and physical. This financial fraud taxonomy generates 4x2x2x5=80 classes of fraud. Sets of 80 fraud signatures can be applied in the design of the fraud detection information system. Fraud intrusion detection systems aim at detecting each of the 80 frauds, based on embedded information in signatures. Literature furnishes information on how to defend firms from the frauds and to implement countermeasures to preclude actualization of the frauds.

The study defines fraud response as the sequence of actions that are effected if a fraud is in action. That is, given the information of fraud responses, the study introduces an information system of detecting financial frauds, based on the aforementioned 80 signatures, and of enabling the planning of responses to preclude the detected fraud, search for the imposter, and recover from the prevented fraud. Such a system is defined effectively as a fraud detection and response system.

2 Introduction

A general support system is however interconnected information resources under the same direct management control which shares common functionality. This is the basic infrastructure of a financial firm owning e-Banking capabilities. A general support system normally includes hardware, software, information, data, applications, communication facilities, and personnel and furnishes support for a variety of clients and / or applications. A general support system, for example, can be a local area network, including smart terminals that support a branch office, an agency backbone, a communications network, or a departmental data processing center, including operating system and utilities. This study is focused on financial incidents induced by Internet-based attacks. The general support system is the only source of any network disruptions at the origin of financial incidents. A source of literature on Internet-based security disruptions is furnished in Ludovic and Cedric [1].

At the same time, institutions, including agencies of the federal government, have applications that have value and require protection. Certain applications, because of the information they contain, process or transmit, or because of their criticality to the missions of the institutions, require special management oversight. These applications are defined as major applications.

Major applications are systems that perform clearly defined functions for which there are readily identifiable security considerations and needs, as for example, in an electronic funds transfer system. As in a general support system, a major application might comprise many individual programs and hardware, software, and telecommunications components. These components can be a single software application or a combination of hardware / software focused on supporting a specific mission-related function. A major application may also consist of multiple individual applications if all are related to a single mission function, like an e-Banking application.

The function of a risk management program is to determine the level of protection currently provided, the level of protection required, and a cost-effective method of furnishing needed protection for a general support system of an institution or a major application. The output of such an activity is a risk-driven security program. The most fundamental element of risk management, in a financial firm, is however, the evaluation of the security position of the firm. Risk management identifies the impact of events on the security position and determines whether or not such impact is acceptable and, if not acceptable, furnishes corrective actions.

The primary purpose for conducting a risk analysis is to evaluate a system risk position of a firm and to identify the most cost-effective security controls for reducing risks. Risk analysis involves a detailed examination of the target system. It includes the threats that may exploit the vulnerabilities of the operating environment, which result in information leakage, information corruption, or denial of system services. Risk analysis activities are planned in terms of the current status and mission of the financial firm.
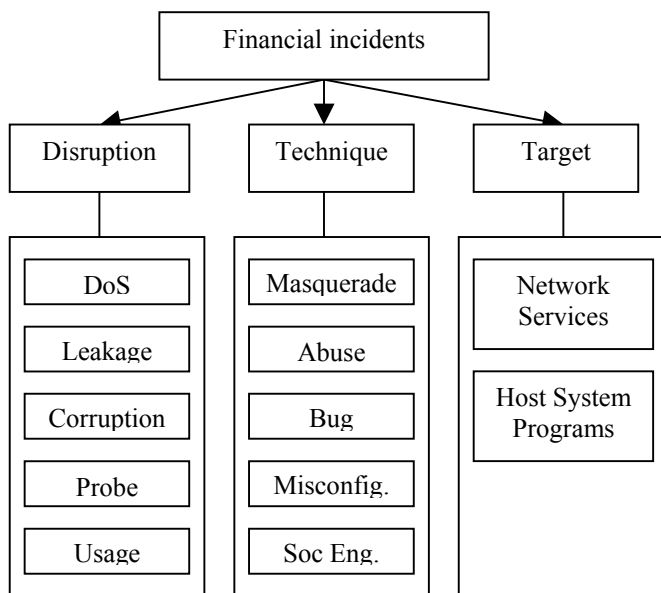
Figure 1:   Threats to the general support system.

3   Methodology

The FAPG cycle organizes the evolution of the Internet-based attack steps that together generate additional risk to the target general support system of the firm or the major application. The target system starts by having vulnerabilities. If the target system does not suffer from any vulnerability conditions, then the system cannot be victim of attacks of attackers. Even if the vulnerability conditions exist but there are no threats capable of exploiting those conditions, then there will still be no risks to the target system. This study posits that the attacks are dormant until some vulnerability conditions and some exploiting threats co-exist in order for the attacks to be started by the attackers. Figure 1 furnishes a fundamental taxonomy of Internet-based attacks and techniques used to compose those attacks. Denning [2], Lippmann et al. Ludovic and Cedric [1] furnish further information on this taxonomy of attacks.

The passage from a dormant stage to an initiation stage may be achieved through the following access conditions, also defined as privilege conditions in Kendall [4]:

- IR: Initiation by Remote Access to Network;
- IL: Initiation by Local Access to Network;
- IP: Initiation by Physical Access to Host;
- IS: Initiation by Super User Account; and
- IU: Initiation by Local User Account.

The passage from the attack initiation stage to planning the attack involves the study and selection of the technique or model to be used in the attack process. As defined in Lippmann et al. e study is focused on the following attack models:

- PM: Planning Attack by Masquerade;
- PA: Planning Attack by Abuse;
- PB: Planning Attack by Exploiting a Bug;
- PC: Planning Attack by Exploiting an Existing Misconfiguration; and
- PS: Planning Attack by Social Engineering.

The passage from the planning stage to executing the attack may involve sequential steps, including testing or elevation of privileges to prepare the sufficient conditions to carry the attack. The taxonomy of attacks employed in designing the FAPG model is focused on the following attack classes:

- EP: Executing Attack by Probe;
- EL: Executing Attack by Information Leakage;
- EC: Executing Attack by Information Corruption;
- ED: Executing Attack by Denial of Service; and
- EU: Executing Attack by Unauthorized Use of Resources.

Figure 2 furnishes the steps needed to provoke some attack scenarios leading to financial incidents. The last stage in the FAPG cycle is the response stage. The appropriate response activities to mitigate risks generated following the execution of an attack are focused on the following classes:

- RM: Response by Managerial Controls;
- RO: Response by Operational Controls; and
- RT: Response by Technical Controls.
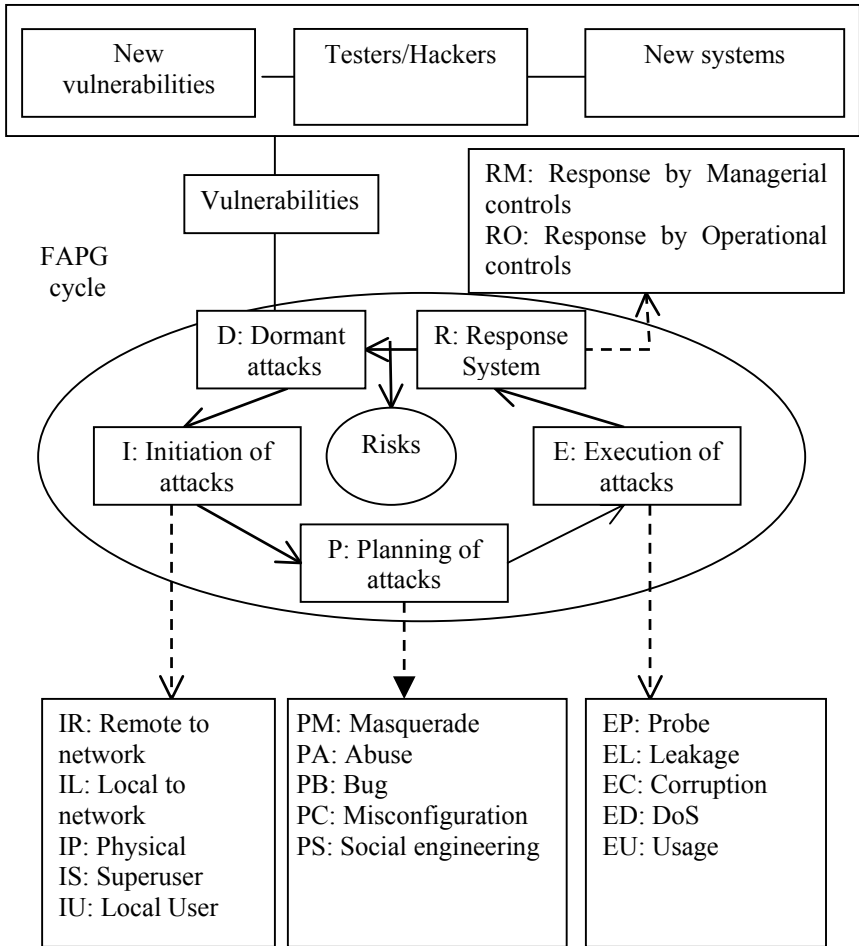
Figure 2:   Attack initiation scenarios.
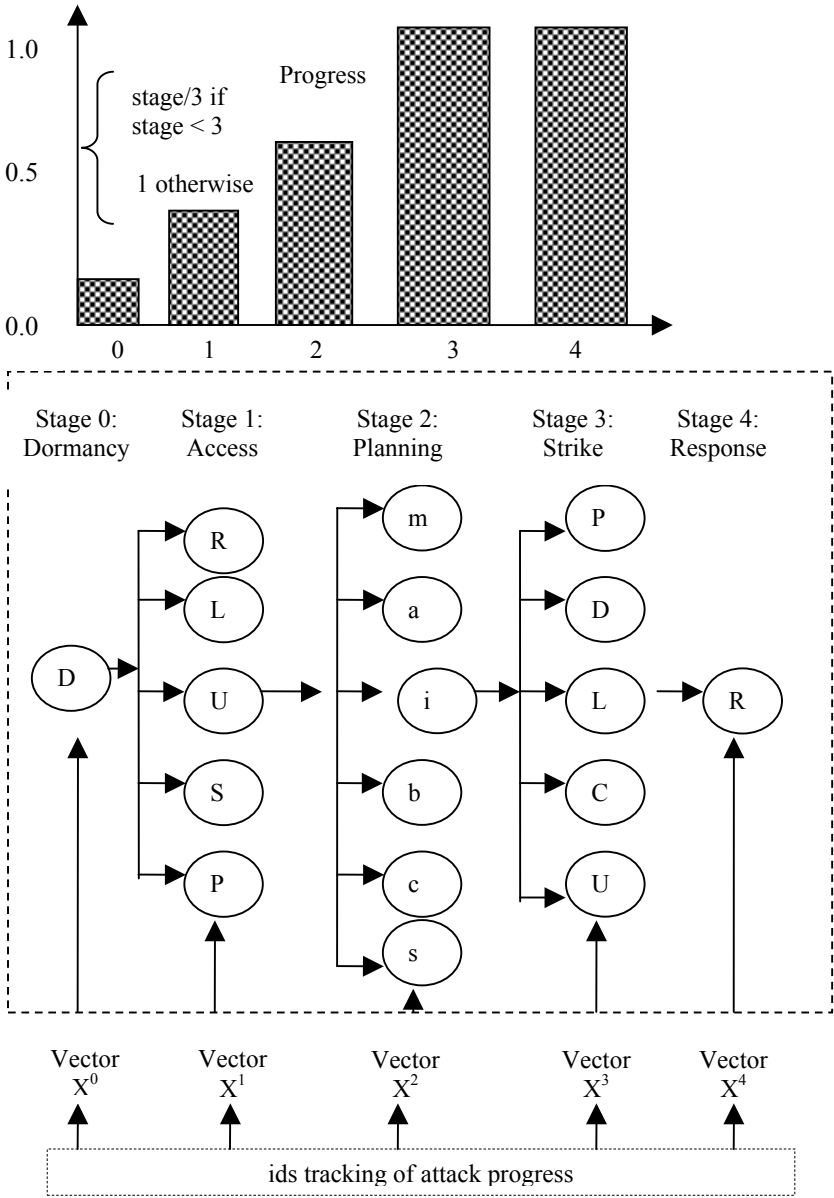
Figure 3:   Detecting financial incidents through detecting security attacks.

4 Analysis

The FAPG captures all data describing the behavior of the general support system of the final institution and the major application, integrating simple log files or well-designed intrusion detection systems (ids) Denning [2] and Smets 5]. The system processes data to analyze all events that lead to financial incidents. Most often, the intrusion detection systems are sufficient to detect financial incidents and prevent the incidents Barrus [6], Porras and Neumann [7] and Ryon [8]. If these systems do not identify these attacks on time, then incident responses cannot be planned earlier to preempt the execution of those attacks. In this scenario, recovery actions are evoked by the firm. The study employed basic belief assignments (bba) to model the problem domain Smets 5], Smets and Kennes [9] and Lindqvist and Jonsson [10].

Assume that the basic risk is given by the following bba, where A denotes an attack and $\neg$ A its negation:

$$m_0 \text{ bba on } \theta=\{A, \neg A\}; m_0(A)=r_0; m_0(\theta)=1.$$

The current risk position of the firm is computed based on evidence obtained from the ids and meta-evidence obtained from the financial management team. Smets [5] and Smets and Kennes [9] further indicate belief functions in the modeling of uncertainty and generating decisions.

The ids generates the following evidence:

- $m^s[D]: 2^\theta [0, 1]$;
- $m^s[I]: 2^\theta \rightarrow [0, 1]$;
- $m^s[P]: 2^\theta \rightarrow [0, 1]$;
- $m^s[E]: 2^\theta \rightarrow [0, 1]$; and
- $m^s[R]: 2^\theta \rightarrow [0, 1]$.

Meta-evidence is defined in the following:

- $m^m[D]: 2^\theta \rightarrow [0, 1]$;
- $m^m[I]: 2^\theta \rightarrow [0, 1]$;
- $m^m[P]: 2^\theta \rightarrow [0, 1]$;
- $m^m[E]: 2^\theta \rightarrow [0, 1]$; and
- $m^m[R]: 2^\theta \rightarrow [0, 1]$.

The residual risks are computed in the following:

- $m^r[D]= m^s[D] \oplus m^m[D]$;
- $m^r[I]= m^s[I] \oplus m^m[I]$;
- $m^r[P]= m^s[P] \oplus m^m[P]$;
- $mr[E]= ms[E] \oplus mm[E]$; and
- $mr[R]= ms[R] \oplus mm[R]$.

The corporate security residual risk is computed in the following:

$$- \quad m^r = m^r[D] \oplus m^r[I] \oplus m^r[P] \oplus m^r[E] \oplus m^r[R].$$

That is, the system residual risk is expressed in the following:

- $m^r: 2^\theta \rightarrow [0, 1]$;
- $m^r(A) = m^r[D] \oplus m^r[I] \oplus m^r[P] \oplus m^r[E] \oplus m^r[R]$ (A); and
- $m^r(\theta) = 1 - m^r(A)$.

The response decision is illustrated in the decision tree furnished in Figure 3. The study assumes that risk owners at financial firms have their own private models that they apply in estimating financial recovery costs (R) and their own reservation values for their real losses (D), in the scenario of a given financial incident induced by an Internet-based attack.
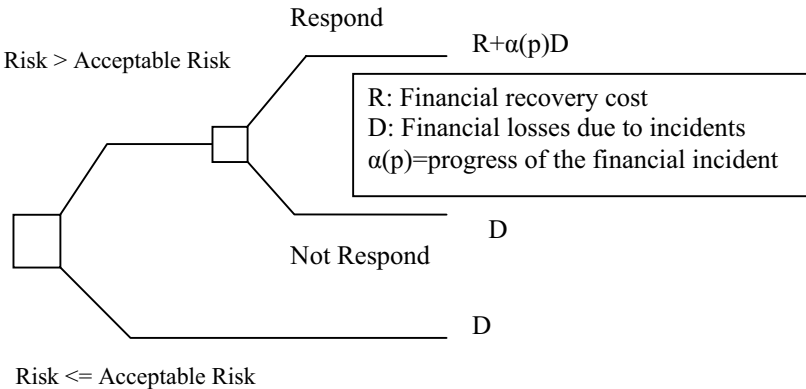
Figure 4:   Responses to financial incidents based on risk.

5   Conclusion

This paper posited a new analytical model for the generation of a risk-driven financial assurance program capable of preventing, detecting, and responding to financial incidents (FAPG) for a general support system. The study reviewed a single general support system with known basic risk, integrating ids evidence and meta-evidence obtained from financial management, in order to estimate current financial risk positions. The study showed the functioning of the FAPG, by generating a risk-driven financial assurance program, for a small general support system in a financial firm. This study was limited to financial incidents induced by Internet-based attacks but introduced a framework for further innovation and research, which will be of interest to chief security officers in the financial services industry.

References

[1]   Ludovic, M. & Cedric, M., Intrusion detection: a bibliography. *Technical Report* SSIR-2001-01, SUPELEC, France, September, 2001.

[2]   Denning, D., *Information Warfare and Security*, Addison Wesley: Reading, MA, 1999.

[3]   Lippmann, R. et al., Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation. *Proceedings of the 2000 DARPA Information Survivability Conference and Exposition*, January, 2000.

[4]   Kendall, K., A database of computer attacks for the evaluation of intrusion detection systems. *Thesis, Master of Engineering in Electrical Engineering and Computer Science*, Massachusetts Institute of Technology, Boston, June, 1999.

[5]   Smets, P. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9, pp. 1-35, 1993.

[6]   Barrus, J., Intrusion detection in real time in a multi-node, multi-host environment. *Thesis, Master of Science*, Naval Postgraduate School, Monterey, CA, September, 1997.

[7]   Porras P. & Neumann, P.G., EMERALD: event monitoring enabling responses to anomalous live disturbances. *Proceedings of the 20th National Information Systems Security Conference*, National Institute of Standards and Technology, October, pp. 353-365, 1997.

[8]   Ryon, L.E., A method for classifying attack implementation based upon its primary objective. *Thesis, Master of Science*, Iowa State University, Ames, Iowa, 2004.

[9]   Smets, P. & Kennes, R., The transferable belief model. *Artificial Intelligence*, 66, pp. 191-234, 1994.

[10]   Lindqvist, U. & Jonsson, E., How to systematically classify computer security intrusions. *Proceedings of the IEEE Symposium on Research in Security and Privacy*, Oakland, CA, May, 1997.

This page intentionally left blank

An innovative interdisciplinary curriculum in financial computing for the financial services industry

A. Joseph & D. Anderson
Pace University, USA

Abstract

Finance is a fast growing field in business and is among the fastest growing in scientific computing, helping to sustain economies that include those of New York City and of the United States. The dynamics of finance have enticed computer scientists, engineers, mathematicians, and physicists. This has helped in the growth of interdisciplinary fields that involve computational finance, financial computing, financial engineering, mathematical finance, and quantitative finance. While most of these interdisciplinary programs are introduced to graduate students at universities, few of them are introduced to the undergraduate students. The frequent model that includes a computer science minor and a financial major requires a finance student to be in a general computer science minor that is open to all students who satisfy the minimum requirements for the minor. This interdisciplinary model does not serve sufficiently the needs of industry and of society. The study introduces an interdisciplinary major/minor curriculum model that seamlessly integrates computer science into finance through free elective credits. The model is that of financial computing that is both discipline and industry oriented in the university. The paper of the study evaluates the financial computing model, indicating how it conforms to the needs of financial firms in industry and of society and that of the international Basel II Capital Accord. This study will benefit educators and researchers in integrating a special and timely curriculum model helpful to the financial services industry.
Keywords: assessment, curriculum models, disciplinary grounding, finance, financial computing and interdisciplinary curriculum.

1 Background

At the core of information technology is computer science, which has transformed an industrial society to an information society, with a knowledge-based economy where information is a commodity and its efficient processing by a financial firm can lead to a competitive advantage. Its impact is likely to affect finance and economics immeasurably (Tsang and Serafin [1]). The financial services industry was one of the first in the civilian sector of the global economy to computerize business. Financial institutions, such as Bear Stearns in the United States, prefer to recruit graduates of computer science, finance, accounting, or some related disciplines. Finance is likely the fastest growing field in business and is among the fastest growing areas in scientific computing (Haugh and Lo [2]). The dynamic nature of finance and the challenging problems inherent within it have attracted many professionals, including computer scientists, engineers, mathematicians, and physicists [1, 2, 10]. This attraction to modern finance has resulted in the growth of many vibrant and related interdisciplinary fields that involve finance. Examples include computational finance, econophysics, financial computing, financial engineering, mathematical finance, and quantitative finance. While most of these interdisciplinary programs are offered at the graduate level, a small but increasing number are offered at the baccalaureate level. Study identified less than 20 such programs internationally.

Bransford et al. [3] reported three major findings about learning that are based on research and "that can beneficially be incorporated into practice."

1. Students come to the classroom with preconceptions about how the world works. If their initial understanding is not engaged, they may fail to grasp the new concepts and information that are taught, or they may learn them for purposes of a test, but revert to their preconceptions outside the classroom.
2. To develop competence in an area of inquiry, students must: (a) have a deep foundation of factual knowledge, (b) understand facts and ideas in…a contextual framework, and (c) organize knowledge in [methods] that facilitate retrieval and application.
3. A "metacognitive" approach to instruction can help students learn to take control of their own learning by defining learning goals and monitoring their progress in achieving them.

They further emphasized that learning transfer from one context to another is critical to understanding and that ultimately the learner needs to be able to transfer learning from the academic setting to the "[daily] setting of home, community, and the [office]." They suggested that schools need to become collaborative and teamwork oriented, rely on tools for problem solving, and promote contextualized reasoning conditioned on abstract logic. Moreover, they outlined that learning transfer is influenced by the following factors: degree of mastery of the original subject, context, relationships between the learnt and

tested material, learners' active involvement in the learning process, frequent and timely performance feedback, learners' self-awareness of their level of learning and assessment of it, ability to build on previous knowledge, ability to understand conceptual change, and cultural practices. Mathison and Freeman [4] reported that the main goal of the many recent developments in interdisciplinary learning is aimed at helping students attain a sufficiently deep understanding of the concepts, so that they can make the necessary connections to their daily lives. They further referenced research that found "forming connections between fields of knowledge is an essential educational need for success in the 21st century". This view is also supported by Mansilla [6].

Interdisciplinarity is not a clearly defined concept. This is evidenced in the Different definitions furnished by researchers [4–6, 8, 14, 15]. Some of these definitions assume names that depend on the level and the method that two or more disciplines are combined. They include crossdisciplinary, multidisciplinary, pluridisciplinary, transdisciplinary, metadisciplinary, integrated, and integrative [4, 5, 14]. Nissani [5], who indicated interdisciplinarity as a multidimensional fluid continuum, furnished the following practical definition: "bringing together in some fashion distinctive components of two or more disciplines." To support his definition, he outlined four types of interdisciplinarity: knowledge, research, education, and theory. He further stated that:

> At any given historical point, the interdisciplinary richness of any two exemplars of knowledge, research, and education can be compared by weighing four variables: the number of disciplines involved, the distance between them, the novelty and creativity involved in combining the disciplinary elements, and the degree of integration.

He expounded on degree of integration by saying that meaningful "integration must satisfy the condition of coherence: the blending of elements is not random, but helps to endow knowledge, research, or instruction with meaningful connections and greater unity." However, he acknowledged that the ranking of interdisciplinary richness is not a measure of quality. Mansilla [6] addressed the issue of quality in an interdisciplinary learning environment, through interdisciplinary understanding and informed assessment of students' performance. She defined interdisciplinary understanding as follows: "the capacity to integrate knowledge and modes of thinking drawn from two or more disciplines to produce a cognitive advancement." Examples of cognitive advancement include creative problem solving and product creation using the knowledge and skills from more than one discipline. Within this definition, the disciplines maintained their distinctive features and their interaction at the boundaries are leveraged to obtain the desired solution. The four main premises supporting this definition of interdisciplinary understanding are the following: performance – accurately and flexibly applying learnt concepts to novel situations; disciplinary grounding – being deeply informed by disciplinary expertise; integration and leverage – blending disciplinary views; and purposefulness. Mansilla [6] further provided the framework for assessing a student's performance that is consistent with her definition of interdisciplinary

understanding. The assessment criteria included disciplinary grounding; integrative leverage; and a critical stance where there is clarity of purpose, reflectivity, and self-critique. Quality interdisciplinary student work must be able to withstand critique when it is evaluated "against its goals." Ivanitskaya et al. [7] indicated that "repeated exposure to interdisciplinary thought [helps] learners to develop more advanced epistemological beliefs, enhanced critical thinking ability and metacognitive skills, and an understanding of the relations among perspectives in different disciplines." Bradbeer [8] said that interdisciplinary study is not easy to achieve because of the problems of functioning in different disciplines and synthesizing disciplines. He indicated that these problems resulted from differences in disciplinary epistemologies, discourses, and traditions of teaching and learning, as well as differences in students' learning styles and techniques. He indicated that helping students to become self-aware active learners was a critical step in enabling them to function across and within different disciplines. Furthermore, he indicated that disciplinary epistemologies, discourses, and traditions of teaching and learning were supportive evidence of disciplines being structures of both knowledge and cultures. He noted that although knowledge construction in a discipline may be unique, learning the knowledge is not: epistemology and culture are separable issues in teaching. Bradbeer [8] indicated that students' learning styles was a factor in their choice of a discipline. However, his investigations of Kolb on learning style and forms of knowledge, and his investigation of research on the Myers Briggs personality types, indicated the possibility of students successfully studying academic disciplines that do not necessarily match their preferred attributes. He further noted that teachers' concepts and practices of teaching and learning are also a hindrance to interdisciplinary learning in higher education. Bradbeer [8] noted research implying that most teachers' idea of teaching is information transfer. This mode of teaching does not promote deep learning.

From research of undergraduate interdisciplinary curricula that combine computer science and finance, the study introduces three basic models: university wide, discipline oriented, and industry oriented. The university wide model involves the finance major taking a computer science minor open to all students in the university. The discipline oriented model may use the major/minor principle of the university wide model with the minor specifically designed to meet the needs of the finance major. The industry oriented model integrates finance with computer science to meet the financial industry need for new graduates. The university wide model is the most common.

Study introduces an interdisciplinary major/minor curriculum model that seamlessly integrates computer science into finance through its free elective credits. It is called financial computing. This model is both discipline and industry oriented. This curriculum is unique and innovative: its capstone course purposefully, theoretically, and experientially integrates finance and computer science where students perform "real world" financial problem solutions under the mentorship of industry experts and entrepreneurs. Study contrasts this model with the existing ones and indicates how it meets the needs of students, industry, and society.

2   Introduction

Education's major economic roles include the public good of knowledge production and the private good of status enhancement (Appold [9]). These two roles "coincide when the skills taught are needed for the performance of tasks…and easily measurable". Objective in the financial computing curriculum is to satisfy the needs of the student and the public. It is expected that there will be challenges. These may include students' engagement in surface learning and faculty preference for information transfer as their main mode of teaching. Teachers will be encouraged to use teaching techniques, procedures, and examinations that promote active and reflective learning to furnish students with tools for recognizing and interpreting concepts within and between disciplinary frameworks. Teachers will also be encouraged to reflect on their teaching, challenge students' learning styles, and help students become self-aware learners. This should facilitate both intradisciplinary and interdisciplinary learning and promote an efficient learning process. This efficiency will make students more functional in the knowledge-based economy, where they can easily access, manipulate, and interpret units of knowledge (or data) in a novel manner and within different contexts so as to generate greater understanding or new knowledge.

The central objective of integrating finance with computer science is to improve the learning of finance students in the context of computing to meet the needs of the student and the public. Many of the problems in modern finance are currently being tackled using the tools of scientific computing as found in physics, engineering, mathematics, and computer science. Some of these problems include the dynamic portfolio optimization problem (Haugh and Lo [2]) and risk management for large portfolios. At the same time, some of the basic concepts in economics with implications in finance are being re-examined using very large financial datasets, advanced algorithms, complex models, and the processing power of computers (Tsang and Serafin [1]). Two examples are rationality and the efficient market hypothesis. The financial industry needs employees with a good foundation in mathematics and computer science and a "strong interest in finance and financial markets" for positions in quantitative modelling and analysis, risk management, and portfolio management (IAFE [10]). Moreover, with the Basel II Capital Accord scheduled for implementation within the next two years (Basel Committee on Banking Supervision [11]), it is anticipated that there will be an increased demand for technically trained graduates with finance backgrounds, especially in the areas of risk management and quantitative modelling. This accord is likely to have a special impact on New York City, the nation's financial center and the central location of Pace University.

Although there is a demand in the financial industry for technically trained finance graduates with strong mathematical and computing skills, the typical finance graduate is inexperienced in computer programming languages, such as C/C++ and Java. In the financial computing curriculum introduced in this study, finance students will learn to program in Java, thereby developing their

programming skills. In addition, they will develop their analytical, quantitative, interpersonal, collaborative, communication, and critical thinking skills as well as an entrepreneurial mindset. The computer science component of the financial computing option of finance will be a 13-credit minor. In fact, students will take eight credits of programming, four credits of data structures and algorithms, and four credits of financial computing. Therefore, the financial computing program will deliver graduates who are likely to help the New York City financial community meet its challenges by hiring local technically trained finance graduates.

3 Curriculum design methodology

Study discloses three basic models of undergraduate computer science and finance integration. They are called the university wide model, the discipline oriented model, and the industry oriented model. The university wide model combines the finance major with a computer science minor where the computer science minor is generic, open to all students within the university that meet its minimum requirements, and tends to favour students with strong mathematical or engineering background. This model may serve the needs of the student, but it does not necessarily serve the needs of industry and the rest of the public. Examples of this model can be found at New York University, Stevens Institute of Technology, and Duke University. The discipline oriented model may or may not use the major/minor principle of the university wide model. If a computer science minor is used by a university, it is specifically designed to meet the needs of the finance major or a group of majors that include finance. Otherwise, the computer science courses are seamlessly integrated into the finance or hybrid finance curriculum. An example of this model is found at Western Michigan University, where finance students take the general computer science minor tailored to non-science students. Most examples of the discipline oriented model are of the integrated nature – integrating mathematics with finance and leveraging it with computer science. These programs tend to target students with strong quantitative skills and adequate computer programming capabilities. Rice University's computational finance minor and University of Michigan's mathematics of finance and risk management are examples. The industry oriented model purposefully integrates finance with computer science usually in a single curriculum without a minor component, and it targets the need in the financial industry for graduates with strong quantitative and computing skills as well as very good business related skills. Three examples of this model can be found in the financial computing curricula at Northwest Missouri State University and Brunel University, as well as in the computational/quantitative finance program at the National University of Singapore. The university wide model is the most common while the discipline oriented is the least because it is an emerging model. A problem with the major/minor component of the university wide and the discipline oriented models is that they do not necessarily simultaneously satisfy the needs of the student and the public. In the university wide model, it is difficult to achieve cognitive advancement, because the

disciplines are combined in a simplistic sense – they are set side by side and usually with no attempt made to assess for interdisciplinary understanding. On the other hand, the industry oriented model may serve the need of the public, but it does not necessarily serve the need of the student, since the student may only be motivated by external forces to respond to public demand.

Our university's original bachelor of business administration (BBA) degree program, with a major in finance and minor in computer science, is an example of the university wide model. It consisted of 60 university core credits, 33 business core credits, 16 finance credits, 17 or more computer science credits, and 6 credits of auxiliary economics courses. The 13 free electives in the finance program were subsumed in the 17 computer science credits, which were generic university wide computer science minor courses. Since the finance program was 128 credits without the computer science minor and at least 132 credits with it, and the computer science minor was generic – open to the university wide population, there was a disincentive for finance majors to take the computer science minor. The proposed curriculum option is a redesign of the original one, because it replaces eight credits of computer science courses that have additional prerequisite requirements with a 4-credit project based financial computing course. Moreover, it reduces the computer science minor to 13 credits, which is the same as the number of free electives, indicated in Figure 1. Therefore, the finance degree program now becomes 128 credits, with or without the computer science minor. This updated computer science minor for finance majors constitutes a financial computing option of the finance degree program. It consists of courses in high level programming languages, such as Java, data structures and algorithms, and financial computing. These courses will provide the finance major practitioner-level skills in the four functional areas of computer science: algorithmic thinking, representation, programming, and design (Denning [12]).

The financial computing course will be the capstone course for the minor; its objectives include the following:

1. Students will acquire a fundamental understanding of the key scientific concepts and mathematical tools used in modern finance to develop and implement financial models that describe financial situations.
2. Students will gain practical understanding of planning, designing, and developing reasonably scaled financial software products.
3. Students will understand the role of creative thinking and innovation in new business creation, gain experience in business plan development, and acquire tools needed for an entrepreneurial mindset.
4. Through participation in software project development teams, students will develop team-building, social, and organizational skills that they can further develop in other classes and in their professional careers.

The course's main components are computing, finance, and experiential entrepreneurship.

It will use financial and business experts, computing professionals, and entrepreneurs to mentor and guide students in their project choice and project development. In financial computing, students will leverage their knowledge and
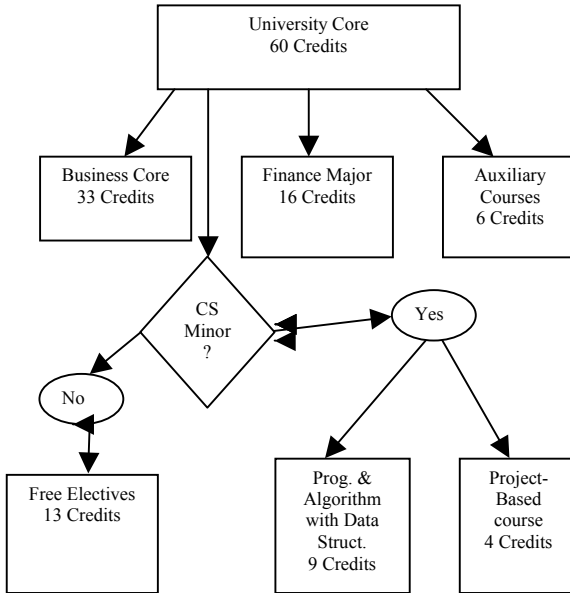
Figure 1: 128-credit finance major with computer science minor.

skills of finance, computer science, and experiential entrepreneurship to develop a creative and innovative financial software product. Students will receive frequent and timely feedback on their performance and progress. The courses in the financial computing option will be taught using a combination of lecture, discussion, cooperative/collaborative learning, problem solving, and project and laboratory instruction, in order to actively involve students' in knowledge generation and skill development. Faculty shall train students in teamwork skills, while leveraging their learning styles to improve understanding and furnish students with the tools to become reflective learners. The assessment of the computer science courses will include written and oral examinations, peer evaluations, portfolios, journals, project demonstration and evaluations, computer program evaluation, project documentation, and project reports. These assessments should illustrate that the students have attained an interdisciplinary understanding: show disciplinary grounding in finance, computer science, and experiential entrepreneurship; show integration of these three disciplines and their use to the advantage of each other; and show knowledge of the capabilities, limitations, and implications of their projects.

4 Implications

Today's employers need employees who are business minded and computer literate. IAFE [10] reported that a growing number of financial firms have recognized that computing and mathematical skills are essential for business

success, and therefore increased their recruitment of students with financial computing related degrees. In a 2001 National Association of Colleges and Employers survey to determine the qualities employers seek most in applicants, the leading ones cited were written and oral communication, honesty/integrity, teamwork skills, interpersonal skills, motivation/initiative, strong work ethic, analytical skills, flexibility/adaptability, computer skills, and self-confidence (Joint Task Force of Computing Curricula [13]). Moreover, leading financial institutions, such as JPMorgan Chase and Bear Stearns, prefer entrepreneurial recruits with strong analytical, quantitative, and communication skills. In today's financial business environment, computing technology support systems are needed to manipulate and process huge volumes of data and to effectively simulate financial situations.

In the proposed curriculum students will integrate financial theory and principles and computing and mathematical science theories and techniques with their knowledge of experiential entrepreneurship and financial products to design and develop creative and innovative financial products for a targeted sector of the financial industry. The knowledge, skills, and mindset developed in this curriculum are those needed to develop and grow in today's financial and supporting information technology systems firms. In addition, the curriculum will prepare finance students for graduate studies in financial computing, where most other students' undergraduate background is in computer science, physics, mathematics, and engineering (IAFE [10]). Furthermore, interdisciplinary major minor curriculum mode of this study integrates computer science with finance into a financial computing curriculum that combines elements of the discipline and industry oriented models.

This integration furnishes the finance major/computer science minor curriculum with a unique characteristic among curriculum models: its capstone course, financial computing, purposefully, theoretically, and experientially integrates finance with computer science and leverages the synthesis with experiential entrepreneurship to obtain an industry orientation. Thus, the model is designed to enable students to achieve cognitive advancement at the boundaries of finance and computer science and maintains its academic focus through its discipline orientation.

5   Conclusion

The financial computing curriculum of the study is likely to offer finance students an excellent foundation in interdisciplinary thinking and understanding; a strong foundation in programming, basic principles of software engineering, and the fundamentals of data structures and algorithms; and solid grounding in teamwork, collaboration, social, and communication skills. It offers these students privileged knowledge in experiential entrepreneurship from industry experts. Thus, the proposed financial computing curriculum model of this study is likely to furnish entrepreneurial graduates who are able to help the New York City financial community meet its impending demand for strong computing, quantitative, analytical, and teamwork skills.

References

[1]   Tsang, E. & Serafin, M., Computational finance. *IEEE Computational Intelligence Society*, August, pp. 8-13, 2004.
[2]   Challenges in Financial Computing; Haugh, M. & Lo, A. http://web.mit.edu/alo/www/Papers/haugh2.pdf.
[3]   Bransford, J., Brown, A., & Cocking, R., (eds). *How People Learn: Brain, Mind, Experience, and School*, National Academies Press: Washington, D.C., pp. 3-66, 1999.
[4]   The Logic of Interdisciplinary Studies; Mathison, S. & Freeman, M., ERIC Document Reproduction Service No. ED418434. http://www.eric.edu.gov/.
[5]   Nissani, M., Fruits, salads, and smoothies: a working definition of interdisciplinarity. *Journal of Educational Thought*, 29(3), pp. 121-128, 1995.
[6]   Mansilla, V., Assessing student work at disciplinary crossroads. *Change*, January/February, pp. 14-21, 2000.
[7]   Invanitskaya, L., Clark, D., Montgomery, G., & Primeau, R., Interdisciplinary learning: process and outcomes. *Innovative Higher Education*, 27(2), pp. 95-111, 2002.
[8]   Bradbeer, J., Barriers to interdisciplinarity: disciplinary discourses and student learning. *Journal of Geography in Higher Education*, 23(3), pp. 381-396, 1999.
[9]   Appold, S., Competing to improve? A difficult terrain. *Proc. of the 1st Int. Conf. on Teaching and Learning in Higher Education*, eds. D. Pan, C. Wang, & K. Mohanan, National University of Singapore: Singapore, pp. 264-269, 2004.
[10]   Frequently Asked Questions; International Association of Financial Engineers (IAFE). http://www.iafe.org/?id=faq.
[11]   International Convergence of Capital Measurement and Capital Standards: A Revised Framework; Basel Committee on Banking Supervision, Bank for International Settlements, Press & Communications, CH-4002 Basel. http://www.federalreserve.gov/boarddocs/press/bcreg/2004/20040626/attachment.pdf.
[12]   Computer Science: The Discipline; Denning, P. http://www.idi.ntnu.emner/dif8916/denning.pdf.
[13]   Computing Curricula 2001 Computer Science Volume Final Report; Joint Task Force on Computing Curricula, IEEE Computer Society & Association for Computing Machinery (ACM). http://sigcse.org/cc2001/cc2001.pdf.
[14]   Jacobs, H., (ed). *Interdisciplinary Curriculum: Design and Implementation*, Association for Supervision and Curriculum Development (ASCD): Alexandria, VA, 1989.
[15]   Loepp, F., Models of curriculum integration. *Journal of Technology Studies: A Refereed Publication of Epsilon Pi Tau*, 25(2), pp. 21-25, 1999.

Critical success factors in planning for Web services in the financial services industry

H. Howell-Barber[1] & J. Lawler[2]
[1]Executive Advisory Board of Ivan G. Seidenberg School of Computers, Science and Information Systems, Pace University, USA
[2]Information Systems Department, Pace University, USA

Abstract

As increasingly more firms in the financial services industry expand their use of Web services, and as others begin to adopt services, understanding the planning requirements for this technology becomes increasingly critical for business managers and technologists. This study explores a generic methodology of a Web services plan that can be used to accelerate accurate project planning, helping to avoid project planning becoming a major project in itself. This study identifies critical success factors that contribute effectively to the planning success of Web services projects in the financial services industry. The study furnishes a methodology model for the features of such a plan, by identifying components that can be reused and refined safely for a small inter-departmental project, a medium cross-departmental project, and a large inter-firm project. Business and methodological factors are indicated to be more important than technological factors in the success of the projects, though technology is reviewed in the study, and implications include planning recommendations, as they relate to Web services. This study will benefit management practitioners, researchers and technologists in the successful planning for Web services in the financial services industry.
Keywords: BPEL4WS, business process, project plan, service description, service-oriented architecture, SOA, UDDL, Web services, XML and WSDL.

1   Background

A Web service communicates using Simple Object Access Protocol (SOAP) messages over HyperText Transfer Protocol (HTTP). Services are published and

described in a Universal Description, Discovery and Integration (UDDI) registry, using Web Services Descriptor Language (WSDL). Clients search the registry for services using SOAP messages. Messages that cross firewalls may be secured with Web Services Security (WS-S). Web services implementations parallel the client/server paradigm (Ma [1]), except that components use text-based Extensible Markup Language (XML) and share open standards and cross-vendor support. SOAP is from the World Wide Web Consortium (W3C), and WSDL, UDDI, and WS-S are from the Organization for the Advancement of Structured Information Standards (OASIS), whose members include all the major software vendors. HTTP, the de facto standard for connecting servers on the Web, is defined by RFC 1945 of the Internet Engineering Task Force (IETF).

A Service Oriented Architecture (SOA) provides loosely coupled services that expose business functions with published discoverable interfaces (Adams et al. [2]). An enterprise SOA leverages business logic and data from existing systems to support flexibility and adaptability of changing business environments of systems. SOA services map to business entities, allowing enterprise integration on the business level (Krafzig et al. [3]). Web services may be implemented as the first step to SOA; however, it is possible to have an SOA without Web services. The additional layer of abstraction in Web services allows authorized users access to information on heterogeneous native platforms. Services, discovered in legacy applications or created anew, may be combined into new services, using Business Process Execution Language for Web services (BPEL4WS), also from OASIS.

Businesses are being pushed to explore SOA architectures to avoid missing competitive advantages, while vendors race to produce or upgrade products that support these specifications. The additional layer of abstraction, new technology, and limited timeframes make planning for Web Services and SOA simultaneously more critical and more complicated. This study explores techniques for handling the added complexity, by highlighting tasks specific to Web services projects, and providing recommendations for using prior project experience to facilitate scheduling activities. The suggestions are vendor-neutral.

2   Introduction

The manager responsible for Web services project initiation must ensure that the business leads the project. Business participation is the most important factor in the success of an SOA (Lawler et al. [4]). Sponsors (business, customer, and technology) will be identified. Advisory groups with representatives from business, customer and technology areas will be established for the project. Stakeholders (individuals not directly involved in the project, but who can affect the outcome of the project) will also be identified by the groups. Regular advisory group and stakeholder meetings will be scheduled by the groups. An exercise to assess the organization's tolerance for change must be completed with the assistance of the advisory group. Remedial actions may be taken if the organization is change-averse.

2.1   SOA governance

A governance board (with business, customer and technology participants) will guide the SOA implementation across a range of Web services projects. It will (a) maintain communication among SOA participants and stakeholders, (b) establish service access rules, (c) define business goals and performance indicators, (d) define an approach for modeling business processes, (e) establish service quality reviews, (f) document assumptions included in service requirements, (g) promote the cultural changes required for SOA success, (j) establish a process for business component identification, (k) establish a process for service prioritization, and (l) establish processes for lifecycle management and versioning (Bieberstein et al. [5]).

3   Methodology model

It is important to think big, but start small. It is useful to identify an entire business segment that can benefit from SOA, but the pilot project should deliver a few Web services in six months or less (Knorr and Rist [6]). Each pilot activity will lay the foundation for advancement toward the implementation of a true SOA. When the first deliverables successfully address an obvious business problem, they help to ensure approval and funding of larger projects.

3.1   Small (pilot) project (intra-departmental)

The pilot project (3–6 months) will address a critical business requirement, while ascertaining a technology and planning approach. For example, it could combine digital images of signed trust documents with customer data in a banking operations area. The seemingly disproportionate number of management tasks in the project plan furnishes a framework for future projects.

3.2   Medium project (inter-departmental)

A medium-sized project (6 months to a year) could involve rolling out a set of processes across operations areas in the same firm. An example is if the same trust documents were made available to the compliance monitoring area along with additional services that provide historical transaction activity. The SOA governance team will begin to exercise its mandate. The plan for the medium project will include additional coordination, requirements gathering (including the creation of a UDDI registry), and technology tasks. A carefully maintained project history will assist future (larger) projects.

3.3   Large project (parent firm and select subsidiaries)

A large project (a year or longer) will lead to the creation of a full SOA. This project could provide an expanded set of processes to selected subsidiaries of the financial services firm. The SOA governance role must be fully activated. While considering all tasks in the extensive list of activity details at

www.hbink.com/webservicesplan, project communications, service security and cross-platform compatibility will be critical. A Web services glossary is furnished for further study. Prior project history, lessons learned, and best practices from earlier projects will be critical.

4   Analysis of model

Table 1 furnishes a high level outline of Web services plan activities. The Group Sequence column suggests an order for the initiation of planning activities. Activities with the same Group Sequence may begin in parallel. Groups 2–9 occur throughout the project lifecycle. Activities ramp up at the beginning of requirements, analysis and design, and development/implementation. Testing requirements are refined during analysis and design and performed during implementation.

Table 1:   Web services planning activity groups.

Activity Group	Group Sequence	Activity Group	Group Sequence
Methodology	0	Requirements	10.a
Project Initiation	1	Security	10.b
Project Process	2	Testing[**]	10.c
Project Communications	3.a	Project Change Management	11
Project Planning	3.b	Analysis and Design	12.b
Role Assignment and Confirmation	3.c [*]	Architecture	12.a
Risk Management	4	Development/Implementation	13
Best Practices	5	Deployment	14.a
Problem Management	6	Management and Monitoring	14.b
Procurement Management	7 [*]	Project History	15
Human Resources	8 [*]	Sunset	16
Training	9 [*]		

[*] Activity ramps up at the beginning of requirements, analysis and design, and development / implementation.
[**] Refined during design; performed during implementation.

4.1   Methodology

Methodology factors are important to the success of an SOA project (Lawler et al. [4]). Including service orientation into project management assumes selection of an established methodology that will be enhanced to include service-related tasks. Complexity and changing business requirements will require iterative development. For the strategically important SOA, the methodology will tend toward the side of the heavyweight processes (Krafzig et al. [3]).

4.1.1   Project initiation
Project initiation must include creation and syndication of a strong business case. Business opportunities and potential benefits will be analyzed, prioritized and used as input to statements of scope, objectives, and goals. The initial

cost/benefits analysis will be performed (with clearly documented assumptions). Sign-off and funding authority for the overall project and major project checkpoints (initiation, requirements, analysis and design, testing, and deployment) will be identified by the project director. Sign-off at each major checkpoint will be required and will include funding for the subsequent activities.

After sign-off of initiation, an experienced manager for the remainder of the project will be assigned by senior managers. Her first activity will be to create a vision statement from the scope, objectives, and goals documents. A kick-off presentation will include the vision statement and, optionally, a prototype to help demonstrate proposed project deliverables.

4.1.2   Project process
The project process must include requirements management, project planning, tracking, oversight, quality assurance, and configuration management, in order to produce repeatable results (Level 2 in the software Capability Maturity Model (CMM)) (Leffingwell and Widrig [7]). Quality is important as mistakes will require changes to a larger number of project artifacts. Therefore, the requirements and analysis and design activities should take at least 60% of the project effort. As projects move toward true SOA, plans will include defined process features (CMM Level 3), such as cross-organization process, a formal training program, integrated software management, product engineering, inter-group coordination, and peer reviews. Measurement and monitoring of the process itself (CMM Level 4) will support inclusion of successful process features into future projects. Establishing a naming standard for project artifacts helps organize the main sections of the project and enables easy referencing by team members and clients throughout the project. An adaptable process with appropriate enforcement mechanisms will help to ensure that the project processes themselves are as non-intrusive as possible (Goto and Colter [8]).

4.1.3   Project communications and project information center
If all software builds took no time, and development was perfect, the limiting factor in project success would probably be communications (Doar [9]). Communication is facilitated by building a standard terminology used across the business and technical communities (Bieberstein et al. [5]). Thus, a communications plan and processes will be established early in the activity sequence. A Project Information Center will furnish a foundation for the communications plan and processes. This single (virtual) source will have secured, role-based access, with an index and pointers to (a) all project communications, (b) project process, (c) project plans and planning archives, (d) startup, requirements, analysis, design and architecture documents, (e) a project glossary with industry specific XML schema (e.g., FinXML, FIXML), and a taxonomy of business terminology specific to the firm and technology terms specific to the project, (f) contact information, (g) development artifacts and version information (checked in daily), (h) project samples, (i) status reports, (j) issues tracking and problem management files, (k) risk management

documents, and (l) change control procedures. Maintenance procedures will include daily backups and the ability to achieve full recovery in half a day.

4.1.4 Project planning

The project manager will create a plan to track a greater number of tasks and corresponding roles (Table 2), even on small projects. The work breakdown structure (WBS) will include a larger number of artifacts to be developed and tracked by her. Greater involvement of the customer community requires coordination of individual activities outside the manager's official organization. Task scheduling will challenge both the manager and activity owners to furnish estimations for tasks they have never performed or observed, using new software, and in new environments. Complex or unfamiliar activities will need higher priority and may be scheduled ahead of their usual sequence to give team participants more time to resolve unexpected problems. Smaller deliverables will furnish beneficial results in a shorter timeframe. For example, deliverables may be scheduled weekly for small projects, every two weeks for medium projects, and monthly for larger projects. The project manager will help himself and future project managers by maintaining a detailed planning archive. Each plan change and reasons for that change will be recorded for future reference and problem avoidance. In successive projects, the project manager may find estimation assistance in his own plans or in the plans of his predecessors. The plan will be well syndicated, with multiple copies in a highly visible location to promote awareness and compliance.

4.1.5 Role assignment and confirmation

Table 2 lists the roles associated with a Web services project [10]. Most traditional roles will be expanded, several roles will be added, and user roles modified to take advantage of new services. Responsibilities of each role will be defined, assigned, and confirmed by the manager. Training requirements will be identified by the manager.

Table 2:   Web services / SOA project roles.

Architect (*)	Project administrator (*)
Business analyst (*)	Project manager (*)
Business testers (*)	Security specialist (*)
Change process manager (*)	Service developer (+)
Configuration manager (*)	Service modeler (+)
Database administrator (*)	Services librarian/governor (+)
Deployment team (*)	SOA architect (+)
Developers (*)	Systems administrator (*)
Facilitators (*)	Technical writer
Governance team (+)	Test manager (*)
Interoperability tester (+)	Tool administration (*)
Knowledge transfer facilitator (*)	Toolsmith (*)
Legacy adaptation specialist (+)	UDDI administrator (+)
Network administrator (*)	User Roles (≠)
Process flow designer (Optional) (+)	Vendor interface (*)

* Expanded Roles
\+ New Roles
≠ Modified Roles

4.1.6   Risk management

Risk management receives greater emphasis because of exceptional challenges associated with objective business process extraction, resistance to change, information gaps within and among business and technology communities, implementation of new vendor products, new and updated industry standards, staffing requirements, environmental complexity, and absence of cooperation across business silos. Risks and possible countermeasures will be listed, evaluated, prioritized, and tracked by management.

4.1.7   Best (and worst) practices - patterns and antipatterns

Patterns are collections of best practices from the combined experiences of industry specialists. Patterns for e-Business are a group of reusable assets that can help speed the process of developing Web-based applications. They include business, integration, application, runtime, and combination patterns [11]. Since more than 80% of projects fail or run over budget, antipatterns are an even greater area to be mined for problem avoidance techniques (Ang et al. [12]). They prevent problems by identifying frequently recurring errors.

4.1.8   Problem management

Though problem management is normally associated with testing, it receives its own heading because problems that occur outside the testing activity must also be managed by management. A process including problem capture, evaluation, categorization, prioritization, resolution, and reporting will be implemented and enforced by management. A history of problems and their resolutions will be maintained to help future teams avoid similar problems.

4.1.9   Procurement management

A structured procurement process will ameliorate risks by helping ensure that vendors and products adequately support project objectives. Vendor selection criteria will include staff quality, responsiveness, short-term support capabilities, and the probability that they will be able to support future requirements. Product selection will include processes for installation and rigorous in-house testing before purchase agreements are signed by management. If specialist consultants are required on the project, there must be clearly stated performance objectives and willingness to make adjustments if the objectives are not met by the consultants.

4.1.10 Human resources

These activities include identification of skills requirements, skills assessments, and identification of training needs. If there is no time to train existing staff, it may be necessary to hire additional staff (from approved vendors), including contract developers or specialist consultants. If this is the case, new staff orientation procedures must be in place (to help them "hit the ground running"). Orientation will include physical access rules, development environment access, equipment, telephone (if functioning on premises), connection capabilities (if not functioning on premises), personnel introductions, business introductions, firm

orientation, and a technology overview. Support and counseling will help ensure optimal performance.

4.1.11 Training

The leaders of the SOA revolution will be business personnel who understand technology and technology personnel who understand the business (McKean [13]). Therefore, appropriate training and cross-training are critical in both business and technology spheres. Training of business analysts to identify and model necessary business services is the next item of importance in the success to an SOA, followed by appropriate technology training.

4.1.12 Requirements

Requirements for a Web services or SOA project must be more clearly defined than for prior implementations. The business and customers should drive the activity. Business analysts must help to determine which services (spanning organizational boundaries) need priority and which processes will be included in the services (including the possibility of combining processes from multiple applications within the same service), determine what data will be included in the services, and (most critically) how the data will be named in the system. Defining the business meaning of transactions and data is the most intractable issue that systems managers face in the system (Sleeper [14]). Proportionately more time will be spent in requirements gathering and specification of services, in order to ensure that business participants agree among themselves on terminology, scope, goals, and priorities. Service Level Agreements (SLA) will be included in the requirements. Acceptance criteria will be clearly defined by management.

4.1.13 Security

Though important in financial services, security will have an extra layer of complication when users from different business groups begin to access the same services. Though identification, authentication, authorization, integrity, confidentiality, auditability, and non-repudiation should be considered as part of requirements, Web service tools may fail to support all requirements (Van de Putte et al. [15]). Because Web Services Security (WS-S–April, 2004) and Security Assertions Markup Language (SAML–March, 2005) are new, team members will start early evaluations to ensure that security products provide the appropriate level of security, while conforming to industry standards. Though WS-S indicates how to use previously existing security technologies within the Web services environment, achieving the right mixture of features takes significant time and effort (Newcomer and Lomow [16]). Finally, to avoid delays in development, testing and deployment, role-based access to project artifacts (old and new), services and services components will be defined by management.

4.1.14 Testing

Testing requirements and test cases should be identified along with requirements for processes and services. The test plan will be finalized during analysis and

design, and will start to be implemented during development and implementation. Though this sequence for testing requirements is not new for Web services, it is important that testing activities occur early and often. Business functionality should be as clear as possible before testing of Web services begins for the system.

4.1.15 Project change management

Appropriate attention to requirements, analysis and design will help decrease the need for changes. However, a major justification for SOA implementation is rapid support for inevitable changes in the business environment. Therefore, there must be a process for gathering, reviewing, prioritizing, and signing-off to project changes that is as rigorous as the requirements and design efforts. A clearly defined process for version control will be followed by management. Change management for underlying legacy components will be included in the process. Changes at the business and process level will be controlled by the governance team. Business participants will be aware of the effect changes will have on project timetables and budget before they sign off on changes.

4.1.16 Analysis and design

As with the requirements, business participation will be more critical than before in non-SOA projects. Analysis of existing applications will identify candidate functions and data. Redundant functions and data will be flagged as candidates for sunset activities after the successful project completion. Where possible, industry schemas will be used by management. Because the WSDL is the contract between the developer of the service and the user of the service, it is important to design the WSDL first before developing the service. The framework for Web services management should also be included in the design.

4.1.17 Architecture

The technology facing members of the team will begin evaluating existing environments against requirements as soon as a first draft of the requirements is available. The architect will furnish feedback to the requirements team regarding what is feasible, given the state of the technology, and will ultimately recommend an environment that will address the finalized requirements. After approvals and corresponding funding, she will partner with the procurement team to acquire, install, and test product upgrades or new products. To prepare for success as the organization moves toward a complete SOA, scalability will be included in the recommendation.

4.1.18 Development/implementation

Development will take a smaller proportion of project effort than with normal projects because of the increased efforts in requirements, analysis, and design. Coding standards are a good idea. At a minimum, team members should agree to code formatting guidelines and artifact naming standards. Development of project artifacts will occur within the framework of an agreed upon methodology that will include continuous integration (regular builds of the software, along with build tests). Strict version control of all artifacts within and across

environments will begin with the first build. SOA-specific activities, such as legacy adaptations, as in service description and registry, should occur in parallel with the usual programming efforts to allow time for discovering and solving problems with the new technology. Error handling will be implemented and a list of error messages compiled by management.

4.1.19 Deployment

While final testing is being conducted, a rollout schedule addressing client orientation and role upgrades will be created, reviewed, and signed-off by management. A roll-out plan with necessary fallback steps will be created by management. Documentation will include a deployment diagram, deployment checklists, release documentation, system administration and general operations requirements (including recovery and failover plans).

4.1.20 Management and monitoring

Monitoring includes logging, tracing messages, security enforcement, and quality of service tracking (as specified in the requirements). Monitoring software will be evaluated and implemented during the testing cycles. Production monitoring will begin with the first deployment, using metrics and report layouts created in parallel with service design. Service level agreements specified in the requirements will be monitored by management. Potential effects on legacy systems will be reviewed by management.

4.1.21 Project history as a reusable asset

The plan will include project evaluation, project turnover, and process improvement (with critical input from the post implementation report). Documenting project history can help to develop better estimates and save planning time by leveraging templates from past projects [17]. This requires a strategy for recording project information across the team.

4.1.22 Sunset

A plan for eliminating duplicated systems, functions, data and overlapping projects, as discovered during analysis and design, will be created and reviewed by management. Redundancies will be eliminated in the process, as successively more services are implemented by management.

5   Implications

Immediate implications of this study include *business benefits*. Successful Service Oriented Architecture (SOA) will benefit from planned communications between business and technology departments that cooperate as partners. Business personnel will develop adequate knowledge of technology themes that will help in the implementation of intelligent changes. Technology personnel will have adequate knowledge of business topics that will help in improvements that are both technically and financially feasible. Training will be included to maintain the necessary knowledge. As a result, change-averse firms will become

capable of further flexibility, as they discern benefits from changes of a Service Oriented Architecture.

Implications include *increased capability and maturity* of the information technology department. Departments that already have a bona fide methodology will expand and improve techniques, in order to manage increased complexity of a Service Oriented Architecture. Departments that do not have a methodology will institute one, in order to manage the projects. Planning will be critical in a methodology. Departments that institute change management processes will have probable and resultant successful Service Oriented Architecture projects.

Final implication of the study includes the *criticality of initiating pilot projects* in Service Oriented Architecture. Departments in information technology of firms have been successful in basic Web services projects and, in the main, have been developing advanced Service Oriented Architecture projects. Such projects furnish a foundation for practitioner and scholarly studies of potential benefits to firms that have not introduced the latter projects. Standards may be learned from best of class practitioner consultants and vendors that have helped firms in Service Oriented Architecture development and implementation of systems. Study indicates competitive advantage for fast follower and first mover firms that invest in the Service Oriented Architecture soon.

6   Conclusion

Appropriate planning will emphasize leadership from the business community. A sequence of plans, with each plan furnishing input to subsequent plans, will facilitate the implementation of Web services and migration to a full Service Oriented Architecture. Plans for medium and large-size developments will inherit successful sections from previous plans, while avoiding the problems discovered in earlier planning. Emphasis on elimination of typical project failure points will allow time for careful investigation and implementation of new development paradigms.

References

[1]   Ma, K. J., Web services: What's real and what's not. *IT Pro*, March / April, p. 15, 2005.
[2]   Adams, H., Gisolfi, D., Snell, J., & Varadan, R., Best practices for Web services. *IBM developerWorks,* 1 November, 2002.
[3]   Krafzig, D., Banke, K. & Slama, D., *Enterprise SOA: Service - Oriented Architecture Best Practices,* Pearson PTR: Upper Saddle River, New Jersey, Online, 2004.
[4]   Lawler, J., Anderson, D., Howell-Barber, H., Hill, J., Javed, N., & Li, Z., A study of Web services strategy in the financial services industry. *Information Systems Education Journal*, 3(3), pp. 1–25, 2005.
[5]   Bieberstein, N., Bose, S., Fiammante, M., Jones, K. & Shah, R., *Service - Oriented Architecture Compass: Business Value, Planning, and*

Enterprise Roadmap, IBM Press (Pearson plc): Upper Saddle River, New Jersey, Online, 2006.

[6] Knorr, E. & Rist, O., 10 steps to SOA. *Infoworld,* 7 November, p. 24, 2005.

[7] Leffingwell, Dean & Widrig, D., *Managing Software Requirements: A Unified Approach,* Addison Wesley: Boston, p. 475, 2003.

[8] Goto, K. & Colter, E., *Workflow that Works, Second Edition,* New Rider's Press: Indianapolis, Indiana, Online, 2004.

[9] Doar, M. B., *Practical Development Environments*, O'Reilly Media, Inc.: Sebastopol, California, Online, 2005.

[10] Web Services Project Roles, *IBM developerWorks*, Online www - 128.ibm.com/developerworks/webservices/library/ws - roles/, 2004.

[11] IBM Patterns for e-Business, *IBM developerWorks,* http://www-128.ibm.com/developerworks/patterns/, 2004.

[12] Ang, J., Cherbakov, L., & Ibrahim, M., SOA anti-patterns. *IBM developerWorks,* November, 2005.

[13] McKean, K., Business-ification of IT. *Infoworld*, 23 May p. 8, 2005.

[14] Sleeper, B., The SOA puzzle: five missing pieces. *Infoworld,* 13 September, p. 42-51, 2004.

[15] Van de Putte, G., Jana, J., Keen, M., Kondepudi, S., Mascarenhas, R., Satish, O., Rudrof, D., Sullivan, K., & Withinbank, P., *Using Web Services for Business Integration*, IBM Redbooks: San Jose, California, p. 33, 2004.

[16] Newcomer, E. & Lomow, G., *Understanding SOA with Web Services,* Addison Wesley Professional: Boston, Online, 2004.

[17] Work Essentials for Project Managers: Using Historical Data to Improve Future Projects, http://office.microsoft.com/en-us/FX012217241033.aspx, 2004.

Section 2
Advanced computing
and simulation

This page intentionally left blank

Integrated equity applications after Sarbanes–Oxley

O. Criner[1] & E. Kindred[2]
[1]*Department of Computer Science, Texas Southern University, USA*
[2]*Software Engineer, USA*

Abstract

Primary among the requirements of the Sarbanes–Oxley legislation are that chief executive officers and chief financial officers certify the accuracy of their corporations' financial statements. This act spawned a thrust to complete total accounting systems with end-to-end financial audit capabilities. The federal government's use of XML and XBRL will eventually be extended to require that all public companies file all forms and reports with them using XML or XBRL. The Securities and Exchange Commission (SEC) is currently accepting XBRL filings from corporations on a voluntary basis. The potential improvement and analytical capability offered by this new environment requires the planning and implementation of new software for computational science research. This paper discusses the technological convergence that allows the implementation of systems that more accurately and rapidly monitor the performance of public companies through their SEC filings and news events.
Keywords: Sarbanes–Oxley, XBRL, XML, computational modelling, accounting data integrity, financial forensic analysis.

1   Background

The Sarbanes–Oxley Act was passed by the United States Congress in the aftermath of several corporate scandals involving large public corporations during and after 2001. This research topic became of interest to the senior author because of his involvement as a juror in one of the high profile trials of that time. Several questions arose during that trial concerning the accuracy of accounting information, accounting procedures that thwarted transparency, the integrity of corporate financial data, and the uses and limitations of mathematical and

computational tools in finance. As evident by the unusually large number of corporations restating their financial results, there must have been a widespread practice of manipulation of records or "cooking of the books". The obvious intent of this manipulation of reporting was to affect equity market prices, since the compensation of many executives was directly connected to the price of the stock. Almost all of the firms involved in the scandals were audited by major public accounting firms several of which were also found culpable in the affairs, because of co-optation of the audit process by a consulting relationship with the audit client. Further complications were caused by the implicit conflict of interest that existed between the equities research analysts and the investment bankers involved in business transactions with the corrupt or failing companies and may have been enablers of the corrupt practices. The regulatory agencies of the Federal and State governments were officially unaware of the crisis in the making although some regulators sensed an approaching economic problem. [1].

In an era of ubiquitous anytime computing, the question of why these companies and their questionable business practices were not identified by the securities police, the Securities and Exchange Commission (SEC), remains. The answer lies in both the inability of the SEC to effectively monitor these companies by timely analysis of the thousands of reports and in the islands of automation that exists throughout most of the business world, specifically the separation of operational from financial accounting systems. This situation brings into question the integrity of all public financial data and, in particular, the prices of publicly traded instruments.
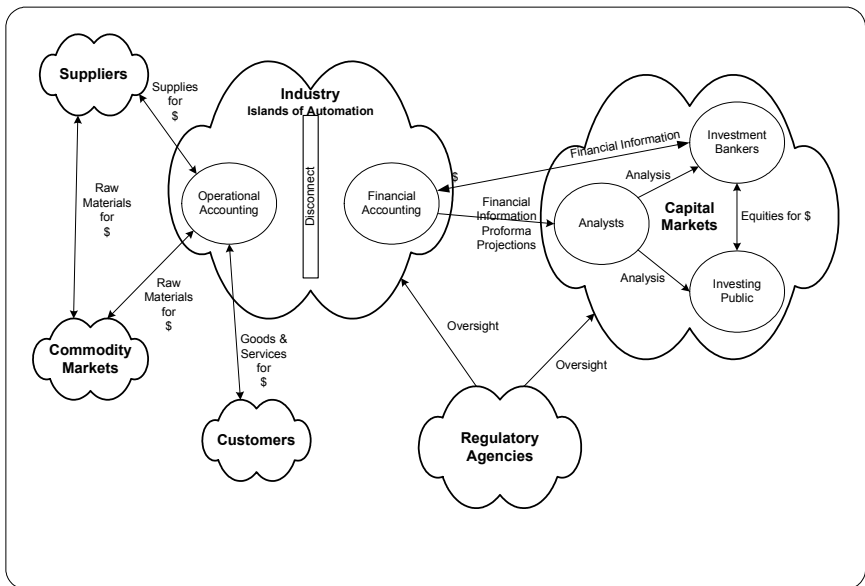
Figure 1:   Islands of accounting automation.

The dichotomy between the operational accounting systems and the financial accounting systems shown in Figure 1 creates the greatest problem of integrity and accuracy in financial data for most firms. Sarbanes–Oxley is intended to ensure that investors and stakeholders have accurate financial information upon which to base financial decisions. The act required that chief executive officers and chief financial officers certify the accuracy of a company's financial statements. It provides for severe penalties for knowingly and wilfully misstating financial statements. Satisfaction of these requirements of the law requires that companies institute new controls and data integration between the two islands. Since there is rarely integration connecting the two, most companies rely upon manual processes (with personal productivity tools) to produce the accounting reports. "There have not been major expenditures for new systems since the Y2K effort, so one can only assume that these data and integration problems have existed for some time." [2].

Since computational finance is predicated upon the assumption that good reliable financial data is availability, the entire research enterprise is threatened if that is not the case. Therefore, it should be of great interest to computational finance practitioners to know the sources and processes of the data generated by so many companies in so many variations of the accounting process, which finally ends in the earnings per share value and other parameters used to specify corporate performance [3].
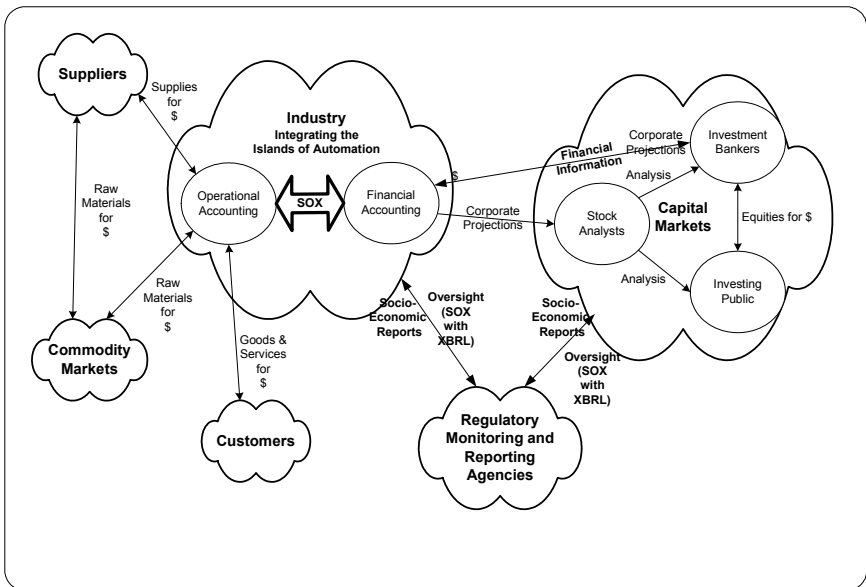
Figure 2:   Islands of automation with a SOX bridge.

Some companies appear to have integrated their operational and financial accounting and information systems. This was the competitive edge strategically

deployed that enabled Wal-Mart to capture so much of the retail market. The bridge between the two islands is motivated by the SOX legislation. But the magnitude of the SOX problem has some companies lobbying to have some relief from the requirements.

The application of scientific methods to business processes requires that the source of the information be somewhat accurate and standard. The facts are that the results of the accounting and auditing processes are approximate and far from precise. Trust in the financial reporting system is fundamental to the capital market system. In order to ensure the trust in the system, there needs to be much more accountability in the monitoring system. This implies that financial reports should be examined differently utilizing the technology of the time. Companies are required to redefine their operating processes so that auditors can assess the effectiveness of their internal controls. Companies must define seam-less systems that integrate and preserve the audit trail for the thousands of processes and the millions of transactions they generate that affect the financial statement

This effort has developed more slowly than we had anticipated in 2002, however, in the near future automated agents will be utilized to mine report databases and examine all corporate reports filed with regulatory bodies [4, 5, 6]. This technology must be incorporated into the analytical capability of the investing public in order for the public to be able to evaluate the viability of a firm for investment. The Extended Business Reporting Language (XBRL), a derivative of XML will be the required format for corporate reporting in the near future. The SEC is required to monitor public corporations to ensure that the investing public is not defrauded, a task at which the agency has not been very effective and has been prevented from being so in large part by the business lobby.

Web Services provide the architectural framework for new integrated applications in financial information. By utilizing the new XBRL language and the infrastructure of XML, it is possible to integrate equities analysis in a totally new framework. When this research was planned, the writers assumed that the business community would embrace XBRL and XML as widely as the federal government. Unfortunately, this is not yet the case and many companies seem to view the Sarbanes–Oxley requirements as an unnecessary expense rather than an opportunity to make a commitment to be a full participant in the world of e-commerce.

2 Computational modelling of financial markets

This paper discusses a computational approach that integrates the financial reporting with the analyses of price time histories with the objective of identifying the signatures of corporate events. By identifying the signatures of corporate events in the data derived from the market, it may be possible to classify the response to such events and to assess their effect upon the prices in the market in a deterministic manner. It is well known that various events affect the price of equity issues and futures prices, e.g., announcements of various government indices, earnings announcements, bankruptcies, and credit rating

changes. Traders and investors are known to wait for the reports before taking some action. We want to take the output of the system and work backwards to determine its cause. This is an inverse problem in dynamic systems [7].

Computational modelling is used in many fields where there are not sufficient data and theory; it is an application of logic, mathematics, computational techniques, and heuristics. Computational scientists usually consider very large complex problems that usually do not yield to a complete mathematical analysis, fit neatly into a theory, or can be examined in a laboratory. The problems considered by computational scientists are not amenable to the traditional scientific method of observation, theory and experimentation. Indeed the usual data that one needs for a well-posed problem generally does not exist nor do many of the equations or inferential schemes.

The "direct" or "forward" approach to problems in science is the situation where there is a "complete description" of a physical system within the confines of some logical system, which provides the rules of inference sufficient to derive additional true statements in the logical system, which correspond to the prediction of some observed events. In most cases, the logical system is expressed in mathematics. However, mathematics is not the only implementation of a logical system with which to study complex phenomena. Computational modelling extends the mathematical analyses beyond the so-called well-posed problems or it may be a completely heuristic set of processes. In inverse problems the issue is to use "the actual results of some measurement to infer the values of the parameters that characterize the system" [7]. In computational finance those measurements include the publicly reported financial data, which is why there is concern as to its accuracy and integrity.

We generalize the logical system in computational modelling to be comprised of five components: (1) Definitions are descriptions of the objects under consideration, (2) Assumptions are true statements that are known about the objects and taken a priori, (3) Rules of inference that describe the process of taking the definitions and assumptions as inputs and concluding a new true statement as output from the process, (4) The collections of theorems or true statements that are derivable from the definitions and assumptions using the rules of inference and the collection of derived true statements, and (5) associative relationships or alternative paths through the inferential process to obtain the same true statements.

Computational finance is a specialization of computational science, in the sense that scientific computing or the solution or investigation of scientific problems is done using computers. Computational finance is the application of computational science to problems and issues of financial systems. Computational science is third modality of knowledge determination inextricably co-equal with experimentation and observation, logical inference and theoretical analyses. In this sense, computational finance is not finance. Corporate finance provides one of the datasets for computational finance, but the computational models of financial systems are much more general.

The XBRL Taxonomies [6] effectively partition the set of corporations into equivalence classes. Each company that uses the standard taxonomy for a

particular group is equivalent to other company in the class from a computational point of view. Practical applications of this technology may be a few years away, however, because of the size and effort required to construct systems that utilize Web Services and data bases implemented with XBRL. Design and development must begin early.

3 The dynamic model - Integrating the price and financial time series

In this paper we describe a computational methodology for integrating the price time history of an equity with its "fundamental" financial reports. We create a new model integrating analysis of equities that is based upon estimating the rates of change of the price time series and the affect of corporate events. This dynamic model of price could be given by an equation of the form

$$\ddot{x}_i(t) = f_i(x_i(t), \dot{x}_i(t)) \tag{1}$$

where $x_i(t), \dot{x}_i(t)$ and $\ddot{x}_i(t)$ are the price, first derivative of the price, and second derivative, respectively, of the price of the ith stock [9]. To include the financial analysis results into this model, we assume that eqn (1) can be rewritten in the form

$$\ddot{x}_i(t) = f_i(x_i(t), \dot{x}_i(t), q_i(t), k_i(t), e_i(t)) \tag{2}$$

where the additional functions $q(t), k(t)$, and $e(t)$ are to be determined using the results of the quarterly, annual, and event reports, respectively, and the so-called "analysts expectations" and news releases.

We seek to determine signatures of the various announcements and events by mining the data available in the SEC EDGAR database. This is done by examining the time series of the price in the neighbourhood of the event. While the usual reaction to announcements or events is a rapid change in the market values of the instrument, we seek to quantify that change in the derivatives of the price and to estimate the effect on the price movement. In this sense we parameterize the functions $q(t), k(t)$, and $e(t)$. The parameterization process in this methodology correlates the time and magnitude of the various events with the first or second derivatives of the price of the stock issue as suggested by the model eqn (2).

Relating these magnitudes to the estimates of $\dot{x}(t)$ and $\ddot{x}(t)$ will provide a measure of the effect. For example, the earnings estimate at time t will be paired with a value of $\dot{x}(t)$ and $\ddot{x}(t)$ to create a function relating earnings to the second derivative of the price. Averaging these measures over time will provide values of the function that are used in the decision algorithms. Figure 3 shows the basic components of the computational model and where the additional components are combined with $x(t), \dot{x}(t), \ddot{x}(t)$, the volume, 10-Q, 10-K, and 8-K reports.
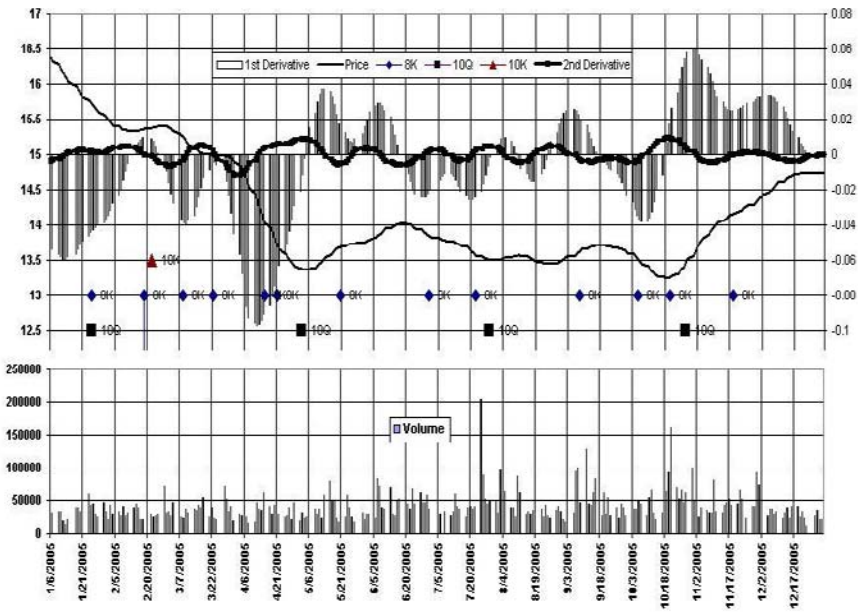
Figure 3:   Basic components of the computational model of Xerox stock.

Figure 4 shows phase diagrams for Xerox stock for three complete years, 2003, 2004, and 2005 and the first two months of 2006. The variables of a dynamic system specify its phase space. The motion of the system corresponds to a trajectory or orbit in the abstract phase space [10]. Since we do not know the specific functional relationships of each variable, we must investigate the manner in which the diagrams vary as a result of specific events that occur. Deterministic dynamical systems are characterized by their phase plane orbits. Clearly these diagrams show that our assumption that the process is a dynamical system has merit. We want to discover or synthesize some process that simulates the system in the time domain. Although the graphs show that the system is highly nonlinear, it is not clear that it is chaotic. If it is chaotic then it may be possible to find an attractor to which the process tends. Many questions arise in this case. Are there attractors for each stock or equivalence class of stocks? Are attractors time dependent or do they depend on other parameters? One issue that should be settled by the capacity to construct phase plane diagrams is that the process is deterministic. This demonstration should set the efficient market hypothesis to rest.

Many other components can be added including analysts' estimates and government reports of the essential economic indicators to model their effect on the price of major companies and industry groups. This procedure will also be helpful in forensic analyses because it will create a time series of the essential components of the company's financial statements [11, 12].
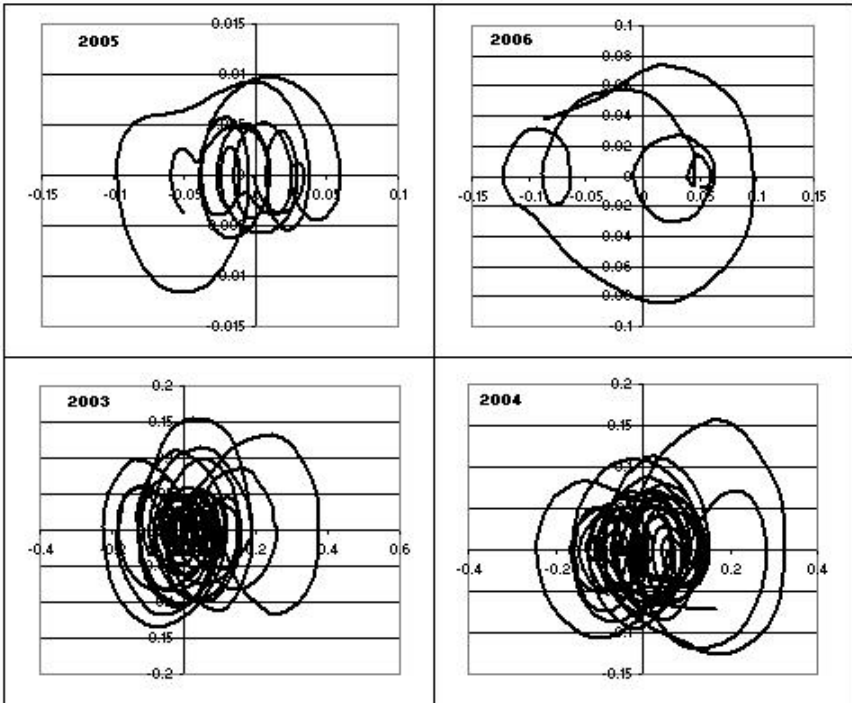
Figure 4:   Phase plane plots of Xerox stock.

4   The technological convergence

With the pervasiveness of the Internet and the availability of CPU cycles and massive storage devices, the technology is now present to implement the underlying infrastructure.   And of course, the lingua franca that binds the disparate entities of the business community is XBRL. With these technological components at our disposal, a re-examination of Figure 1 yields Figure 2 – Integrating the Islands.

Looking at most industries, there exists ample opportunity for real-time or near real-time data collection in their operations and supply chain as evidenced in retail sales by Wal-Mart and Home Depot. Transaction data can be collected from customers and suppliers for sales and inventory using point of sale (POS) technologies like barcodes and radio frequency identification (RFID). Location and condition within the supply chain can be tracked using the global positioning system (GPS), automated weighing systems for bulk supplies, and remote sensing for environmentally sensitive resources. This data is then fed to the operational accounting system, which hopefully optimizes its operating practices and minimizes its operating costs with statistical improvement methods like Six Sigma.

The transaction, supply, and operations data are then transferred to the financial accounting system, which optimizes its budgetary, capitalization, and investment activities based on "known" corporate requirements. The condition of the company can now be reported internally to the corporate executives and externally to industry analysts and investment bankers in the Capital Markets, to the Investing Public, and to the various Regulatory Agencies. The numbers will be traceable and have meaning and fulfil the theoretical goal of accounting – to represent the operations of the business. The technologies for this market communication are the industry-specific taxonomies implemented in XBRL.

Once these reports are collected by the Regulatory Agencies in a format that can be data-mined, the monitoring function can be automated utilizing a collection of techniques suggested in this paper. Data-mining results can be used for highlighting potential problems and by the Capital Markets and the Investing Public for investment decisions. Results can be disseminated by Web Service-based applications. By no means is the widespread implementation of these technologies trivial but, with an evolutionary approach, financial information will have real meaning and integrity will return to our markets.

On top of this highly networked infrastructure lies a plethora of computationally intensive techniques:

- Data-mining – to draw relationships between quantified data
- Red Flag Analysis – to identify stellar or troubled companies and industries
- Natural Language Processing – to quantify prose reports
- Chaos – to graphically represent interpretations of complex datasets
- Heuristics – to build systems incorporating expert domain knowledge
- Grid Computing – to provide the computational capacity on the desktop or across the enterprise

5 The grand challenge of computational finance

The scientist always asks, "How good is the data with which I am working?" This is not a statistical question but a question about the process of measurement. How is this data being created? Regardless of the sophistication of my analytical tools, the old computer science dictum "garbage-in-garbage-out" still applies.

Secondly, it is now possible, or will be in the near future, to analyze every formally traded financial instrument (stocks, bonds, futures, options and other formally traded derivatives), in the light of the publicly available socio-economic data to determine causal relationships among them, for example:

- Land and water use and commodity production
- Non-renewable resources and population growth
- Environmental preservation and economic growth

It is imperative that all computational science be conducted in an environment of high quality data. The ubiquity of these data from the media, the web, and the press assaults the senses and cries out for computational solutions.

References

[1]   Levitt, A. & Dwyer, P., *Take on the Street, What Wall Street and Corporate America Don't Want You to Know*; *What You Can Do to Fight Back*, Pantheon: New York, 2002.

[2]   Get Ready for an Increased Number of Financial Restatements, Visage Solutions Web Site, http://www.visagesolutions.com/

[3]   Berenson, A., *The Number: How the Drive for Quarterly Earnings Corrupted Wall Street and Corporate America*, Random House, New York, 2003.

[4]   Bovee, M., Kogan, A., Nelson, K., Srivastava, R. & Vasarhelyi, M., Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL), *Journal of Information Systems*, 19(1), pp. 19-41, 2005.

[5]   Lawler, J., et al, A Study of Web Services Strategy in the Financial Services Industry, *Proc. ISECON*, v21, 2004.

[6]   Leinnemann, C, Schlottmann, F., Seese, D., & Stuempert, T., Automatic Extraction and Analysis of Financial Data from the EDGAR Database, *Proc. 2nd Annual Conference on World Wide Web Application*, Johannesburg, 2000.

[7]   Tarantola, A., *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial and Applied Mathematics, Philadelphia, 2005.

[8]   XBRL International Web Site, www.xbrl.org

[9]   Criner, O., Optimal control strategies for portfolios of managed futures, *Computational Finance and Its Applications*, WIT Press, Southampton, UK, pp. 189, 2004.

[10] Saaty, T., & Bram, J., *Nonlinear Mathematics*, Dover Publications, New York, Chapter 4, 1964.

[11] Looking for Trouble: The SEC Upgrades Technology to Be a Better Watchdog, http://www.WallStreetandTech.com/showArticle.jhtml? articleID=41600001

[12] Apostolou, N., & Crumbley, D., Forensic Investing: Red Flags, http://www.bus.lsu.edu/accounting/faculty/napostolou/forensic.html

C++ techniques for high performance financial modelling

Q. Liu
School of Management,
University of Electronic Science and Technology of China,
Chengdu, Sichuan, People's Republic of China

Abstract

In this paper, several C++ techniques, such as eliminating temporary objects, swapping vectors, utilizing the Matrix Template Library (MTL), and computing at compile-time, are shown to be highly effective when applied to the design of high performance financial models. Primarily, the idea emphasized is to achieve high performance numerical computations by delaying certain evaluations and eliminating many compiler-generated temporary objects. In addition, the unique features of the C++ language, namely function and class templates, are applied to move certain run-time testing into the compiling phase and to decrease the memory usage and speed up performance. As an example, those techniques are used in implementing finite difference methods for pricing convertible bonds; the resulted code turns out to be really efficient.
Keywords: C++, high performance, financial modelling, C++ template, Matrix Template Library, vector swapping, compile-time computation, convertible bond.

1   Introduction

What do Adobe Systems, Amazon.com, Bloomberg, Google, Microsoft Windows OS and Office applications, SAP's database, and Sun's HotSpot Java Virtual Machine have in common? They are all written in C++ (Stroustrup [1]). Still, when people talk about high performance numerical computations, Fortran seems to be the de facto standard language of choice.

To the author's knowledge, C++ is actually widely used by Wall Street financial houses; as an example Morgan Stanley is mentioned by Stroustrup [1] on his website. Techniques developed in the past few years, such as expression

template (Furnish [2]) and compile-time computation or meta-arithmetic (Alexandrescu [3]), has made C++ a strong candidate for high performance numerical computations.

In this article I discuss four aspects of C++, namely trying to get rid of unnecessary temporary objects, swapping vectors for objects re-use, taking advantage of the performance gain provided by the Matrix Template Library [4], and doing compile-time computations, which are used in combination to achieve high performance numerical computation for financial modelling. Sample codes throughout this paper are taken directly from the library of a real-world convertible bond pricing model implementing finite difference methods.

2   Watching for temporary objects

C++ programs use quite a few temporary objects, many of which are not explicitly created by programmers (Stroustrup [5], Meyers [6], and Sutter [7]). Those temporary objects will drag down the performance tremendously if not eliminated. A few examples will make this point clear.

A typical step in the pricing process, or commonly known as diffusion on Wall Street, takes a list of stock prices and a list of bond prices, which are probably represented as vectors of doubles in C++ (or vector<double>) as in the following code (with some parameters omitted for simplicity), and returns a list of new bond prices:

```
typedef vector<double> VecDbl;

VecDbl diffuse(VecDbl stocks_in, VecDbl bonds_in) {
      VecDbl bonds_out;
      …
      return bonds_out;
}
```

What is wrong with this simple, innocent piece of code? Use too many unnecessary temporary objects!

Let's analyze this carefully. First of all, the list of stock and bond prices are passed into the function by-value, as is commonly known in C++. When a function is called, a temporary copy of the object that passes by-value is created by the compiler. In the above code, two temporary objects, one for the list of stock prices and the other for the bond prices, are created (and then destroyed when the function returns). In a typical situation, the list of stock prices may have a length of a few hundred, so it is expensive (in terms of computing time) to create and destroy such a temporary object.

Further, inside the function, a local object of type vector<double> is used to store the values of the new bond prices temporarily. Finally, for the function to return a vector<double> object, one more object may have to be created by the copy constructor, if the function is used as in the following code:

```
VecDbl my_vd;
…
my_vd = diffuse(stocks_in, bonds_in);
```

Note that the additional object created here could be eliminated by doing the function call and the object instantiation in a single step:

```
VecDbl my_vd( diffuse(stocks_in, bonds_in) );
```

Therefore, depending on how the function is used, one may force C++ to construct yet another object! As a result, this simple function creates at least three unnecessary yet expensive objects, which can hardly be efficient.

To fix the problems in the code, we pass function arguments by-reference or by-pointer. Note that normally the list of stock prices is not changed through out the whole diffusion, but the prices of the bond are modified at every step (so the list of bond prices is used as both input and output as in the following):

```
void diffuse(const VecDbl & stocks_in, VecDbl & bonds_io) {
        VecDbl bonds_local; // for implicit finite difference method
        …
}
```

Because no temporary object needs to be created when function arguments are passed by-reference, no temporary object is created in the modified code above. Let's say that a typical diffusion takes about a thousand steps, so a total of about two thousand objects of vector<double> is eliminated by this simple modification!

For the explicit finite difference method (Hull [8]), even the local object inside the function can be eliminated by the following trick:

```
while (iter != last) {                  // last == end() -1
        val_plus = *iter_next++;     // value of next element

        *iter++ = Up * val_plus + Mid * val + Down * val_minus;

        val_minus = val;                // value of previous element
        val = val_plus;                 // value of current element
}
```

Note that two iterators, one points to the current and the other to the next element, are used to keep track of the elements in the vector. Therefore, a further savings of a thousand objects is achieved.

To most C++ programmers, the above is probably obvious. C++ does have more subtle surprises for us in term of temporary objects. Look at the following standard piece of code seeing in many textbooks:

```
for (int k = 0; k < N; k++) {…}
```

Could one see any problem?

Temporary object, of course! It may not be obvious, but the postfix increment operator actually creates an unused temporary object. Thus, the prefix increment operator shall be used here instead, which does not create a temporary object. The savings in this peculiar case is probably negligible, but any performance conscious coder should take home the point.

As a rule of thumb, prefix increment is preferred over postfix increment; unary operator, such as +=, is preferred over its binary counterpart, +, whenever possible. Those may not seem to be any big deal, but in order to achieve high performance numerical computation, one has to pay special attentions to those numerical operators. This point will become even more prominent in the following sections.

3 Re-using vector objects by swapping

Typically, a two-dimensional array of size 200x1000 (roughly the number of price points times the number of steps) for derivatives prices is used in finite difference methods (Clewlow and Strickland [9]). In another word, there is equivalently one individual vector<double> object for each step of diffusion. Normally we are only interested, however, in the final price slide at the valuation date. Therefore, is the two-dimensional array necessary?

Not at all. Since each step of diffusion involves only two neighbouring states, two vector<double> objects are actually enough:

```
for (int step = 1; step <= 1000; ++step) {
        std::swap(bonds, prev_bonds);
        diffuseOneStep(…, prev_bonds, bonds);
}
```

Note that by swapping and re-using the two objects, a two-dimensional array is no longer necessary. Swapping of two vector<double> objects can be very efficiently implemented (Stroustrup [5]). Not only the construction of almost a thousand more objects is avoided, but also the resource required for the code is much lighter (run-time resource for two objects instead of for a thousand objects).

4 Using the Matrix Template Library (MTL)

The Matrix Template Library is a free, high performance numerical C++ library maintained currently by the Open Systems Laboratory at Indiana University (MTL website [4]). MTL is based extensively on the modern idea of generic programming ([4] and Stroustrup [5]) and designed using the same approaches as the well-known Standard Template Library (STL). It is interesting to know that as MTL has demonstrated that "C++ can provide performance on par with

Fortran" [4], but it may still be surprising to some that "There are even some applications where the presence of higher-level abstractions can allow significantly higher performance than Fortran" [4].

As a library for linear algebra operations, MTL offers extensive algorithms and utility functions. Only one example of using MTL for financial modelling will be shown here to make the point, however. The following line of code is taken almost directly from the convertible bond model mentioned in the Introduction (with slight modifications to simplify the presentation):

```
mtl::add(mtl::scaled(mtl::scaled(stocks, cr), df), bonds);    //y += x
```

where cr and df are scalar variables, and the variables stocks and bonds are of type mtl::dense1D<double> (similar to vector<double>) as provided by MTL.

What the single line does is this: multiply every stock price in the vector by cr, then multiply the results by df, and finally add the results to bonds. Without using MTL algorithms, at least three loops would be necessary if the operators for addition, multiplication, and assignment were defined conventionally. This would be expensive, for it is well-known that it is optimal to perform more operations in one loop iteration (Dowd and Severance [10]). Further, more loops also mean many more temporary objects needed to be created to store the intermediate results of the arithmetic operations, which will slow down the computation even more (Furnish [2]). One could of course hand-code the one loop that does all the operations in one shot, but that misses the point here, since in so doing, which is ugly and error-prone, we lose the beauty of writing simple, arithmetic-like code.

MTL, however, does all the operations in one loop. Let's now see how MTL achieves this incredible feat. The function mtl::scaled prepares a multiplication of a vector by a scalar, but does not actually execute the multiplication. Then the result is scaled once more by another mtl::scaled. Again the multiplication is not executed. Finally mtl::add does two multiplications and one addition in one loop (for each element in the vector). Further note that the mtl::add here utilizes the unary operator += instead of the conventional binary operator + and then assignment operator; as a result, the temporary object needed by operator + is avoided.

5   Compile-time computation

Loosely speaking, compile-time computation is also known as static polymorphism, meta-programming, or meta-arithmetic, made possible by the C++ template mechanism. High performance is achieved by moving certain computation from run-time to compile-time, delaying certain computation or eliminating unnecessary temporary objects (Furnish [2] and Alexandrescu [3]). Further performance enhancement can be gained by coupling meta-programming with the C++ inline facility and the so-called lightweight object optimization.

What one can do with meta-programming is only limited by one's imagination, as Alexandrescu has aptly demonstrated in his excellent book [3]. Again one very simple example will be shown here just to make the point.

Convertible bonds are complicated financial contracts with many parameters. To pass all those parameters to the pricing code, a map with keys and values as strings are used. The values could actually be int, double, string, or some other types. The code has to convert all the values stored in strings to their proper type efficiently. How could this be done?

One could of course use a series of if-test's to determine the various types at run-time. That is not efficient, however. Or one could handle each value individually, but that is not elegant and error-prone. C++ meta-programming in fact enables us to do better and do something as the following:

```
ReturnType val_lv; // ReturnType can be int, double, etc.
findParam(key_in, params_in, val_lv);
```

Where given a return type, the program will choose the right function to use at compile-time. The findParam functions are explicitly defined for each possible return type as in the following fashion:

```
typedef map<string,string> StrPair;

template<class OutType>        // template function
void findParam(const string & key_, const StrPair & map_, OutType &
val_out ) {}

template<> void findParam(const string & key_, const StrPair & map_, int
& val_out ) {              // specialize int type
    ParamFinderImpl<int>::findParam(stoi, key_,map_, val_out);
}

template<> void findParam(const string & key_, const StrPair & map_,
double & val_out ) {              // specialize double type
    ParamFinderImpl<double>::findParam(stof, key_,map_, val_out);
}
```

Note stoi converts a string to an int, while stof to a double. Here the template function specialization, or template<> (Stroustrup [5]), is utilized. Further, since the template parameter in findParam<int>, for example, can be deduced from the type of the relevant function argument, the <int> does not have to be specified when to be either defined or called.

As a result, the client code for using findParam is very simple and uniform. More importantly, since choosing the right version of findParam's is done at compile-time according to the return types specified by the client, the program is more efficient. Furthermore, some of the functions could be inlined to improve the performance additionally.

For completeness, the definition of the class ParamFinderImpl is shown below:

```
template<class OutType>
struct ParamFinderImpl {
      typedef bool (*p2f)( const string & s, OutType & val_out );

      static void findParam(p2f func_, const string & key_, const StrPair
& map_, OutType & val_out ) {
            StrPair::const_iterator pmi;

            if ((pmi = map_.find( key_ )) != map_.end() )
                  func_( pmi->second, val_out );
      }
};
```

6   Performance estimation

The convertible bond from Ayache et al.. [11] (see Table 1 below for details) is used in the performance test. The AFV model (Ayache et al. [11]) is implemented with the Crank-Nicolson method. The diffusion is done daily; in another word, there are 1826 time-steps in the diffusion. The state variable (stock or bond price) is divided into 281 points.

Table 1: Convertible bond data used in performance estimation.

Valuation date	01/01/2005 (mm/dd/yyyy)
Maturity	01/01/2010
Conversion ratio	1
Convertible	01/01/2005 to 01/01/2010
Call price	110
Callable	01/01/2007 to 01/01/2010
Call notice period	0
Put price	105
Putable	On 01/01/2008 (one day only)
Coupon rate	8%
Coupon frequency	Semi-annual
First coupon date	07/01/2005
Par	100
Hazard rate, p	0.02
Volatility	0.2
Recovery rate, R	0.0
Partial default	η=0.0
Risk-free interest, r	0.05

The C++ code is compiled using Microsoft Visual Studio .NET 2003 with optimization flag /O2. The program is executed on a Lenovo Laptop (240 MB memory and 1500 MHz Pentium Processor).

For ten runs, the main diffusion loop takes an average of 0.244 seconds to finish. Roughly speaking, four bonds could be priced in about one second, or two hundred bonds done in less than one minute. With such high speed, traders would be able to do portfolio-based optimization in real-time. This is believed to be quite efficient.

Acknowledgement

References

[1]   Stroustrup, B., C++ Applications.
       public.research.att.com/~bs/applications.html
[2]   Furnish, G., Disambiguated glommable expression templates. *Computers in Physics*, **11(3)**, pp. 263-269, 1997.
[3]   Alexandrescu, A., *Modern C++ Design*, Addison-Wesley: Boston, 2001.
[4]   MTL, The Matrix Template Library. www.osl.iu.edu/research/mtl/
[5]   Stroustrup, B., *The C++ Programming Language*, Special ed., Addison-Wesley, 2000.
[6]   Meyers, S., *More Effective C++: 35 New Ways to Improve Your Programs and Designs*, Addison-Wesley, 1996.
[7]   Sutter, H., *Exceptional C++: 47 Engineering Puzzles, Programming Problems, and Solutions*, Addison-Wesley, 2000.
[8]   Hull, J. C., *Options, Futures, and Other Derivatives*, 5th ed., Prentice Hall: Upper Saddle River, New Jersey, 2003.
[9]   Clewlow, L. & Strickland, C., *Implementing Derivatives Models*, John Wiley & Sons: New York, 1998.
[10]   Dowd, K. & Severance, C. R., *High Performance Computing*, 2nd ed., O'Reilly & Associates: Cambridge, 1998.
[11]   Ayache, E., P., Forsyth, A. & Vetzal, K. R., The valuation of convertible bonds with credit risk. *Journal of Derivatives* **11**, pp. 9-29, 2003.

Solving nonlinear financial planning problems with 10^9 decision variables on massively parallel architectures

J. Gondzio & A. Grothey
School of Mathematics, University of Edinburgh

Abstract

Multistage stochastic programming is a popular technique to deal with uncertainty in optimization models. However, the need to adequately capture the underlying distributions leads to large problems that are usually beyond the scope of general purpose solvers. Dedicated methods exist but pose restrictions on the type of model they can be applied to. Parallelism makes these problems potentially tractable, but is generally not exploited in today's general purpose solvers.

We apply a structure-exploiting parallel primal-dual interior-point solver for linear, quadratic and nonlinear programming problems. The solver efficiently exploits the structure of these models. Its design relies on object-oriented programming principles, treating each substructure of the problem as an object carrying its own dedicated linear algebra routines. We demonstrate its effectiveness on a wide range of financial planning problems, resulting in linear, quadratic or non-linear formulations.

Also coarse grain parallelism is exploited in a generic way that is efficient on any parallel architecture from ethernet linked PCs to massively parallel computers. On a 1280-processor machine with a peak performance of 6.2 TFlops we can solve a quadratic financial planning problem exceeding 10^9 decision variables.
Keywords: asset and liability management, interior point, massive parallelism, structure exploitation.

1 Introduction

Decision making under uncertainty is an important consideration in financial planning. A promising approach to the problem is the multistage stochastic

programming version of the asset liability management model as reported in [1–4]. Its advantages include the ability to model the dynamic features of the underlying decision problem by allowing the rebalancing of the portfolio at different times as well as capturing possible dynamic effects of the asset distributions. Unfortunately realistic models tend to cause an explosion in dimensionality due to two factors: firstly the size of the problem grows exponentially with the number of portfolio rebalancing dates (or stages). Further a considerable number of realizations are required to capture the conditional distribution of asset returns with a discrete approximation. For T stages and p realizations the dimension of the resulting problem will be of order p^T.

The last decade has seen a rapid improvement of methods to solve large scale stochastic programs. However most of these are only applicable in a very special setting. Nested Benders Decomposition approaches [5, 6] are limited to LP formulations. Linear algebra approaches such as [7, 8] are usually limited to very special structures resulting for example from constraints on the allowed type of recurrence relation.

In this paper we discuss our experiences with the modern, general structure exploiting interior point implementation OOPS (Object-Oriented Parallel Solver) [9, 10]. We show that our approach makes the solution of general large nonlinear financial planning problems feasible. Furthermore it allows for fast computation of efficient frontiers and can exploit parallel computer architectures.

In the following Section 2 we state the asset liability management model that we are concerned with and present various nonlinear extensions. In Section 3 we give a brief description of the Object-Oriented Parallel Solver OOPS, while in Section 4 we report numerical results on the various problem formulations.

2 Asset liability management via stochastic programming

We are concerned with finding the optimal way of investing into assets $j = 1, \ldots, J$ over several time-periods $t = 0, \ldots, T$. The returns of the assets at each time-period are assumed to be uncertain but with a known joint distribution. An initial amount of cash b is invested at $t = 0$ and the portfolio may be rebalanced at discrete times $t = 1, \ldots, T$. The objective is to maximize the expectation of the final value of the portfolio at time $T + 1$ while minimizing the associated risk measured by the variance of the final wealth.

The uncertainty in the process is described by an *event tree*: each node of the event tree at depth t corresponds to a possible outcome of events at time t. Associated with every node i in the event tree are returns $r_{i,j}, 1 \leq j \leq J$ for each of the assets and the probability p_i of reaching this node. For every node, children of the node are chosen in such a way, that their combined probabilities and asset returns reflect the (known) joint distribution of all assets at the next time period, given the sequence of events leading to the current node. The question how to best populate the event tree to capture the characteristics of the joint distribution of asset returns is an active research area, we refer the reader to [11].

We use the following terminology: Let L_t be the set of nodes in the event tree corresponding to time stage t. L_T is the set of final nodes (leaves) and $L = \bigcup_t L_t$ the complete node set. An $i \in L$ denotes any node in the tree, with $i = 0$ corresponding to the root and $\pi(i)$ denotes the predecessor (parent) of node i. Let v_j be the value of asset j, and c_t the transaction cost. It is assumed that the value of the assets will not change throughout time and a unit of asset j can always be bought for $(1+c_t)v_j$ or sold for $(1-c_t)v_j$. A unit of asset j held in node i (coming from node $\pi(i)$) will generate extra return $r_{i,j}$. Denote by $x_{i,j}^h$ the units of asset j held at node i and by $x_{i,j}^b, x_{i,j}^s$ the transaction volume (buying, selling) of this asset at this node, respectively. Similarly $x_{t,j}^h, x_{t,j}^b, x_{t,j}^s$ are the random variables describing the holding, buying and selling of asset j at time stage t. The inventory constraints capture system dynamics: the variables (asset holdings) associated with a particular node and its parent are related

$$(1 + r_{i,j})x_{\pi(i),j}^h = x_{i,j}^h - x_{i,j}^b + x_{i,j}^s, \quad \forall i \neq 0, j. \tag{1}$$

We assume that we start with zero holding of all assets but with funds b to invest. Further we assume that one of the assets represents cash, i.e. the available funds are always fully invested. Cash balance constraints describe possible buying and selling actions within a scenario while taking transaction costs into account:

$$
\begin{aligned}
\sum_j (1 + c_t)v_j x_{i,j}^b + l_i &= \sum_j (1 - c_t)v_j x_{i,j}^s + C_i \quad \forall i \neq 0 \\
\sum_j (1 + c_t)v_j x_{0,j}^b &= b,
\end{aligned}
\tag{2}
$$

where l_i are liabilities to pay at node i and C_i are cash contributions paid at node i. Further restrictions on the investment policy such as regulatory constraints or asset mix bounds can be easily expressed in this framework.

Markowitz portfolio optimization problem [12] combines two objectives of the investor who wants to: (i) maximize the final wealth, and (ii) minimize the associated risk. The final wealth y is expressed as the expected value of the portfolio at time T converted into cash [13]

$$y = \mathbb{E}\left((1 - c_t) \sum_{j=1}^{J} v_j x_{T,j}^h\right) = (1 - c_t) \sum_{i \in L_T} p_i \sum_{j=1}^{J} v_j x_{i,j}^h. \tag{3}$$

The risk is measured with the variance of return:

$$\mathrm{Var}\left((1 - c_t) \sum_{j=1}^{J} v_j x_{T,j}^h\right) = \sum_{i \in L_T} p_i (1 - c_t)^2 [\sum_j v_j x_{i,j}^h]^2 - y^2. \tag{4}$$

These two objectives are combined into a single concave quadratic function of the following form

$$f(x) = \mathbb{E}(F) - \lambda \mathrm{Var}(F), \tag{5}$$

where F denotes the final portfolio converted into cash (3) and λ is a scalar expressing investor's attitude to risk. Thus in the classical (multistage) Markowitz model we would maximize (5) subject to constraints (1), (2) and (3).

The need to well approximate the continuous joint distribution of asset returns leads to large event trees and subsequently very large problems. These models however display a regular structure which can be exploited by our solution methodology.

2.1 Extensions of asset liability management problem

There are several disadvantages associated with the standard mean-variance formulation of the asset liability model as described in the previous section. It has been observed, for example, that the mean-variance model does not satisfy the second order stochastic dominance condition [14]. Furthermore, by using variance to measure risk, this model penalizes equally the overperformance and the underperformance of the portfolio. A portfolio manager is interested in minimizing the risk of loss hence a semi-variance (downside risk) seems to be a much better measure of risk.

To allow more flexibility for the modelling we introduce two more (nonnegative) variables s_i^+, s_i^- per scenario $i \in L_t$ as the positive and negative variation from the mean and add the constraint

$$(1 - c_t) \sum_{j=1}^{J} v_j x_{i,j}^h + s_i^+ - s_i^- = y, \quad i \in L_T \qquad (6)$$

to the model. The variance can be expressed as

$$\mathrm{Var}(X) = \sum_{i \in L_t} p_i (s_i^+ - s_i^-)^2 = \sum_{i \in L_t} p_i ((s_i^+)^2 + (s_i^-)^2), \qquad (7)$$

since $(s_i^+)^2, (s_i^-)^2$ are not both positive at the same time. Using (6) we can easily express the semivariance $\mathrm{sVar}(X) = \mathbb{E}[(X - \mathbb{E}X)^2_-] = \sum_{i \in L_t} p_i (s_i^+)^2$ to measure downside risk. The standard Markowitz model can be written as

$$\max_{x,y,s \geq 0} y - \rho \Big[\sum_{i \in L_T} p_i ((s_i^+)^2 + (s_i^-)^2) \Big] \quad \text{subject to} \quad (1), (2), (3), (6). \qquad (8)$$

In this paper we are concerned with its extensions (we implicitly assume constraints (1), (2), (3), (6) in all of these):

• Downside risk (measured by the semi-variance) is constrained:

$$\max_{x,y,s \geq 0} y \quad \text{s.t.} \quad \sum_{i \in L_t} p_i (s_i^+)^2 \leq \rho. \qquad (9)$$

• Objective in a form of a logarithmic utility function captures risk-adversity:

$$\max_{x,y,s \geq 0} \sum_{i \in L_t} p_i \log\Big(\sum_{j=1}^{J} v_j x_{i,j}^h\Big) \quad \text{s.t.} \quad \sum_{i \in L_t} p_i (s_i^+)^2 \leq \rho. \qquad (10)$$

- Following Konno et al. [15] objective function takes skewness into account and captures the investors preference towards positive deviations in the case of non-symmetric distribution of returns for some assets

$$\max_{x,y,s \geq 0} y + \gamma \sum_{i \in L_t} p_i (s_i^+ - s_i^-)^3 \quad \text{s.t.} \quad \sum_{i \in L_t} p_i ((s_i^+)^2 + (s_i^-)^2) \leq \rho. \quad (11)$$

All these extensions have attractive modelling features but they inevitably lead to nonlinear programming formulations. It is worth noting that to date only a few algorithms have ever been considered for these formulations [7, 8] and it is not obvious if they can be extended to the general settings. Our approach can easily handle all these models.

2.2 Efficient frontier

The standard Markowitz objective function $f(x) = \mathbb{E}(F) - \lambda \text{Var}(F)$, uses the risk-aversion parameter λ to trade off the conflicting aims of maximizing return while minimizing risk. However a risk-aversion parameter is not an intuitive quantity, a better picture of the possible options would be gained from the complete trajectory $(\text{Var}(F, \lambda), \mathbb{E}(F, \lambda))$ for all values of λ, that is knowing how much extra expected return could be gained from an increase in the still acceptable level of risk. This $(\text{Var}(F, \lambda), \mathbb{E}(F, \lambda))$ trajectory is known as the *efficient frontier*.

The efficient frontier can be calculated by repeatedly solving the ALM model for different values of λ. However it would be desirable if this computation could be sped up by the use of warm-starts; after all we seek to solve a series of closely related problems. Unfortunately both proposed solution approaches for multistage stochastic programming, namely decomposition and interior point methods suffer from a perceived lack of efficient warmstarting facilities. We will show that OOPS comes with a warm starting facility that allows a significant decrease in computational cost when calculating the efficient frontier.

3 Object-oriented parallel solver (OOPS)

Over the years, interior point methods for linear and nonlinear optimization have proved to be a very powerful technique. We review basic facts of their implementation in this section and show how OOPS uses the special structure in stochastic programming problems to enable the efficient (and possible parallel) solution of very large problem instances.

Consider the nonlinear programming problem

$$\min \quad f(x) \quad \text{s.t.} \quad g(x) + z = 0, \ z \geq 0$$

where $f : \mathcal{R}^n \rightarrow \mathcal{R}$ and $g : \mathcal{R}^n \rightarrow \mathcal{R}^m$ are assumed sufficiently smooth. Interior point methods proceeed by replacing the nonnegativity constraints with

logarithmic barrier terms in the problem objective to get

$$\min \quad f(x) - \mu \sum_{j=1}^{n} \ln z_j \quad \text{s.t.} \quad g(x) + z = 0,$$

where $\mu \geq 0$ is a barrier parameter. First order stationary conditions of this problem are

$$
\begin{aligned}
\nabla f(x) - \nabla g(x)^T y &= 0 \\
g(x) + z &= 0, \\
YZe &= \mu e,
\end{aligned}
$$

where $Z = \text{diag}\{z_1, \ldots, z_n\}$. Interior point algorithms for nonlinear programming apply Newton method to solve this system of nonlinear equations and gradually reduce the barrier parameter μ to guarantee convergence to the optimal solution of the original problem. The Newton direction is obtained by solving the system of linear equations:

$$
\begin{bmatrix}
Q(x,y) & A(x)^T & 0 \\
A(x) & 0 & I \\
0 & Z & Y
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
\Delta y \\
\Delta z
\end{bmatrix}
=
\begin{bmatrix}
-\nabla f(x) - A(x)^T y \\
-g(x) - z \\
\mu e - YZe,
\end{bmatrix}, \quad (12)
$$

where $Q(x,y) = \nabla^2 f(x) + \sum_{i=1}^{m} y_i \nabla^2 g_i(x) \in \mathcal{R}^{n \times n}$ and $A(x) = \nabla g(x) \in \mathcal{R}^{m \times n}$ are the Hessian of Lagrangian and the Jacobian of constraints, respectively. After substituting $\Delta z = \mu Y^{-1} e - Ze - ZY^{-1} \Delta y$ in the second equation we get

$$
\begin{bmatrix}
-Q(x,y) & A(x)^T \\
A(x) & \Theta_D
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
-\Delta y
\end{bmatrix}
=
\begin{bmatrix}
\nabla f(x) + A(x)^T y \\
-g(x) - \mu Y^{-1} e
\end{bmatrix}, \quad (13)
$$

where $\Theta_D = ZY^{-1}$ is a diagonal matrix. Interior point methods need to solve several linear systems with this *augmented system* matrix at every iteration. This is by far the dominant computational cost in interior point implementations.

In many important applications (such as stochastic programming) the augmented system matrix displays a nested block structure. Such a structure can be represented by a matrix tree, that closely resembles the event tree of the corresponding stochastic program. Every node in the matrix tree represents a particular block-component of the augmented system matrix. OOPS exploits this structure by associating with each node of the event/matrix tree a linear algebra implementation that exploits the corresponding block matrix structure in operations such as matrix factorizations, backsolves and matrix-vector-products. It also enables the exploitation of parallelism, should several processors be available to work on a node.

In effect, all linear algebra operations required by an interior point method are performed in OOPS recursively by traversing the event tree, where several processors can be assigned to a particular node, if required. More details can be found in [9, 10, 16].

4 Numerical results

We will now present the computational results that underpin our claim that very large nonlinear portfolio optimization problems are now within scope of a modern structure exploiting implementation of general mathematical programming algorithms like OOPS.

We have used OOPS to solve the three variants (9), (10) and (11) of the Asset and Liability Management problem. All test problems are randomly generated using a symmetric scenario tree with 3-4 periods and between 24-70 realizations per time stage (Blocks). The data for the 20-40 assets used are also generated randomly. Statistics of the test problems are summarized in Table 1. As can be seen problem sizes increase to just over 10 million decision variables. Computational results for the three ALM variants (9), (10), (11) are collected in Table 2. Computations were done on the SunFire 15K at Edinburgh Parallel Computing Centre (EPCC), with 48 UltraSparc-III processors running at 900MHz and 48GB of shared memory. Since the parallel implementations relies solely on MPI we expect these results to generalize to a more loosely linked network of processors such as PCs linked via Ethernet. We used an optimality tolerance of 10^{-5} throughout.

All problems can be solved in a reasonable time and with a reasonable amount of interior point iterations - the largest problem needing just over 7 hours on a single 900MHz processor. OOPS displays good scalability, achieving a parallel efficiency of up to 0.96 on 8 processors. With the event of multi-core architectures even for desktop PCs, this shows that large nonlinear portfolio management problems are tractable even on modest computing hardware.

4.1 Comparison with CPLEX 9.1

We wish to make the point that a structure exploiting solver is an absolute requirement to solve very large stochastic nonlinear programming problems. To demonstrate this we have compared OOPS with the state-of-the-art commercial solver CPLEX 9.1. Since CPLEX has only the capability to solve QPs and we do not have a parallel CPLEX license, we compare CPLEX with OOPS for the QP model (8) on a single 3GHz processor with 2GB of memory. Results are reported in Table 3.

As can bee seen OOPS needs consistently less memory than CPLEX (which actually fails to solve problem C70 due to running out of memory - the time for this

Table 1: Asset and liability management problems: problem statistics.

Problem	Stages	Blk	Assets	Total Nodes	Constraints	Variables
ALM1	3	70	40	4971	208,713	606,322
ALM2	4	24	25	14425	388,876	1,109,525
ALM3	4	55	20	169456	3,724,953	10,500,112

Table 2: Results for nonlinear ALM variants.

Problem	1 proc		2 procs		4 procs		8 procs	
	iter	time (s)	time (s)	pe	time (s)	pe	time (s)	pe
variant (9): semi-variance								
ALM1	35	568	258	1.10	141	1.01	92	0.76
ALM2	30	1073	516	1.04	254	1.05	148	0.91
ALM3	43	18799	9391	1.00	4778	0.98	2459	0.96
variant (10): logarithmic utility								
ALM1	25	448	214	1.05	110	1.02	72	0.78
ALM2	31	1287	618	1.04	306	1.05	179	0.90
ALM3	60	24414	12480	0.98	6275	0.97	3338	0.91
variant (11): skewness								
ALM1	50	820	390	1.05	208	1.02	130	0.79
ALM2	43	1466	715	1.03	396	0.93	207	0.89
ALM3	62	23664	11963	0.99	6131	0.97	3097	0.96

Table 3: Comparison of OOPS with CPLEX 9.1.

Problem	Constraints	Variables	Blk	CPLEX 9.1		OOPS	
				time	memory	time	memory
C33	57,274	168,451	33	292	497MB	344	156MB
C50	130,153	382,801	50	1361	1.3GB	828	345MB
C70	253,522	745,651	70	(5254)	OoM	1627	664MB

Table 4: Dimensions and solution statistics for very large problems.

T	Blk	J	Scenarios	Constraints	Variables	Iter	Time	Procs	Mach
7	128	6	12,831,873	64,159,366	153,982,477	42	3923	512	BG/L
7	64	14	6,415,937	96,239,056	269,469,355	39	4692	512	BG/L
7	128	13	12,831,873	179,646,223	500,443,048	45	6089	1024	BG/L
7	128	21	16,039,809	352,875,799	1,010,507,968	53	3020	1280	HPCx

problem has been extrapolated from the number of nonzeros in the factorization as reported by CPLEX). The smallest problem C33 is solved slightly faster by CPLEX, while for larger problems OOPS becomes much more efficient than CPLEX.

4.2 Massively parallel architecture

In this section we demonstrate the parallel efficiency of our code running on a massively parallel environment. We have run the QP model (8) on two supercomputers: the BlueGene/L service at Edinburgh Parallel Computing Centre (EPCC) in co-processor mode, consisting of 1024 IBM-PowerPC-440 processors

Table 5: Parallel efficiency of OOPS.

Procs	Mem	time	Cholesky	Solves	MatVectProd
16	426MB	2587 (1.00)	1484 (1.00)	956 (1.00)	28.8 (1.00)
32	232MB	1303 (0.99)	743 (1.00)	485 (0.98)	18.0 (0.80)
64	132MB	688 (0.94)	377 (0.98)	270 (0.88)	13.0 (0.55)
128	84MB	348 (0.93)	187 (0.99)	139 (0.86)	9.0 (0.40)
256	56MB	179 (0.90)	93 (0.99)	73 (0.82)	5.8 (0.31)
512	46MB	94 (0.86)	47 (0.98)	39 (0.76)	3.9 (0.23)

Table 6: Warmstarting OOPS on efficient frontier problems for a series of λ.

Constraints	Variables	Procs	0.001	0.01	0.05	0.1	0.5	1	5	10
533,725	198,525	1	14	14	14	14	15	18	18	17
			14	5	5	6	5	5	9	10
5,982,604	16,316,191	32	23	24	23	25	22	24	23	24
			24	11	13	11	13	12	12	14
70,575,308	192,478,111	512	52	45	43	44	42	44	46	46
			52	13	15	15	16	16	23	25

running at 700Mhz and 512MB of RAM each. The second machine was the 1600-processor HPCx service at Daresbury, with 1GB of memory and 1.7GHz for every processor.

Results for these runs are summarized in Table 4. As can be seen OOPS is able to solve a problem with more than 10^9 variables on HPCx in less than one hour. Table 5 also gives the parallel efficiency for a smaller problem scaling from 16-512 processors on BlueGene. OOPS achieves a parallel efficiency of 86% on 512 processors as compared to 16 processors, with the dominant factorization part of the code even achieving 98% parallel efficiency.

4.3 Efficient frontier

Finally we have run tests calculating the efficient frontier for several large problems with up to 192 million decision variables on BlueGene. For every efficient frontier calculation the mean-variance model was solved for 8 different values of the risk-aversion parameter λ using OOPS' warmstarting facilities [16]. Results are gathered in Table 6. For every problem instance, the first line gives iteration numbers for computing points on the efficient frontier from coldstart, while the bottom line gives the iteration count for the warmstarted method. The last two large problems have been solved using 32 and 512 processors (procs), respectively. As can be seen OOPS' warmstart was able to save 45–75% percent of total iterations across the different problem sizes, demonstrating that warmstarting capabilities for truly large scale problems are available for interior point methods.

5 Conclusion

We have presented a case for solving nonlinear portfolio optimization problems by general purpose structure exploiting interior point solver. We have concentrated on three variations of the classical mean-variance formulations of an Asset and Liability Management problem each leading to a nonlinear programming problem. While these variations have been recognized for some time for their theoretical value, received wisdom is that these models are out of scope for mathematical programming methods. We have shown that in the light of recent progress in structure exploiting interior point solvers, this is no longer true. Indeed nonlinear ALM problems with several of millions of variables are within grasp of the next generation of Desktop PCs, while massively parallel machines can tackle problems with over 10^9 decision variables.

References

[1] Consigli, G. & Dempster, M., Dynamic stochastic programming for asset-liability management. *Annals of Operations Research*, **81**, pp. 131–162, 1998.

[2] Mulvey, J. & Vladimirou, H., Stochastic network programming for financial planning problems. *Management Science*, **38**, pp. 1643–1664, 1992.

[3] Zenios, S., Asset/liability management under uncertainty for fixed-income securities. *Annals of Operations Research*, **59**, pp. 77–97, 1995.

[4] Ziemba, W.T. & Mulvey, J.M., *Worldwide Asset and Liability Modeling*. Publications of the Newton Institute, Cambridge University Press: Cambridge, 1998.

[5] Birge, J.R., Decomposition and partitioning methods for multistage stochastic linear programs. *Operations Research*, **33**, pp. 989–1007, 1985.

[6] Ruszczynski, A., Decomposition methods in stochastic programming. *Mathematical Programming B*, **79**, pp. 333–353, 1997.

[7] Blomvall, J. & Lindberg, P.O., A Riccati-based primal interior point solver for multistage stochastic programming. *European Journal of Operational Research*, **143**, pp. 452–461, 2002.

[8] Steinbach, M., Hierarchical sparsity in multistage convex stochastic programs. *Stochastic Optimization: Algorithms and Applications*, eds. S. Uryasev & P.M. Pardalos, Kluwer Academic Publishers, pp. 363–388, 2000.

[9] Gondzio, J. & Grothey, A., Parallel interior point solver for structured quadratic programs: Application to financial planning problems. *Technical Report MS-03-001*, School of Mathematics, University of Edinburgh, Edinburgh EH9 3JZ, Scotland, UK, 2003. Accepted for publication in *Annals of Operations Research*.

[10] Gondzio, J. & Grothey, A., Solving nonlinear portfolio optimization problems with the primal-dual interior point method. *Technical Report MS-04-001*, School of Mathematics, University of Edinburgh, Edinburgh EH9

3JZ, Scotland, UK, 2004. Accepted for publication in *European Journal of Operational Research*.

[11] Høyland, K., Kaut, M. & Wallace, S.W., A heuristic for moment-matching scenario generation. *Computational Optimization and Applications*, **24(2/3)**, pp. 169–186, 2003.

[12] Markowitz, H.M., *Portfolio Selection: Efficient Diversification of Investments*. John Wiley & Sons, 1959.

[13] Steinbach, M., Markowitz revisited: Mean variance models in financial portfolio analysis. *SIAM Review*, **43(1)**, pp. 31–85, 2001.

[14] Ogryczak, W. & Ruszczynski, A., Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization*, **13(1)**, pp. 60–78, 2002.

[15] Konno, H., Shirakawa, H. & Yamazaki, H., A mean-absolute deviation-skewness portfolio optimization model. *Annals of Operational Research*, **45**, pp. 205–220, 1993.

[16] Gondzio, J. & Grothey, A., Reoptimization with the primal-dual interior point method. *SIAM Journal on Optimization*, **13(3)**, pp. 842–864, 2003.

This page intentionally left blank

Section 3
Derivatives pricing

This page intentionally left blank

Mean-variance hedging strategies in discrete time and continuous state space

O. L. V. Costa[1], A. C. Maiali[1] & A. de C. Pinto[2]
[1]*Escola Politécnica - Universidade de São Paulo, Brazil*
[2]*Fundação Getulio Vargas - EAESP, Brazil*

Abstract

In this paper we consider the mean-variance hedging problem of a continuous state space financial model with the rebalancing strategies for the hedging portfolio taken at discrete times. An expression is derived for the optimal self-financing mean-variance hedging strategy problem, considering any given payoff in an incomplete market environment. To some extent, the paper extends the work of Černý [1] to the case in which prices may assume any value within a continuous state space, a situation that more closely reflects real market conditions. An expression for the "fair hedging price" for a derivative with any given payoff is derived. Closed-form solutions for both the "fair hedging price" and the optimal control for the case of a European call option are obtained. Numerical results indicate that the proposed method is consistently better than the Black and Scholes approach, often adopted by practitioners.
Keywords: discrete-time mean-variance hedging, options pricing, optimal control.

1 Introduction

The problem of hedging options has systematically been the focus of attention from both researchers and practitioners alike. The complex nature of most derivatives has led academics to often simplify the conditions under which trading occurs, proposing models which, albeit computational and mathematically treatable, do not capture all of the peculiarities of these instruments. When modelling the dynamics of an asset price, its derivatives and the corresponding hedging process, the choices of state space and time parameter are determined so as to simplify the model's complexity. However, with respect to hedging, the situation that more closely follows what is observed in real market conditions is the use

of discrete times for representing portfolio rebalancing instants, and continuous state spaces for values possibly assumed by prices. Indeed, decisions regarding rebalancing the hedged position naturally occur at discrete times, whereas the smallest possible price variation ("market ticks") can be more adequately modelled within a continuous state space framework. It is, therefore, the purpose of this work to solve, for a given option, the mean-variance hedging problem of a continuous state space financial model with the rebalancing strategies for the hedging portfolio taken at discrete times.

Most studies of mean-variance hedging to date have considered the case of rebalancing strategies taken at continuous time. For discrete-time rebalancing, various intertemporal mean-variance criteria were analysed by Schäl [2] in the case of a constant investment opportunity set. A solution for the general problem with one asset and non-stochastic interest rate, which does not have a fully recursive structure, was presented by Schweizer [3]. This difficulty was overcome by the work of Bertsimas [4], who presented a fully recursive dynamic programming solution for the case of one basis asset and non-stochastic interest rate. Černý [1] proposed a general and simple recursive solution for the hedging problem with stochastic interest rate and an arbitrary number of basis assets.

The purpose of this work is to extend the work of Černý [1] to the case where the dynamics of a risky asset price is represented by an Itô diffusion with constant parameters. This approach allows us to obtain expressions for both the *fair hedging price* (mean-value process) of the option to be hedged, and the optimal control to be applied at any rebalancing instant. In particular, we derive closed-form solutions for the case of European vanilla call options which eliminate the recursiveness of previous models, thus producing considerable computational gains.

The paper is organized as follows. Section 2 presents the basic model and the proposed method which produces non-recursive expressions for the mean value process of an option with any given payoff and its corresponding optimal control at any rebalancing instant. Section 3 applies the methodology described in Section 2 to the case of a European vanilla call option deriving closed-form expressions for the option value and for the amount of underlying asset to be bought or sold for hedging purposes, i.e. the optimal control. Numerical results comparing hedging strategies suggested by the optimal self-financing mean-variance hedging proposed in this paper and that by the Black and Scholes (B&S) [5] approach are presented in Section 4. Finally, a summary and brief conclusions are presented in Section 5.

2 Discrete time, continuous state space mean-variance hedging strategy

Let $t \in [0, T^c]$ represent a particular time instant in a continuous-time model, and $\tau \in \{0, 1, \cdots, T\}$ represent the corresponding time instant in a discrete-time model. Consider that the time interval between two consecutive discrete-time instants is Δt, and that, for a particular τ whose corresponding continuous-

time instant is t, we have that $T - \tau = n$, with n being given by $n = (T^c - t)/\Delta t$.

Let $S(t)$ denote the price of a dividend-paying asset at time t. We assume that $S(t)$ follows a geometric Brownian motion described, in the continuous time setting notation, by the stochastic process below:

$$S(t + \Delta t) = S(t)e^{\sigma \Delta W^P(t+\Delta t) + (\mu - \rho - \frac{1}{2}\sigma^2)\Delta t}, \tag{2.1}$$

and in the discrete-time setting notation, by:

$$S(\tau + 1) = S(\tau)e^{\sigma \Delta W^P(\tau+1) + (\mu - \rho - \frac{1}{2}\sigma^2)\Delta t}. \tag{2.2}$$

The parameter μ represents the asset's expected rate of return; ρ, the asset's dividend yield; and σ, the volatility, all assumed to be constant. $W^P(\cdot)$ is a Wiener process under the probability measure P.

In a discrete-time setting, consider a market free of arbitrage opportunities composed of a risky asset S and a risk-free asset S^0, whose value at discrete time τ is $S^0(\tau)$. The risk-free interest rate, r, is assumed to be constant, for all $\tau \in \{0, 1, \cdots, T\}$, with S^0 and r being related by $S^0(\tau + 1) = S^0(\tau)e^{r\Delta t}$, with $S^0(0) = 1$.

Let H be a non-attainable derivative, maturing at time $\tau = T$, whose underlying asset is S. The derivative payoff is $H(T)$. Assume that a position in H must be hedged at discrete time instants $\tau, \tau + 1, \ldots, T - 1$, called *rebalancing instants*.

Let V be a self-financing portfolio composed of these two assets. The value of the portfolio at time τ is $V(\tau)$, with $V(0)$ being the initial wealth. An optimal hedging strategy, $\{u(\tau)\}_{\tau=0,\cdots,T-1}$ (optimal control law), can be obtained by solving the mean-variance hedging problem, which gives the best approximation by means of self-financing trading strategies, with the optimality criterion being the expected squared replication error.

Defining $E_\tau^P[\cdot]$ as the conditional expectation operator w.r.t. probability measure P given the filtration $\mathcal{F}_\tau$, the value function to be minimized at time 0, $\tilde{\mathcal{J}}_T$, is given by:

$$\tilde{\mathcal{J}}_T(0) = \min_{V(0), u_0, \ldots, u_{T-1}} E_0^P[(V(T) - H(T))^2], \tag{2.3}$$

with $V(0)$ being $\mathcal{F}_0$-measurable, and u_τ $\mathcal{F}_\tau$-measurable, $\tau = 0, 1, \cdots, T - 1$.

Let $\Delta X(\cdot)$, the discounted gain process of S, be given by:

$$\Delta X(\tau + 1) = \frac{S(\tau + 1)}{S^0(\tau + 1)} + \frac{\delta(\tau + 1)}{S^0(\tau + 1)} - \frac{S(\tau)}{S^0(\tau)}, \tag{2.4}$$

with $\delta(\tau)$ corresponding to the dividends paid for holding the risky asset S between discrete-time instants τ and $\tau + 1$.

The value $V(\tau)$, $\mathcal{F}_\tau$-measurable, evolves according to the optimal control law, i.e. it is the portfolio generated by the control policy $\{u(\tau)\}_{\tau=0,\cdots,T-1}$. At time $\tau = 0$, the value of this portfolio is $V(0)$. It can be shown that:

$$V(\tau + 1) = e^{r\Delta t}V(\tau) + S^0(\tau+1)u(\tau)\Delta X(\tau+1). \qquad (2.5)$$

Under these conditions, the solution of the optimisation problem defined in (2.3) is, as shown in Černý [1], given by

$$\tilde{u}(\tau) = -\frac{E_\tau^P\left\{k(\tau+1)\Delta X(\tau+1)\left(\frac{V(\tau)}{S^0(\tau)} - \frac{H(\tau+1)}{S^0(\tau+1)}\right)\right\}}{E_\tau^P\{k(\tau+1)(\Delta X(\tau+1))^2\}}, \tau = 0,\cdots,T-1,$$
$$(2.6)$$

$$V(0) = H(0), \qquad (2.7)$$

where:

$$H(\tau) = S^0(\tau)E_\tau^P\left\{m_{T,\tau}^{P\to Q}\frac{H(T)}{S^0(T)}\right\}, \qquad (2.8)$$

$$m_{T,\tau}^{P\to Q} = \prod_{j=\tau}^{T-1} m_{j+1,j}^{P\to Q}, \qquad (2.9)$$

$$m_{j+1,j}^{P\to Q} = \frac{k(j+1) - \frac{E_j^P\{k(j+1)\Delta X(j+1)\}}{E_j^P\{k(j+1)(\Delta X(j+1))^2\}}k(j+1)\Delta X(j+1)}{E_j^P\{k(j+1)\} - \frac{(E_j^P\{k(j+1)\Delta X(j+1)\})^2}{E_j^P\{k(j+1)(\Delta X(j+1))^2\}}}, \quad (2.10)$$

$$\frac{k(\tau)}{R_f^2(\tau)} = E_\tau^P\{k(\tau+1)\} - \frac{(E_\tau^P\{k(\tau+1)\Delta X(\tau+1)\})^2}{E_\tau^P\{k(\tau+1)(\Delta X(\tau+1))^2\}}, \qquad (2.11)$$

$$k(T) = 1. \qquad (2.12)$$

Extending the work of Černý [1] to the case where the price of a risky asset price is represented by a lognormal geometric brownian motion with constant parameters, as in (2.2), we obtain explicit expressions for both the mean-value process, $H(\tau)$, of the option to be hedged, and the optimal control, $\tilde{u}(\tau)$, to be applied at the rebalancing instant τ. The main results are given by Theorems 2.1 and 2.2 stated below. Full proofs can be found in Maiali [6].

In what follows we use the following notation:

1. $E_{l,\tau}^Q\{\cdot\}$ is the conditional expectation operator, as defined before. The subscript l is used just to explicitly show the dependence of the operator on l, which will be introduced due to the change from the probability measure P to Q, with Q being a probability measure whose Radon-Nikodým derivative with respect to P will depend on l. The same holds for $E_{l,\tau}^S\{\cdot\}$ and $E_{l,\tau}^T\{\cdot\}$.

2. $\mathbb{I}_A(x)$ represents the indicator function of x w.r.t. the set A.

3. $C_{p,l}$ is the l-th element of the set $C_p, 1 \leq l \leq \binom{n}{p}$, whose elements are subsets formed by p elements, $0 \leq p \leq n$, taken from the set $\{1,\cdots,n\}$.

4. $\sigma_l(\tau + j - 1) = \sigma \mathbb{I}_{C_{p,l}}(j)$.

Theorem 2.1 *Let $H(\tau)$ and $m_{T,\tau}^{P \to Q}$ be given by (2.8) and (2.9), respectively. Then, $H(\tau)$ can be written as:*

$$H(\tau) = e^{-r(T-\tau)\Delta t} \sum_{p=0}^{n} \left(a_0^{n-p} a_1^p \sum_{l=1}^{\binom{n}{p}} E_{l,\tau}^Q \{H(T)\} \right), \qquad (2.13)$$

where:

$$a_0 = \frac{e^{(r-\mu)\Delta t} - e^{\sigma^2 \Delta t}}{1 - e^{\sigma^2 \Delta t}}, \qquad a_1 = 1 - a_0 = \frac{1 - e^{(r-\mu)\Delta t}}{1 - e^{\sigma^2 \Delta t}}, \qquad (2.14)$$

with Q being a probability measure whose Radon-Nikodým derivative is given by:

$$\frac{dQ}{dP} = \exp \left(\sum_{j=1}^{n} \left(\sigma \mathbb{I}_{C_{p,l}}(j) \Delta W^P(\tau + j) - \frac{1}{2} (\sigma \mathbb{I}_{C_{p,l}}(j))^2 \Delta t \right) \right)$$

$$= \exp \left(\sum_{j=1}^{T-\tau} \left(\sigma_l(\tau + j - 1) \Delta W^P(\tau + j) - \frac{1}{2} \sigma_l^2(\tau + j - 1) \Delta t \right) \right). \quad (2.15)$$

Theorem 2.2 *Let $\Delta X(\tau+1)$, $V(\tau)$, $k(\tau+1)$, and $H(\tau+1)$ be given respectively by (2.4), (2.5), (2.11), and (2.13). Then, the optimal control $\tilde{u}(\tau)$, given by (2.6), can be written as:*

$$\tilde{u}(\tau) = \frac{e^{-r(T-\tau)\Delta t} \sum_{p=0}^{n} a_0^{n-p} a_1^p \sum_{l=1}^{\binom{n}{p}} (e^{(\mu-r)\Delta t} E_{l,\tau}^S \{H(T)\} - E_{l,\tau}^T \{H(T)\})}{S(\tau)(e^{(2\mu-2r+\sigma^2)\Delta t} - 2e^{(\mu-r)\Delta t} + 1)}$$

$$- \frac{V(\tau)(e^{(\mu-r)\Delta t} - 1)}{S(\tau)(e^{(2\mu-2r+\sigma^2)\Delta t} - 2e^{(\mu-r)\Delta t} + 1)}, \qquad (2.16)$$

where S and T are probability measures whose Radon-Nikodýn derivatives are given by:

$$\frac{dS}{dP} = \exp \left(\sum_{j=1}^{n} \left(\Lambda_l(\tau + j - 1) \Delta W^P(\tau + j) - \frac{1}{2} \Lambda_l^2(\tau + j - 1) \Delta t \right) \right),$$

$$\Lambda_l(\tau + j - 1) = \begin{cases} \sigma_l(\tau + j - 1) & j = 2, \cdots, T - \tau \\ \sigma & j = 1, \end{cases} \qquad (2.17)$$

$$\frac{dT}{dP} = \exp\left(\sum_{j=1}^{n}\left(\Gamma_l(\tau+j-1)\Delta W^P(\tau+j) - \frac{1}{2}\Gamma_l^2(\tau+j-1)\Delta t\right)\right),$$

$$\Gamma_l(\tau+j-1) = \begin{cases} \sigma_l(\tau+j-1) & j=2,\cdots,T-\tau \\ 0 & j=1. \end{cases} \tag{2.18}$$

3 Application: European call options

Here we apply the results obtained in the previous section to the case in which the derivative to be hedged is a European vanilla call option. We derive closed-form solutions for both the mean-value process, $H(\tau)$, of the option to be hedged, and the optimal control, $\tilde{u}(\tau)$, to be applied at rebalancing instant τ. It should be noted that their final expressions are extensions of the B&S formulae. These closed-form solutions eliminate the recursiveness of previously proposed models, thus producing considerable computational gains. Similar procedures would lead to closed-form solutions for the case of European vanilla put options.

Numerical analyses are presented in Section 4. As in the previous section, the main results are presented in the form of theorems, with their full proofs being found in Maiali [6].

Theorem 3.1 *Consider an European vanilla call option whose payoff is given by* $H(T) = (S(T) - K)^+$. *Equations (2.13) and (2.16) can be written as:*

$$H(\tau) = \sum_{p=0}^{n}\binom{n}{p}a_0^{n-p}a_1^p[e^{[(\mu-r-\rho)(T-\tau)+\sigma^2 p]\Delta t}S(\tau)N(d_R)$$

$$- e^{-r(T-\tau)\Delta t}KN(d_Q)]), \tag{3.1}$$

where:

$$d_Q = \frac{\ln(\frac{S(\tau)}{K}) + \left(\mu - \rho - \frac{1}{2}\sigma^2\right)(T-\tau)\Delta t + \sigma^2 p\Delta t}{\sigma\sqrt{(T-\tau)\Delta t}},$$

$$d_R = d_Q + \sigma\sqrt{(T-\tau)\Delta t}, \tag{3.2}$$

and

$$\tilde{u}(\tau) = \frac{e^{-r(T-\tau)\Delta t}\sum_{p=0}^{n}a_0^{n-p}a_1^p\sum_{l=1}^{\binom{n}{p}}(e^{(\mu-r)\Delta t}E_{l,\tau}^S\{H(T)\} - E_{l,\tau}^T\{H(T)\})}{S(\tau)(e^{(2\mu-2r+\sigma^2)\Delta t} - 2e^{(\mu-r)\Delta t} + 1)}$$

$$-\frac{V(\tau)(e^{(\mu-r)\Delta t}-1)}{S(\tau)(e^{(2\mu-2r+\sigma^2)\Delta t}-2e^{(\mu-r)\Delta t}+1)}, \tag{3.3}$$

where:

$$E^S_{l,\tau}\{H(T)\} = S(\tau)e^{(\mu-\rho)(T-\tau)\Delta t+\sigma^2\Delta t(\varphi_{p,l}+1)}N(d_U) - KN(d_S), \tag{3.4}$$

$$E^T_{l,\tau}\{H(T)\} = S(\tau)e^{(\mu-\rho)(T-\tau)\Delta t+\sigma^2\Delta t\varphi_{p,l}}N(d_V) - KN(d_T), \tag{3.5}$$

$$d_S = \frac{\ln(\frac{S(\tau)}{K}) + \left(\mu-\rho-\frac{1}{2}\sigma^2\right)(T-\tau)\Delta t + \sigma^2\Delta t(\varphi_{p,l}+1)}{\sigma\sqrt{(T-\tau)\Delta t}}, \tag{3.6}$$

$$d_U = d_S + \sigma\sqrt{(T-\tau)\Delta t}, \tag{3.7}$$

$$d_T = \frac{\ln(\frac{S(\tau)}{K}) + \left(\mu-\rho-\frac{1}{2}\sigma^2\right)(T-\tau)\Delta t + \sigma^2\Delta t\varphi_{p,l}}{\sigma\sqrt{(T-\tau)\Delta t}}, \tag{3.8}$$

$$d_V = d_T + \sigma\sqrt{(T-\tau)\Delta t}, \tag{3.9}$$

$$\varphi_{p,l} = \begin{cases} 0 & \text{if} \quad p=0 \\ p-1 & \text{if} \quad p\neq 0; 1\leq l\leq \binom{n-1}{p-1} \\ p & \text{if} \quad p\neq 0; \binom{n-1}{p-1} < l \leq \binom{n}{p}. \end{cases} \tag{3.10}$$

4 Numerical results

Here the results obtained in Section 3 are applied to European call options maturing in 6 and 12 months. Consider that $r = 17\%$ per annum (present level of Brazilian interest rates), that the current value of the underlying asset is $S = 100$, and that it pays no dividend ($\rho = 0$).

Results for three different strikes are compared, $K = 95$, $K = 100$, and $K = 115$, corresponding to in-the-money, at-the-money and out-of-the-money options, respectively. For each possible situation (maturity date and strike) we observe the effects of different expected rates of return, μ, with $\mu = 10\%$ and $\mu = 20\%$, different volatilities, σ, with $\sigma = 20\%$ and $\sigma = 40\%$, and different number of rebalancing instants, n, with $n = 6$ and $n = 10$. Paths of the underlying asset are simulated according to (2.1). For each path there is a payoff, $H(T)$, which is compared with the value of the hedging porfolio at maturity, $V(T)$. The hedging error, expressed as the present value of the square root of the mean-squared difference between the option's payoff and hedging portfolio at maturity, is calculated relative to the option's current value. The procedure is repeated for two hedging methods: (i) the dynamic programming approach (DP) proposed in Section 3; and (ii) the B&S approach (delta-hedging). Results for the error incurred by both methods, as well as the relative error of DP with respect to B&S, are presented for each combination of parameters. Results for in-, at- and out-of-the-money call options maturing in 6 and 12 months are given in Tables 1, 2 and 3, respectively. Hedging errors for both methods (columns "error DP" and "error

Table 1: $K = 95$ (in-the-money); $r = 17\%, S = 100$.

n	μ	σ	T = 6 months			T = 12 months		
			error B&S	error DP	rel. error	error B&S	error DP	rel. error
	10%	20%	13.07%	12.93%	-1.11%	10.33%	10.20%	-1.24%
6	10%	40%	35.94%	35.50%	-1.21%	28.24%	27.59%	-2.31%
	20%	20%	8.88%	8.84%	-0.41%	6.01%	5.95%	-1.09%
	20%	40%	32.18%	31.95%	-0.71%	25.24%	24.88%	-1.43%
	10%	20%	12.49%	12.17%	-2.50%	10.30%	9.86%	-4.32%
10	10%	40%	37.90%	37.40%	-1.33%	31.77%	31.04%	-2.32%
	20%	20%	8.30%	8.29%	-0.15%	5.70%	5.67%	-0.45%
	20%	40%	33.28%	33.19%	-0.26%	27.23%	27.07%	-0.60%

Table 2: $K = 100$ (at-the-money); $r = 17\%, S = 100$.

n	μ	σ	T = 6 months			T = 12 months		
			error B&S	error DP	rel. error	error B&S	error DP	rel. error
	10%	20%	30.31%	30.06%	-0.84%	17.57%	17.37%	-1.12%
6	10%	40%	48.33%	47.79%	-1.11%	34.31%	33.57%	-2.15%
	20%	20%	21.71%	21.67%	-0.19%	11.64%	11.55%	-0.71%
	20%	40%	46.53%	46.27%	-0.56%	32.23%	31.84%	-1.20%
	10%	20%	31.32%	30.75%	-1.82%	18.95%	18.26%	-3.66%
10	10%	40%	52.45%	51.85%	-1.16%	39.28%	38.43%	-2.19%
	20%	20%	21.65%	21.65%	-0.02%	11.57%	11.55%	-0.24%
	20%	40%	49.22%	49.10%	-0.24%	35.16%	35.00%	-0.46%

B&S") as well as the DP error relative to that of the B&S approach (column "relative error") are presented.

It can be observed that, in all cases, whenever the expected rate of return, μ, assumes values close to the risk-free rate r (e.g. $r = 17\%$ and $\mu = 20\%$), the results from the B&S model approach, but are consistently worse than those obtained by the DP model, as it should be expected, since in a B&S risk-neutral setting, an Itô diffusion with rate μ corresponds to a risk-neutral diffusion with rate r. Conversely, whenever μ and r are apart (e.g. $r = 17\%$ and $\mu = 10\%$), the DP model behaves considerably better, as the assumptions of the B&S model no longer hold.

Since both models are linear approximations for $H(\cdot)$, the results indicate that, irrespective of the moneyness of the option to be hedged, for a small number of rebalancing instants (e.g. $n = 6$), and high volatility (e.g. $\sigma = 40\%$), both methods

Table 3: $K = 115$ (out-of-the money); $r = 17\%$, $S = 100$.

			T = 6 months			T = 12 months		
n	μ	σ	error B&S	error DP	rel. error	error B&S	error DP	rel. error
	10%	20%	65.65%	65.17%	-0.73%	37.95%	37.60%	-0.95%
6	10%	40%	67.99%	67.40%	-0.87%	46.00%	45.21%	-1.71%
	20%	20%	79.94%	79.84%	-0.13%	33.31%	33.24%	-0.20%
	20%	40%	74.37%	74.20%	-0.22%	47.76%	47.44%	-0.66%
	10%	20%	67.72%	66.59%	-1.68%	42.17%	40.93%	-2.95%
10	10%	40%	70.81%	70.07%	-1.05%	51.79%	50.71%	-2.08%
	20%	20%	86.55%	86.54%	-0.01%	36.27%	36.27%	-0.02%
	20%	40%	80.19%	80.11%	-0.10%	54.01%	53.85%	-0.31%

produce significant hedging errors. Nevertheless, even in this situation, it can be observed that the proposed method outperforms the B&S model. It should be noted that, as n increases, although results produced by the DP model converge to those obtained by the B&S model (following the assumption of infinitesimal rebalancing instants from the latter), the proposed method consistently incurs less hedging errors than those obtained the B&S approach, apart from results for small n, in which case both models behave poorly.

The situation that indicates the best relative performance of the proposed method is the case of small volatilities (see results for $\sigma = 20\%$ in Tables 1, 2 and 3), as the payoff of the option becomes less unpredictable.

5 Summary and concluding remarks

In this work we have analysed the mean-variance hedging problem of a continuous state space financial model with the rebalancing strategies for the hedging portfolio taken at discrete times. We have derived an expression for the optimal self-financing mean-variance hedging strategy problem, considering any given payoff in an incomplete market environment. As an application of the proposed method, we have obtained closed-form solutions for the value European vanilla call options and for the amount of the corresponding underlying asset to be bought or sold for hedging purposes (optimal control law).

The results showed that the proposed solution is consistently better than the B&S delta-hedging approach for all possible combinations of parameters considered. As expected, the proposed method presents relatively better results, especially when the market structure does not follow their basic assumptions. The method is flexible enough with regard to the determination of optimal hedging strategies to be applied to a broad variety of European-style derivatives and stochastic price processes of their underlying asset. In particular, our current

research is concentrated towards: (i) obtaining closed-forms solutions for other instruments; and (ii) modelling asset prices whose dynamics are represented by jump-diffusions and/or stochastic volatility models.

Acknowledgments

References

[1] Černý, A., Dynamic programming and mean-variance hedging in discrete time. *Applied Mathematical Finance*, **1(11)**, pp. 1–25, 2004.

[2] Schäl, M., On quadratic cost criteria for option hedging. *Mathematics of Operations Research*, **1(19)**, pp. 121–131, 1994.

[3] Schweizer, M., Variance-optimal hedging in discrete time. *Mathematics of Operations Research*, **1(20)**, pp. 1–32, 1995.

[4] Bertsimas, D., Kogan, L. & Lo, A.W., Hedging derivative securities in incomplete market: An ϵ-arbitrage approach. *Operations Research*, **3(49)**, pp. 372–397, 2001.

[5] Black, F. & Scholes, M., The pricing of options and corporate liabilities. *Journal of Political Economy*, **(81)**, pp. 637–654, 1973.

[6] Maiali, A.C., *Stochastic optimal control at discrete time and continuous state space applied to derivatives*. Ph.D. thesis, Escola Politécnica - Universidade de São Paulo, 2006.

[7] Pham, H., Rheinländer, T. & Schweizer, M., Mean-variance hedging for continuous processes: New results and examples. *Finance and Stochastics*, **(2)**, pp. 173–198, 1998.

[8] Laurent, J.P. & Pham, H., Dynamix programming and mean-variance hedging. *Finance and Stochastics*, **1(3)**, pp. 83–110, 1999.

[9] Schweizer, M., Mean-variance hedging for general claims. *The Annals of Applied Probability*, **1(2)**, pp. 171–179, 1992.

[10] Schweizer, M., Approximation pricing and the variance-optimal martingale measure. *The Annals of Applied Probability*, **1(24)**, pp. 206–236, 1996.

The more transparent, the better – evidence from Chinese markets

Z. Wang
School of Management, Xiamen University, People's Republic of China

Abstract

The Chinese stock markets, including the Shanghai Stock Exchange and the Shenzhen Stock Exchange, increased the real-time public dissemination of limit order book from the 3 best ask and bid quotes to 5 best on December 8, 2003. This change in transparency regime allows me to assess the effect of pre-trade transparency on the two markets. The most striking finding is that the effect of an increase in pre-trade transparency on the two different markets is quite similar. I find that the informational efficiency of price improves significantly, the market liquidity increases significantly, the volatility of price decreases and the component of asymmetric information in the bid-ask spread reduces after the two Exchanges adopt this action to improve transparency.
Keywords: market transparency, limit order book, bid-ask spread, liquidity, volatility.

1   Introduction

O'Hara [11] defined market transparency as the ability of market participants to observe information about the trading process. Madhavan [9] divided transparency into pre- and post-trade dimensions. Pre-trade transparency refers to the wide dissemination of current bid and ask quotations, depths (bid sizes and ask sizes), and possibly also information about limit orders away from the best prices, as well as other pertinent trade related information such as the existence of large order imbalances. Post-trade transparency refers to the public and timely transmission of information on past trades, including execution time, volume, price, and possibly information about buyer and seller identifications.

   Previous theoretical research finds that transparency affects market quality, including liquidity, trading costs, and the speed of price discovery. Models by

Chowdhry and Nanda [3], Madhavan [7, 8], Pagano and Röell [12], and Baruch [1] among others, reach mixed conclusions regarding the effects of transparency. Hence, empirical evidence on transparency and its effects on the quality of markets are absolutely necessary. Since changes in transparency regimes are rare, analysis of each event becomes more crucial in our ability to evaluate prevailing theory accurately.

Chinese stock markets, including Shanghai Stock Exchange and Shenzhen Stock Exchange, enhanced the level of pre-trade transparency On December 8, 2003. The two markets extend real-time public dissemination of the depth and limit order prices form up to three price levels above and below the current market to five. The system also required that all depth should be automatically displayed. This change provides me a unique opportunity to study the impact of an increase in pre-trade transparency on the two different markets. Beyond the rarity of such a change in transparency regime, the Chinese stock markets, as rapidly developing emerging markets, their protocol change is of special interest for us.

I examine how this increase of transparency in the two Chinese stock markets affects the market quality, including the informational efficiency of prices, market liquidity, the component of asymmetric information in the bid-ask spread and volatility. My empirical results strongly support the prediction suggested by Glosten [4] and Baruch [1]at higher transparency will improve market quality.

Even though the theoretical literature provides conflicting predictions on the effect of market transparency, China Securities Commission has repeatedly emphasized the need for increased pre-trade transparency. My research is an empirical study to provide support for such a policy.

2 Brief review of related empirical work

Empirical papers on investigation into the impact of limit-order book transparency on informational efficiency and liquidity is rare. The following two papers are representative.

Boehmer et al. [2] studied pre-trade transparency by looking at the introduction of NYSE's OpenBook service that provides limit-order book information to traders off the exchange floor on January 24, 2002. They found that traders attempt to manage limit-order exposure: They submit smaller orders and cancel orders faster. Specialists' participation rate and the depth they add to the quote decline. Liquidity increases in that the price impact of orders declines, and they found some improvement in the informational efficiency of prices. These results suggest that an increase in pre-trade transparency affects investors' trading strategies and can improve certain dimensions of market quality.

By contrast, Madhavan et al. [10] examined the natural experiment affected by the Toronto Stock Exchange when it publicly disseminated the limit order book on both the traditional floor and on its automated trading system on April 12, 1990. They found that the increase in transparency reduces liquidity. In particular, execution costs and volatility increase after the limit order book is

publicly displayed. They also showed that the reduction in liquidity is associated with significant declines in stock prices.

3 Research design

3.1 Event windows

I use event study to examine the effect of the change of pre-trade transparency on the market quality. As we know, it is important for event study to pinpoint the exact event date. While the investors knew that the transparency regime would change before December 8, 2003, which is the implementation date of increasing pre-trade transparency, trading strategies cannot be implemented without this information. Therefore, the effects we wish to investigate are best examined around the implementation date.

Since traders cannot use the information in the limit order book prior to December 8, there is no need to eliminate a long window before the event in order to obtain the steady state of traders' strategies. I choose the full 2 trading weeks (10 trading days) prior to the introduction week as the pre-event period (November 17 through November 28). The choice of an appropriate post-event period is more complex. While traders are able to see limit-order book information beginning December 8, learning how to use this information probably takes some time. This is true both for traders who want to use it just to optimize the execution of their orders and for traders who plan to use it to design profitable trading strategies. Furthermore, once such strategies are in place, other traders may experience poorer execution of their limit orders, prompting more traders to change their strategies until a new equilibrium emerges.

To allow for adjustment to an equilibrium state and to examine this adjustment, I use three post-event periods rather than one. As with the pre-event period, I use 2 weeks as the length of a post-event period to capture a reasonably stationary snapshot of the trading environment. More specifically, for each of the first 3 months after the introduction of the new disclosure regime I use the middle 2 full weeks of trading: December 15–26, January 12–February 3, (the Spring Festival holiday is included in this period,) February 16–27 (The four windows are named as November, December, January and February respectively hereafter). These three post-event periods enable us to examine how the new equilibrium emerges over time.

3.2 Data sources and sample

The data in this study are from CCER China Tick Data Database (provided by the Sinofin Information Services), and contain every trade and quote, with associated prices, volumes, and bid and ask sizes. The data are time stamped to the nearest second.

The sample includes all component stocks of the Shanghai Stock Exchange 180 Component Index and the Shenzhen Stock Exchange Component 100 Index. Since the two markets adjust their components of index twice a year, Shanghai

Stock Exchange on June and December, and Shenzhen Stock Exchange on May and November respectively. 18 stocks in Shanghai and 7 stocks in Shenzhen are rule out. In addition, 2 stocks in Shenzhen are picked out due to data error. After these procedures, 162 stocks in Shanghai (named as Shanghai 180 hereafter) and 91 stocks (named as Shenzhen 100 hereafter) are remained in the sample. Since the sample from the Shanghai market is almost twice as that from the Shenzhen market, I divided the sample of Shanghai Stock Exchange into two groups according to the median of share trading volume from July 1 to November 30 of 2003 (named as Group 1 and Group 2 respectively hereafter), and conducted the analysis separately for each group in order to comparing the effect on the two different market.

4 Empirical findings and analysis

4.1 Informational efficiency of prices

Both Glosten [4] and Baruch [1] predicted that improved transparency would lead to increased informational efficiency of prices. I implement the test of this hypothesis based on the variance decomposition procedure in Hasbrouck [5]. Using information about trade size and execution price for all transactions, Hasbrouck proposed a vector autoregression model to separate the efficient (random walk) price from deviations introduced by the trading process (e.g., short-term fluctuations in prices due to inventory control or order imbalances in the market). More specifically, the variance of log transaction prices, $V(p)$, is decomposed into the variance of the efficient price and the variance of the deviations induced by the trading process, $V(s)$. Because the expected value of the deviations is assumed by the procedure to be zero, the variance is a measure of their magnitude.

The ratio of $V(s)$ to $V(p)$, $VR(s/p)$, reflects the proportion of deviations from the efficient price in the total variability of the transaction price process. If the pre-trade transparency increasing allows traders to better time their trading activity to both take advantage of displayed liquidity and provide liquidity in periods of market stress, the proportion of deviations from the efficient price should be smaller after the event. Table 1 shows median changes between the pre- and post-event periods for $VR(s/p)$. All values in the table are negative, and the changes are significantly different from zero in the December and February post-event periods. The changes are not significantly different from zero in the January post-event period. I presume that the reason should be that this period includes a long Spring Festival holiday, and more information cumulated in the holiday must have been priced when the market reopen after the holiday.

The result of test points to significant improvement in informational efficiency under the new pre-trade transparency regime. At the very least, the evidence demonstrates that increasing the transparency of limit order book does not lead to deterioration in the efficiency of prices.

Table 1:   Change in informational efficiency.

	Dec–Nov		Jan–Nov		Feb–Nov	
ΔVR(s/p)	Median	P Value	Median	P Value	Median	P Value
Shanghai 180	-1.29E-03***(0.000)		-3.99 E-04	(0.107)	-1.32 E-03***(0.000)	
Group 1	-1.29E-03***(0.005)		-3.99 E-04	(0.656)	-1.32E-03***(0.000)	
Group 2	-1.65E-03***(0.002)		-9.45 E-04***(0.006)		-8.41 E-04***(0.000)	
Shenzhen100	-1.29E-03***(0.000)		-3.46 E-04	(0.264)	-1.07 E-03***(0.000)	

The p-value in parentheses is a Wilcoxon signed rank test against the hypothesis of a zero median. ***, **, * indicate significance at the 1%, 5%, and 10% level respectively.

4.2 Liquidity

I will examine in this section how the changing of transparency creates a new state of liquidity provision in the market. I define relative spread as $(P_{a1} - P_{b1})/P_m$; proportional effective spread as $|P_t - P_m|/P_m$; market depth 1 as $V_{a1}P_{a1} + V_{b1}P_{b1}$; and market depth 2 as $\frac{1}{3}\sum_{i=1}^{3}(V_{ai}P_{ai} + V_{bi}P_{bi})$. Where P_t is the trade price of a security at time t, P_{ai} is the ith best (lowest) ask quote, and P_{bi} is the ith best (highest) bid quote. V_{ai} is the share volume corresponding to the ith best ask quote, V_{bi} is the share volume corresponding to the ith best bid quote, and $P_m = \frac{1}{2}(P_{a1} + P_{b1})$ is the midpoint of the first best quote. I measure the spread by both the relative spread and proportional effective spread, and the depth by both market depth 1 and market depth 2. Then I compare the differences of median between pre- and post-event periods.

Table 2 reports the effect of the event on the market liquidity. All values in the Panel A and Panel B are negative and significantly different from zero. It shows that the spread decreases significantly after increasing the pre-trade transparency. By contrast, changes in market depth (see Panel C and Panel D) are all positive and significantly different from zero.

Because there is much evidence that liquidity is affected by attributes such as volume, I run a multivariate test to examine the change in liquidity conditional on three control variables. The controls are the average daily dollar volume, intra-day volatility expressed as the average daily range of transaction prices (high minus low), and the average transaction price of the stock (to control for price level effects).

The econometric specification assumes that the liquidity measure for stock i in period t (where $t \in \{$pre, post$\}$), L_{it}, can be expressed as the sum of a stock-specific mean (β_0), an event effect (α), a set of control variables, and an error term (η_{it}):

$$L_{i,t} = \beta_0 + \alpha Dummy_t + \beta_1 AvgVol_{i,t} + \beta_2 HiLow_{i,t} + \beta_3 Avg\Pr c_{i,t} + \eta_{it} \quad (1)$$

Table 2:   Change in liquidity.

	Dec–Nov		Jan–Nov		Feb–Nov	
	Median	P value	Median	P value	Median	P value
	Panel A					
Δrelative spread						
Shanghai 180	-9.93E-05	(0.135)	-2.25E-04***(0.000)		-3.54 E-04***(0.000)	
Group 1	-8.36E-05	(0.152)	-2.90 E-04***(0.000)		-4.44 E-04***(0.000)	
Group 2	-9.98E-05	(0.425)	-1.92 E-04** (0.012)		-3.1 E-04*** (0.002)	
Shenzhen 100	-0.00012***(0.001)		-2.13 E-04***(0.000)		-3.62 E-04***(0.000)	
	Panel B					
Δproportional effective spread						
Shanghai 180	-5E-05***	(0.002)	-1.1 E-04***	(0.000)	-2.17 E-04***(0.000)	
Group 1	-3.62E-05	(0.173)	-1.05E-04**	(0.013)	-2.34 E-04***(0.000)	
Group 2	-5.31E-05***(0.002)		-1.18E-04***(0.000)		-1.99 E-04***(0.000)	
Shenzhen 100	-6.38E-05***(0.000)		-9.35E-05***(0.000)		-1.92 E-04***(0.000)	
	Panel C				Unit: 100 Yuan	
Δmarket depth 1						
Shanghai 180	293.23 ***(0.000)		113.07	(0.390)	739.16*** (0.000)	
Group 1	146.41 ***(0.000)		178.10***(0.002)		691.11*** (0.000)	
Group 2	490.91 ***(0.001)		-83.07	(0.201)	836.24*** (0.000)	
Shenzhen 100	373.12 ***(0.000)		184.40***(0.004)		1039.49***(0.000)	
	Panel D				Unit: 100 Yuan	
Δmarket depth 2						
Shanghai 180	1171.33***(0.000)		370.87	(0.633)	2989.08***(0.000)	
Group 1	587.65*** (0.000)		745.97***(0.005)		2877.40***(0.000)	
Group 2	1723.11***(0.000)		-518.29	(0.141)	3639.57***(0.000)	
Shenzhen 100	1490.81***(0.000)		681.03**(0.013)		3764.30***(0.000)	

Where $Dummy_t$ is an indicator variable that takes the value zero in the pre-event period and one in the post-event period, $AvgVol$ represents dollar volume, $HiLow$ is intra-day volatility, and $AvgPrc$ is the price. By assuming that the errors are uncorrelated across securities and over the two periods (although we do not require them to be identically distributed), I can examine differences between the post- and pre-event periods and eliminate the firm-specific mean:

$$\Delta L_i = \alpha + \beta_1 \Delta AvgVol_i + \beta_2 \Delta HiLow_i + \beta_3 \Delta Avg \Pr c_i + \varepsilon_i \qquad (2)$$

where Δ denotes a difference between the post- and pre-event periods.

I estimate the eqn (2) using OLS and compute test statistics based on White's heteroskedasticity-consistent standard errors. Table 3 reports only the results that are significant. Panel A presents the intercepts and *p*-values from the regressions using the change to relative spread as the liquidity variable. The intercepts for all three post-event periods are all negative and significant, indicating some decrease in spread in the post-event period. Panel B reports the intercepts and p-values from regressions using the change to market depth 2 as the liquidity variable. The intercepts for December and February are positive and significant.

The empirical results of these two tests support the prediction of Glosten [4] and Baruch [1], which claimed that greater transparency would improve liquidity.

Table 3:   Analysis of liquidity—multivariate test.

| Δrelative spread | Panel A | | | | | |
| | Dec−Nov | | Jan−Nov | | Feb−Nov | |
	α	P value	α	P value	α	P value
Shanghai 180	-1.095E-03**(0.024)		-3.431E-04* (0.052)		-7.941E-04**(0.036)	
Shenzhen 100	-2.048E-04* (0.082)		-9.926E-04**(0.040)		-7.896E-04**(0.049)	
Δmarket depth 2	Panel B					
Shanghai 180	1316.64	(0.155)	-3318.48	(0.167)	5387.88***	(0.004)
Shenzhen 100	2542.22*	(0.055)	366.96	(0.636)	4983.32***	(0.000)

The p-value in parentheses is a t test against the hypothesis of a zero median.
***, **, * indicate significance at the 1%, 5%, and 10% level respectively.

4.3 Asymmetric information

Finding spread width decreases following increasing the transparency of the limit order book suggests that the adverse selection component of the spread may have decreased as well. To investigate changes in adverse selection, I use the model developed in Lin et al. [6] to decompose the component of asymmetric information:

Table 4:   Component of asymmetric information.

November	Shanghai 180	Shenzhen 100
Mean of λ(median)	0.2189(0.2032)	0.1429251(0.1401)
Mean of Adjusted R Square(Median)	0.0587(0.0562)	0.03573(0.0257)
t statistic	15.8285(16.2285)	11.101042(10.1119)
The proportion of stocks significant at 1%	98.15%	96.70%
December	Shanghai 180	Shenzhen 100
Mean of λ(median)	0.2119(0.2181)	0.1421(0.1367)
Mean of Adjusted R Square(Median)	0.0449(0.0375)	0.0286(0.0217)
t statistic	14.4798(14.6035)	12.9490(12.3768)
The proportion of stocks significant at 1%	95.68%	100%
January	Shanghai 180	Shenzhen 100
Mean of λ(median)	0.2037(0.2109)	0.1202(0.1239)
Mean of Adjusted R Square(Median)	0.0556(0.0517)	0.0208(0.0166)
t statistic	19.5341(19.5036)	11.4550(11.1334)
The proportion of stocks significant at 1%	95.06%	97.80%
February	Shanghai 180	Shenzhen 100
Mean of λ(median)	0.1994(0.2026)	0.1260(0.1249)
Mean of Adjusted R Square(Median)	0.0520(0.0549)	0.0241(0.0178)
t statistic	21.6457(22.7031)	13.6048(12.7745)
The proportion of stocks significant at 1%	98.77%	98.90%

$$\Delta Q_{t+1} = \lambda z_t + e_{t+1} \tag{3}$$

where,

$$Q_t = \ln\left[\frac{1}{2}(P_{a1} + P_{b1})\right] \tag{4}$$

$$\Delta Q_{t+1} = Q_{t+1} - Q_t \tag{5}$$

$$z_t = \ln p_t - Q_t \tag{6}$$

λ is the asymmetric information parameter. I first estimate λ for every single stock at every period, after that, I calculate the mean and median of all stock in Shanghai market and Shenzhen market respectively.

Table 4 shows that the components of asymmetric information present the trend of decrease in both two markets. The component of asymmetric information of Shanghai 180 (Shenzhen 100) decreases by 9.78% (13.41%) from November through February.

Table 5 shows median changes between the pre- and post-event periods for the adverse selection component. We can find that the adverse selection component decrease significantly (except for December) following the transparency increases. These findings result in our supporting the hypothesis that transparency increases will reduce the asymmetric component of the spread.

Table 5:   Change in component of asymmetric information.

$\Delta\lambda$	Dec–Nov		Jan–Nov		Feb–Nov	
	Median	P value	Median	P value	Median	P value
Shanghai 180	-7.05 E-03	(0.230)	-6.73 E-03*	(0.099)	-2 E-02***	(0.000)
Shenzhen 100	-2.3 E-04	(0.438)	-2.04 E-02***	(0.001)	-1.58 E-02**	(0.016)

4.4 Volatility

I measure the volatility by standard deviation of returns. Table 6 displays median changes between the pre- and post-event periods for return volatility. It shows that the volatility first increases on December and then has a significant decrease on both January and February for all stocks. It seems reasonable to infer that the change in transparency is associated with less volatility in both markets.

Table 6:   Change in volatility.

$\Delta\sigma$	Dec–Nov		Jan–Nov		Feb–Nov	
	Median	P value	Median	P value	Median	P value
Shanghai 180	1.33 E-05****	(0.000)	-1.46 E-04***	(0.000)	-3.69 E-04***	(0.000)
Group 1	2.03 E-04***	(0.000)	-1.74 E-04***	(0.000)	-3.87 E-04***	(0.000)
Group 2	7.91E-05***	(0.003)	-1.27 E-04***	(0.000)	-2.81 E-04***	(0.000)
Shenzhen 100	1.03 E-04***	(0.000)	-8.04E-05***	(0.004)	-2.45 E-04***	(0.000)

The extant literature documents a positive relationship between price volatility and trading frequency, which in turn may result from exogenous events such as news announcements. I use the following model to examine the event effect after controlling for the volume of trade.

$$\Delta\sigma_i = \beta_0 + \beta_1 \Delta N\_Trade_i \tag{7}$$

where $\Delta\sigma_i$ denotes the difference of standard deviation of returns for firm i between the pre- and post-event periods, ΔN_Trade_i is the difference of number

of transaction for firm $i,$ and β_0 capture the event effect. Table 7 reports the estimates of β_0 and β_1 from the regression model even though I focus on the β_0.

Table 7:   Analysis of volatility—multivariate test.

	Dec–Nov	Jan–Nov	Feb–Nov
Shanghai 180			
β_0 (t statistic)	4.54 E-03***(4.42)	-4.70 E-04(-0.722)	-2.82E-05**(-2.03)
β_1 (t statistic)	-8.16E-07**(-2.78)	-1.85E-07(-1.13)	-2.49E-07 (-1.66)
Adjusted R square	4.02%	0.17%	4.08%
Shenzhen 100			
β_0 (t statistic)	6.74E-05 (-1.34)	-1.20E-04(0.47)	-2.27E-04**(-2.41)
β_1 (t statistic)	-2.55E-08 (-1.30)	-3.55E-09(-0.03)	-2.75E-08 (-1.16)
Adjusted R square	0.76%	-1.12%	0.38%

The p-value in parentheses is a t test against the hypothesis of a zero median. ***, **, * indicate significance at the 1%, 5%, and 10% level respectively.

We can find β_0 is positive on December and then becomes negative on January and February for the two markets. That means, consistent with my earlier results, that the volatility increases at first post-event period and then decreases for both Shanghai market and Shenzhen market stocks. The empirical results of these two tests seem to support the prediction that the volatility decreases following the transparency increases.

5   Conclusions

Transparency is a topic of considerable importance to investors, academics, and regulators. Previous theoretical research often presents contradictory views of transparency. The most interest is that empirical evidence from different markets regarding *pre*-trade transparency support different predictions. This study analyzes empirically the impact of an increase in pre-trade transparency, focusing on the two emerging markets.

Consistent with the common presumption among many policy makers and regulators, my results provide empirical support for the view that improved pre-trade transparency of a limit-order book will improve the market quality.

The most striking finding of my paper is that the effect of pre-trade transparency increases on the two different markets is quite similar. They change at the same pace following the transparency increases. I find some improvement in informational efficiency, an increase in displayed liquidity in the book, and a decline in the price volatility after the two Exchanges adopt action to improve transparency. The equilibrium effects on the state of the market, both in terms of

liquidity and informational efficiency, seem to suggest that increased transparency is a win−win situation.

Acknowledgements

References

[1] Baruch, Shmuel, 2005, Who benefits from an open limit-order book? *Journal of Business* **78**, 1267-1306.

[2] Boehmer. E., Saar. G. and Yu.L. 2005, Lifting the Veil: An Analysis of Pre-Trade Transparency at the NYSE. *Journal of Finance* **60** (2), 783-815.

[3] Chowhdry, Bhagwan, and Vikram Nanda, 1991, Multimarket trading and market liquidity, *Review of Financial Studies* **4**, 483−511.

[4] Glosten, Lawrence R., 1999, Introductory comments: Bloomfield and O'Hara, and Flood, Huisman, Koedijk, and Mahieu, *Review of Financial Studies* **12**, 1−3.

[5] Hasbrouck, J., 1993, Assessing the quality of a security market: a new approach to transaction-cost measurement. Review of Financial Studies **6**,191-212.

[6] Lin, J. C., Sanger, G., Booth, G., 1995, Trading size and components of the bid-ask spread. *Review of Financial Studies* **8**, 1153-1183.

[7] Madhavan, Ananth N., 1995, Consolidation, fragmentation, and the disclosure of trading information, *Review of Financial Studies* **8**, 579−603.

[8] Madhavan, Ananth N., 1996, Security prices and market transparency, *Journal of Financial Intermediation* **5**, 255−283.

[9] Madhavan, Ananth N., 2000, Market microstructure: A survey. *Journal of Financial Markets* **3**,205-258.

[10] Madhavan, A., Porter, D. and Weaver.D. 2005, Should Securities Markets be Transparent? Journal of financial Markets **8,** 265-287.

[11] O'Hara, M., 1995, Market microstructure theory. Basil Blackwell, Cambridge, MA.

[12] Pagano, Marco, and Ailsa Röell, 1996, Transparency and liquidity: A comparison of auction and dealer markets with informed trading, *Journal of Finance* **51**, 579−611.

Herd behaviour as a source of volatility in agent expectations

M. Bowden & S. McDonald
School of Economics, University of Queensland, Brisbane, Australia

Abstract

Herd Behaviour is often cited as one of the forces behind excess volatility of stock prices as well as speculative bubbles and crashes in financial markets. This paper examines if social interaction and herd behaviour, modelled within a multi-agent framework, can explain these characteristics. The core of the model is based on the social learning literature which takes place in a small world network. We find that when the network consists entirely of herd agents then expectations become locked in an information cascade. Herd agents receive a signal, compare it with those agents with whom they are connected, and then adopt the majority position. Adding one expert agent enables the population to break the cascade as information filters from that agent to all other agents through contagion. We also find that moving from an ordered to a small world network dramatically increases the level of volatility in agent expectations and it quickly reaches a higher level (at which point increasing the randomness of the network has little effect). Increasing the influence of the experts, by increasing the number of connections from these agents, also increases volatility in the aggregate level of expectations. Finally it is found that under certain network structures herd behaviour will lead to information cascades and potentially to the formation of speculative bubbles.
Keywords: social learning, herd behaviour, small world networks, information contagion, volatility, information cascades.

1   Introduction

Herd Behaviour is probably one of our most basic instincts and one we easily assume. Further when individuals are influenced by this it creates a first order effect [1]. Intuitively this results in herd behaviour having a potentially

significant impact on economic variables whether it is voting patterns, crime, fashion or prices in financial markets. In this paper a multi-agent model of herd behaviour is constructed to analyse the dynamic process of expectation formation. In this model agent's expectations are formed from simple decision making rules within the self organisational framework [2, 3]. The core of the model is based on the social learning framework initially developed by Bikhchandani et al. [4] (here after referred to as BHW). The social learning takes place in a social network consistent with the work on small worlds by Watts [5].

The basic model consists entirely of herd agents who receive a signal, compare it with the expectations of other agents with whom they are connected and adopt the majority position. In the absence of heterogeneous decision making rules agents enter into an information cascade, learning stops and agents become fixed upon a given set of expectations. Heterogeneous decision making is introduced with the adding of expert agents, who are similar to the fashion leaders and experts discussed in BHW [4]. We find that the addition of one expert agent will be enough to enable the population to break the cascade, with information regarding changes in the state of the world filtering to the herd agents from the expert agents through contagion.

We also find that in an ordered network volatility in the aggregate level of agent expectations appears to increase linearly, but less than one to one, with the number of expert agents. Moving from an ordered to a small world network dramatically increases the level of volatility and it quickly reaches a higher level. At this point increasing the randomness of the network has little effect while increasing the number of experts has minimal effect. Increasing the number of connections has a significant effect independent of the small world properties. This provides some insight behind changes in the volatility in agent expectations over time.

Lastly we consider whether the structure of the social network can lead to instances when information cascades form in the presence of heterogeneous decision makers. We find that increasing the number of connections between herd agents creates an information cascade. This may explain the situation where agents continue to hold a view on the market (for example that the market remains in a bull run) despite evidence to the contrary. It can also provide a reason for their sudden collapse in confidence in a bull market where the state of the world had already changed but this information did not filter to herd agents until network connections decreased.

There are a number of approaches to modelling the process of expectation formation. For example Lux [6] and Brock et al. [7] use non linear dynamics to determine supply and demand and then close the model through an exogenous market maker. A second approach is through a Markov switching process [8]. A third approach introduces the concept of the social network whereby agents only communicate with, and see the actions and sometimes payoffs of, those agents in which they have a connection with. Therefore, in formulating their decision, agents use the experience of this subset of society, and possibly their own experiences, in updating their posterior using Bayes Law [9, 10]. The paper is

also related to the literature on Word-of-Mouth particularly Banerjee and Fudenberg [11] and Ellison and Fudenberg [12].

Conceptually this paper uses a similar approach to [9, 10] in analysing the impact of network architecture on both the long run and dynamic properties of the agent expectations. The point of differentiation is this model introduces the concepts of small worlds, which is then extended to examine the implications and influence of expert agents by varying the number and strength of connections from expert to other agents.

The paper is organised as follows. Section two outlines the model. The third and fourth sections examine the long run equilibrium and dynamic properties. The fifth section draws some conclusions and suggests areas of further work.

2   The model

The centrepiece of a model of herd behaviour is the coordination mechanism. It comprises of an observable signal, a social network and decision making rules. Consider the following. There are $i \in I = \{1,...,N\}$ agents. At the beginning of each round $t \in [1,,...,T]$ each agent receive a private binary signal $x \in X = \{0,1\}$ on the state of the world where 0 (1) represents an expectation that the stock that will fall (rise) in price in the next period. As an example this signal could take the form of a private belief based on learning from prices. Each agent i would then undertake a process to establish a view on how the market will perform in the next period. They do this by considering the signal they receive, as well as the most recent view taken by each of the other agents with which they have a connection. Agent i's signal is then adjusted in light of the discussions with connected agents and this becomes their view. It is this view that is presented to the market with the private signal never released.

2.1 Generating the signal

Agents do not know the true state of the world. Instead they form a posterior belief through a Bayesian learning process. Agents receive a private binary signal with a probability dependent on the state of the world $v \in V = \{0,1\}$. The agent's posterior probability that the true state of the world is $V = 1$ is given by:

$$P(V=1|X=1) = \frac{P(X=1|V=1) \cdot P(V=1)}{P(X=1|V=1) \cdot P(V=1) + P(X=1|V=0) \cdot P(V=0)} \tag{1}$$

The value of both the conditional likelihood function and the prior will need to be determined. There would be a variety of factors that would be considered in formulating a view on the future direction of an individual security (or even a market as represented by an index). It is also likely that these factors will differ between agents. Take the extreme positions of a fundamental verse a herd trader. For the former V is likely to represent if a stock is over or undervalued according to fundamental value, while for the latter V is more likely to represent whether

the market is in a bull or bear run. To complicate matters agents may not follow their own beliefs. For example agents may believe that stocks are overpriced but that the price will continue to rise in the next period [13].

In order to focus on the effects of social learning and network structure, rather than the Bayesian learning process a simplified framework is employed whereby agents have the following conditional likelihood functions and priors:

$$P(X = 1|V = 1) = P(X = 1|V = 0) = q > 0.5 \qquad (2)$$

$$P(X = 0|V = 0) = P(X = 0|V = 1) = 1 - q \qquad (3)$$

$$P(V = 1) = P(V = 0) = 0.5 \qquad (4)$$

2.2 The social network

The network consists of: a population of agents I in some finite social space; and a list of connections between agents initially defined as either 1 or 0. For any two individuals i and j a connection exists if $X(i, j) = 1$, otherwise $X(i, j) = 0$. In latter sections the strength between certain agents will be varied to replicate the case where the views of these agents (such as experts) hold more sway than other agents (thereby introducing the concept of 'social distance').

To develop the small world network each agent i is selected in turn along with the edge to the nearest neighbour in a clockwise sense. The connection is deleted and replaced with a random connection with a pre-determined probability p. Each agent goes thought this process until all agents have been assessed. The process then repeats itself for the next nearest neighbour if $k = 4$ and so on (see fig. 1 which is based on the work by Watts [5]). There is no social justification for a model that replaces one connection with another connection at random. However, in the world of stock market trading agents are just as likely to source information from unknown analysts via the web as to talk to neighbours, so the random approach may not be far from reality.

| $k = 2$ and $p = 0$ | $k = 4$ and $p = 0$ | $k = 2$ and $p > 0$ |
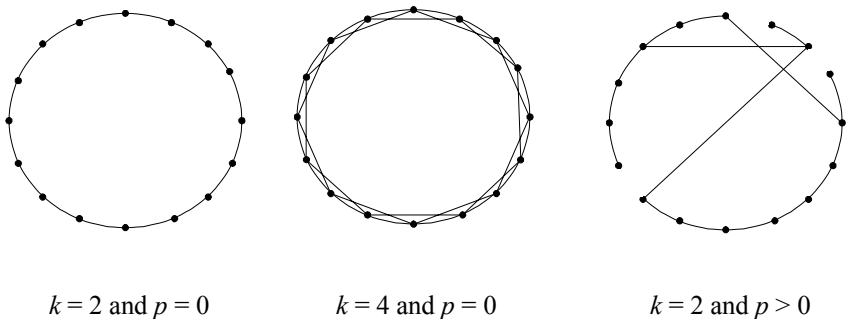
Figure 1:   Ring, small world and random graphs.

2.3 Decision making rule

In the first round each agent receives a signal according to eqn (1) and follows that signal. Therefore, the network does not impact on the expectations of agents in the first round. This is justified as the focus is on the stability of long run equilibria and the dynamics of steady state. At the end of the first round $t = 1$ agents have adopted an expectation x_i. Let X_i be the set of opinions of those agents connected to i. In the case of a ring lattice with $k = 2$ $X_i = (x_{i-1}, x_i, x_{i+1})$, where: x_{i-1} represents the expectation formed by $I - 1$ at time t, x_i represents the signal received by i at time t and x_{i+1} represents the expectation formed by $i + 1$ at time $t - 1$. The prior probability of V can now be updated by forming the posterior of V given the knowledge gained through conversation according to:

$$P_i(V|X_i) = \frac{P(X_i|V) \cdot P(V)}{P(X_i)} \tag{5}$$

Returning to the case of a ring lattice with $k = 2$, if both agents $I - 1$ and $i + 1$ formed an expectation that $V=0$ and i receives a signal $x = 1$ then:

$$P_i(V = 1|x_{i-1} = 0; x_i = 1; x_{i+1} = 0) = \frac{P(X_i|V) \cdot P(V)}{P(|X_i)} \tag{6}$$

$$=$$

$$\frac{P(x_i = 1|V = 1) \cdot P(V = 1|x_{i-1} = 0; x_{i+1} = 0)}{P(x_i = 1|V = 1) \cdot P(V = 1|x_{i-1} = 0; x_{i+1} = 0) + P(x_i = 1|V = 0) \cdot P(V = 0|x_{i-1} = 0; x_{i+1} = 0)} \tag{7}$$

Faced with this scenario and assuming that agents give equal weight to all X_i then, as $P(x_i = 1|V = 0) \cdot P(V = 0|x_{i-1} = 0; x_{i+1} = 0) > P(x_i = 1|V = 1) \cdot P(V = 1|x_{i-1} = 0; x_{i+1} = 0)$, they will ignore their own signal and update their prior so that the true state of the world is 0. The dynamic model becomes:

$$P_{i,t}(V_t|X_{i,t}) = \frac{P(X_{i,t}|V_t) \cdot P(V_t)}{P(X_{i,t})} \qquad X_{i,t} \subset X_t \tag{8}$$

where $X_t = \{x_{1,t}; ....; x_{i-1,t}; x_{i,t}; x_{i+1,t-1}; .....; x_{n,t-1}\}$

Agents update their decision sequentially but make repeated decisions. Further, in updating their prior, herd agents do not take into account their expectation formed in the previous round only the signals they receive from other agents. Essentially the agent starts each time period with a blank sheet of paper and a new signal. This can be justified in instances where the past does not matter (such as fads or fashion) or is captured in the state of the world and consequently in the signal obtained by the agents. For example, stock market prices incorporate past information with the only concern to agents being the future direction of the price.

This process does not mimic the types of conversations, and social learning, that occurs when individuals meet (for example there will be an element of joint decision making rather than agent i conferring with agent $I + 1$ prior to formulating a decision, then in turn $I + 1$ confers with i). However, what this approach does do is emphasise the effects of 'Chinese Whispers' where, because the communication is by word of mouth, hard evidence is not always provided [12]. The decision process also incorporates a form of 'public weighting' appropriate to such models.

3 Long run equilibrium

Consistent with the results of BHW [4] when the network consists entirely of herd agents, information becomes blocked and all learning ceases. For the purpose of undertaking the numerical analysis the following parameter values are used unless specified otherwise: $N = 200$, $q = 0.7$ and $k = 2$. In order to test the robustness of these results simulations are also run with $N = 100$ and $q = 0.6$ and 80 with no noticeable changes to the results.

We now examine the probability that a network consisting of 200 agents can avoid an information cascade after 900 rounds. 100 trials were run for each increment of q (noting that $q = 50$ represents the case where agents are following a random walk).
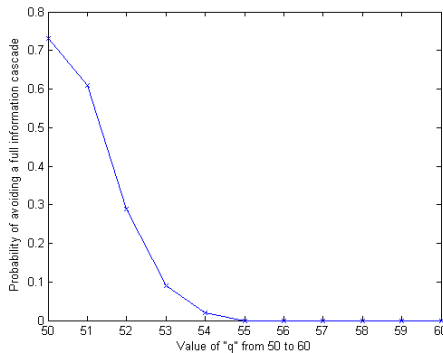
Figure 2:   Probability of avoiding cascades.

It confirms that an information cascade forms with a probability of one even for low q (i.e. $q = 50 + \varepsilon$). As agents follow their own signal in the first round the probability that agents cascade on the wrong state of the world is negligible. This is consistent with the results of Ellison and Fudenberg [12] which also adopts an exogenous initial state with agents making repeated decisions.

Expert agents add another dimension to the decision making process. Experts tend to be high precision individuals that are more inclined to use their own information rather than those that they come into contact with [4]. For the purpose of numerical analysis the expert agents are spaced evenly within the

network (so if there is one expert agent and $N = 200$, the 100th agent is an expert). Within the framework of BWH [4] this is equivalent to high precision individuals that make their decision later in the sequence. It is shown that, with the inclusion of one expert, agents always herd around the correct state of the world.

For $t < 300$, v is set exogenously to 0. As can be seen from fig. 3 agents quickly herd around $x = 0$. At $t = 301$ v is changed to 1 representing a structural change in the system. Within a short period of time agents switch their belief of v to 1 (i.e. all but a few agents hold that $x = 1$ at any point in time). At t = 601 v is again changed and the same result occurs.
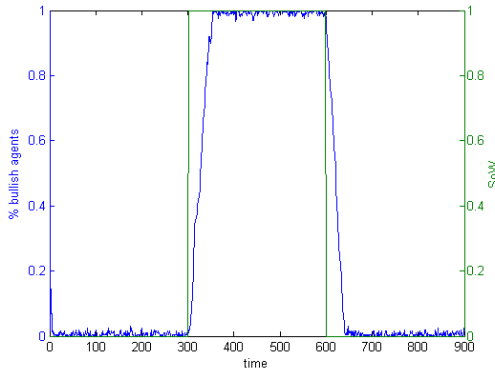
Figure 3:   One expert agent.

This outcome of the model has some similarity with that of BWH [4], in that the presence of an expert, when they appear later in the sequence, has the potential to break information cascades. In our model experts always break cascades, with the herd switching to the correct state of the world in finite time. Experts ensure that information always flows to all agents through contagion as they make decisions over time. Therefore, when the average number of connections are low ($k = 2$), the presence of expert agents means that there is no long term mispricing. There is some delay between the change in the state of the world and the ensuing shift in agent expectations. This may result in overshooting of prices. Nevertheless the agents' response to changes in the state of the world is quite rapid. Our simulations have shown that increasing the number of expert agents only shortens this lag. These results are consistent with Banerjee and Fudenberg [11] and Bala and Goyal [9].

4   Dynamic properties

4.1   Small world properties of the social network

Firstly we consider the level of volatility as you increase the level of randomness p and the number of expert agents. Volatility is measured as the standard

deviation with $\sigma = 1$. As can be seen from fig. 4, when p is approximately equal to 0 the number of experts affects the level of volatility in a linear fashion; steadily increasing from 0.1 when one expert is present to 0.05 when 10 expert agents are present. As the level of p increases the level of volatility rises sharply before reaching a plateau for $p > 1$ (as emphasised in fig. 4b which focuses on the range in p from 0 to 3). At this point, increases in either the number of expert agents, or the level of randomness (but holding k constant and equal to 2) has very little effect on the level of volatility. Assuming that $p > 1$ for all social networks then there is an inherent level of volatility in agent expectations. If individuals trading decisions are influenced by their expectations then this inherent level of volatility may in turn induce volatility in financial prices.

a

b
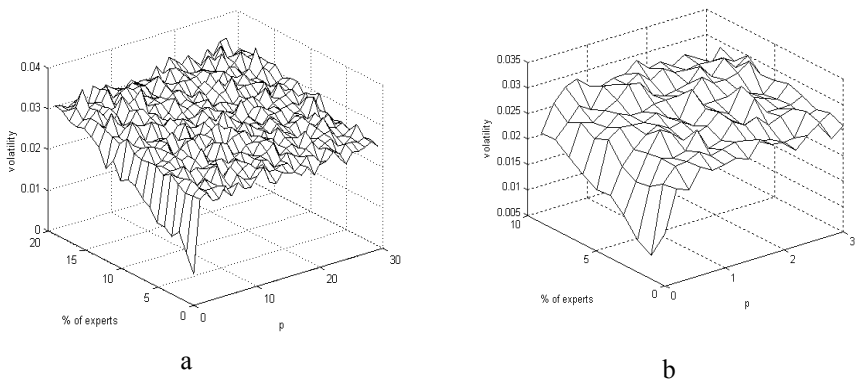
Figure 4:   Volatility vs. the number of experts and p.

4.2 The power of expert agents

As noted earlier expert agents are high precision individuals that tend to use their own information. However, experts also tend to have an increased influence over other agents. Experts are important because they provide valuable information particularly where that information is difficult to obtain or process or drawing conclusions is subjective. Two types of experts are considered in this paper. The first are experts that are well respected in the general community and are connected to many other agents in the network, such as Warren Buffet or Allan Greenspan. These are represented in the model as agents who have one way connections with many agents. The second type of agent is one whom is recognised locally as an expert. A good example of such an expert might be the local financial planner. In the model these agents have the same number of connections as the herd agent but the strength of their connections is increased.

   As can be seen from fig. 5, as you increase the number of connections from the experts volatility dramatically increases (three percent of agents are experts). Unlike the previous case where the number of expert agents is increased, the effect of increasing the number of connections persists for $p > 1$. Further the

volatility associated with this increases is in addition to the volatility due to the small world effect. These results suggest that volatility will be highest at times when experts are having the greatest effect, as measured by the number of connections, even though the average number of connections between all agents is not high. Doubling the strength of the connections from these experts increases the level of volatility, however, further increases have little effect (results not shown here). Therefore, any variation in the volatility of expectations can only be coming from an increase in the number of connections from expert agents.
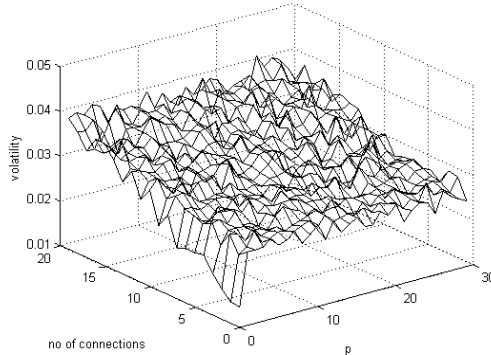
Figure 5:   Volatility as you increase the number of connections from expert agents.

A number of questions arise from this result: when is the influence of experts strongest and is volatility high during these periods? Intuitively, connections from experts are high (low) when faith in the market is strong (weak). At this point agents are at their most receptive to news about the stock market. If this is the case then prices might be most volatile when markets are rising.

4.3 Information cascades and bubbles

In the scenarios considered thus far herd behaviour increases the level of volatility in the market but does not lead to long run and significant mispricing. In what follows the number of connections between herd agents k is increased from two to four. Five percent of all agents are expert agents. It is found that when the network is ordered, agents enter into an information cascade (see fig. 6a). However, for $p \geq 1$ the cascade is broken and volatility decreases significantly (fig. 6b). When k is greater than four agents are always in an information cascade with a result similar to fig. 6a (not shown here).

It is therefore possible that under certain network structures herd behaviour can lead to information cascades. This locking of expectations could lead to the formation of speculative bubbles. Intuitively, as long as the average number of connections between agents is low information can flow within the social network. As the number of connections increase information flows become congested as the actions of other agents dominate their own private signal. The

surprising result here is that the number of connections per agent does not need to be large before information becomes blocked.

a   b
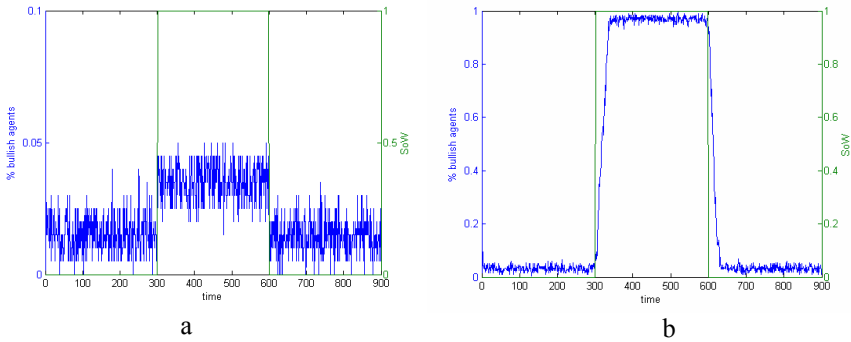
Figure 6:   Emergence of an information cascade.

Interestingly speculative bubbles in financial markets are characterised by excessive reporting in the media. It also dominates social discussions between neighbours or within the workplace. This could also explain the "bandwagon effect", where people exhibit herd behaviour out of fear of missing out on opportunities.

5   Conclusion

In this paper it is found that social interaction and herd behaviour, modelled in a multi-agent based framework, can explain the underlying volatility in agent expectations. It can also explain the variation in the level of volatility over time. Herd behaviour is often cited as one of the forces behind speculative price bubbles and crashes in stock markets. It is found that under certain network structures, where the number of connections between agents is increased, herd behaviour will lead to information cascades that have the potential to provide an explanation for the formation of speculative bubbles.

There are a number of potentially testable theories which arise from the work in this paper. Does volatility in agent expectations increase when communication from experts rises? Also, do bubbles occur during times when the number of connections between agents is high and is volatility high or low during these periods? There are also a number of extensions to the model including determining the impact of changing expectations on prices by incorporating a pricing mechanism. There is also growing empirical evidence that analysts herd in their recommendations, particularly inexperienced analysts [14, 15]. It would be useful to analyse this behaviour within the framework of a social network by linking the experts together in their own sub network.

Acknowledgement

References

[1]   Devenow, A. & Welch, I., Rational Herding in Financial Economics. *European Economic Review*, 40(3-5), pp. 603-15, 1996.

[2]   Foster, J., Competitive Selection, Self-Organisation and Joseph A. Schumpeter. *Journal of Evolutionary Economics*, 10(3), pp. 311-28, 2000.

[3]   Judd, K.L. & Tesfatsion, L., (eds). *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics,* Handbooks in Economics Series: North-Holland, forthcoming.

[4]   Bikhchandani, S., Hirshleifer, D. & Welch, I., A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades. *Journal of Political Economy*, 100(5), pp. 992-1026, 1992.

[5]   Watts, D.J., Small Worlds: *The Dynamics of Networks between Order and Randomness*, Princeton University Press: New Jersey 1999.

[6]   Lux, T., The Socio-Economic Dynamics of Speculative Markets: Interacting Agents, Chaos, and the Fat Tails of Return Distributions. *Journal of Economic Behavior and Organization*, 33(2), pp. 143-65, 1998.

[7]   Brock, W., Hommes, C. & Wagener, F., Evolutionary Dynamics in Markets with Many Trader Types. *Journal of Mathematical Economics*, 41(1-2), pp. 7-42, 2005.

[8]   Kijima, M. & Uchida, Y., A Markov model for valuing asset prices in a dynamic bargaining market. *Quantitative Finance*, 5(3), p. 277–88, 2005.

[9]   Bala, V. & Goyal, S., Learning from Neighbours. *Review of Economic Studies*, 65(3), pp. 595-621, 1998.

[10] Gale, D. & Kariv, S., Bayesian Learning in Social Networks. *Games and Economic Behavior*, 45(2), pp. 329-46, 2003.

[11] Banerjee, A. & Fudenberg, D., Word-of-Mouth Learning. *Games and Economic Behavior*, 46(1), pp. 1-22, 2004.

[12] Ellison, G. & Fudenberg, D., Word-of-Mouth Communication and Social Learning. *Quarterly Journal of Economics*, 110(1), pp. 93-125, 1995.

[13] Vissing-Jorgensen, A. Perspectives On Behavioral Finance: Does "Irrationality" Disappear With Wealth? Evidence From Expectations And Actions, 2 June 2003, Northwestern University, NBER and CEPR, US.

[14] Hwang, S. & Salmon, M., Market Stress and Herding, *Journal of Empirical Finance*, 11(4), pp. 585-616, 2004.

[15] Wylie, S., Fund Manager Herding: A test of the Accuracy of Empirical Results Using U.K. Data, *Journal of Business*, 78(1), pp. 381-403, 2005.

This page intentionally left blank

A Monte Carlo study for the temporal aggregation problem using one factor continuous time short rate models

Y. C. Lin
Queen Mary College, University of London, UK

Abstract

For most continuous time models formulated in finance, there is no closed form for the likelihood function and estimation of the parameters on the basis of discrete data will be based on an approximation rather than an exact discretization. For example, the Euler method introduces discretization bias because it ignores the internal dynamics that can be excessively erratic. We view the approximation as a difference equation and note that the solution of the continuous time model does not satisfy this difference equation. The effectiveness of the approximation will depend on the rate at which the underlying process is sampled. We investigate how much it matters: can we get significantly different estimates of the same structural parameter when we use say hourly data as compared with using monthly data under given discretization? If yes, then that discretization when applied to a data set in hand, as is done in practice, cannot be said to give robust results. We compare numerically the application of methods by Yu and Phillips (2001), Shoji and Ozaki (1998) and Ait-Sahalia (2002) in the maximum likelihood estimation of the unrestricted interest rate model proposed by Chan et al. (1992). We find that reducing the sampling rate yield large biases in the estimation of the parameters. The Ait-Sahalia method is shown to offer a good approximation and has the advantage of reducing some of the temporal aggregation bias.
Keywords: the discretization method.

1   Introduction

The purpose of the paper is to evaluate the performance of different discretization approximation to a structure continuous time model formulated as

a stochastic differential equation and show that the fact that the discretization approximation depends on the time interval. For most models formulated in continuous time, there is no closed form for the likelihood function and estimation of the parameters of the model on the basis of discrete data needs to be based on an approximate rather than an exact discretization. This has been one of the main issues in formulating and estimating interest rate diffusion models. For example, the discretization method used by Chan et al. [6] (CKLS, hereafter) is based on the Euler method. However, the Euler method introduces discretization bias because it ignores the internal dynamics that can be excessively erratic. It therefore motives the main emphasis will be on how to use the accurate restrictions to the data (the solution of the stochastic models) to study the econometric properties. Our model is specified as a simple first order stochastic differential equation system but we allow this system to be driven by a constant elasticity of volatility. This model is called the CKLS model in the literature of interest rates. The model considered represented some of the well known and most frequently used models in practice (Merton, 1973; Vaslek, 1977; CIR SD, 1985, the geometric Brownian motion (GBM) process of Black and Scholes, 1973). Our starting point is to view the discretization as a difference equation and to note that the solution of the continuous time model does *not* satisfy this difference equation when the discretization is not exact. This has major implications for estimation. With discrete time sampling, we must simulate a large number of sample paths along which the process is sampled very finely; otherwise, ignoring the difference generally results in inconsistent estimates, unless the discretization happens to be an exact one. This is the time aggregation problem inherent in the dichotomy between the time scale of the continuous time model and that of the observed data. As a result, the effectiveness of the discretization will vary depending on the rate at which the underlying process is sampled. Since the rate at which we sample the data matters when the discretization is approximated, we investigate how much it matters: can we get significantly different estimates of the same structural parameter when we use say hourly data as compared with using say monthly data under given discretization? If the answer is "yes", then that discretization when applied to a data set in hand, as is done in practice, cannot be said to give robust results. By Monte Carlo simulations and empirical study our aim is to investigate which approximation discretization is most robust to temporal aggregation for the interest rates we usually consider. We compare numerically the application of methods by Yu and Phillips [12], Shoji and Ozaki [11] and Ait-Sahalia [1] in the maximum likelihood estimation of the unrestricted interest rate model proposed by Chan et al. [6]. In this paper we look at the effects of systematic sampling, where we use observations picked out every n periods. For all estimation methods considered in this paper, we find that reducing the sampling rate will yield large biases in the estimation of the parameters. The Ait-Sahalia method is shown to offer a good approximation and has the advantage of reducing some of the temporal aggregation bias.

2   The model and the estimation methods

Following Chan et al. [6] (hereafter CKLS) a one dimensional continuous time specification of the interest rate is considered:

$$dx(t) = (\alpha + \beta x(t))dt + \sigma x(t)^r \, dB(t). \tag{2.1}$$

where $\{x(t), t > 0\}$ is the interest rate variable, $\alpha, \beta, \sigma,$ and γ are unknown structural parameters, $\{B_t, t \geq 0\}$ is a standard Brownian motion.

In practice, one could simulate a discretized process with a discretization step Δ. Then one might consider the estimator based on this approximated process. The Euler approximation to (2.1) is given by

$$x(t + \Delta) = x(t) + [\alpha + \beta x(t)]\Delta + u_{t+\Delta}^e \tag{2.2}$$

where $u_{t+\Delta}^e = \sigma x_t^\gamma \Delta B(t) = \sigma x_t^\gamma (B(t + \Delta) - B(t))$ is the disturbance term.

In principle, we can obtain more and more accurate discretization scheme including further stochastic terms from the stochastic Taylor expansion to the approximation scheme (2.2). This is because these stochastic terms contain additional information about the sample path of the Brownian motion. Despite this possibility, we need to stress the importance of the discretization scheme because neglect errors introduced as a result of time aggregation. Moreover, the approximation scheme (2.2) will not allow us to derive the exact maximum likelihood estimator. The Gaussian estimators will be consistent and asymptotically normal provided $\Delta \to 0$ or $N \to \infty$. The size of the approximation error in the discretized process is a function of the length of the discrete time interval. In other words, the approximation error is smaller for shorter time intervals. It is well known that ignoring this bias in the estimation process would give rise to inconsistent estimates of the model's parameters.

On the other hand, (2.1) could be interpreted as representing the integral equation:

$$x(t + \Delta) - x(0) = \int_0^{t+\Delta} [\alpha + \beta x(s)]ds + \int_0^{t+\Delta} \sigma x^\gamma dB(s), \ (t > 0). \tag{2.3}$$

For any initial value $x(0)$, the solution to model (2.1) is thus given by

$$x(t + \Delta) = \frac{\alpha}{\beta}(e^{\beta\Delta} - 1) + e^{\beta\Delta}x(t) + \int_0^\Delta e^{\beta(\Delta-\tau)}\sigma x^\gamma (t + \tau)dB(\tau). \tag{2.4}$$

Equation (2.4) is the exact discrete model. But, (2.4) cannot be used for estimation because the last term on the right hand involves the level of the process. Along the line of Bergstrom's method [3], Nowman [10] to assume that the volatility of the interest rate change at the beginning of the unit observation period and then remains constant and then apply the Bergstrom's method to estimate the parameters of interest. Let t' be the smallest integer greater than or equal to t, Nowman considers the following SDE:

$$dx(t) = (\alpha + \beta x(t))dt + \sigma x(t'-1)^r \, dB(t), t' \le t < t'+1. \qquad (2.5)$$

Then, following Bergstrom ([3], Theroem 2) the form of the corresponding exact discrete model of (2.1) can be expressed as:

$$x(t+1) = \tfrac{\alpha}{\beta}(e^\beta - 1) + e^\beta x(t) + \eta_t, t = 1,...,T, \qquad (2.6)$$

where $\eta_t (t = 1,...,T)$ is assumed to follow a normal and satisfies the conditions:

$$E(\eta_s \eta_t) = 0 \ s \ne t, \qquad (2.7)$$

$$E(\eta_t^2) = \int_{t-1}^{t} e^{2(t-\tau)\beta} \sigma^2 x^{2\gamma}(t-1)d\tau = \tfrac{\sigma^2}{2\beta}(e^{2\beta} - 1)x^{2\gamma}(t-1). \qquad (2.8)$$

Comparing to the approximation scheme (2.4), equation (2.6) allows us to use the exact maximum likelihood estimator. This should be help to reduce some of the temporal aggregation bias.

Also along the line of Bergstrom's method [3], Yu and Phillips employ the Dsmbis, Dubins-Schwarz (DDS) theorem and apply the time change formula to cover the residual processes to follow a Normal density. Let the last term in (2.5) be $M(\Delta)$ and it will be a continuous martingale with quadratic variation:

$$[M]_\Delta = \sigma^2 \int_0^\Delta e^{2\beta(\Delta-\tau)} x^{2\gamma}(t+\tau)d\tau.$$

Applying DDS theorem, Yu and Phillips transform $M(\Delta)$ to DDS Brownian motion. This method produces an exact discrete Gaussian model. Comparing to the Nowman's method, which is to equate the observation interval with the unit interval and to consider the exact discrete model on the sequence of the equi-spaced observations, the Yu and Phillips's method will cause a sequence to be non-equispaced observations.

Shoji and Ozaki [11] use the Ito formula to transform (2.1) as a diffusion process with a nonlinear drift term but a constant diffusion term. They use the local linear technique to approximate that new process. Basically, by the method of Shoji and Ozaki we will have a linear SDE as an approximation to any continuous diffusion, which allows us to derive the exact discretization of the continuous diffusion. The exact representation allows us to use the Bergstrom methods to estimate the parameters of a continuous time systems from discrete data.

Alternative estimation method that efficiently takes account of the time aggregation bias is Ait-Sahalia's method [1]. Comparing to the Shoji and Ozaki method, to simulate the discreted time observations of the process that is the solution of the locally linearization, Ait-Shalia approximates the unknown transition density function by Gaussian. Let $\theta = [\alpha, \beta, \sigma, \gamma]$. Also based on

the Ito formula, Ait-Shalia considers the new process $y(t)$ is observed at the time points $\{t = i\Delta, 0 \le i \le n\}$ for Δ is fixed and defines the increment of $y(t)$ as

$$z(t) = \Delta^{-1/2}(y(t) - y(0) - h(y(0); \theta)), \qquad (2.9)$$

where

$$h(y(t); \theta) = \alpha\sigma^{-1}(\sigma(1-\gamma)y(t))^{\frac{-\gamma}{1-\gamma}} + \beta(1-\gamma)y(t) - \tfrac{1}{2}\gamma(1-\gamma)y^{-1}(t).$$

Then, Ait-Sahalia [1] constructs the random variable $z(t)$ so that its density p_z can be close to a standard normal density. Following Ait-Sahalia [1] one can use the Hermite series expansion up to the J th term to approximate the density function p_z for fixed $\Delta, y(0), \theta$. One then can construct the approximation to the unknown density function for the diffusion process $x(t)$. Ait-Sahalia [1] proves that the density of the random variable $z(t)$ is close to the standard normal density and the approximation is close to the true density function of $x(t)$ when J $\to \infty$ but the sampling interval Δ remains fixed. Further, more and more accurate approximation to the true density can be obtained provided the order of approximation J gets larger and larger in this scheme. Comparing to the Euler scheme, we note that the sampling interval is not assumed that $\Delta \to 0$ in order to calculate the parameters explicitly.

In conclusion, when the sampling time interval is sufficiently small, one could expect that the approximation path for (2.1) by the Euler scheme would be close to the true trajectories such that these estimates of the parameters could converge to the true one. However, when the discretization step is observed equidistantly, then the estimates will show different performance depending on the frequency of the data. This is the problem of temporal aggregation in continuous time econometrics. To overcome this problem we would like to derive a discrete time model that will correspond exactly to the underlying continuous time process, in the sense that it generates exactly the same data at discrete points as does the continuous time model. We thus examine this problem of temporal aggregation by discussing several discretization schemes for the stochastic process (2.1) and estimation of the parameters of these discretized models. Basically, we extend the Monte Carlo results in Shoji and Ozaki [11] and Cleur [5]. Both studies only consider the effect of varying the frequency of the data on the estimation of parameters. But, Cleur [7] does not discuss the existence of the exact discretization of the diffusion equation in (2.1) that takes into account time aggregation bias. It is well known that ignoring this bias in the estimation process would give rise to inconsistent estimates of the model's parameters. In our empirical studies we will focus on the strategy of discretizing equation (2.1), which is the correct representation of the diffusion equation (2.1), by solving the stochastic differential equation and then discretizing the solution to this stochastic differential equation. See Nowman [10] and Yu and Phillips [12].

3 Monte Carlo results

Our results are follows.

1. When the frequency of the data is lower, for example using the monthly data, the estimates appear to converge toward a value away from the corresponding true value, particularly, inaccuracy of the estimates of α and β are quite impressive. This asymptotic bias is becoming increasing evident for all methods.

2. For the low frequency data (monthly or weekly data), the estimate of the parameters is biased and the rise in the frequency of the data will lead to an increase in the bias. In the simulation of daily data, the discretization bias is small by the use of Ait-Sahalia's J=3 method. This implies that discretization bias may not be very important as expected for Ait-Sahalia's J=3 method. This provides some evidences that high frequency data may not be particularly important.

3. In all cases, the biases are serious for empirically relevant of α We also find that the bias in for the estimator of the parameters α and β will translate into a serious bias for the diffusion parameters σ and γ Instead of the CIR model, we use the CKLS model to estimate parameters $[\alpha, \beta, \sigma, \gamma]$. We still use the CIR SR type process to generate the hourly data. Our outcome shows the estimates of σ and γ are sensitive to changes in α For example, using Ait-Shalais's method, γ is always downward biased and this is consistent with the upward bias in estimated α In magnitude, the downward bias for γ stays within the 2%. By contrast, σ is substantially upward biased. For the α=6.0 case, the percentage bias for σ in the worse case is large than 40% (using Ait-Sahalia's (J=2) method). To examine whether the bias of the estimator of γ is affected by other parameters, we show that the bias in the estimator of γ is indeed affected by the parameter α

4. We compare the MSEs between these three estimation methods by using 36000 simulated data. Ait-Sahalia J=3 method appears to be more efficient than other two methods. Hence, for a small sample size, the Ait-Sahalia's method would have efficiency gain because that method will produce a less bias and a less increase in standard errors.

5. After 1000 replications of the estimation procedure, we perform the Kolmogorov-Smirnov test to compare the distribution of the 1000 estimates. Our aim is to examine if these 1000 estimates come from the same distribution for two different sampling frequencies. The null hypothesis is that two samples come form the same distribution. We compare the distributions for hourly / daily, hourly / weekly, and hourly / monthly. Because hourly data can provide much precise confidence intervals, we can investigate the distorted effects by comparing if the sampling distribution of estimates for other sampling frequencies is far from that of estimates for hourly data. Hence, 84.8% for the Yu and Phillips method for hourly / daily data should be compared to one. Obviously, for the Yu and Phillips method and Ait-Sahalia's J=2 method, the rejection rates are too large. For example, for hourly / daily data under both methods the empirical rejection rates are

one almost for all cases. This means that the distributions for hourly and monthly data are not the same due to the effects of systematic sampling. We expect the distorted effects should increase as the extent of the data frequency decreases. Hence, for the shoji and Ozaki method and Ait-Sahalia's J=3 method, rejection rates are reasonable because the rates increase to one as the data move from hourly /daily to hourly / monthly. However, the test results reflect that fact that for the Yu and Phillips method and Ait-Sahalia's J=2 method, the serious distorted effects will occur even using high frequency data and therefore these two methods cannot effectively eliminate the biases.

In addition, as expected in the parameter estimation, our test results also show that, for Ait-Sahalia's J=3 method, the distorted effects are not as strong as the Yu and Phillips method. Hence, although Ait-Sahalia's J=3 method does not completely eliminate this sort bias it still can be expected relatively powerful on reduction of bias. Although we do not report here, it will be easy to find the reduction is not so obvious when using lower frequency monthly data, and the reduction will be much small the smaller the sample size and the greater the frequency of sampling. Also we show that there is little reduction in bias in using the higher frequency weekly data over and above monthly data, and there could be a substantial reduction in bias from using daily data for Ait-Sahalia's J=3 method.

The results by using the Kolmogorov-Smirnov test are consistent with the results using the Mann Whitney rank sum test to examine if the variances for two sampling frequencies are equivalent and the usual F test to examine if the means are equivalent. Also we report the CDF value for the Mann Whitney rank sum test and the usual F test. All of our cases in Tables are one, which means that we reject the null hypothesis that two samples come from the same distribution.

4   Empirical results

Six series of daily and monthly interest rates are used in the empirical study, including the Canada rates, the Germany rates and the US rates. Our goal is to determine the robustness of discretization methods to different sampling intervals. In addition to estimating the models using the entire daily and monthly samples, we also use the sampling scheme in the simulation to augment weekly and monthly observations with daily data. Then, we repeat the estimations using these observations. We estimate the real daily rates and real monthly rate. However, we take every 5 daily observations to be the weekly data and every 4 weekly data to be the monthly data, which forms our augmented monthly data in our Monte Carlo study. By using augmented monthly data, we show the Ait-Sahalia J=3 method produces estimates that are similar to the ones by using the real monthly data. Bu, this is not the case for the Yu and Phillips method. The Yu and Phillips method will produce seriously biased estimates when estimating α For example, using sampling scheme in our Monte Carlo study, the Yu and Phillips method will provide an estimate of α of 49.4331 for the Germany case, while it is 18.4542 for real monthly data. However, Ait-Sahalia's

J=3 method will provide a small estimate for *α* of 4.4694, which is more consistent with the estimate, 4.0620, for the real monthly data. For *σ* and *γ,* the performances of the Yu and Phillips method for the augmented monthly data and the real monthly data are similar to each other. This is because in the Yu and Phillips method the Nowman's procedure is used to estimate *σ* and *γ* This result shows that Ait-Sahalia's J=3 method has better performances than the Yu and Phillips method, consistent with the findings from the Monte Carlo study. Furthermore, by using augmented monthly data, Ait-Sahalia's J=3 method will produces a small estimate of *α* and a larger estimate of *β* comparing to the real monthly data. Also all methods show that there is a more distorted effect in the estimate of *α* comparing to the estimates of *β,* once again consistent with the findings from the Monte Carlo study. However, contrary to the findings in the Monte Carlo study, Ait-Sahalia's J=2 method does not results in more distorted effects comparing to the Shoji and Ozaki method and the Yu and Phillips method.

5   Conclusions

In this paper we compare the estimation performances for the continuous time short arte models. We investigate which approximation discretization is most robust to temporal aggregation for the interest rates we usually consider. We compare numerically the application of methods by Yu and Phillips [12], Shoji and Ozaki [11] and Ait-Sahalia [1] in the maximum likelihood estimation of the unrestricted interest rate model proposed by Chan et al. [6]. We find that reducing the sampling rate yield large biases in the estimation of the parameters. The Ait-Sahalia method is shown to offer a good approximation and has the advantage of reducing some of the temporal aggregation bias.

References

[1]   Ait-Sahalia, Y. (2002)   Maximum likelihood estimation of discrete sampled diffusion: A closed form approximation approach. *Economertica* 70, 223-262.

[2]   Bergstrom, A.R. (1983) Gaussian estimation of structure parameters in high – order continuous time dynamic models. *Econometrica* 51, 117-151.

[3]   Bergstrom, A.R. (1984) Continuous time stochastic models and issues of aggregation over time. In: Z. Griliches & M.D. Intriligator (eds), *Handbook of Econometrics*, pp1145-1212. Amsterdam: North-Holland.

[4]   Bergstrom, A.R. (1985) The estimation of parameters in non-stationary higher-order continuous time dynamic models, *Econometric Theory*, 1 369-385.

[5]   Bergstrom, A.R. (1986) The estimation of open higher-order continuous time dynamic models with mixed stock and flow data, *Econometric Theory*, 2 350-373.

[6]   Chan, K.C.G., G. Andrew Karolyi, Francis A Longstaff, and Anthony B Sanders (1992) An empirical comparison of alternative models of the short- term interest rate, Journal of finance 47, 1209-1227.

[7]   Cleur, E. M. (2001) Maximum likelihood estimates of a class of one-dimensional stochastic differential equation models from discrete data, *Journal of Time Series Analysis*, Vol. 22, No. 5, 505-516.

[8]   Cox, John C., Jonathan E. Ingersoll, and Stephen A.Ross (1985a), A Theory of the Term Structure of Interest Rates, *Econometrica,* vol. 53, 385-407.

[9]   McCrorie, J.R. (2000a) Deriving the Exact Discrete Analog of a Continuous Time System. *Econometric Theory* 16, 998-1015.

[10]   Nowman, K. (1997) Gaussian estimation of a single – factor continuous time models of the term structure of interest rate, *Journal of Finance*, 52, 1695 -1703.

[11]   Shoji, I. and Ozaki, T. (1997) Comparative study of estimation methods for continuous time stochastic processes, *Journal of Time Series Analysis*, 18, 485-506.

[12]   Yu, J. and P.C.B. Phillips (2001) Gaussian estimation of continuous time models of the short interest rate, *Econometrics Journal*, vol. 4, 2, 210-224.

This page intentionally left blank

Contingent claim valuation with penalty costs on short selling positions

O. L. V. Costa & E. V. Queiroz Filho
Departamento de Engenharia de Telecomunicações e Controle,
Escola Politécnica da Universidade de São Paulo, Brazil

Abstract

In this paper we consider a discrete-time finite sample space financial model with penalty costs on short selling positions. We start by presenting a necessary and sufficient condition for the non-existence of arbitrage opportunities. This reduces to the existence of a martingale measure for the case in which the penalties are zero. Next we consider the problem of contingent claim valuation. Our main result states that, under certain conditions, for every contingent there will be a seller price and a buyer price, with a perfect portfolio replication for each of them. Again when the penalty costs on short selling positions are zero, our conditions coincide with the traditional condition for the market to be complete. An explicit and constructive procedure for obtaining hedging strategies, not necessarily in the binomial framework, is presented.
Keywords: transaction costs, perfect replication, bid and ask option pricing.

1 Introduction

The general theory for contingent claim valuation considers that the prices for a buying position and a short selling position in a security are the same. However in practice these values are not the same, due to penalty costs on short selling positions. This penalty can been seen as a premium risk charged on a short selling position or on the way in which the bid and ask process affects the prices.

The subject of pricing derivatives with transaction costs and portfolio selection under transaction costs is of practical importance, and has been in evidence over the last years. Two types of transaction costs are considered; fixed costs, which are paid whenever there is a change of position, and proportional costs, which are charged according to the volume traded. Several different approaches to the prob-

lem of pricing derivatives with transaction costs and the portfolio choice problem under transaction costs can be found in the literature. In Edirisinghe et al. [1], the authors approach the problem by finding the least-cost replication strategy for hedging the pay off of contingent claims. In Davis et al. [2] the authors state the problem as an stochastic optimal control problem. In Leland [3] the author presents an alternative replicating strategy which depends on the size of transaction costs and frequency of revisions. As a result it presents a modified volatility that is incorporated in the Black-Scholes formula. In Boyle and Vorst [4] the authors construct a replicating strategy for the call option by embedding a given binomial model with proportional transaction costs into a complete Cox-Ross-Rubinstein [5] model with such costs. This approach can be seen as a discrete-time variant of the result derived by Leland [3], and was somewhat extended in Melnikov and Petrachenko [6]. Several other papers studied the problem considering, for instance, the theory of cones, mean-variance techniques, minimizing an expected discount loss function, etc (see, for instance, [7–10]). In comparison to these papers, our work gives an explicit and constructive procedure for obtaining hedging strategies, not necessarily in the binomial framework.

In this paper we consider a discrete-time finite sample space financial model with different prices for a buying and short selling position in the value of the portfolio. This is done by introducing penalty factors for the short selling position. Due to this, strategies are broken into H_i^+ and H_i^-, denoting the long and short positions respectively. In additional we introduce the concept of maximal trading strategy (see Definition 2.2). Moreover, unlike the standard theory, in our case the non-existence of arbitrage does not necessarily implies the non-existence of dominant strategies neither that the law of one price holds (see Pliska [11], p.10). Our definition of contingent claim consistently realizable (see Definition 2.7) rules out these situations, so that logical pricing can be obtained.

The paper is organized in the following way. Section 2 presents the model, the main definitions, and some preliminary results. In section 3 we present some results and definitions for the single -period case. In section 4 we present the multi-period case with an special attention to Theorem 4.3 that defines an algorithm to pricing contingent claims under penalty costs. The paper is concluded in section 5 analyzing the binomial case.

2 Definitions and preliminaries

Let m be a positive integer. The real m-dimensional vector space will be denoted by $\mathbb{R}^m$ and for $x \in \mathbb{R}^m$ we shall write x_i for the ith component of the vector x. We write $x \geq 0$ to denote that all components of x are positive, that is, $x_i \geq 0$ for $i = 1, \ldots, m$. The transpost of a vector or a matrix will be denoted by $\prime$. The vector formed by 1 in all components will be represented by e, and the vector with 1 in the ith component and 0 elsewhere by b^i. For a finite sample space Ω define $\mathbb{P}(\Omega)$ as the set of probability measures over Ω.

Let κ and N be positive integer numbers. We consider the following elements for the financial market.

i) Initial date $t = 0$ and terminal date $t = T$, with trading possible at any time t between these two dates.

ii) A finite sample space Ω with κ elements, that is, $\Omega = \{\omega_1, \ldots, \omega_\kappa\}$.

iii) A probability $P \in \mathbb{P}(\Omega)$ with $P(\omega) > 0$ for each $\omega \in \Omega$.

iv) A bank account process $B(t)$, $t = 0, \ldots, T$, with $B(0) = 1$, and $B(t)$, $t = 1, \ldots, T$, random variables on Ω with $B(t) \geq 1$.

v) A price process $S = \{S(t); t = 0, \ldots, T\}$, where $S(t)$ are N-dimensional positive random variables on Ω. $S_i(t)$ represents the price value of the ith security at time t.

vi) A filtration $\mathcal{F} = \{\mathcal{F}_t; t = 0, \ldots, T\}$ where each $\mathcal{F}_t$ represents the σ-field generated by the random vectors $\{S(0), \ldots, S(t)\}$ and random variables $\{B(0), \ldots, B(t)\}$.

vii) The penalty cost factors $\{\alpha_i(t); i = 0, 1, \ldots, N, t = 0, 1, \ldots, T\}$. These factors are related to the penalty costs that should be paid when holding a short selling position. When they are zero, there is no penalty cost, and the model reduces to the standard model.

Now, define the discounted price process $S^* = \{S^*(t); t = 0, \ldots, T\}$ as: $S_i^*(t) := \frac{S_i(t)}{B(t)}$, $i = 1 \ldots, N$; and for $t = 0, \ldots, T$, define $\Delta B(t) := B(t) - B(t-1)$, $\Delta S(t) := S(t) - S(t-1)$ and $\Delta S^*(t) := S^*(t) - S^*(t-1)$.

A trading strategy $H = (H(1), \ldots, H(T))$ describes an investor's portfolio from time $t = 0$ up to time $t = T$. Each $H(t)$ is a $(N+1, 2)$-dimensional random matrix with all components positive. Here it will be more convenient to represent the components of $H(t)$ as follows: $H_i^+(t) \geq 0$, $i = 0, \ldots, N$ denotes the $N+1$ components of the first column of the matrix $H(t)$, and $H_i^-(t) \geq 0$, $i = 0, \ldots, N$ denotes the $N+1$ components of the second column of the matrix $H(t)$. The elements $H_i^+(t)$ represent the buying position at the security i, while $H_i^i(t)$ represents the short selling position at the security i. We assume that each trading position $H(t)$, $t = 1, \ldots, T$, is $\mathcal{F}_{t-1}$-measurable, so that it is established by taking into account only the information available up to time $t - 1$.

Associated to a trading strategy H we have the value process $V := (V(0), \ldots, V(T))$ describing the total value of the portfolio at each time t. This can be written, at time $t = 0$, as

$$V(0) = (H_0^+(1) - H_0^-(1))B(0) + \sum_{i=1}^{N}(H_i^+(1) - H_i^-(1))S_i(0). \qquad (1)$$

and at times $t = 1, \ldots, T$, as

$$V(t) = (H_0^+(t) - H_0^-(t))B(t) + \sum_{i=1}^{N}(H_i^+(t) - H_i^-(t))S_i(t),$$
$$- \left(\alpha_0 H_0^-(t)B(t-1) + \sum_{i=1}^{N}\alpha_i H_i^-(t)S_i(t-1) \right). \qquad (2)$$

The quantity $V(t)$ represents the value of the portfolio at time t just before any change of ownership positions take place at that time. The penalty costs due to short selling positions are represented by $\alpha_0 H_0^-(t)B(t-1)$ and $\alpha_i H_i^-(t)S_i(t-1)$. The assumption here is that these costs are fixed at time $t-1$ as a percentual (α_i) of the value of the security ($B(t-1)$ or $S_i(t-1)$). If no short selling position is hold at the security i, $H_i^-(t) = 0$ and no cost is paid.

The value of the portfolio at time $t+1$ just after the change of ownership positions is

$$(H_0^+(t+1) - H_0^-(t+1))B(t) + \sum_{i=1}^{N}(H_i^+(t+1) - H_i^-(t+1))S_i(t). \quad (3)$$

We consider in this paper self-financing trading strategies, so that no money is added or withdrawn from the portfolio between times $t=0$ to time $t=T$. Any change in the portfolio's value is due to a gain or loss in the investments, and penalty costs due to the short selling positions. Thus (3) must coincide with $V(t)$, that is,

$$V(t) = (H_0^+(t+1) - H_0^-(t+1))B(t),$$
$$+ \sum_{i=1}^{N}(H_i^+(t+1) - H_i^-(t+1))S_i(t). \quad (4)$$

From eqns (1), (2) and (4) we have for $t = 0, \ldots, T-1$

$$V(t+1) = V(t) + (H_0^+(t+1) - H_0^-(t+1))\Delta B(t+1),$$
$$+ \sum_{i=1}^{N}(H_i^+(t+1) - H_i^-(t+1))\Delta S_i(t+1),$$
$$- \left(\alpha_0 H_0^-(t+1)B(t) + \sum_{i=1}^{N}\alpha_i H_i^-(t+1)S_i(t)\right). \quad (5)$$

We recall that the discounted process $V^*(t)$ is defined as: $V^*(t) := \frac{V(t)}{B(t)}$. From eqns (1), (2) and (4) we have for $t = 0, \ldots, T-1$,

$$V^*(t+1) = V^*(t) + \sum_{i=1}^{N}(H_i^+(t+1) - H_i^-(t+1))\Delta S_i^*(t+1),$$
$$- \left(\alpha_0 H_0^-(t+1)\frac{B(t)}{B(t+1)} + \sum_{i=1}^{N}\alpha_i H_i^-(t+1)\frac{S_i(t)}{B(t+1)}\right). \quad (6)$$

The following proposition is easily shown.

Proposition 2.1 *The following assertions are equivalent:*
  i) the trading strategy H is self-financing;

ii) the value process $V := (V(0), \dots, V(T))$ associated to a trading strategy H satisfies eqns (1) and (5);

iii) the value process $V := (V(0), \dots, V(T))$ associated to a trading strategy H satisfies eqns (1) and (6).

From eqn (2) we notice that any reasonable trading strategy H will be such that $H_i^+(t) \times H_i^-(t) = 0$ since otherwise, the investor would be holding, at the same time, a buying and short selling position at the security i, incurring on unnecessary payment of taxes. Therefore we introduce the following definition:

Definition 2.2 *We say that the trading strategy H is maximal if $H_i^+(t) \times H_i^-(t) = 0$ for every $i = 0, \dots, N$ and $t = 1, \dots, T$.*

We have the following result:

Proposition 2.3 *For any self-financing trading strategy H with associated value process V we can define a self-financing maximal trading strategy $\widetilde{H}$ with associated value process $\widetilde{V}$ such that $V(0) = \widetilde{V}(0)$ and $V(t) \le \widetilde{V}(t)$ for $t = 1, \dots, T$.*

Defining an appropriate recursive trading strategy $\widetilde{H}$ it is easy to verify that $H_i^+(t), H_i^-(t)$ are $\mathcal{F}_{t-1}$-measurable, $\widetilde{H}_i^+(t) \times \widetilde{H}_i^-(t) = 0$ so that $\widetilde{H}$ is a maximal trading strategy, $\widetilde{H}$ is self-financing, and that $V(t) \le \widetilde{V}(t)$, with $V(0) = \widetilde{V}(0)$.

Next we recall the definition of an arbitrage opportunity.

Definition 2.4 *We say that there is an arbitrage opportunity if for some self-financing maximal trading strategy H we have*

i) $V(0) = 0$,

ii) $V(T) \ge 0$, and

iii) $E(V(T)) > 0$.

The next proposition shows that we do not need to require the trading strategy to be maximal in the definition of an arbitrage.

Proposition 2.5 *There is a arbitrage opportunity if and only if for some self-financing trading strategy H we have i), ii) and iii) in Definition 2.4 verified.*

The main concern of this paper will be the problem of valuation of a contingent claim. We recall (see [11]) that a contingent claim is a random variable X representing a payoff at the final time T. We shall need the following definitions.

Definition 2.6 *We say that a contingent claim X is realizable if there exists a self-financing trading strategy H with associated value process V such that $X(\omega_j) = V(T, \omega_j)$ for every $j = 1, \dots, \kappa$. We say in this case that H is a replicating trading strategy for X.*

Definition 2.7 *We say that a contingent claim X is consistently realizable if X is realizable and for any replicating self-financing trading strategy H for X with associated value process V and any self-financing trading strategy $\overline{H}$ with associated value process $\overline{V}$ such that $X(\omega_j) \le \overline{V}(T, \omega_j)$ for every $j = 1, \dots, \kappa$ we have that if $X(\omega_j) < \overline{V}(T, \omega_j)$ for some $j = 1, \dots, \kappa$ then $\overline{V}(0) > V(0)$.*

Definition 2.8 *We say that a contingent claim X is maximally consistently realizable if X it is consistently realizable and there is a maximal replicating trading strategy H for X.*

3 The single-period case

In this section we present some results and definitions from the single-period model that we shall use later:

Theorem 3.1 *There are no arbitrage opportunities if and only if there exists a probability measure $\pi \in \mathbb{P}(\Omega)$ and a real number r such that*

i) $0 \le r \le \alpha_0 E_\pi[\frac{1}{B(1)}]$,

ii) $E_\pi[\Delta S_i^*] \le r S_i(0) \le E_\pi[\Delta S_i^* + \alpha_i \frac{S_i(0)}{B(1)}]$, *for* $i = 1, \ldots, N$,

iii) $\pi_j := \pi(\omega_j) > 0$, *for* $j = 1, \ldots, \kappa$.

Define the matrix A_1 as:

$$A_1 = \begin{pmatrix} 1 & S_1^*(1, \omega_1) & \ldots & S_N^*(1, \omega_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & S_1^*(1, \omega_\kappa) & \ldots & S_N^*(1, \omega_\kappa) \end{pmatrix}.$$

Theorem 3.2 *If $B(1) = 1 + r_f$ and A_1 has an inverse then every contingent claim X is maximally realizable. Moreover there exists a unique maximal trading strategy H that replicates X.*

Now, define the set $\mathbb{J} := \{a = (a_0, a_1, \ldots, a_N); a_i = +$ or $-\}$, and for $a \in \mathbb{J}$, $pos(a) = \{1 \le i \le N; a_i = +\}$, and $neg(a) = \{1 \le i \le N; a_i = -\}$. Note that the number of elements of $\mathbb{J}$ is 2^{N+1}.

Definition 3.3 *For $a \in \mathbb{J}$ such that $a_0 = +$, set*

$\Theta^a := \{\pi \in \mathbb{P}(\Omega);$

a) $\pi_j > 0, j = 1, \ldots, \kappa$,

b) $E_\pi[\frac{S_i(1)}{B(1)}] = S_i^*(0)$ *for* $i \in pos(a)$,

c) $E_\pi[\frac{S_i(1) + \alpha_i S_i(0)}{B(1)}] = S_i^*(0)$ *for* $i \in neg(a)\}$,

and for $a \in \mathbb{J}$ such that $a_0 = -$, set

$\Theta^a := \{\pi \in \mathbb{P}(\Omega);$

a) $\pi_j > 0, j = 1, \ldots, \kappa$,

b) $E_\pi[\frac{S_i(1) - \alpha_0 S_i(0)}{B(1)}] = S_i^*(0)$ *for* $i \in pos(a)$,

c) $E_\pi[\frac{S_i(1) + (\alpha_i - \alpha_0) S_i(0)}{B(1)}] = S_i^*(0)$ *for* $i \in neg(a)\}$.

Theorem 3.4 *If $B(1) = 1 + r_f$, A_1 has an inverse and for every $a \in \mathbb{J}$, $\Theta^a \ne \emptyset$ then every contingent claim X is maximally consistently realizable. Moreover the maximal replicating trading strategy H is unique.*

4 The multi-period case

For the multi-period case we follow the approach adopted in [11] by considering the information structure described by the sequence $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_T$ of partitions of Ω, with $\mathcal{P}_0 = \{\Omega\}$, $\mathcal{P}_T = \{\{\omega_1\}, \ldots, \{\omega_\kappa\}\}$, and satisfying the property that each $\mathcal{A} \in \mathcal{P}_t$ is equal to the union of some elements in $\mathcal{P}_{t+1}$ for every $t < T$ (see [11]). Let us write $\mathcal{P}_t = \{\mathcal{A}(t, 1), \ldots, \mathcal{A}(t, l_t)\}$, and we recall that $\mathcal{A}(t, i) \cap \mathcal{A}(t, j) = \emptyset, i \ne j$ and $\cup_{j=\ell}^{l_t} \mathcal{A}(t, \ell) = \Omega$.

For each $\mathcal{A}(t, \ell)$ let $\nu(t, \ell)$ be the number of sets $\mathcal{A}(t + 1, j)$ such that $\mathcal{A}(t + 1, j) \subseteq \mathcal{A}(t, \ell)$ (recall that $\mathcal{A}(t, \ell)$ is the union of some elements in $\mathcal{P}_{t+1}$). For each $\mathcal{A}(t + 1, j)$ such that $\mathcal{A}(t + 1, j) \subseteq \mathcal{A}(t, \ell)$, consider a representative element $\omega \in \mathcal{A}(t + 1, j)$, and define the set $\Lambda(t, \ell)$ formed by these elements. We set

$$\Lambda(t, \ell) := \{\varpi_1(t, \ell) \ldots, \varpi_{\nu(t,\ell)}(t, \ell)\}.$$

The following result can be proved following the same steps as in Pliska [11], pages 95-96, in conjunction with Theorem 3.1 but, we shall omit the details.

Theorem 4.1 *There are no arbitrage opportunities if and only if there exists a probability measure π and real number $r(t, \ell)$, $t = 0, \ldots, T - 1$, $\ell = 1, \ldots, l_t$, such that for every $\omega \in \mathcal{A}(t, \ell)$,*

   i) $0 \leq r(t, \ell) \leq \alpha_0 E_\pi[\frac{1}{B(t+1)}|\mathcal{F}_t](\omega)$,

   ii) for $i = 1, \ldots, N$,

$$E_\pi[\Delta S_i^*(t + 1)|\mathcal{F}_t](\omega) \leq r(t, \ell)S_i(t)(\omega),$$

$$\leq E_\pi\left[\Delta S_i^*(t + 1) + \alpha_i \frac{S_i(t)}{B(t+1)}|\mathcal{F}_t\right](\omega),$$

   iii) $\pi_j := \pi(\omega_j) > 0, \text{for } j = 1, \ldots, \kappa.$

For each $t = 0, \ldots, T - 1$ and $\ell = 1, \ldots, l_t$, define the matrices

$$A_1(t, \ell) =$$

$$\begin{pmatrix} 1 & S_1^*(1, \varpi_1(t, \ell)) & \ldots & S_N^*(1, \varpi_1(t, \ell)) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & S_1^*(1, \varpi_{\nu(t,\ell)}(t, \ell)) & \ldots & S_N^*(1, \varpi_{\nu(t,\ell)}(t, \ell)) \end{pmatrix}.$$

We have the following result, extending Theorem 3.2 to the multi-period case.

Theorem 4.2 *If for every $t = 0, \ldots, T - 1$ and $\ell = 1, \ldots, l_t$ we have $B(t + 1, \omega) = 1 + r_f(t, \ell)$ for every $\omega \in \mathcal{A}(t, \ell)$, and $A_1(t, \ell)$ has an inverse, then every contingent claim X is maximally realizable. Moreover there exists a unique maximal trading strategy H that replicates X.*

Proof. The basic idea in this proof is to move backward in time from $t = T$ to $t = 0$, and apply Theorem 3.2 for each single period t to $t + 1$ and each node $\mathcal{A}(t, \ell)$, $\ell = 1, \ldots, l_t$. From eqn (2) we have

$$X^* = V^*(T) = (H_0^+(T) - H_0^-(T)) + \sum_{i=1}^{N}(H_i^+(T) - H_i^-(T))S_i^*(T),$$

$$- \left(\alpha_0 H_0^-(t)\frac{B(T-1)}{B(T)} + \sum_{i=1}^{N}\alpha_i H_i^-(T)\frac{S_i(T-1)}{B(T)}\right). \quad (7)$$

Applying Theorem 3.2 for each one single period node $\ell = 1, \ldots, l_{T-1}$, and recalling that $B(T, \omega) = 1 + r_f(T - 1, \ell)$ for every $\omega \in \mathcal{A}(T - 1, \ell)$, we obtain

a unique maximal trading strategy $H(T)$ such that verifies (7). To have a self-financing trading strategy, we evaluate from (4)

$$V(T-1) = (H_0^+(T) - H_0^-(T))B(T-1),$$

$$+ \sum_{i=1}^{N} (H_i^+(T) - H_i^-(T))S_i(T-1). \tag{8}$$

By doing this, we obtain the value of $V(T-1)$ for each node $\ell = 1, \ldots, l_{T-1}$. We repeat now the procedure as in (7) to obtain a unique maximal trading strategy $H(T-1)$ that maximally replicates $V(T-1)$, and as in (8) to obtain the values of $V(T-2)$ for each node $\ell = 1, \ldots, l_{T-2}$. We carry on doing this up to time $t = 0$, ■

For each $t = 0, \ldots, T-1$, $\ell = 1, \ldots, l_t$, define $\Theta^a(t, \ell)$ as in Definition 3.3, replacing $\mathbb{P}(\Omega)$ by $\mathbb{P}(\mathcal{A}(t, \ell))$, $S_i(1)$ by $S_i(t+1)$, and $S_i(0)$ by $S_i(t)$. We have the following result, extending Theorem 3.4 to the multi-period case.

Theorem 4.3 *If for every $t = 0, \ldots, T-1$ and $\ell = 1, \ldots, l_t$ we have $B(t+1, \omega) = 1 + r_f(t, \ell)$ for every $\omega \in \mathcal{A}(t, \ell)$, $A_1(t, \ell)$ has an inverse, and for every $a \in \mathbb{J}$, $\Theta^a(t, \ell) \neq \emptyset$ then every contingent claim X is maximally consistently realizable. Moreover there exists a unique maximal self-financing trading strategy H that replicates X.*

Proof. Following the same idea as in the proof of Theorem 4.2, we move backward in time from $t = T$ to $t = 0$, and apply Theorem 3.4 for each one single period node $\ell = 1, \ldots, l_t$. For the single period $t = T-1$ to $t = T$ and each node $\ell = 1, \ldots, l_{T-1}$, we have from Theorem 3.4 that X is maximally consistently realizable. From (8) we get the values of $V(T-1)$ so that the strategy is self-financing. By repeating the same procedure for the single period $t = T-2$ to $t = T-1$ and each node $\ell = 1, \ldots, l_{T-2}$, we obtain from Theorem 3.4 that $V(T-1)$ is maximally consistently realizable. We carry on doing this up to the last single period $t = 0$ to $t = 1$. ■

Under the assumptions of Theorem 4.3 we have that there will be a seller price and a buyer price for each contingent claim. The seller price, denoted by $V_s(0)$, is obtained by applying to X the backward algorithm as presented in Theorem 4.2. The buyer price, denoted by $V_b(0)$, is obtained by applying the backward algorithm to $-X$, and taking $V_b(0) = -V(0)$.

Let us call $X_s(0)$ and $X_b(0)$ the seller and buyer prices respectively of X at time $t = 0$. If $X_s(0) > V_s(0)$ then one could sell the contract in the market at the price $X_s(0)$, and buy a replicant portfolio at the value $V_s(0)$, making a risk-free profit of $X_s(0) - V_s(0)$. At the final time T the portfolio will provide exactly the right value to settle the obligation on the contingent claim. Thus we have shown that if the market seller price $X_s(0)$ is bigger than $V_s(0)$, there will exist an arbitrage opportunity. Moreover, from the fact that X is consistently realizable and has a unique replicant trading strategy, no other portfolio will have a final value greater or equal to X with a initial value lower than $V_s(0)$. Thus the pricing $V_s(0)$ is logically consistent.

A similar conclusion can be constructed to the case $X_b(0) < V_b(0)$.

If $V_b(0) < X_b(0) \leq X_s(0) < V_s(0)$ then no arbitrage nor perfect hedging can be made.

5 Example

Let us consider the binomial model, which consists of a single risky security satisfying $S(t) = u^{N(t)}d^{t-N(t)}S(0)$, $t = 1, \ldots, T$, where $0 < d < 1 < u$ and $N = \{N(t); t = 1, \ldots, T\}$ is a binomial process with parameter p, $0 < p < 1$. The interest is assumed to be constant, so that $B(t) = (1 + r_f)^t$, $t = 0, 1, \ldots, T$. Let us obtain conditions that guarantee that every contingent claim X is maximally consistently realizable. It is easy to see that in this case $\mathbb{J} = \{(+, +), (+, -), (-, +), (-, -)\}$, and we have the following possibilities.

i) $(+, +)$; in this case,

$$\pi_1 = \frac{1 + r_f - d}{u - d}, \; \pi_2 = \frac{u - (1 + r_f)}{u - d}.$$

ii) $(+, -)$; in this case,

$$\pi_1 = \frac{1 + r_f - \alpha_1 - d}{u - d}, \; \pi_2 = \frac{u - (1 + r_f - \alpha_1)}{u - d}.$$

iii) $(-, +)$; in this case,

$$\pi_1 = \frac{1 + r_f + \alpha_0 - d}{u - d}, \; \pi_2 = \frac{u - (1 + r_f + \alpha_0)}{u - d}.$$

iv) $(-, -)$; in this case,

$$\pi_1 = \frac{1 + r_f + \alpha_0 - \alpha_1 - d}{u - d}, \; \pi_2 = \frac{u - (1 + r_f + \alpha_0 - \alpha_1)}{u - d}.$$

From above it is clear that the condition which guarantees that every contingent claim X is maximally consistently realizable is that $u > 1 + r_f + \alpha_0$ and $d < 1 + r_f - \alpha_1$. If this is satisfied, we have $0 < \pi_1 < 1$, $0 < \pi_2 < 1$ for all the four cases above.

Let us consider the following numerical example. Suppose that $S(0) = 5$, $u = \frac{4}{3}$, $d = \frac{8}{9}$, $\alpha_0 = \alpha_1 = \frac{1}{30}$, $r_f = \frac{1}{9}$. For this case we have $1 + r_f + \alpha_0 = \frac{103}{90} < u = \frac{4}{3}$, and $1 + r_f - \alpha_1 = \frac{97}{90} > d = \frac{8}{9}$, and every contingent claim X is maximally consistently realizable. Let us consider the following option: $X = \max\{S(2) - 5, 0\}$. By applying the backward procedure described in Theorem 4.2 we obtain that the seller price for X is $V_s(0) = 1.3272$, with the following heading strategy: $H_0^+(0) = 0$, $H_0^-(0) = 2.796$, $H_1^+(0) = 0.8246$, $H_1^-(0) = 0$, and for the case in which the risky security goes up, $H_0^+(1) = 0$, $H_0^-(1) = 3.932$, $H_1^+(1) = 1.0$, $H_1^-(1) = 0$, $V(1) = 2.2977$, while for the case in which it goes down, $H_0^+(0) = 0$, $H_0^-(0) = 1.4563$, $H_1^+(0) = 0.4687$, $H_1^-(0) = 0$ and $V(1) = 0.4652$.

By repeating the procedure now for $-X$ obtain that the buyer price for X is $V_b(0) = 0.9355$, with the following heading strategy: $H_0^+(0) = 2.6926$, $H_0^-(0) = 0$, $H_1^+(0) = 0$, $H_1^-(0) = 0.7256$, and for the case in which the risky security goes up, $H_0^+(1) = 4.23$, $H_0^-(1) = 0$, $H_1^+(1) = 0$, $H_1^-(1) = 1.0$, $V(1) = 1.9667$, while for the case in which it goes down, $H_0^+(0) = 1.5562$, $H_0^-(0) = 0$, $H_1^+(0) = 0$, $H_1^-(0) = 0.4688$ and $V(1) = 0.3542$. As expected, $V_b(0) = 0.9355 < V_s(0) = 1.3272$.

Acknowledgment

References

[1] Edirisinghe, C., Naik, V. & Uppal, R., Optimal replication of options with transactions costs and trading restrictions. *Journal of Financial and Quantitative Analysis*, **28(1)**, pp. 117138, 1993.

[2] Davis, M.H.A., Panas, V.G. & Zariphopoulou, T., European option pricing with transactions costs. SIAM *J Control Optim*, **34**, pp. 470493, 1993.

[3] Leland, H.E., Option pricing and replication with transactions costs. *The Journal of Finance*, **40(5)**, pp. 12831301, 1985.

[4] Boyle, P.P. & Vorst, T., Option replication in discrete time with transactions costs. *The Journal of Finance*, **47(1)**, pp. 271293, 1992.

[5] Cox, J.C., Ross, S.A. & Rubinstein, M., Option pricing: A simplified approach. *Journal of Financial Economics*, **7**, pp. 229263, 1979.

[6] Melnikov, A.V. & Petrachenko, Y.G., On option pricing in binomial market with transaction costs. *Finance and Stochastics*, **9**, pp. 141149, 2005.

[7] Bertsimas, D., Kogan, L. & Lo, A.W., Hedging derivative securities and incomplete markets: An ϵ arbitrage approach. *Operations Research*, **49(3)**, pp. 372397, 2001.

[8] Liu, H., Optimal consumption and investment with transactions costs and multiple risky assets. *The Journal of Finance*, **49(1)**, pp. 289331, 2004.

[9] Cvitanic, J., Minimizing expected loss of hedging in incomplete and constrained markets. *SIAM J Control Optim*, **38(4)**, pp. 10501066, 2000.

[10] Stettner, L., Option pricing in discrete-time incomplete market models. *Mathematical Finance*, **10**, pp. 305321, 2000.

[11] Pliska, S.R., *Introduction to Mathematical Finance*. Blackwell Publishers, 1997.

Geometric tools for the valuation of performance-dependent options

T. Gerstner & M. Holtz
Institut für Numerische Simulation, Universität Bonn, Germany

Abstract

In this paper, we describe several methods for the valuation of performance-dependent options. Thereby, we use a multidimensional Black–Scholes model for the temporal development of the asset prices. The martingale approach then yields the fair price as a multidimensional integral whose dimension is the number of stochastic processes in the model. The integrand is typically discontinuous, though, which makes accurate solutions difficult to achieve by numerical approaches. However, using tools from computational geometry we are able to derive a pricing formula which only involves the evaluation of smooth multivariate normal distributions. This way, performance-dependent options can efficiently be priced even for high-dimensional problems as is shown by numerical results.
Keywords: option pricing, multivariate integration, hyperplane arrangements.

1 Introduction

Performance-dependent options are financial derivatives whose payoff depends on the performance of one asset in comparison to a set of benchmark assets. Here, we assume that the performance of an asset is determined by the relative increase of the asset price over the considered period of time. The performance of the asset is then compared to the performances of a set of benchmark assets. For each possible outcome of this comparison, a different payoff of the derivative can be realized.

We use a multidimensional Black–Scholes model, see, e.g., Karatzas [1] for the temporal development of all asset prices required for the performance ranking. The martingale approach then yields a fair option price as a multidimensional integral whose dimension is the number of stochastic processes used in the model. In the so-called full model, the number of processes equals the number of assets.

In the reduced model, the number of processes can be smaller. Unfortunately, in either case there is no direct closed-form solution for these integrals. Moreover, the integrands are typically discontinuous which makes accurate numerical solutions difficult to achieve.

The main contribution of this paper is the derivation of closed-form solutions to these integration problems. For the reduced model, two novel tools from computational geometry are used. These tools are a fast enumeration method for the cells of a hyperplane arrangement and an algorithm for its orthant decomposition. The resulting closed-form solutions only involve the evaluation of smooth multivariate normal distributions which can be efficiently computed using numerical integration schemes which we illustrate in various numerical results.

2 Performance-dependent options

We assume that there are n assets involved in total. The considered asset gets assigned label 1 and the $n - 1$ benchmark assets are labeled from 2 to n. The price of the i-th asset varying with time t is denoted by $S_i(t), 1 \leq i \leq n$. All stock prices at the end of the time period T are collected in the vector $\mathbf{S} = (S_1(T), \ldots, S_n(T))$.

2.1 Payoff profile

First, we need to define the payoff of a performance-dependent option at time T. To this end, we denote the relative price increase of stock i over the time interval $[0, T]$ by $\Delta S_i := S_i(T)/S_i(0)$. We save the performance of the first asset in comparison to a given strike price K (often $K = S_1(0)$) and in comparison to the benchmark assets at time T in a ranking vector $\mathrm{Rank}(\mathbf{S}) \in \{+, -\}^n$ defined by

$$\mathrm{Rank}_1(\mathbf{S}) = \begin{cases} + & \text{if } S_1 \geq K, \\ - & \text{else} \end{cases} \quad \text{and } \mathrm{Rank}_i(\mathbf{S}) = \begin{cases} + & \text{if } \Delta S_1 \geq \Delta S_i, \\ - & \text{else} \end{cases}$$

for $i = 2, \ldots, n$. In order to define the the payoff of the performance-dependent option we require bonus factors $a_{\mathbf{R}}$ which determine the bonus for each possible ranking $\mathbf{R} \in \{+, -\}^n$, see Section 5 for example profiles. In all cases we set $a_{\mathbf{R}} = 0$ if $\mathbf{R}_1 = -$ which corresponds to the option characteristic that a non-zero payoff only occurs if the stock price if above the strike.

The payoff of the performance-dependent option at time T is then defined by

$$V(S_1, T) = a_{\mathrm{Rank}(\mathbf{S})} \left(S_1(T) - K \right). \tag{1}$$

In the following, we aim to determine the fair price $V(S_1, 0)$ of such an option at the current time $t = 0$.

2.2 Multivariate Black–Scholes model

We assume that the stock prices are driven by $d \leq n$ stochastic processes modeled by the system of stochastic partial differential equations

$$dS_i(t) = S_i(t) \left(\mu_i dt + \sum_{j=1}^{d} \sigma_{ij} dW_j(t) \right) \tag{2}$$

for $i = 1, \ldots, n$, where μ_i denotes the drift of the i-th stock, σ the $n \times d$ volatility matrix of the stock price movements and $W_j(t), 1 \leq j \leq d$, the corresponding Wiener processes. The matrix $\sigma\sigma^T$ is assumed to be positive definite. If $d = n$, we call the corresponding model full, if $d < n$, the model is called reduced.

By Itô's formula we get the explicit solution of (2) by

$$S_i(T) = S_i(\mathbf{X}) = S_i(0) \exp \left(\mu_i T - \bar{\sigma}_i + \sqrt{T} \sum_{j=1}^{d} \sigma_{ij} X_j \right) \tag{3}$$

for $i = 1, \ldots, n$ with $\bar{\sigma}_i := \frac{1}{2}(\sigma_{i1}^2 + \ldots + \sigma_{id}^2) T$ and $\mathbf{X} = (X_1, \ldots, X_d)$ being a $N(\mathbf{0}, \mathbf{I})$-normally distributed random vector.

3 Pricing formula in the full model

We now derive the price of a performance-dependent option as a multivariate integral in the case that the number of stochastic processes d equals the number of assets n.

3.1 Martingale approach

Using the usual Black–Scholes assumptions, the option price $V(S_1, 0)$ is given by the discounted expectation

$$V(S_1, 0) = e^{-rT} E[V(S_1, T)] \tag{4}$$

of the payoff under the unique equivalent martingale measure. To this end, the drift μ_i in (3) is replaced by the riskless interest rate r for each stock i. Plugging in the density function $\varphi(\mathbf{x}) := \varphi_{\mathbf{0},\mathbf{I}}(\mathbf{x})$ of the random vector $\mathbf{X}$ (note that $\mathbf{S} = \mathbf{S}(\mathbf{X})$), we get that the fair price of a performance-dependent option with payoff (1) is given by the d-dimensional integral

$$V(S_1, 0) = e^{-rT} \int_{R^d} \sum_{\mathbf{R} \in \{+,-\}^n} a_{\mathbf{R}} (S_1(T) - K) \chi_{\mathbf{R}}(\mathbf{S}) \varphi(\mathbf{x}) \, d\mathbf{x} \tag{5}$$

where the characteristic function $\chi_{\mathbf{R}}(\mathbf{S})$ is defined to be equal to one if $\mathrm{Rank}(\mathbf{S}) = \mathbf{R}$ and zero else.

3.2 Pricing formula

Now, we aim to derive an analytical expression for the computation of (5) in terms of smooth functions. To proof our main theorem we need the following two lemmas. For the first Lemma, we denote by $\varphi_{\mu,\mathbf{C}}(\mathbf{x})$ the Gauss kernel with mean μ and covariance matrix $\mathbf{C}$ and by $\Phi(\mathbf{C}, \mathbf{b})$ the multivariate normal distribution corresponding to $\varphi_{\mathbf{0},\mathbf{C}}$ with limits $\mathbf{b} = (b_1, \ldots, b_d)$.

Lemma 3.1 *Let* $\mathbf{b}, \mathbf{q} \in \mathbb{R}^d$ *and* $\mathbf{A} \in \mathbb{R}^{d \times d}$ *with full rank, then*

$$\int_{\mathbf{Ax} \geq \mathbf{b}} e^{\mathbf{q}^T \mathbf{x}} \varphi(\mathbf{x}) d\mathbf{x} = e^{\frac{1}{2} \mathbf{q}^T \mathbf{q}} \Phi(\mathbf{A}\mathbf{A}^T, \mathbf{A}\mathbf{q} - \mathbf{b}).$$

Proof: A simple computation shows that $e^{\mathbf{q}^T \mathbf{x}} \varphi(\mathbf{x}) = e^{\frac{1}{2}\mathbf{q}^T \mathbf{q}} \varphi_{\mathbf{q},I}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$. Using the substitution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y} + \mathbf{q}$ we obtain

$$\int_{\mathbf{Ax} \geq \mathbf{b}} e^{\mathbf{q}^T \mathbf{x}} \varphi(\mathbf{x}) d\mathbf{x} = e^{\frac{1}{2}\mathbf{q}^T \mathbf{q}} \int_{\mathbf{y} \geq \mathbf{b} - \mathbf{Aq}} \varphi_{\mathbf{0},\mathbf{AA}^T}(\mathbf{y}) \, d\mathbf{y}$$

and thus the assertion. $\qquad \square$

For the second Lemma, we define a comparison relation for two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with respect to the ranking $\mathbf{R}$ by $\mathbf{x} \geq_{\mathbf{R}} \mathbf{y} :\Leftrightarrow R_i(x_i - y_i) \geq 0$ for $1 \leq i \leq n$.

Lemma 3.2 *We have* $Rank(\mathbf{S}) = \mathbf{R}$ *exactly if* $\mathbf{AX} \geq_{\mathbf{R}} \mathbf{b}$ *with*

$$\mathbf{A} := \sqrt{T} \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \sigma_{11} - \sigma_{21} & \cdots & \sigma_{1d} - \sigma_{2d} \\ \vdots & & \vdots \\ \sigma_{11} - \sigma_{n1} & \cdots & \sigma_{1d} - \sigma_{nd} \end{pmatrix}, \quad \mathbf{b} := \begin{pmatrix} \ln \frac{K}{S_1(0)} - rT + \bar{\sigma}_1 \\ \bar{\sigma}_1 - \bar{\sigma}_2 \\ \vdots \\ \bar{\sigma}_1 - \bar{\sigma}_n \end{pmatrix}.$$

Proof: Using (3) we see that $Rank_1 = +$ is equivalent to

$$S_1(T) \geq K \iff \sqrt{T} \sum_{j=1}^{d} \sigma_{1j} X_j \geq \ln \frac{K}{S_1(0)} - rT + \bar{\sigma}_1$$

which yields the first row of the system $\mathbf{AX} \geq_{\mathbf{R}} \mathbf{b}$. Moreover, for $i = 2, \ldots, n$, the outperformance criterion $Rank_i = +$ can be written as

$$S_1(T)/S_1(0) \geq S_i(T)/S_i(0) \iff \sqrt{T} \sum_{j=1}^{d}(\sigma_{1j} - \sigma_{ij})X_j \geq \bar{\sigma}_1 - \bar{\sigma}_i$$

which yields rows 2 to n of the system. $\qquad \square$

Now we can state the following pricing formula which, in a slightly more special setting, is originally due to Korn [2].

Theorem 3.3 *The price of a performance-dependent option with payoff (1) is for the model (2) in the case $d = n$ given by*

$$V(S_1, T) = \sum_{\mathbf{R} \in \{+,-\}^n} a_{\mathbf{R}} \left(S_1(0)\, \Phi(\mathbf{A_R A_R^T}, -\mathbf{d_R}) - e^{-rT} K \Phi(\mathbf{A_R A_R^T}, -\mathbf{b_R}) \right)$$

where $(\mathbf{b_R})_i := \mathbf{R}_i \mathbf{b}_i$, $(\mathbf{d_R})_i := \mathbf{R}_i \mathbf{d}_i$ and $(\mathbf{A_R})_{ij} := \mathbf{R}_i \mathbf{A}_{ij}$ with $\mathbf{A}$ and $\mathbf{b}$ defined as in Lemma 3.2. Furthermore, $\mathbf{d} := \mathbf{b} - \sqrt{T}\mathbf{A}\sigma_1$ with σ_1^T being the first row of the volatility matrix σ.

Proof: The characteristic function $\chi_{\mathbf{R}}(\mathbf{S})$ in the integral (5) can be eliminated using Lemma 3.2 and we get

$$V(S_1, 0) = e^{-rT} \sum_{\mathbf{R} \in \{+,-\}^n} a_{\mathbf{R}} \int_{\mathbf{Ax} \geq_{\mathbf{R}} \mathbf{b}} (S_1(T) - K)\varphi(\mathbf{x})d\mathbf{x}. \qquad (6)$$

By (3), the integral term can be written as

$$S_1(0)e^{rT - \bar{\sigma}_1} \int_{\mathbf{Ax} \geq_{\mathbf{R}} \mathbf{b}} e^{\sqrt{T}\sigma_1^T \mathbf{x}} \varphi(\mathbf{x})d\mathbf{x} - K \int_{\mathbf{Ax} \geq_{\mathbf{R}} \mathbf{b}} \varphi(\mathbf{x})d\mathbf{x}.$$

Application of Lemma 3.1 with $\mathbf{q} = \sqrt{T}\sigma_1$ shows that the first integral equals

$$e^{\frac{1}{2}\mathbf{q}^T\mathbf{q}} \int_{\mathbf{y} \geq_{\mathbf{R}} \mathbf{b} - \mathbf{Aq}} \varphi_{0,\mathbf{AA^T}}(\mathbf{y})\, d\mathbf{y} = e^{\bar{\sigma}_1} \int_{\mathbf{y} \geq \mathbf{d_R}} \varphi_{0,\mathbf{A_R A_R^T}}(\mathbf{y})\, d\mathbf{y} = e^{\bar{\sigma}_1} \Phi(\mathbf{A_R A_R^T}, -\mathbf{d_R}).$$

By a further application of Lemma 3.1 with $\mathbf{q} = \mathbf{0}$ we obtain that the second integral equals $K\Phi(\mathbf{A_R A_R^T}, -\mathbf{b_R})$ and thus the assertion holds. $\qquad \square$

4 Pricing formula in the reduced model

The pricing formula of Theorem 3.3 allows a stable and efficient valuation of performance-dependent options in the case of moderate-sized benchmarks. If the number n of benchmark assets is large, the high number 2^n of terms and the high dimension of the required normal distributions prevents an efficient application of the pricing formula, however. In this Section, we will derive a similar pricing formula for the reduced model which incorporates less processes than companies ($d < n$). This way, substantially fewer rankings have to be considered and much lower-dimensional integrals have to be computed.

4.1 Geometrical view

Lemma 3.2 and thus representation (6) remains also valid in the reduced model. Note, however, that $\mathbf{A}$ is now an $(n \times d)$-matrix which prevents the direct application of Lemma 3.1. At this point, a geometrical point of view is advantageous to illustrate the effect of performance comparisons in the reduced model.
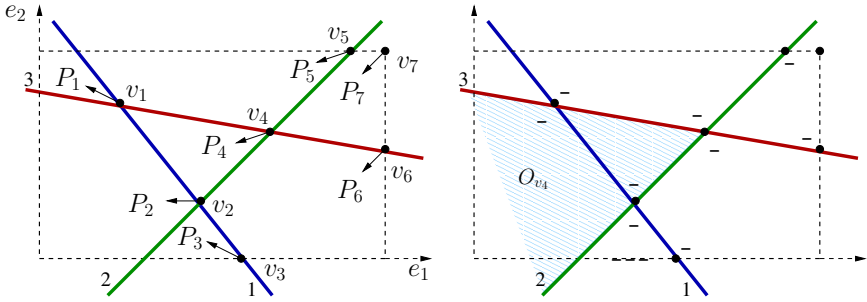
Figure 1: Illustration of the mapping between intersection points $\{\mathbf{v}_1, \ldots, \mathbf{v}_7\}$ and polyhedral cells $P_j := P_{\mathbf{v}_j}$ for a hyperplane arrangement $\mathcal{A}_{3,2}$ (left) and corresponding reflection signs $s_{\mathbf{v},\mathbf{w}}$ as well as the orthant $O_{\mathbf{v}_4}$ (right).

The matrix $\mathbf{A}$ and the vector $\mathbf{b}$ define a set of n hyperplanes in the space $\mathbb{R}^d$. Its dissection into different cells is called a hyperplane arrangement and denoted by $\mathcal{A}_{n,d}$. Each cell in $\mathcal{A}_{n,d}$ is a (possibly open) polyhedron P which can uniquely be represented by a ranking vector $\mathbf{R} \in \{+, -\}^n$. Each element of the ranking vector indicates on which side of the corresponding hyperplane the polyhedral cell is located. Each polyhedron has the representation $P = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{A}\mathbf{x} \geq_{\mathbf{R}} \mathbf{b}\}$.

As the number of cells in the hyperplane arrangement $\mathcal{A}_{n,d}$ is much smaller than 2^n if $d < n$ (see Edelsbrunner [3]), we can significantly reduce the number of integrals which have to be computed by identifying all cells in the hyperplane arrangement. This way, (6) can be rewritten as

$$V(S_1, 0) = e^{-rT} \sum_{P \in \mathcal{A}} a_{\mathbf{R}} \int_P (S_1(T) - K)\varphi(\mathbf{x})d\mathbf{x}. \qquad (7)$$

4.2 Tools from computational geometry

Looking at (7), two problems remain: first, it is not easy to identify which ranking vectors appear in the hyperplane arrangement; second, the integration region is now a general polyhedron which requires involved integration rules. To resolve these difficulties, we need some more utilities from computational geometry.

First, we choose a set of linearly independent directions $\mathbf{e}_1, \ldots, \mathbf{e}_d \in \mathbb{R}^d$ to impose an order on all points in $\mathbb{R}^d$. Thereby, we assume that no hyperplane is parallel to any of the directions. Moreover, we suppose that the hyperplane arrangement is non-degenerate which means that exactly d hyperplanes intersect in each vertex. Using the directions $\mathbf{e}_i$, an artificial bounding box which encompasses all vertices can be defined (see Figure 1, left). This bounding box is only needed for the localization of the polyhedral cells in the following Lemma and does not implicate any approximation.

Lemma 4.1 *Let the set $\mathcal{V}$ consist of all interior vertices, of the largest intersection points of the hyperplanes with the bounding box and of the largest corner point of*

the bounding box. Furthermore, let $P_{\mathbf{v}} \in \mathcal{A}_{n,d}$ be the polyhedron which is adjacent to the vertex $\mathbf{v} \in \mathcal{V}$ and which contains no other vertex which is larger than $\mathbf{v}$ with respect to the direction vectors. Then the mapping $\mathbf{v} \mapsto P_{\mathbf{v}}$ is one-to-one and onto.

Such a mapping is illustrated in Figure 1 (left). The proof of Lemma 4.1 can be found in our paper [4]. Using Lemma 4.1, an easy to implement optimal order algorithm can be constructed to enumerate all cells in a hyperplane arrangement.

Note that by Lemma 4.1 each vertex $\mathbf{v} \in \mathcal{V}$ corresponds to a unique cell $P_{\mathbf{v}} \in \mathcal{A}_{n,d}$ and thus to a ranking vector $\mathbf{R}$. We can, therefore, also assign bonus factors to vertices by setting $a_{\mathbf{v}} := a_{\mathbf{R}}$. Next, we assign to each vertex $\mathbf{v}$ an associated orthant $O_{\mathbf{v}}$. An orthant is defined as an open region in $\mathbb{R}^d$ which is bounded by $k \leq d$ hyperplanes. To find the orthant associated with the vertex $\mathbf{v}$, we look at k backward (with respect to the directions $\mathbf{e}_i$) points by moving $\mathbf{v}$ backwards on each of the k intersecting hyperplanes. The unique orthant which contains $\mathbf{v}$ and all backward points is denoted by $O_{\mathbf{v}}$. By definition, there exists a $(k \times d)$-submatrix $\mathbf{A}_{\mathbf{v}}$ of $\mathbf{A}$ and a k-subvector $\mathbf{b}_{\mathbf{v}}$ of $\mathbf{b}$ such that the orthant $O_{\mathbf{v}}$ can be characterised as the set

$$O_{\mathbf{v}} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{A}_{\mathbf{v}}\mathbf{x} \geq_{\mathbf{R}} \mathbf{b}_{\mathbf{v}}\}, \tag{8}$$

where $\mathbf{R}$ is the ranking vector which corresponds to $\mathbf{v}$. Furthermore, given two vertices $\mathbf{v}, \mathbf{w} \in \mathcal{V}$, we define the reflection sign $s_{\mathbf{v},\mathbf{w}} := (-1)^{r_{\mathbf{v},\mathbf{w}}}$ where $r_{\mathbf{v},\mathbf{w}}$ is the number of reflections on hyperplanes needed to map $O_{\mathbf{w}}$ onto $P_{\mathbf{v}}$ (see Figure 1, right). Finally, let $\mathcal{V}_{\mathbf{v}}$ denote the set of all vertices of the polyhedron $P_{\mathbf{v}}$.

Lemma 4.2 *It is possible to algebraically decompose any cell of a hyperplane arrangement into a signed sum of orthant cells by*

$$\chi(P_{\mathbf{v}}) = \sum_{\mathbf{w} \in \mathcal{V}_{\mathbf{v}}} s_{\mathbf{v},\mathbf{w}} \chi(O_{\mathbf{w}}),$$

where χ is the characteristic function of a set. Moreover, all cells of a hyperplane arrangement can be decomposed into a signed sum of orthants using exactly one orthant per cell.

The first part of Lemma 4.2 is originally due to Lawrence [5], the second part can be found in [4].

4.3 Pricing formula

Now, we are finally able to give a pricing formula for performance-dependent options also for the reduced model.

Theorem 4.3 *The price of a performance-dependent option with payoff (1) is for the model (2) in the case $d \leq n$ given by*

$$V(S_1, 0) = \sum_{\mathbf{v} \in \mathcal{V}} c_{\mathbf{v}}(S_1(0)\Phi(\mathbf{A}_{\mathbf{v}}\mathbf{A}_{\mathbf{v}}^T, -\mathbf{d}_{\mathbf{v}}) - e^{-rT}K\Phi(\mathbf{A}_{\mathbf{v}}\mathbf{A}_{\mathbf{v}}^T, -\mathbf{b}_{\mathbf{v}}))$$

with $\mathbf{A_v}, \mathbf{b_v}$ as in (8) and with $\mathbf{d_v}$ being the corresponding subvector of $\mathbf{d}$. The weights $c_\mathbf{v}$ are given by

$$c_\mathbf{v} := \sum_{\mathbf{w}\in\mathcal{V}:\ \mathbf{v}\in P_\mathbf{w}} s_{\mathbf{v},\mathbf{w}} a_\mathbf{w}.$$

Proof: By Lemma 4.1 we see that the integral representation (7) is equivalent to a summation over all vertices $\mathbf{v} \in \mathcal{V}$, i.e.

$$V(S_1,0) = e^{-rT} \sum_{\mathbf{v}\in\mathcal{V}} a_\mathbf{v} \int_{P_\mathbf{v}} (S_1(T) - K)\varphi(\mathbf{x})d\mathbf{x}.$$

By Lemma 4.2 we can decompose the polyhedron $P_\mathbf{v}$ into a signed sum of orthants and obtain

$$V(S_1,0) = e^{-rT} \sum_{\mathbf{v}\in\mathcal{V}} a_\mathbf{v} \sum_{\mathbf{w}\in\mathcal{V}_\mathbf{v}} s_{\mathbf{v},\mathbf{w}} \int_{O_\mathbf{w}} (S_1(T) - K)\varphi(\mathbf{x})d\mathbf{x}.$$

By the second part of Lemma 4.2 we know that only $c_{n,d}$ different integrals appear in the above sum. Rearranging the terms leads to

$$V(S_1,0) = e^{-rT} \sum_{\mathbf{v}\in\mathcal{V}} c_\mathbf{v} \int_{O_\mathbf{v}} (S_1(T) - K)\varphi(\mathbf{x})d\mathbf{x}.$$

Since now the integration domains $O_\mathbf{v}$ are orthants, Lemma 3.1 can be applied exactly as in the proof of Theorem 3.3 which finally implies the Theorem. □

5 Numerical results

In this Section, we present numerical examples to illustrate the use of the pricing formula from Theorem 4.3. In particular, we compare the efficiency of our algorithm to the standard pricing approach (denoted by STD) of quasi-Monte Carlo simulation of the expected payoff (4) based on Sobol point sets, see, e.g., Glasserman [6]. We systematically compare the numerical methods

- Quasi-Monte Carlo integration based on Sobol point sets (QMC),

- Product integration based on the Clenshew Curtis rule (P), and

- Sparse Grid integration based on the Clenshew Curtis rule (SG)

for the evaluation of the multivariate cumulative normal distributions (see Genz [7]). The Sparse Grid approach is based on [8]. All computations were performed on an Intel(R) Xeon(TM) CPU 3.06GHz processor. We consider a reduced Black–Scholes market with $n = 30$ assets and $d = 5$ processes. Thereby, we investigate two different choices for the bonus factors $a_\mathbf{R}$ in the payoff function (1):
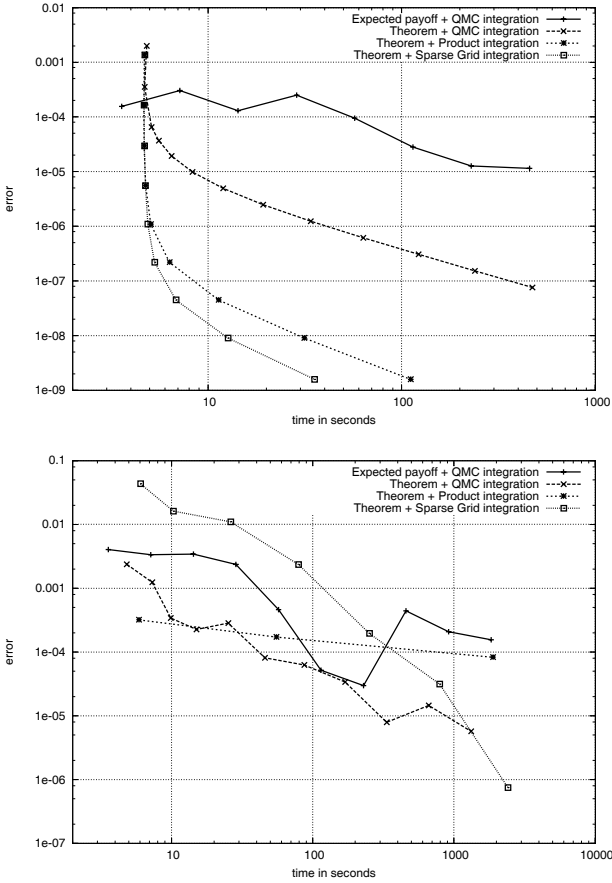
Figure 2: Errors and timings of the different numerical approaches to price the performance-dependent options of Examples 5.1 (top) and 5.2 (bottom).

Example 5.1 *Ranking-dependent option:*

$$a_{\mathbf{R}} = \begin{cases} m/(n-1) & \text{if } \mathbf{R}_1 = + \\ 0 & else, \end{cases}$$

where m denotes the number of outperformed benchmark assets. If the company ranks first there is a full payoff $(S_1(T) - K)^+$. If it ranks last the payoff is zero.

Example 5.2 *Outperformance option:*

$$a_{\mathbf{R}} = \begin{cases} 1 & \text{if } \mathbf{R} = (+, \ldots, +) \\ 0 & else. \end{cases}$$

A payoff only occurs if $S_1(T) \geq K$ and if all benchmark assets are outperformed.

In both cases, we use the following model parameters: $K = 100$, $S_1(0) = 100$, $T = 1$, $r = 5\%$; σ is a 30×5 volatility matrix whose entries are uniformly distributed in $[-1/d, 1/d]$.

Depending on the specific choice of bonus factors, it turns out that often many weights $c_\mathbf{v}$ are zero in the formula of Theorem 4.3 which reduces the number of required normal distributions. Furthermore, all vertices $\mathbf{v}$ located on the boundary of the bounding box correspond to orthants which are defined by $k < d$ intersecting hyperplanes. For these vertices, only a k-dimensional normal distribution has to be computed. In Example 5.1, we have 41 integrals with maximum dimension 2 while in Example 5.2, 31 integrals with maximum dimension 5 arise.

The convergence behaviour of the four different approaches (STD, QMC, P, SG) to price the options from the Examples 5.1 and 5.2 is shown in Figure 2. There, the time is displayed which is needed to obtain a given accuracy. One can see that the standard approach (STD) quickly achieves low accuracies. The convergence rate is slow and clearly lower than one, though. The integration scheme suffers under the irregularity of the integrand which is highly discontinuous and not of bounded variation. The QMC scheme clearly outperforms the STD approach in all examples. It exhibits a convergence rate of about one and leads to significantly smaller errors. As expected, the product integration approach (P) performs only really well in the Example 5.1 which is of low intrinsic dimension. The combination of Sparse Grid integration with our pricing formula (SG) leads to the best convergence rates. However, for higher dimensional problems as Example 5.2, this advantage is only visible if very accurate solutions are required. In the pre-asymptotic regime, the QMC scheme leads to smaller errors.

Acknowledgement

References

[1] Karatzas, I., *Lectures on the Mathematics of Finance*, volume 8 of *CRM Monograph Series*. American Mathematical Society: Providence, R.I., 1997.

[2] Korn, R., A valuation approach for tailored options. Working paper, 1996.

[3] Edelsbrunner, H., *Algorithms in Combinatorial Geometry*. Springer, 1987.

[4] Gerstner, T. & Holtz, M., The orthant decomposition of hyperplane arrangements, 2006. In preparation.

[5] Lawrence, J., Polytope volume computation. *Math Comp*, **57(195)**, pp. 259–271, 1991.

[6] Glasserman, P., *Monte Carlo Methods in Financial Engineering*. Springer, 2003.

[7] Genz, A., Numerical computation of multivariate normal probabilities. *J Comput Graph Statist*, **1**, pp. 141–150, 1992.

[8] Gerstner, T. & Griebel, M., Numerical integration using sparse grids. *Numerical Algorithms*, **18**, pp. 209–232, 1998.

Optimal exercise of Russian options in the binomial model

R. W. Chen[1] & B. Rosenberg[2]
[1] *Department of Mathematics, University of Miami, Coral Gables, Florida, USA*
[2] *Department of Computer Science, University of Miami, Coral Gables, Florida, USA*

Abstract

The Russian option is a two-party contract which creates a liability for the option seller to pay the option buyer an amount equal to the maximum price attained by a security over a specific time period, discounted for the option's age. The Russian option was proposed by Shepp and Shiryaev. Kramkov and Shiryaev first examined the option in the binomial model. We improve upon their results and give a near-optimal algorithm for price determination.

Specifically, we prove that the optimal exercising boundary is monotonic and give an $O(N)$ dynamic programming algorithm to construct the boundary, where N is the option expiration time. The algorithm also computes the option's value at time zero in time $O(N)$ and the value at all of the $O(N^3)$ nodes in the binomial model in time $O(N^2)$.

Keywords: Russian option, binomial model, dynamic programming.

1 Introduction

The Russian Option is a two-party contract which creates a liability for the option seller to pay the option buyer an amount equal to the maximum price attained by a security over a specific time period, discounted for the option's age. For an $N + 1$ step time period $0, 1, 2, \ldots, N$, the option seller's liability at time step n, $0 \leq n \leq N$, is,

$$L(n) = \beta^n \max_{0 \leq t \leq n} s_t$$

where s_t is the security price at time t and β is the discount factor. In this paper we consider the value of this option under a standard binomial model of security prices, and give efficient algorithms for value calculation.

The Russian option was proposed by Shepp and Shiryaev [1]. At this time it is not traded. Their work gives the optimal expected present value and the optimal exercise strategy under the Black-Scholes market model. Kramkov and Shiryaev [2] first examined the option in the binomial model of Cox et al. [3]. They present an $O(N^2)$ algorithm for calculating option price at the first time step.

This work gives an $O(N)$ algorithm determining the option price at all time steps as well as optimal execution and the execution boundary. We also present an $O(N^2)$ algorithm for general determination of option value given option structure and security price history up to time n, $0 \leq n \leq N$.

2 Definitions and basic facts

The binomial model, introduced by Cox et al. [3], assumes discrete price announcements at equal time intervals with each price related to the previous price by either an up-step or down-step, according to a random process. For s_i the security price at time i, the price process is given by,

$$s_{i+1} = u^{\epsilon_i} s_i, \quad \epsilon_i \in \{1, -1\},$$

with $u > 1$. The probability of an up-step, $\epsilon_i = 1$, is p, independent of i. The probability of a down-step is $q = 1 - p$. The existence of a risk-free bond is also assumed,

$$b_{i+1} = (1 + r)b_i,$$

where $r > 0$ is the bond's interest rate.

The rational markets theory stipulates that the price sequence s_i is a martingale,

$$E(s_{i+1} \mid s_i) = (1 + r)E(s_i).$$

This determines the martingale measure for the random process,

$$p = \frac{u(1 + r) - 1}{u^2 - 1}.$$

Note that this implies $u \geq (1 + r)$, that is, that the risky security must return at least the risk-free rate in order that the martingale measure exist.

The option value and liability depends only on the current time step n, the current security price s_n, and the maximum value s_n^* attained by the security in the time period 0 up to n. The current and maximum price can be expressed as integers j and k such that $s_n = u^j s_0, s_n^* = u^k s_0$. Without loss of generality we assume $s_0 = 1$. Hence the process can be modeled as a graph V whose nodes are 3-tuples (n, j, k), indicating time step n, current price u^j and maximum price u^k, and whose edges indicate up-steps and down-steps labeled with probabilities p and q, respectively.

$$(n, j+1, \max(k, j+1))$$

$$p \nearrow$$

$$(n, j, k)$$

$$q \searrow$$

$$(n, j-1, k)$$
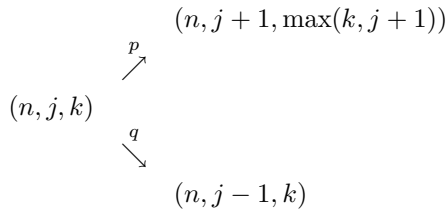
Figure 1: Example subgraph of V.

The liability at any node $(n, j, k) \in V$ is $L(n, j, k) = \beta^n u^k$. At each time step but the last, the option's owner can either exercise and receive the liability or hold. The expected value of the option at node (n, j, k) is therefore given by the backwards recurrence,

$$E(n, j, k) = \max(\beta^n u^k, \alpha(pE(n+1, j+1, \max(k, j+1)) + qE(n+1, j-1, k)))$$

where $\alpha = 1/(1 + r)$ is the discount to present value for one time step. At the last time step, the owner must exercise. This gives the boundary condition $E(N, j, k) = \beta^N u^k$, for all j and k.

This recurrence defines values only for those nodes reachable in the graph V starting from node $(0, 0, 0)$. These are called *accessible nodes.* Inaccessible nodes are of no importance and their values are ignored.

Lemma 1 (Node accessibility) *The nodes (n, j, k) is accessible if and only if,*

$$0 \le k \le n \le N, \quad -n \le j \le k, \text{ and } j + n = 2(k + i)$$

for some non-negative integer i.

Proof: Let e_u be the number of up-steps and e_d be the number of down steps,

$$n = e_u + e_d, \ j = e_u - e_d, \text{ therefore } n + j = 2e_u = 2(k + i).$$

The integer i is the number of up-steps which do not contribute to attaining the maximum k. Given appropriate n, j, k and i, access the node (n, j, k) by first taking k up-steps, then $n - k - i$ down-steps, and finally i up-steps. Since $n + j \le 2n$, then $k + i \le n$. Therefore the construction is well defined.

We assume that $\beta < 1$, else the incentive to hold the option is too strong. The recurrence insures that for accessible nodes, $E(n, j, k) \ge \beta^n u^k$. If it is not true that for accessible nodes this inequality is strict when $j = k$, then the incentive to hold the option would be too weak. Assuming the contrary,

$$\beta^n u^k < E(n, k, k)$$
$$= \alpha(pE(n+1, k+1, k+1) + qE(n+1, k-1, k))$$
$$< \alpha(p\beta^{n+1}u^{k+1} + q\beta^{n+1}u^k).$$

Using the martingale measure for p, this reduces to the following constraints on β,

$$\frac{(1+u)(1+r)}{u(2+r)} < \beta < 1$$

We have thus proved the following lemma,

Lemma 2 (Option viability) *Let* $(1 + u)(1 + r)/(u(2 + r)) < \beta < 1$, p *be the martingale measure, and* $(n, k.k)$ *be an accessible node. Then* $\beta^n u^k < E(n, k, k)$ *for* $n < N$.

For the remainder of this paper, we will assume that β is as required for option viability and p is the martingale measure.

Lemma 3 (Option monotonicity) *Assuming nodes are accessible,* $E(n, j, k) \leq E(n, j', k)$ *if* $j \leq j'$ *and* $E(n, j, k) \leq E(n, j, k')$ *if* $k \leq k'$.

Proof: Use induction starting at $n = N$ and working towards smaller n. Note that for (n, j, k) and (n, j', k) to both be accessible, $j' - j$ must be even.

3 Analysis of the Russian option

We first prove some technical theorems and they apply them to determine the exercise boundary. Finally, an efficient algorithm is given to determine the boundary.

3.1 Induction theorems concerning option value

Theorem 1 (First induction theorem) *Suppose* (n, j, k) *is accessible and* l *is an integer satisfying* $0 \leq k + 2l \leq n$. *Then* $(n, j + 2l, k + 2l)$ *is accessible and* $u^{2l} E(n, j, k) = E(n, j + 2l, k + 2l)$.

Proof: We begin by proving that if (n, j, k) is accessible and l is an integer $0 \leq k + 2l \leq n$ then $(n, j + 2l, k + 2l)$ is accessible.

Reduce to the case $l = 1$. Hence $k + 2 \leq n$. Since (n, j, k) is accessible, $0 \leq k \leq n \leq N$, $-n \leq j \leq k$ and $n + j = 2(k + i)$ for a non-negative integer i. In fact, because $k + 2 \leq n, i$ must be positive. Therefore $0 \leq k + 2 \leq n \leq N$, $-n \leq j + 2 \leq k + 2$ and $n + j + 2 = 2(k + 2 + i')$ where $i' = i - 1 \geq 0$.

We now prove the equality. Reduce to the case $l = 1$ and proceed by induction on n. For $n = N$, the result is immediate, since $E(N, j, k) = \beta^N u^k$.

Assume the theorem for $n + 1$. We first consider the case $E(n, k, k)$. The option viability lemma allows us to insert and remove the $\max()$ operation in the

following calculation,

$$
\begin{aligned}
u^2 E(n, k, k) &= u^2 \max\big(\beta^n u^k, \alpha(pE(n+1, k+1, k+1) \\
&\qquad + qE(n+1, k-1, k))\big) \\
&= u^2\big(\alpha(pE(n+1, k+1, k+1) + qE(n+1, k-1, k))\big) \\
&= \alpha\big(pE(n+1, k+3, k+3) + qE(n+1, k+1, k+2)\big) \\
&= \max\big(\beta^n u^{k+2}, \alpha(pE(n+1, k+3, k+3) \\
&\qquad + qE(n+1, k+1, k+3))\big) = E(n, k+2, k+2).
\end{aligned}
$$

We consider the final case, $E(n, j, k)$ where $j < k$,

$$
\begin{aligned}
u^2 E(n, j, k) &= u^2 \max\big(\beta^n u^k, \alpha(pE(n+1, j+1, \max(k, j+1)) \\
&\qquad + qE(n+1, j-1, k))\big) \\
&= u^2 \max\big(\beta^n u^k, \alpha(pE(n+1, j+1, k) + qE(n+1, j-1, k))\big) \\
&= \max\big(\beta^n u^{k+2}, \alpha(pE(n+1, j+3, k+2) \\
&\qquad + qE(n+1, j+1, k+2))\big) \\
&= \max\big(\beta^n u^{k+2}, \alpha(pE(n+1, j+3, \max(k+2, j+3)) \\
&\qquad + qE(n+1, j+1, k+2))\big) = E(n, j+2, k+2).
\end{aligned}
$$

This completes the induction and the proof.

Theorem 2 (Second induction theorem) *Suppose (n, j, k) is accessible and l is an integer satisfying $(N - n) \geq l \geq 0$. Then $(n+l, j+l, k+l)$ is accessible and $(\beta u)^l E(n, j, k) \geq E(n+l, j+l, k+l)$.*

Proof: We begin by proving that if (n, j, k) is accessible and l is an integer $0 \leq l \leq N - n$ then $(n+l, j+l, k+l)$ is accessible.

Reduce to the case $l = 1$. Hence $n + 1 \leq N$. Since (n, j, k) is accessible, $0 \leq k \leq n \leq N$, $-n \leq j \leq k$ and $n + j = 2(k + i)$ for a non-negative integer i. Therefore $0 \leq k + 1 \leq n + 1 \leq N$, $-n \leq j + 1 \leq k + 1$ and $n + 1 + j + 2 = 2(k + 1 + i)$.

We now prove the inequality. We reduce to the case $l = 1$ and proceed by induction on n. The similarity with the proceeding proof allows us to omit some steps.

For $n = N - 1$,

$$
\begin{aligned}
\beta u E(N-1, j, k) &= \beta u \max\big(\beta^{N-1} u^k, \alpha(pE(N, j+1, \max(k, j+1)) \\
&\qquad + qE(N, j-1, k))\big) \\
&\geq \beta u(\beta^{N-1} u^k) = E(N, j+1, k+1).
\end{aligned}
$$

Assume the theorem for $n + 1$. We first consider the case $E(n, k, k)$,

$$\beta u E(n, k, k) = \beta u \max \big(\beta^n u^k, \alpha(pE(n + 1, k + 1, k + 1)$$
$$+ qE(n + 1, k - 1, k)) \big)$$
$$\geq \alpha \big(pE(n + 2, k + 2, k + 2) + qE(n + 2, k, k + 1) \big)$$
$$= E(n + 1, k + 1, k + 1).$$

We consider the final case, $E(n, j, k)$ where $j < k$,

$$\beta u E(n, j, k) = \beta u \max \big(\beta^n u^k, \alpha(pE(n + 1, j + 1, \max(k, j + 1)) +$$
$$qE(n + 1, j - 1, k)) \big)$$
$$\geq \max \big(\beta^{n+1} u^{k+1}, \alpha(pE(n + 2, j + 2, k + 1)$$
$$+ qE(n + 2, j, k + 1)) \big) = E(n + 1, j + 1, k + 1).$$

This completes the induction and the proof.

Theorem 3 (Third induction theorem) *Suppose (n, j, k) is an accessible node with $k > 0$. Then $(n, j-2, k-1)$ is accessible and $uE(n, j-2, k-1) \leq E(n, j, k)$.*

Proof: We begin by proving that if (n, j, k) is accessible and $k > 0$ then $(n, j - 2, k - 1)$ is accessible.

Since (n, j, k) is accessible, $0 \leq k \leq n \leq N$, $-n \leq j \leq k$ and $n+j = 2(k+i)$ for a non-negative integer i. Since $k > 0$ then $n + j - 2 \geq 0$. Therefore $0 \leq k - 1 \leq n \leq N$, $-n \leq j - 2 \leq k - 1$ and $n + j - 2 = 2(k - 1 + i)$.

We now prove the inequality. The proof is by induction on n. For $n = N$ the result is immediate.

Assume the theorem for $n+1$. We first consider the case $E(n, j, k)$ where $j < k$,

$$uE(n, j - 2, k - 1) = u \max \big(\beta^n u^{k-1}, \alpha(pE(n + 1, j - 1, k - 1)$$
$$+ qE(n + 1, j - 3, k - 1)) \big)$$
$$\leq \max \big(\beta^n u^k, \alpha(pE(n + 1, j + 1, k)$$
$$+ qE(n + 1, j - 1, k)) \big) = E(n, j, k).$$

For $j = k$,

$$uE(n, k - 2, k - 1) = u \max \big(\beta^n u^{k-1}, \alpha(pE(n + 1, k - 1, k - 1)$$
$$+ qE(n + 1, k - 3, k - 1)) \big)$$
$$\leq \max \big(\beta^n u^k, \alpha(puE(n + 1, k - 1, k - 1)$$
$$+ qE(n + 1, k - 1, k)) \big),$$

using the first induction theorem,

$$uE(n+1, k-1, k-1) < u^2 E(n+1, k-1, k-1) = E(n+1, k+1, k+1),$$

so,

$$uE(n, k-2, k-1) \leq \max\bigl(\beta^n u^k, \alpha(pE(n+1, k+1, k+1)$$
$$+ qE(n+1, k-1, k))\bigr) = E(n, k, k).$$

This completes the induction and the proof.

3.2 Determining the exercise boundary

In this section we show that the value of a Russian option obtains its liability value once the difference between the peak security price and current security price, $k - j$, differ by at least an integer h_n, this integer depending on n. This integer is called the *exercise boundary*. The examination of the exercise boundary leads to an optimal strategy for exercise of Russian options.

Lemma 4 *Suppose* (n, j, k) *and* (n, j', k') *are accessible and* $k - j \leq k' - j'$. *Then* $E(n, j, k) = \beta^n u^k$ *implies* $E(n, j', k') = \beta^n u^{k'}$.

Proof: Various cases are argued. First, assume $k - j = k' - j'$. Since j and j' must agree with $n \bmod 2$, $2|(j' - j)$. The result follows by using the first induction theorem with $l = (j - j')/2$.

  Now assume $k - j < k' - j'$. If $2|(k' - k)$ use the first induction theorem with $l = (k' - k)/2$,

$$u^{2l} E(n, j, k) = E(n, j + 2l, k + 2l) = E(n, j + 2l, k').$$

Note $k - j = k' - (j + 2l) < k' - j'$ so $j' < j + 2l$. Using option monotonicity,

$$E(n, j', k') \leq E(n, j + 2l, k') = \beta^n u^{k+2l} = \beta^n u^{k'}.$$

The definition of $E(n, j', k')$ implies a lower bound $\beta^n u^{k'} \leq E(n, j', k')$. Hence equality holds.

  Now assume $k - j < k' - j'$ and $2|(k' - k + 1)$. If $k > 0$ apply the third induction theorem, then the first induction theorem with $l = (k' - k + 1)/2$,

$$u^{2l} E(n, j, k) \geq u^{2l} uE(n, j - 2, k - 1)$$
$$= uE(n, j - 2 + 2l, k - 1 + 2l) = uE(n, j - 2 + 2l, k').$$

Note $k - j = (k' - 2l + 1) - j < k' - j'$ so $j' \leq j - 2 + 2l$. Using option monotonicity,

$$E(n, j', k') \leq E(n, j - 2 + 2l, k') \leq u^{2l-1} E(n, j, k) = \beta^n u^{k+2l-1} = \beta^n u^{k'}.$$

Matching the lower bound on $E(n, j', k')$. Hence equality holds.

If $k = 0$ we must assume $n \geq 2$. For the remaining cases, $n = 0, 1$ the theorem is trivial. We apply the first induction theorem with $l = 1$ and the third induction theorem,

$$u^2 E(n, j, 0) = E(n, j + 2, 2) \geq u E(n, j, 1).$$

We apply the first induction theorem with $l = (k' - 1)/2$ and, since $k - j = -j < k' - j' = 2l + 1 - j'$ implies $j' - 2l \leq j$, we can apply option monotonicity,

$$E(n, j', k') = E(n, j', 2l + 1) = u^{2l} E(n, j' - 2l, 1)$$
$$\leq u^{2l} E(n, j, 1) \leq u^{2l+1} E(n, j, 0) = \beta^n u^{k'},$$

Matching the lower bound on $E(n, j', k')$. Hence equality holds.

This concludes consideration of all cases.

Definition 1 *The exercise boundary at n is the least integer h_n such that $E(n, k - h_n, k)$ obtains its liability value $\beta^n u^k$, if such an integer exists. The execution boundary is the maximal sequence of execution boundaries at n starting from some n_o and continuing in consecutive n up to N.*

The consequence of the previous lemma is that if the execution boundary at n exists, then $E(n, j, k) = \beta^n u^k$ whenever $k - j \geq h_n$.

Lemma 5 *If h_n exists then $h_{n'}$ exists for all $n \leq n' \leq N$ and $h_{n'} \leq h_n$.*

Proof: Directly from the second induction theorem.

Lemma 6 *If h_n exists and $n < N$, then h_{n+1} exists and $0 \leq h_n - h_{n+1} \leq 1$.*

Proof: For $h_n = 1$ or 0 there is nothing to show. We assume $h_n \geq 2$.

Since $k - j \geq 2$ and (n, j, k) is accessible, so are $(n, j, k-1)$ and $(n+1, j+1, k-1)$. It is sufficient to show that if $E(n, j, k) = \beta^n u^k$ and $E(n, j, k - 1) > \beta^n u^{k-1}$ then $E(n + 1, j + 1, k - 1) > \beta^{n+1} u^{j-1}$.

Arguing by contradiction, assume $E(n + 1, j + 1, k - 1) = \beta^{n+1} u^{k-1}$. By option monotonicity, $E(n + 1, j - 1, k - 1) = E(n + 1, j + 1, k - 1)$ and,

$$E(n, j, k - 1) = \max\big(\beta^n u^{k-1}, \alpha(p E(n + 1, j + 1, k - 1)$$
$$+ q E(n + 1, j - 1, k - 1))\big)$$
$$= \alpha(p E(n + 1, j + 1, k - 1) + q E(n + 1, j - 1, k - 1))$$
$$= \alpha \beta^{n+1} u^{k-1} < \beta^n u^{k-1},$$

where the last inequality is justified by $\beta < 1 \leq 1 + r = \alpha^{-1}$. The contradiction completes the proof.

Theorem 4 (Execution boundary) *Let,*

$$n_o = \min\{\, n \mid E(n, j, k) = \beta^n u^k \text{ for some accessible } (n, j, k) \,\}.$$

The set is non-empty hence the execution boundary exists and is,

$$h_{n_o} \geq h_{n_o+1} \geq \ldots \geq h_{N-1} = 1 > h_N = 0$$

where $0 \leq h_n - h_{n+1} \leq 1$.

Proof: Since $E(N, j, k) = \beta^N u^k$ the set is non-empty. It is easy to show from the definition of $E(N-1, k-1, k)$ and the inequality $\alpha\beta < 1$ that $E(N-1, k-1, k) = \beta^{N-1} u^k$. Hence h_{N-1} is at least, and at most, 1.

3.3 Efficient algorithms for optimal exercise

Lemma 7 (Canonical node) *Let* $\pi(i, j)$ *equal 0 or 1 depending on whether* i *and* j *agree modulo 2 or not, respectively. For every accessible node* (n, j, k) *there is an accessible node* $\kappa(n, j, k)$, *said to be canonical, defined by,*

$$\kappa(n, j, k) = (n, \pi(n, \delta) - \delta, \pi(n, \delta)) \text{ where } \delta = k - j.$$

Furthermore, $E(n, j, k)/E(\kappa(n, j, k)) = u^{k-\pi(n,\delta)}$, *where* $k - \pi(n, \delta)$ *is an even, non-negative integer. Conversely, for each value of* δ, $0 \leq \delta \leq n$, *there is a canonical node.*

Proof: Either δ or $\delta - 1$ agrees with n modulo 2, so at most one of $(n, -\delta, 0)$ and $(n, 1 - \delta, 1)$ can be accessible. Rearranging one of the accessibility conditions, $\delta = n - k - 2i$ for some non-negative integer i. Setting $i = \lfloor \delta/2 \rfloor$ and $k = \pi(n, \delta)$ gives any δ provided $0 \leq \delta \leq n$.

Starting from an arbitrary accessible node (n, j, k), use the first induction theorem to shift j and k down by an even integer l such that $k - l$ is either 0 or 1. Since $\delta = k - j$ is invariant, $k - l = \pi(n, \delta)$, so $l = k - \pi(n, \delta)$. This proves the lemma.

The practical consequence of this lemma is that for the purpose of tabulating values of E we can arrange nodes in a triangular table table indexed by n, $0 \leq n \leq N$, and δ, $0 \leq \delta \leq n$. As an improvement, the table can be truncated by returning a calculated value for E whenever δ is greater than or equal to the exercise boundary.

Theorem 5 *The algorithm given (see Figure 2) is an* $O(N^2)$ *dynamic programming algorithm determining* $E(n, j, k)$ *for all accessible* (n, j, k). *Since there are* $\Omega(N^2)$ *nodes to determine, the algorithm is optimal. The algorithm gives the optimal exercise strategy. It is possible to give only the optimal exercise strategy using this algorithm in* $O(N)$ *time.*

Proof: The algorithm's correctness and efficiency are easy to show.

```
getValue(n, j, k)
    delta := k - j ;
    if n >= n_o and delta >= h[n]
      then return beta^n * u^k ;
    l = k - pi(n,delta) ;
    return u^l * E[n,delta] ;

initValues(N)
    h[N] = 0 ;
    n_o = N ;
    for n = N-1 downto 0
      for delta = 0 to n
          k = pi(n,delta) ;
          j = k - delta ;
          e = alpha *
              ( p * getValue(n+1,j+1,max(j+1,k))
                  + q * getValue(n+1,j-1,k) ) ;
          if e < ( beta^n * u^k ) then
              h[n] = delta ;
              n_o = n ;
              break ; // next n
          E[n,delta] = e ;
      // end for delta
    // end for n
```

Figure 2: Dynamic programming algorithm for $E(n, j, k)$.

When the option reaches its liability value, that is, it touches the exercise boundary, exercise the option. Since by the maximum, the option is worth more exercised than held.

Only the option boundary is needed to decide the optimal exercise strategy. In an appendix we show that h_{n_o} is independent of N, and only a function of the market structure: α, β and u. Hence a variation of the algorithm which terminates once n_o has been found runs in time $O(N)$.

4 Conclusions

We have given a near optimal algorithm for the pricing of Russian options under the binomial model. We have also given some insight into the price process which these options follow. For such options to be traded, a risk-neutral hedging strategy must be found, and this is an interesting area for future research.

References

[1]   Shepp, L. A. & Shiryaev, A. N., *The Russian Option: Reduced Regret*. Ann. Appl. Probab., **3**, pp. 631–640, 1993.
[2]   Kramkov, D. O. & Shiryaev, A. N., *On the Rational Pricing of the "Russian Option" for the Symmetrical Binomial Model of a (B,S)-Market*. Theory Probab. Appl., **39**, pp. 153–162, 1994.
[3]   Cox, J. C., Ross, R. A., & Rubinstein, M., *Option Pricing: A Simplified Approach*. J. Financial Economics, **7**, pp. 229–263, 1979.
[4]   Duffie, J. D. & Harrison, J. M., *Arbitrage Pricing of Russian Options and Perpetual Lookback Options*. Ann. Appl. Probab., **3**, pp. 641–651, 1993.

This page intentionally left blank

Exotic option, stochastic volatility and incentive scheme

J. Tang & S. S.-T. Yau
Department of Mathematics, Statistics and Computer Science,
University of Illinois at Chicago, USA

Abstract

This paper examines the impact of incentive fee on exotic option pricing when the volatility is a stochastic process and is correlated with the underlying asset price. Since high water mark (HWM) is the benchmark employed by incentive schemes in the hedge fund industry, we first develop the HWM lookback option-pricing framework in stochastic volatility model. This provides an improvement to previous works in constant volatility model. We also explore option prices through Monte Carlo (MC) simulation and variance reduction technique. We further demonstrate that our discrete simulation to HWM option pricing is more practical than models assuming continuous collection of incentive fees. Numerical examples illustrate how the stochastic volatility models and incentive scheme influence option pricing.
Keywords: lookback option, stochastic volatility models, high water mark, risk neutral, Monte Carlo simulation, variance reduction.

1   Introduction

Over the last few years, hedge funds have been experiencing significant growth in both the number of hedge funds and the amount of assets under management. Based on the estimates by Securities and Exchange Commission, there are currently around 8,000 hedge funds in the United States managing around $1 trillion in assets. Hedge fund assets are growing faster than mutual fund assets and have roughly one quarter of the assets of mutual funds. They often provide markets and investors with substantial benefits, such as enhancing liquidity, contributing to market efficiency by taking speculative and value-driven trading

position, and offering investors an important risk management tool by providing valuable portfolio diversification.

Compensation schemes, which align manager interests with investor interests, play an important role in financial market. Hedge fund industries usually employ a never negative incentive fee (NNIF) [4] structure, and use a high water mark (HWM) as the benchmark, which increases over time to make up for previous failures to exceed the target. Fung and Hsieh [6] provide a rationale for the organization of hedge funds and demonstrate the incentive fee paid to successful managers can be significantly higher than the fixed management fee. Carpenter [3] and Basak, Pavlova, and Shapiro [1] examine effects of the incentive compensation on the optimal dynamic investment strategies. Goetzmann, Ingersoll and Ross [7] utilize an option approach to calculate the present value of the fees charged by money managers.

One of the factors that provide an explanation for the recent success of exotic options is their significant hedging role, which meets the hedgers' needs in cost effective ways. The exotic option price derived from the Black-Scholes model [2] under constant volatility assumption could be wildly wrong since most derivative markets exhibit persistently varying volatilities. Li's [11] study of the HWM lookback option in the constant volatility model, under the assumption of incentive fee collected continuously, is not very practical since the fee is usually collected monthly or quarterly in practice. In this paper, we first use MC method to study the price of path dependent HWM lookback option in a stochastic volatility model, in which the stock price and volatility are instantaneously correlated. Then, the framework of the HWM option pricing is set up with stochastic volatility and HWM lookback option is simulated by Monte Carlo discretion and variance reduction technique. Finally, some numerical examples and results are given.

2 HWM option pricing framework

Consider a time interval $[0, T]$ and fix a two-dimensional standard Brownian Motion process $W = \left(W^{(1)}, W^{(2)}\right)$ on a complete filtered probability space (Ω, $\mathcal{F}$, P). Let the filtration F = { $\mathcal{F}_t$: $0 \leq t \leq T$ } be the $P -$ augmentation [16] of the natural filtration of W. Hence the uncertainty in this setup is generated by the process W and the flow of information is represented by the filtration F. We say $W_t^{(1)}$ and $W_t^{(2)}$ are correlated Standard Brownian Motions with correlation ρ if $E\left(W_t^{(1)} W_t^{(2)}\right) = \rho.t$.

Now assume an arbitrage-free financial market consisting of two traded assets in which trading takes place continuously over the period $[0, T]$: one locally risk-free asset B with risk-free interest rate r, and one risky asset of price S (called the primitive asset). We define the time t prices of the asset of the fund as the solution to the following stochastic differential equation

$$dS_t = (r - D)S_t dt + \sigma_t S_t \cdot dZ_t, \ S_t < H_t \qquad (1)$$

where D is the basic management fee, σ is the volatility process, to be discussed in a moment. Z for $0 \le t \le T$ is a standard Brownian Motion (SBM). The correlation between volatility process and the return process of the primitive asset is represented by a constant $\rho \in [0,1]$. H_t is the HWM at time t.

We consider two different dynamics for the volatility process σ. The first is the Geometric Brownian Motion Process (GBMP) [10, 13],

$$d\sigma_t = \alpha\sigma_t dt + \theta\sigma_t dW_t^{(\sigma)}, \qquad 0 \le t \le T \tag{2}$$

where the appreciation rate α and the volatility of the volatility θ are constants. Obviously, σ_t/σ_0 is lognormal with parameters $\left(\alpha - \theta^2/2\right)T$ and $\theta\sqrt{T}$. The second is the Square Root Mean Reverting Process (SRMRP) [9].

$$dv_t = k\left(\bar{v} - v_t\right)dt + \theta\sqrt{v_t}\,dW_t^{(\sigma)}, \qquad 0 \le t \le T \tag{3}$$

where v is square of σ, $\bar{v}$ is the long-run mean variance, and k represents the speed of mean reversion. Feller [5] has shown that the density of v_t at time $t > 0$ conditioned on v_0 at $t = 0$ follows a non-central chi-square distribution with $4k\bar{v}/\theta^2$ degrees of freedom.

Since Z_t and $W_t^{(\sigma)}$ are correlated SBMs with correlation ρ, for the sake of better simulation of Z_t in later section, we can write $Z_t = \sqrt{1 - \rho^2}\,W_t^{(s)} + \rho.W_t^{(\sigma)}$ just by the property of SBM, where $W_t^{(s)}$ is a SBM independent of $W_t^{(\sigma)}$, for detail, see [15]. Then eqn (1) can now be written as follows:

$$dS_t = (r - D)S_t dt + \sigma_t S_t \cdot \left(\sqrt{1 - \rho^2}\,dW_t^{(s)} + \rho dW_t^{(\sigma)}\right), \quad S_t < H_t \tag{4}$$

In the simplest case, the HWM is the highest level the asset value that has reached in the past. For some incentive contracts, the HWM grows at the rate of interest or other contractually stated rate G_t, thus evolution of H_t is locally deterministic as Goetzmann, Ingersoll and Ross [7] point out.

$$dH_t = G_t H_t dt, \qquad S_t < H_t \tag{5}$$

where G_t, the contractual growth rate of the HWM, is usually zero or r. When the primitive asset value reaches a new high, the HWM is reset to this higher level.

Following the arguments in Hull and White [10], there are three state variables, S, σ and H, of which S is traded. When the fund's assets are below the HWM and the volatility is a GBMP, the option price V_t satisfies the following partial differential equation (PDE)

$$\frac{\partial V}{\partial t} + \frac{1}{2}S^2\sigma^2\frac{\partial^2 V}{\partial S^2} + \frac{1}{2}\theta^2\sigma^2\frac{\partial^2 V}{\partial \sigma^2} + \rho\theta S\sigma^2\frac{\partial^2 V}{\partial S\partial\sigma}$$
$$+ GH\frac{\partial V}{\partial H} + S(r - D)\frac{\partial V}{\partial S} + \alpha\sigma\frac{\partial V}{\partial \sigma} = rV, \qquad S < H \tag{6}$$

if the volatility is a SRMRP, the PDE can be written as

$$
\frac{\partial V}{\partial t} + \frac{1}{2}S^2 v \frac{\partial^2 V}{\partial S^2} + \frac{1}{2}\theta^2 v \frac{\partial^2 V}{\partial v^2} + \rho\theta S v \frac{\partial^2 V}{\partial S \partial v}
$$
$$
+ GH \frac{\partial V}{\partial H} + S(r-D)\frac{\partial V}{\partial S} + k(\bar{v}-v)\frac{\partial V}{\partial v} = rV, \qquad S < H,
$$
(7)

The payoff function is

$$
V(S, v, H, T) = \Lambda(S, H, T), \tag{8}
$$

where $\Lambda(S, H, T)$ is defined in the contract. Another condition applies along the boundary $S_t = H_t$. When the asset value rises above H_t to $H_t + \varepsilon_H$, the HWM is reset to $H_t + \varepsilon_H$, and an incentive fee of $q \cdot \varepsilon$, where q = the rate of incentive fee, is paid to the manager reducing the asset value to $H_t + \varepsilon_H(1-q)$. Therefore, the option price before any adjustments of the incentive fee and HWM is $V(H_t + \varepsilon_H, v_t + \Delta v, H_t, t + \Delta t)$, and the option price after the adjustments of the incentive fee and HWM is $V(H_t + \varepsilon_H(1-q), v_t + \Delta v, H_t + \varepsilon_H, t + \Delta t)$. As we know that the option price is continuous. It gives

$$
V(H_t + \varepsilon_H, v_t + \Delta v, H_t, t + \Delta t) = V(H_t + \varepsilon_H(1-q), v_t + \Delta v, H_t + \varepsilon_H, t + \Delta t)
$$

or omitting higher orders of ε_H, Δv and Δt, we have

$$
V_t + \frac{\partial V_t}{\partial S_t}\varepsilon_H + \frac{\partial V_t}{\partial v_t}\Delta v + \frac{\partial V_t}{\partial t}\Delta t =
$$
$$
V_t + \frac{\partial V_t}{\partial S_t}(1-q)\varepsilon_H + \frac{\partial V_t}{\partial v_t}\Delta v + \frac{\partial V_t}{\partial H_t}\varepsilon_H + \frac{\partial V_t}{\partial t}\Delta t .
$$

giving the boundary condition

$$
q\frac{\partial V_t}{\partial S_t} = \frac{\partial V_t}{\partial H_t} \qquad \text{on} \qquad S_t = H_t. \tag{9}
$$

Hence eqn (6) or eqn (7) together with eqn (8) and eqn (9) give the solution of the option price with the HWM provision in different stochastic volatility models.

From a probability view, the current value of a floating strike lookback put option with payoff $(M_T - S_T)$ is the discounted expectation of the payoff under the risk neutral measure.

$$
V(S, M, \sigma, 0) = e^{-rT} E[M_T - S_T], \tag{10}
$$

where $M_t = \max_{0 \le u \le t}\{S_u\}$. Define $I_n = \int_0^t (S_\tau)^n d\tau$ and $M_n = (I_n)^{1/n}$, we consider a lookback option whose value depends on M_n and then take the limit as $n \to \infty$. Recall that as n tends to infinity and by stochastic calculus, we have $M_t = \lim_{n \to \infty} M_n = \max_{0 \le \tau \le t} S_\tau$ [14]. Then we derive the stochastic differential

equation satisfied by M_n, we get $dM_n = \frac{1}{n}\frac{S^n}{(M)^{n-1}}dt$, thus M_n is a deterministic variable [14], as there are no random terms on the right hand side. Since the HWM lookback put is a path-dependent option, its value V is not simply a function of S, σ, H and t, but also on M. If the volatility is a SRMRP, we actually have

$$
\frac{\partial V}{\partial t} + \frac{1}{2}S^2 v \frac{\partial^2 V}{\partial S^2} + \frac{1}{2}\theta^2 v \frac{\partial^2 V}{\partial v^2} + \rho\theta S v \frac{\partial^2 V}{\partial S \partial v} + GH\frac{\partial V}{\partial H}
$$
$$
\frac{1}{n}\frac{S^n}{(M_n)^{n-1}}\frac{\partial V}{\partial M} + S(r-D)\frac{\partial V}{\partial S} + k(\bar{v}-v)\frac{\partial V}{\partial v} = rV, \quad 0 \le S < H
$$
(11)

We now take the limit $n \to \infty$. Since $S \le \max S = M$, in this limit the coefficient of $\frac{\partial V}{\partial M}$ tents to zero. Thus in this limit, for a HWM lookback put with payoff $(H_T - S_T)$, the option price satisfies the PDEs

$$
\frac{\partial V}{\partial t} + \frac{1}{2}S^2 v \frac{\partial^2 V}{\partial S^2} + \frac{1}{2}\theta^2 v \frac{\partial^2 V}{\partial v^2} + \rho\theta S v \frac{\partial^2 V}{\partial S \partial v} + GH\frac{\partial V}{\partial H}
$$
$$
+ S(r-D)\frac{\partial V}{\partial S} + k(\bar{v}-v)\frac{\partial V}{\partial v} = rV, \quad 0 \le S < H
$$
(12)

$$
V(S, H, \sigma, T) = H_T - S_T,
$$
(13)

$$
V(S, H, \sigma, t) = e^{-r(T-t)} E(H_T - S_T),
$$
(14)

$$
q\frac{\partial V}{\partial S} = \frac{\partial V}{\partial H}, \quad \text{on} \quad S = H.
$$
(15)

3   HWM lookback option price simulation algorithm

Suppose an option has payoff $\Lambda_T \equiv \Lambda_T(\omega)$ at time T, where Λ_T may depend on the state $\omega \in \Omega$. Assuming that no arbitrage exists, under the martingale measure P associated with the accumulator numernaire, the option value V_t at time $t < T$ is

$$
V_t = E[\Lambda_T e^{-r(T-t)}],
$$
(16)

which can be solved using plain MC method. A standard reference for applications of MC methods in finance is Jäckel [11]. Eqn (16) is an integral over the state space Ω,

$$
V_t = E[\Lambda_T e^{-r(T-t)}] = e^{-r(T-t)}\int_\Omega \Lambda_T(\omega)dP(\omega),
$$
(17)

which can be approximated by constructing a set $\{\hat{\omega}^n\}_{n=1,..,N}$ of discrete sample paths randomly selected under a measure $\hat{P}$, a discrete approximation to the measure P. Then the approximation $\hat{V}_t$ to V_t is

$$\hat{V}_t = e^{-r(T-t)} \frac{1}{N} \sum_{n=1}^{N} \Lambda_T(\hat{\omega}^n) \tag{18}$$

In our implementation, the processes σ_t or v_t and S_t can be discretized by Euler scheme. For the simplest case, let growth rate G of the HWM and the basic management fee be zero, we have MC simulation algorithm of HMW lookback put option price for the SRMRP as

for $i = 1:N$ /* sample path

  for $j = 1:M$ /* time step

    Initialize HWM_0;   /* HWM_P is the temporary HWM of the Pth fee paying

                        /* cycle for each sample path.

    Initialize $H_{i,1} = HWM_0$;   /* initial value of HWM

    if $j <$ the pay day and $S_{i,j} \le H_{i,j}$

      Set $S_{i,j+1} = S_{i,j} \exp\left((r - \frac{1}{2}\sigma_{i,j}^2)\Delta t + \sigma_{i,j}\left(\sqrt{1-\rho^2}\,\Delta W_{i,j}^{(s)} + \rho\Delta W_{i,j}^{(\sigma)} \right) \right)$;

       Set $v_{i,j+1} = v_{i,j} + k(\bar{v} - v_{i,j})\cdot\Delta t + \theta\sqrt{v_{i,j}}\cdot\Delta W_{i,j}^{(\sigma)}$;

    if $j <$ the pay day and $S_{i,j} > H_{i,j}$

         Set $H_{i,j} = S_{i,j}$;

    if $j =$ the day to pay incentive fee q and $H_{i,j} > HWM_P$ of last paying cycle

          Set $S_{i,j+1} = S_{i,j} - q(H_{i,j} - HWM_P$ of last paying cycle$)$;

          Set $P = P+1$;

    end if

    Set $\hat{V}_i = e^{-r(T-t)}(H_{i,M} - S_{i,M})$;

  end for j

end for i

Average the discounted values over the sample paths $\hat{V} = \frac{1}{N}\sum_{i=1}^{N}\hat{V}_i$;

Compute the standard deviation $\sigma_{\hat{V}} = \sqrt{\frac{1}{(N-1)}\sum_{i=1}^{N}(\hat{V}_i - \hat{V})^2}$;

Compute the standard error $\varepsilon = \sigma_{\hat{V}} \big/ \sqrt{N}$;

4   Examples and numerical results

Now we present some numerical examples to demonstrate the effects of incentive scheme and different stochastic volatility models by the plain MC

simulation. We then utilize antithetic variate (AV) variance reduction technique only for S, not for σ or v, since the estimator is not monotone as a function of the uniforms used to generate them. The experiments are performed on a desktop PC with a Pentium4@3.4GH$_z$ CPU, and the codes are written in Matlab with a Matlab 6.1 compiler.

Within the expiring time $T = 0.5$ year, we compare three situations in each table below: none incentive fee collected (None), fee collected two times (Twice), and fee collected four times (Quarterly). Between tables, option prices with respect to different volatility dynamics are compared. For the simplest case, let growth rate G of the HWM and basic MF be zero. The parameters are $S_0 = H_0 = 100$, $r = 0.05$, $q = 0.20$, and number of periods = 180. Standard errors are in parentheses. In Table 1, the value of constant volatility = 0.15. For the GBMP in Table 2 and 3, we take $\sigma_0 = 0.15$, $\alpha = 0.05$, $\theta = 0.08$. For the SRMRP in Table 4 and 5, we use $v_0 = 0.0225$, $k = 1.5$, $\bar{v} = 0.0225$.

Table 1: Estimated HWM lookback option price with constant volatility.

Number of draws / Payment frequency	1,000	5,000	10,000	100,000
(Plain MC)				
None	6.7703	6.9720	7.0124	7.0092
	(0.1738)	(0.0781)	(0.0565)	(0.0179)
Twice	7.2202	7.3837	7.4535	7.4912
	(0.1819)	(0.0813)	(0.0589)	(0.0187)
Quarterly	7.0960	7.2813	7.3326	7.3575
	(0.1788)	(0.0799)	(0.0577)	(0.0183)
(AV)				
None	7.0638	7.0348	7.0402	7.0057
	(0.0720)	(0.0337)	(0.0239)	(0.0075)
Twice	7.5564	7.5247	7.5241	7.4904
	(0.0724)	(0.0341)	(0.0241)	(0.0076)
Quarterly	7.4038	7.3819	7.3864	7.3564
	(0.0728)	(0.0340)	(0.0241)	(0.0076)

As shown from these results, the option prices of the SRMRP are lower than those of the GBMP or the constant volatility. In both GBMP and SRMRP, the option price is an increasing function of the correlation ρ. It is also worth noticing that the more frequently the incentive fee is paid, the lower the option price, and the price is the lowest when nothing paid. One possible explanation is the price of the underlying asset reduces a portion when the incentive fee is

collected, and it is much difficult for the asset price to reach a new high. Finally, antithetic variate method can reduce the standard error by a factor of about 2.

Table 2:   Estimated HWM lookback option price with GBMP and $\rho = 0$.

Number of draws / Payment frequency	1,000	5,000	10,000	100,000
(Plain MC)				
None	7.1819	7.0649	7.1354	7.1217
	(0.1823)	(0.0802)	(0.0570)	(0.0181)
Twice	7.6057	7.4745	7.5723	7.5996
	(0.1897)	(0.0831)	(0.0593)	(0.0189)
Quarterly	7.5116	7.3761	7.4604	7.4697
	(0.1867)	(0.0818)	(0.0583)	(0.0186)
(AV)				
None	7.2014	7.1426	7.1445	7.1207
	(0.0797)	(0.0343)	(0.0243)	(0.0077)
Twice	7.6753	7.6275	7.6262	7.6038
	(0.0797)	(0.0346)	(0.0245)	(0.0077)
Quarterly	7.5468	7.4935	7.4922	7.4701
	(0.0801)	(0.0348)	(0.0246)	(0.0078)

Table 3:   Estimated HWM lookback option price with GBMP and $\rho = 0.2$.

Number of draws / Payment frequency	1,000	5,000	10,000	100,000
(Plain MC)				
None	7.2125	7.1116	7.1802	7.1528
	(0.1798)	(0.0798)	(0.0569)	(0.0180)
Twice	7.6529	7.5318	7.6243	7.6352
	(0.1874)	(0.0828)	(0.0592)	(0.0188)
Quarterly	7.5621	7.4343	7.5123	7.5051
	(0.1839)	(0.0581)	(0.0581)	(0.0185)
(AV)				
None	7.1746	7.1776	7.1843	7.1513
	(0.0811)	(0.0354)	(0.0252)	(0.0079)
Twice	7.6606	7.6680	7.6696	7.6380
	(0.0820)	(0.0362)	(0.0257)	(0.0081)
Quarterly	7.5372	7.5337	7.5356	7.5042
	(0.0816)	(0.0360)	(0.0256)	(0.0080)

Table 4:   Estimated HWM lookback option price with SRMRP and $\rho = 0$.

Number of draws / Payment frequency	1,000	5,000	10,000	100,000
(Plain MC) None	7.0733 (0.1815)	6.9314 (0.0797)	6.9895 (0.0566)	6.9719 (0.0180)
Twice	7.5008 (0.1888)	7.3445 (0.0826)	7.4280 (0.0588)	7.4534 (0.0188)
Quarterly	7.4045 (0.1857)	7.2448 (0.0814)	7.3140 (0.0578)	7.3213 (0.0185)
(AV) None	7.0726 (0.0856)	7.0015 (0.0355)	6.9979 (0.0250)	6.9703 (0.0079)
Twice	7.5519 (0.0827)	7.4895 (0.0358)	7.4830 (0.0252)	7.4567 (0.0080)
Quarterly	7.4218 (0.0831)	7.3541 (0.0360)	7.3474 (0.0253)	7.3211 (0.0080)

Table 5:   Estimated HWM lookback option price with SRMRP and $\rho = 0.2$..

Number of draws / Payment frequency	1,000	5,000	10,000	100,000
(Plain MC) None	7.1547 (0.1773)	7.0248 (0.0785)	7.0773 (0.0559)	7.0502 (0.0177)
Twice	7.6073 (0.1849)	7.4543 (0.0817)	7.5312 (0.0583)	7.5421 (0.0186)
Quarterly	7.5123 (0.1814)	7.3531 (0.0803)	7.4161 (0.0573)	7.4094 (0.0182)
(AV) None	7.1024 (0.0802)	7.0840 (0.0349)	7.0837 (0.0247)	7.0482 (0.0077)
Twice	7.5981 (0.0815)	7.5836 (0.0359)	7.5784 (0.0254)	7.5441 (0.0080)
Quarterly	7.4756 (0.0811)	7.4472 (0.0357)	7.4424 (0.0253)	7.4078 (0.0079)

References

[1]   Basak, S., Pavlova, A. and Shapiro, A., Offsetting the incentives: risk shifting and benefits of benchmarking in money management. *Working Paper 430303*, MIT Sloan School of Management, 2003.

[2]   Black, F. and Scholes, M., The pricing of options and corporate liabilities. *Journal of Political Economy,* 81, pp. 637 654, 1973.
[3]   Carpenter, J. N., Does option compensation increase managerial risk appetite? *Journal of Finance,* 55, pp. 2311 2331, 2000.
[4]   Elton, E. J., Gruberand, M. J. and Blake, C. R., Incentive fees and mutual funds. *Journal of Finance*, 58, pp. 779 804, 2003.
[5]   Feller, W., Two singular diffusion problems. *Annals of Mathematics,* 54, pp. 173 182, 1951.
[6]   Fung, W., and Hsieh, D., A primer on hedge funds. *Journal of Empirical Finance,* 6, pp. 309 331, 1999.
[7]   Goetzmann, W. N., Ingersoll, J., and Ross, S.A., High water marks and hedge fund management contracts. *Journal of Finance*, 58, pp. 1685 1717, 2003.
[8]   Heath, D. and Platen, E., A variance reduction technique based on integral representations. *Quantitative Finance,* 2, pp. 362 369, 2002.
[9]   Heston, S. I., A closed form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies,* 6, pp. 327 343, 1993.
[10]   Hull, J. and White, A., The pricing of options on Assets with stochastic volatilities. *Journal of Finance,* 42, pp. 281 300, 1987.
[11]   Jäckel, P., Monte Carlo methods in finance. *Wiley Finance Series,* New York: Wiley, 2002.
[12]   Li, Z., Path dependent option: the case of high water mark provision for hedge funds. *Ph.D. Thesis,* University of Illinois at Chicago, 2002.
[13]   Wiggins, J. B., Option values under stochastic volatilities. *Journal of Financial Economics,* 19, pp. 351 377, 1987.
[14]   Wilmott, P., Howison, S. and Dewynne, J., *The Mathematics of Financial Derivatives*. Cambridge University Press, Cambridge, UK, 1995.
[15]   Tang, J. H., Exotic option, stochastic volatility and incentive scheme. *Ph.D. Thesis,* University of Illinois at Chicago, 2005.
[16]   Duffie, D., Dynamic asset pricing theory. Princeton University Press: Princeton and Oxford, pp 323 330, 2001.

Applying design patterns for web-based derivatives pricing

V. Papakostas, P. Xidonas, D. Askounis & J. Psarras
*School of Electrical and Computer Engineering,
National Technical University of Athens, Greece*

Abstract

Derivatives pricing models have been widely applied in the financial industry for building software systems for pricing derivative instruments. However, most of the research work on financial derivatives is concentrated on computational models and formulas. There is little guidance for quantitative developers on how to apply these models successfully in order to build robust, efficient and extensible software applications. The present paper proposes an innovative design of a web-based application for real-time financial derivatives pricing, which is entirely based on design patterns, both generic and web-based application specific. Presentation tier, business tier and integration tier patterns are applied. Financial derivatives, underlying instruments and portfolios are modelled. Some of the principal models for evaluating derivatives (Black–Scholes, binomial trees, Monte Carlo simulation) are incorporated. Arbitrage opportunities and portfolio rebalancing requirements are detected in real time with the help of a notification mechanism. The novelty in this paper is that the latest trends in software engineering, such as the development of web-based applications, the adoption of multi-tiered architectures and the use of design patterns, are combined with financial engineering concepts to produce design elements for software applications for derivatives pricing. Although our design best applies to the popular J2EE technology, its flexibility allows many of the principles presented to be adopted by web-based applications implemented with alternative technologies.
Keywords: financial applications, financial derivatives, pricing models, design patterns, J2EE patterns, web-based applications, multi-tiered architectures.

1 Introduction

Financial derivatives have become extremely popular among investors for hedging and speculating. Their growing use has triggered an increased interest in financial engineering and the emergence of several computational models for evaluating them and determining their characteristics.

Numerous software systems and applications have been developed for implementing such models. Some are in-house applications for large financial institutions and investment banks while others are available as software products. Despite the plethora of computational models that are present in the relevant literature, the existence of books and publications on the design and construction of software systems for implementing them is limited. Even these are usually constrained to conventional object-oriented design, circumstantial use of design patterns and traditional programming languages like C++ or Visual Basic.

The objective of the present paper is to propose an innovative design of a web-based application for real-time financial derivatives and portfolios pricing. The modelled application quotes derivatives and underlying assets prices from market data feeds and applies pricing models for computing derivatives and portfolios theoretical values and characteristics. In addition to rendering pricing information on web pages, it can send notifications (e.g. emails) when prices or attributes of derivatives or portfolios satisfy certain conditions (e.g. permit arbitrage or require portfolio rebalancing).

Design patterns play central role in our design, upon which it is almost entirely based. Both generic [5] and web-based applications specific patterns (J2EE patterns [1]) are applied. The proposed design aims to facilitate the introduction of new derivative instruments, additional valuation models and alternative market data feeds to the system on subsequent phases after its initial release.

2 Background work

Joshi [7] and Duffy [3] apply the concepts of object-oriented programming and adopt design patterns for evaluating financial derivatives. London [9] assembles a number of pricing models implemented in the C++ programming language. Zhang and Sternbach [12] model financial derivatives using design patterns. van der Meij *et al* [11] describe the adoption of design patterns in a derivatives pricing application. Marsura [10] presents a complete application for evaluating derivatives and portfolios using objects and design patterns. Eggenschwiler and Birrer [2], and Gamma and Eggenschwiler [4] describe the use of objects and frameworks in financial engineering. Koulisianis *et al* [8] present a web-based application for derivatives pricing implemented with the PHP technology, using the *Problem Solving Environment* methodology.

3   Derivatives pricing models

It is possible to estimate the value that a financial derivative contract should theoretically have from the underlying asset price and the contract characteristics. If the difference between the market price and the theoretical value of the contract is significant, an investor can achieve guaranteed profit (arbitrage). For this reason, derivatives pricing has become the field of extensive study for the past three decades.

A number of pricing models (analytical and numerical methods) have emerged and applied for derivatives pricing [6]. Black–Scholes equation provides analytical formulas for calculating theoretical prices of European call and put options on non-dividend paying stocks. Binomial trees are particularly useful in cases that an option holder has the potential for early exercise. Monte Carlo simulation is primarily applied when the derivative price depends on the history of the underlying asset price or on multiple stochastic variables.

4   Multi-tiered architecture

The present paper proposes the design of an application for derivatives pricing that is web-based. The use of the internet introduces certain complexity into our model. A multi-tiered architecture has been adopted for our design. Each tier in a multi-tiered architecture is responsible for a subset of the system functionality [1]. It is logically separated from its adjacent tiers and loosely connected to them. It is important to emphasise that a multi-tiered architecture is logical and not physical. This means that multiple tiers may be deployed on a single machine or a single tier may be deployed on multiple machines, especially if it contains CPU intensive components.

5   Design patterns

5.1 Presentation tier

5.1.1 Front Controller
The *Font Controller* pattern forms the initial point of communication for handling user requests, aiming to reduce the administration and deployment tasks for the application [1]. One *Front Controller* is used for all user requests. It is incarnated by the *FrontController* class, which is a servlet.

5.1.2 Context Object
The *Context Object* pattern encompasses state in a protocol independent way, in order to be utilized by different parts of the application [1]. One *Context Object* is used for each type of user request. Requests for futures pricing are modelled by the *FuturePricingRequestContext* class, requests for options pricing by the *OptionPricingRequestContext* class, etc. The *Factory* pattern is applied for their creation.
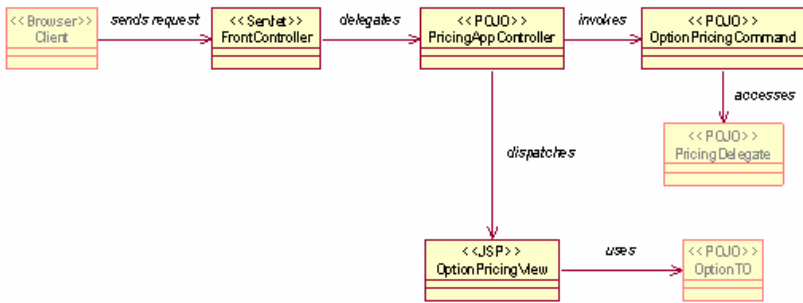
Figure 1: Presentation tier design patterns.

5.1.3 Application Controller

The *Application Controller* pattern centralizes the invocation of actions for handling requests (action management) and the dispatch of response data to the proper view (view management) [1]. Our design suggests the use of the *ApplicationController* interface for modelling *Application Controller* functionality. The *PricingAppController* class, which implements this interface, coordinates *Commands* and *Views* related to pricing requests. The *ManagementAppController* class does the same for requests related to instrument management, such as adding a new financial instrument to the application. The *Factory* pattern is again applied for their creation.

5.1.4 View Helper

The *View Helper* pattern uses views to encapsulate the code that formats responses to user requests and helpers to encapsulate the logic that views require in order to obtain response data [1]. In our design, *Views* are incarnated by a number of JSP pages. *PortfolioDetailsView* displays information related to portfolios definition, *OptionPricingView* displays the results of an option pricing request, etc. *Business Delegates* are used as *Helpers*.

5.1.5 Command

The *Command* pattern encapsulates the action required as the result of a request through the invocation of the corresponding functionality [5]. One *Command* is used for each type of user request. Requests for futures definition invoke class *FutureDefinitionCommand*, requests for portfolio pricing class *PortfolioPricingCommand*, etc. As a result, there is one-to-one correspondence among *Context Objects* and *Commands*. The *Factory* pattern is applied for their creation.

5.1.6 Factory

The *Factory* pattern is responsible for the creation of objects that implement an interface or extend an abstract class. In our design, a number of classes adopt this

pattern, such as *RequestContextFactory* for the creation of *Context Objects*, *ApplicationControllerFactory* for the creation of *Application Controllers*, *CommandFactory* for the creation of *Commands,* etc. *Factories* can be configured declaratively through the use of XML files.

5.1.7 Singleton

The *Singleton* pattern defines classes that are allowed to have only one instance per application [5]. Each class that adopts the *Factory* pattern in our design adopts the *Singleton* pattern in addition.

5.2 Business tier

5.2.1 Business Delegate

The *Business Delegate* pattern encapsulates access to business services, aiming to reduce interconnection between components of the presentation and business tiers [1]. In our design, one *Business Delegate* is defined for each *Session Façade*. The *PricingDelegate* class provides centralised access to the *PricingFacade* class, the *ManagementDelegate* class to the *ManagementFacade* class and the *NotificationDelegate* class to the *NotificationFacade* class.

5.2.2 Service Locator

The *Service Locator* pattern centralises the lookup of services and components [1]. One *Service Locator*, which is incarnated by the *ServiceLocator* class, is used. It also adopts the *Singleton* pattern.
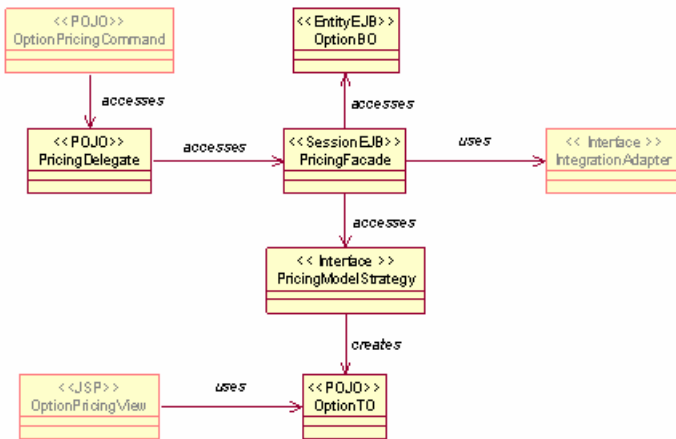
Figure 2:   Business tier design patterns.

5.2.3 Session Façade

The *Session Façade* pattern encapsulates components of the business tier and exposes coarse-grained services to remote clients, aiming to reduce the number

of remote method invocations among components of the presentation and business tiers [1]. Services related to derivatives and portfolios pricing are aggregated to the *PricingFacade* class. Services related to derivatives and portfolios management are encapsulated in the *ManagementFacade* class. Services for the notification mechanism are accumulated in the *NotificationFacade* class.

5.2.4 Application Service

The *Application Service* pattern underlies components that encapsulate business logic, aiming to leverage related services and objects (*Business Objects*) [1]. Our design adopts the layer strategy in regard to the use of *Application Services*. The *PricingAppService* and *NotificationAppService* classes, which reside on the higher layer, expose pricing and notification services respectively. They require pricing modelling related functionality, which is provided by the *PricingModelStrategy* interface, which resides on the lower layer, along with the *BlackScholesAppService*, *BinomialTreeAppService* and *MonteCarloAppService* classes that implement it. Financial instruments volatility is calculated on a daily basis by the *VolatilityAppService* class.
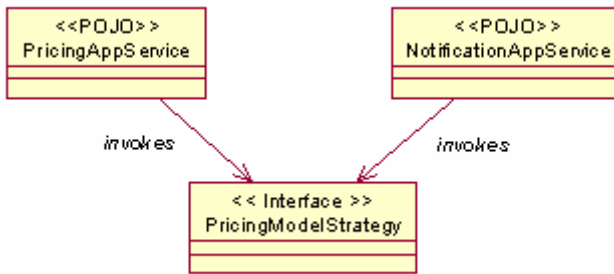
Figure 3:   Application Service layering.

5.2.5 Business Object

The *Business Object* pattern encapsulates and administers business data, behaviour and persistence, aiming at the creation of objects with high cohesion [1]. Our design contains a hierarchy of *Business Objects* that correspond to portfolios and financial instruments. They consist of abstract classes *FinancialInstrumentBO*, *DerivativeBO* and concrete classes *PortfolioBO*, *StockBO*, *IndexBO*, *CurrencyBO*, *FutureBO*, *OptionBO*, *EuropeanOptionBO*, and *AmericanOptionBO*. This way, new derivative types can be added with minor modifications.

5.2.6 Composite Entity

The *Composite Entity* pattern aggregates a set of related *Business Objects* into one coarse-grained entity bean, allowing for the implementation of parent objects that manage dependent objects [1]. In our design, the *PortfolioBO* class, which

represents a portfolio, is defined as parent object and the *PortfolioInstrument* class, which represents a financial instrument that is member of a portfolio, as dependent object. Although a *PortfolioInstrument* object is linked to a *FinancialInstrumentBO* object, it is a separate object. It holds information such as quantity and (call/put) position of a specific instrument in a portfolio.
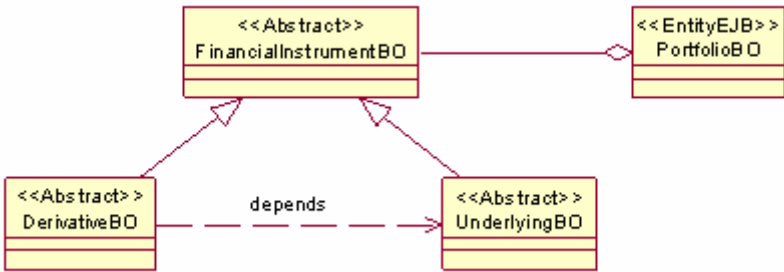
Figure 4:   Business Objects for financial instruments.
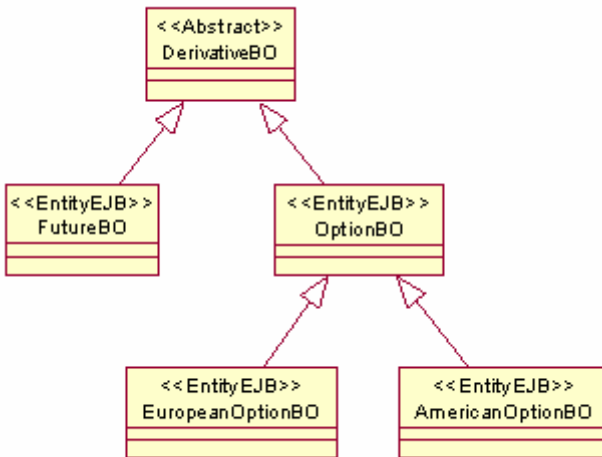
Figure 5:   Hierarchy of Business Objects for derivatives.

5.2.7 Transfer Object

The *Transfer Object* (or *Data Transfer Object*) pattern carries multiple data across application tiers [1]. Our design adopts the multiple transfer objects strategy in regard to the use of *Transfer Objects*. One *Transfer Object* is defined for each *Business Object*. This leads to a hierarchy of *Transfer Objects* that correspond to financial instruments and portfolios. They consist of classes

FinancialInstrumentTO, *DerivativeTO*, *PortfolioTO*, *StockTO*, *IndexTO*, *CurrencyTO*, *FutureTO*, *OptionTO*, *EuropeanOptionTO*, and *AmericanOptionTO*.

5.2.8 Strategy

The *Strategy* pattern encapsulates a family of algorithms and makes them interchangeable [5]. Considering our design, such algorithms are the pricing models for derivatives. The *PricingModelStrategy* interface adopts this pattern. It is implemented by the *BlackScholesAppService*, *BinomialTreeAppService* and *MonteCarloAppService* classes, which contain the algorithms for the Black–Scholes, binomial trees and Monte Carlo models respectively. This way, new pricing models can be introduced with minor modifications.
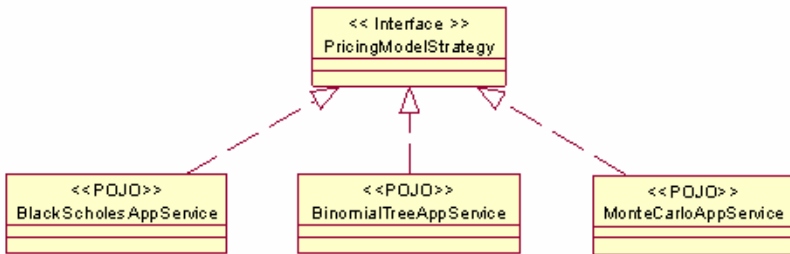
Figure 6:   Strategy.

5.2.9 Observer

The *Observer* pattern defines an one-to-many correspondence between an observable object (*Observable* or *Publisher*) and one or more observer objects (*Observers* or *Subscribers*). When the observable object changes state, all the observer objects are automatically notified [5].

The *Observer* pattern is applied on a very significant feature of our proposed design: the notification mechanism. Notifications are sent when the states of derivatives instruments or portfolios conform to certain predefined rules. For example, when the difference between the market and theoretical price of a derivative becomes large enough to permit arbitrage or when the delta of a portfolio in respect to one its underlying instruments exceeds a certain value. In such cases, users should be notified immediately, in order to take advantage of the arbitrage opportunity or perform portfolio rebalancing.

For simplicity, the *NotificationAppService* class is defined as observable object and not each *Business Object* separately. The *NotificationAppService* class is triggered at constant intervals (e.g. every 60 seconds) by a system timer. It monitors derivatives and portfolios states, sending notifications to observer objects. Observer objects implement the *InstrumentListener* interface. The *EmailAppService* and *SocketAppService* classes, which send notifications via email and TCP/IP respectively, have been defined as observers. Additional observers can be easily introduced.
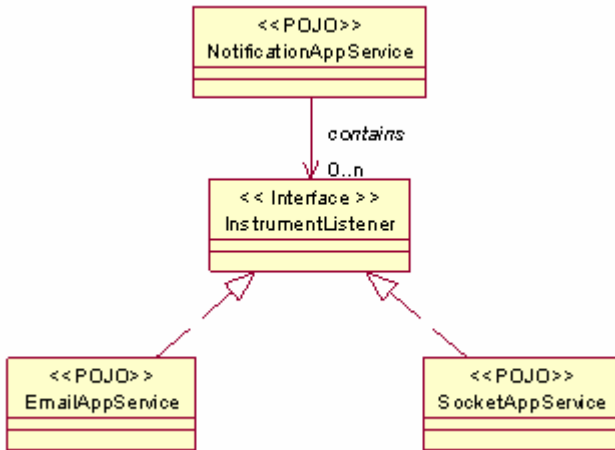
Figure 7:   Observer.

5.3 Integration tier

5.3.1 Integration Adapter

The *Adapter* pattern converts the interface of an object or system to another interface that a client is capable of using [5]. The *Integration Adapter* pattern is a special case of the *Adapter* pattern which refers to the integration with third-party systems that perform similar functionality but provide different interfaces, such as market data feeds. In our design, the *IntegrationAdapter* interface adopts this pattern. It is implemented by the *HTMLAdapter*, *XMLAdapter* and *SOAPAdapter* classes, which consume market data available in HTML, XML and SOAP format respectively. These classes may be further sub-classed to allow data consumption from different providers. This way, additional market data feeds may be introduced with minor modifications.
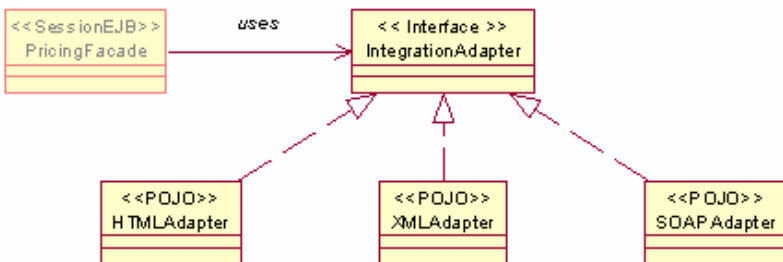
Figure 8:   Integration tier design patterns.

6 Conclusions

The present paper aims to combine the theory behind financial derivatives pricing with the latest trends in software engineering, such as the development of web-based applications, the adoption of multi-tiered architectures and the use of design patterns, in order to propose an innovative design of a web-based application for real-time derivatives pricing. Our design is entirely based on the adoption of design patterns, both generic and web-based applications specific, and incorporates some of the principal models for derivatives pricing (Black–Scholes model, binomial methods, Monte Carlo simulation). The introduction of new types of derivatives instruments, additional pricing models and alternative market data feeds is substantially facilitated by our model.

References

[1]   Alur, D., Crupi, J., & Malks, D., *Core J2EE Patterns: Best Practices and Design Strategies*, Second Edition, Prentice Hall, 2003.
[2]   Birrer, A., & Eggenschwiler, T., Frameworks in the financial engineering domain: an experience report, *Proceedings ECOOP '93*, Springer-Verlag: Berlin, LNCS 707, pp. 21-35, 1993.
[3]   Duffy, D., *Financial Instrument Pricing Using C++*, Wiley, 2004.
[4]   Eggenschwiler, T., & Gamma, E., ET++SwapsManager: Using object technology in the financial engineering domain, *Proceedings OOPSLA '92*, ACM SIGPLAN, **27(10)**, pp. 166-177, 1992.
[5]   Gamma, E., Helm, R., Johnson, R., & Vlissides, J., *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, 1995.
[6]   Hull, J., *Options, Futures and Other Derivatives*, Fifth Edition, Prentice Hall, 2003.
[7]   Joshi, M., *C++ Design Patterns and Derivatives Pricing*, Cambridge, 2004.
[8]   Koulisianis, M., Tsolis, G., & Papatheodorou, T., A web-based problem solving environment for solution of option pricing problems and comparison of methods, *Proceedings of the International Conference on Computational Science (Part I)*, pp. 673-682, 2002.
[9]   London, J., *Modeling Derivatives in C++*, Wiley, 2005
[10]   Marsura P., *A Risk Management Framework for Derivative Instruments*, M.Sc. Thesis, University of Illinois, Chicago, 1998.
[11]   van der Meij, M., Schouten, D., & Eliëns, A., Design patterns for derivatives software, *ICT Architecture in the BeNeLux*, Amsterdam, 1999.
[12]   Zhang, J. Q., & Sternbach, E., Financial software design patterns, *Journal of Object-Oriented Programming*, **8(9)**, pp. 6-12, 1996.

Section 4
Forecasting, advanced computing and simulation

This page intentionally left blank

Applications of penalized binary choice estimators with improved predictive fit

D. J. Miller[1] & W.-H. Liu[2]
[1] Department of Economics, University of Missouri, USA
[2] National Defense Management College, National Defense University, Taiwan, Republic of China

Abstract

This paper presents applications of penalized ML estimators for binary choice problems. The penalty is based on an information theoretic measure of predictive fit for binary choice outcomes, and the resulting penalized ML estimators are asymptotically equivalent to the associated ML estimators but may have a better in-sample and out-of-sample predictive fit in finite samples. The proposed methods are demonstrated with a set of Monte Carlo experiments and two examples from the applied finance literature.
Keywords: binary choice, information theory, penalized ML, prediction.

1 Introduction

The sampling properties of the maximum likelihood (ML) estimators for binary choice problems are well established. Much of the existing research has focused on the properties of estimators for the response coefficients, which is important for model selection and estimating the marginal effects of the explanatory variables. However, the use of fitted models to predict choices made by agents outside the current sample is very important in practice but has attracted less attention from researchers. In some cases, the ML estimators may exhibit poor in-sample and out-of-sample predictive performance, especially when the sample size is small. Although several useful predictive goodness-of-fit measures have been proposed, there are no standard remedies for poor in-sample or out-of-sample predictive fit.

As noted by Train [1], there is a conceptual problem with measuring the in-sample predictive fit – the predicted choice probabilities are defined with respect to the relative frequency of choices in repeated samples and do not indicate the actual

probability that a respondent takes a particular action. Consequently, researchers should focus on the out-of-sample (rather than in-sample) predictive fit of an estimated binary choice model. Accordingly, Miller [2] derives a penalized ML estimator with improved out-of-sample predictive fit by adding a measure of in-sample predictive fit to the log-likelihood function. The purpose of this paper is to compare the ML and penalized ML estimators using examples from applied financial research.

2 ML and penalized ML binary choice estimators

2.1 ML Estimation of the binary choice model

For $i = 1, \ldots, n$ independent agents, we observe $Y_i = 1$ if agent i takes a particular action and $Y_i = 0$ otherwise. The binary decision process is represented by a latent utility model, $Y_i^* = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$, where Y_i^* is the unobserved net utility associated with taking the action, $\mathbf{x}_i$ is a k-vector of individual–specific explanatory variables, $\mathbf{x}_i \boldsymbol{\beta}$ is the conditional mean component of Y_i^* that is common to all agents with characteristics $\mathbf{x}_i$, and ε_i is the mean-zero idiosyncratic error component of latent utility. The agent takes the action ($Y_i = 1$) if their net utility is positive ($Y_i^* > 0$), and the conditional probability that the agent takes the action is

$$\Pr\left[Y_i = 1 \mid \mathbf{x}_i\right] = \Pr\left[Y_i^* > 0 \mid \mathbf{x}_i\right] = \Pr\left[\varepsilon_i > -\mathbf{x}_i \boldsymbol{\beta} \mid \mathbf{x}_i\right] = F_\varepsilon\left(\mathbf{x}_i \boldsymbol{\beta}\right) \quad (1)$$

where the last equality follows if the latent error distribution is symmetric about zero. The two most commonly used model specifications for F_ε are the Normal $(0, \sigma^2)$ CDF (normit or probit model) and the Logistic$(0, \sigma)$ CDF (logit model). The response coefficients $\boldsymbol{\beta}$ are only defined up to scale, and the parameters are commonly identified under the normalization $\sigma = 1$.

Given probability model F_ε, the log-likelihood function is

$$\ell\left(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x}\right) = \sum_{i=1}^{n} Y_i \, \ln\left[F_\varepsilon\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right] + \sum_{i=1}^{n} (1 - Y_i) \, \ln\left[1 - F_\varepsilon\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right] \quad (2)$$

The associated necessary conditions for the ML estimator of $\boldsymbol{\beta}$ are

$$\frac{\partial \ell\left(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x}\right)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i' \left[\frac{Y_i \, f_\varepsilon\left(\mathbf{x}_i \boldsymbol{\beta}\right)}{F_\varepsilon\left(\mathbf{x}_i \boldsymbol{\beta}\right)} - \frac{(1 - Y_i) \, f_\varepsilon\left(\mathbf{x}_i \boldsymbol{\beta}\right)}{1 - F_\varepsilon\left(\mathbf{x}_i \boldsymbol{\beta}\right)}\right] = \mathbf{0} \quad (3)$$

where $f_\varepsilon\left(\mathbf{x}_i \boldsymbol{\beta}\right)$ is the PDF for the latent error process evaluated at $\mathbf{x}_i \boldsymbol{\beta}$. In general, the ML estimation problem does not have a closed-form (explicit) solution for the estimator of $\boldsymbol{\beta}$ (denoted $\widehat{\boldsymbol{\beta}}$), and numerical optimization tools must be used to compute the ML estimates for a given sample.

Under standard regularity conditions, the ML estimator is $\sqrt{n}$-consistent such that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ as $n \to \infty$ where $\boldsymbol{\beta}_0$ is the true parameter vector (up to arbitrary

scale). The ML estimators are also asymptotically normal as $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \overset{d}{\to}$
$\mathrm{N}\left(\mathbf{0}, \boldsymbol{\Delta}_0^{-1}\boldsymbol{\Xi}_0\boldsymbol{\Delta}_0^{-1}\right)$ where

$$\boldsymbol{\Delta}_0 \equiv \lim_{n\to\infty} \mathrm{E}\left[-n^{-1}\left.\frac{\partial^2\ell(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] \tag{4}$$

$$\boldsymbol{\Xi}_0 \equiv \lim_{n\to\infty} \mathrm{E}\left[n^{-1}\left.\frac{\partial\ell(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\left.\frac{\partial\ell(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x})}{\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] \tag{5}$$

If the binary choice model is correctly specified, the information matrix equality holds such that $\boldsymbol{\Xi}_0 = -\boldsymbol{\Delta}_0$ and the ML estimator is asymptotically efficient where $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \overset{d}{\to} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Delta}_0^{-1}\right)$.

The predicted values for each Y_i in a fitted binary choice model are derived from the estimated choice probabilities under the step function

$$\widehat{Y}_i = \begin{cases} 0 & \text{if} \quad F_\varepsilon\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\right) < 0.5 \\ 1 & \text{if} \quad F_\varepsilon\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\right) \geq 0.5 \end{cases} \tag{6}$$

where $F_\varepsilon\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\right)$ is the estimated choice probability conditional on $\mathbf{x}_i$. The standard diagnostic tool for describing in-sample predictive fit of a binary choice model is the prediction success table (see Maddala [3])

Actual	Predicted Outcomes	
Outcomes	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$
$Y_i = 1$	φ_{11}	φ_{10}
$Y_i = 0$	φ_{01}	φ_{00}

Although prediction success tables are typically reported as counts of correct or incorrect predictions, the rows of the tables in this study are stated as the conditional frequency of predicted outcomes given the actual outcomes

$$\varphi_{00} = \frac{\sum_{i=1}^n (1 - Y_i)(1 - \widehat{Y}_i)}{n_0} \quad \text{and} \quad \varphi_{01} = \frac{\sum_{i=1}^n (1 - Y_i)\widehat{Y}_i}{n_0} \tag{7}$$

$$\varphi_{10} = \frac{\sum_{i=1}^n Y_i(1 - \widehat{Y}_i)}{n_1} \quad \text{and} \quad \varphi_{11} = \frac{\sum_{i=1}^n Y_i\widehat{Y}_i}{n_1} \tag{8}$$

where $n_0 = \sum_{i=1}^n (1 - Y_i)$ is the number of observed zeroes, $n_1 = \sum_{i=1}^n Y_i$ is the number of observed ones, and $n_0 + n_1 = n$.

2.2 An information theoretic measure of predictive fit

To form a penalty function for predictive fit, Miller [2] considers the case of ideal in-sample predictive success for which the predicted outcomes $\widehat{Y}_i$ match the observed outcomes Y_i for all i. The ideal conditional outcomes for the prediction success table are denoted $\varphi_{00}^0 = \varphi_{11}^0 = 1$ and $\varphi_{01}^0 = \varphi_{10}^0 = 0$. Further, the goodness of in-sample predictive fit for an estimated model relative to the ideal case is measured as the difference between the estimated conditional distributions $\boldsymbol{\varphi}_j \equiv (\varphi_{j0}, \varphi_{j1})$ and the ideal distributions $\boldsymbol{\varphi}_j^0 \equiv (\varphi_{j0}^0, \varphi_{j1}^0)$ for $j = 0, 1$. From information theory, one plausible measure of this difference is the Kullback–Leibler cross-entropy or directed divergence functional (see Kullback and Leibler [4])

$$I\left(\boldsymbol{\varphi}_j^0, \boldsymbol{\varphi}_j\right) = \varphi_{j0}^0 \ln\left(\frac{\varphi_{j0}^0}{\varphi_{j0}}\right) + \varphi_{j1}^0 \ln\left(\frac{\varphi_{j1}^0}{\varphi_{j1}}\right) = -\ln\left(\varphi_{jj}\right) \geq 0 \qquad (9)$$

for each j. Under this divergence criterion, $I\left(\boldsymbol{\varphi}_j^0, \boldsymbol{\varphi}_j\right) = 0$ if the estimated conditional distributions coincide with the ideal case, $\varphi_{00} = \varphi_{11} = 1$ (i.e., zero predictive divergence). Otherwise, $I\left(\boldsymbol{\varphi}_j^0, \boldsymbol{\varphi}_j\right)$ increases as the observed and ideal cases diverge (i.e., there are more prediction errors).

Further, to make the penalty function suitable for estimation purposes, Miller [2] replaces the step function in eqn. (6) with a smooth approximation, $g(z, \theta) : [0, 1] \to [0, 1]$, that is continuously differentiable when θ is finite, monotonically increasing, and converge to the step function as $\theta \to \infty$. The associated approximation to the elements of the prediction success table are formed by replacing $\widehat{Y}_i$ with $g\left(F_\varepsilon(\mathbf{x}_i\boldsymbol{\beta}), \theta\right)$ in eqns. (7) and (8) above, and the approximated elements in the table are denoted φ_{jh}^a. The approximated predictive divergence functional is

$$I\left(\boldsymbol{\varphi}_j^0, \boldsymbol{\varphi}_j^a\right) = -\ln\left(\varphi_{jj}^a\right) \geq 0 \qquad (10)$$

for each j. The properties of the penalized ML estimator hold for any $g(z, \theta)$ that satisfies these conditions, and the empirical examples presented in the next section are based on the scaled hyperbolic tangent function

$$g\left(z, \theta\right) = \frac{1 + \tanh(\theta(z - 0.5))}{2} \qquad (11)$$

2.3 Sampling properties of the penalized ML estimator

Formally, the penalized ML objective function is

$$M(\boldsymbol{\beta}, \eta) = \ell\left(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x}\right) + \eta \sum_{j=0}^{1} \ln\left(\varphi_{jj}^a\right) \qquad (12)$$

and the penalized ML estimator is denoted $\widetilde{\boldsymbol{\beta}}_\eta$. The parameter $\eta \geq 0$ controls the trade-off between the log-likelihood and the predictive fit of the estimated binary

choice model. As η increases, predictive fit becomes more important in the estimation problem, and the penalized ML estimates are more strongly adjusted. The necessary conditions are

$$\frac{\partial \ell\left(\boldsymbol{\beta};\mathbf{Y},\mathbf{x}\right)}{\partial \boldsymbol{\beta}} + \frac{\eta}{n_1 \varphi_{11}^a} \sum_{i=1}^{n} \frac{\partial g_i}{\partial F_i}\frac{\partial F_i}{\partial \boldsymbol{\beta}}Y_i - \frac{\eta}{n_0 \varphi_{00}^a} \sum_{i=1}^{n} \frac{\partial g_i}{\partial F_i}\frac{\partial F_i}{\partial \boldsymbol{\beta}}(1 - Y_i) = \mathbf{0} \quad (13)$$

where $g_i \equiv g\left(F_\varepsilon(\mathbf{x}_i\boldsymbol{\beta}),\theta\right)$ and $F_i \equiv F_\varepsilon(\mathbf{x}_i\boldsymbol{\beta})$. Note that eqn. (13) reduces to the standard ML necessary condition in eqn. (3) when $\eta = 0$. For $\eta > 0$, the necessary conditions for the penalized ML estimation problem may be numerically solved for $\widetilde{\boldsymbol{\beta}}_\eta$.

The necessary conditions stated in eqn. (13) may also be used to prove the following claims about the large-sample properties of $\widetilde{\boldsymbol{\beta}}_\eta$ for finite $\eta \geq 0$:

- **Proposition 1**: $\widetilde{\boldsymbol{\beta}}_\eta$ is $\sqrt{n}$-consistent such that $\widetilde{\boldsymbol{\beta}}_\eta \xrightarrow{p} \boldsymbol{\beta}^0$.
- **Proposition 2**: $\widetilde{\boldsymbol{\beta}}_\eta$ is asymptotically equivalent to $\widehat{\boldsymbol{\beta}}$.

Formal proofs are based on the differences in stochastic order of the terms in eqn. (13) where the log-likelihood term is $O_p(n)$ and the penalty terms are $O_p(1)$ (assuming $n_1/n = O(1)$). Thus, the penalty terms have smaller stochastic order than the log-likelihood component and do not affect the first-order asymptotic properties of the ML estimator.

2.4 Predictive properties of the penalized ML estimator

In small samples, the penalty in eqn. (12) only adjusts the estimated binary choice probabilities that are *local* or limited to a small neighborhood about the 0.5 threshold in the smoothed step function, $g(z,\theta)$. The penalized ML procedure is also *adaptive* and only corrects some of the ML prediction errors without inducing other in-sample prediction errors. To prove that the method may improve predictive fit, Miller [2] provides the following existence theorem:

- **Proposition 3**: There exists some $\eta > 0$ such that $\widetilde{\boldsymbol{\beta}}_\eta$ has weakly smaller approximated in-sample predictive divergence than $\widehat{\boldsymbol{\beta}}$.

He also demonstrates the locally adaptive character of $\widetilde{\boldsymbol{\beta}}_\eta$ by showing that the fitted binary choice probabilities are increased if $Y_i = 1$ and (i) η increases (predictive fit becomes more important), (ii) $F_\varepsilon(\mathbf{x}_i\widetilde{\boldsymbol{\beta}}_\eta)$ is closer to 0.5 (observations closer to the threshold are better candidates for adjustment), (iii) n_1 decreases (smaller samples require stronger adjustment), and (iv) φ_{11}^a decreases (less favorable predictive success for observations of $Y_i = 1$ require stronger adjustment). Finally, Miller [2] shows how to use a cross-validation (CV) estimator of the penalty weight parameter η. The value of η selected under the CV criterion is denoted $\widetilde{\eta}$ and is $O_p\left(n^{1/3}\right)$ such that $\widetilde{\boldsymbol{\beta}}_{\widetilde{\eta}}$ has the same first-order asymptotic properties as $\widetilde{\boldsymbol{\beta}}_\eta$ stated in Propositions 1 and 2.

3 Examples

In this section, two examples from the applied finance literature are used to illustrate the performance of the penalized ML logit estimator (with alternative values of $\eta > 0$) relative to ML logit ($\eta = 0$). Other plausible estimators are the ML probit estimator as well as semiparametric estimators such as the maximum score estimator introduced by Manski [5, 6] and the smoothed maximum score estimator developed by Horowitz [7]. Although the maximum score estimators are expected to have good predictive fit because the objective functions are the count of correctly predicted $Y_i = 1$ outcomes, the ML logit estimator has the best predictive fit among these traditional alternatives.

3.1 Example 1: mortgage data

The first example is based on data from Dhillon, Shilling, and Sirmans [8]. The dependent variable represents the decision of a mortgage applicant to accept a fixed rate or adjustable rate mortgage (ARM) (i.e., $Y_i = 1$ if ARM), and the data include $n = 78$ observations ($n_0 = 32$ and $n_1 = 46$). The set of explanatory variables includes the fixed interest rate, the difference between the fixed and variable rates, the Treasury yield spread, the ratio of points paid on adjustable versus fixed rate mortgages, the ratio of maturities on adjustable versus fixed rate mortgages, and the net worth of the applicant. The predictive success table for the fitted ML logit model is presented in the upper left corner of table 1. Although n is relatively small, the ML logit model provides reasonably good predictive fit for the fixed rate cases (83% correct) and the ARM cases (72% correct). The prediction success results for the optimal penalized ML estimator are stated in the lower left corner of table 1. Under $\widetilde{\eta} = 11$, the prediction success rates increase to over 93% for the fixed rate case and over 81% for the ARM case. The prediction success tables for other values of η are also presented in table 1, and the fitted penalized ML model achieves perfect predictive fit as η increases above 100.

To illustrate the locally adaptive character of the penalized ML estimator, the fitted ML logit (solid line) and penalized ML logit choice probabilities (circles) are presented in figure 1. The observations are the ordered ML logit predictions $F_\varepsilon(\mathbf{x}_i\widehat{\boldsymbol{\beta}})$ so that outcomes below the 0.5 threshold are $\widehat{Y}_i = 0$ and outcomes above the line are $\widehat{Y}_i = 1$. The penalized ML logit predicted values (circles) are vertically shifted away from the solid line to reflect the locally adaptive changes in the ML logit probabilities. Note that the adjustments are small in cases with strong predictions (i.e., $F_\varepsilon(\mathbf{x}_i\widehat{\boldsymbol{\beta}}) < 0.2$ or $F_\varepsilon(\mathbf{x}_i\widehat{\boldsymbol{\beta}}) > 0.8$), and most of the adjustments to the ML logit outcomes are restricted to outcomes in a neighborhood of 0.5. In the figure, the five observations marked with 'plus' symbols were initially predicted as $\widehat{Y}_i = 0$ under the ML logit model but were corrected to $\widetilde{Y}_i = 1$ under the penalized ML procedure. Further, the three 'minus' cases were initially predicted as $\widehat{Y}_i = 1$ but were corrected under the penalized ML logit model. These eight corrected predictions account for the gain in predictive fit reported in table 1 (0.8261 + 5/46

Table 1: Prediction success tables for Examples 1 and 2.

	Example 1: Mortgage Data			Example 2: Credit Data		
	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	η	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	η
$Y_i = 1$	0.8261	0.1739	0	0.9029	0.0971	0
$Y_i = 0$	0.2812	0.7188		0.6300	0.3700	
$Y_i = 1$	0.9348	0.0652	25	0.9943	0.0057	200
$Y_i = 0$	0.1250	0.8750		0.2267	0.7733	
$Y_i = 1$	0.9348	0.0652	75	1.0000	0.0000	500
$Y_i = 0$	0.0312	0.9688		0.1233	0.8767	
$Y_i = 1$	1.0000	0.0000	101	1.0000	0.0000	3223
$Y_i = 0$	0.0000	1.0000		0.0000	1.0000	
$Y_i = 1$	0.9348	0.0652	$\widetilde{\eta} = 11$	0.9771	0.0229	$\widetilde{\eta} = 88$
$Y_i = 0$	0.1875	0.8125		0.3100	0.6900	
	$n_1 = 46$	$n_0 = 32$		$n_1 = 700$	$n_0 = 300$	

= 0.9348 for $Y_i = 1$ and 0.7188 + 3/32 = 0.8125 for $Y_i = 0$). Also, note that there are four observations among these outcomes that were correctly predicted and were not adjusted due to the adaptive character of the penalized ML estimator.

3.2 Example 2: credit data

Credit scoring models are used to predict the potential success or failure of a borrower to repay a loan given the type of loan and information about the borrower's credit history. Hand and Henley [9] note that lenders increasingly rely on statistical decision tools for credit scoring due to the large increase in loan applications and the limited number of experienced credit analysts. Fahrmeir and Tutz [10] provide a set of credit scores assigned by experienced loan analysts to $n = 1,000$ (with $n_1 = 700$ and $n_0 = 300$) individual loan applicants in southern Germany. The dependent variable is the credit risk of a loan applicant ($Y_i = 1$ for a good credit risk), and the explanatory variables include an indicator of the applicant's relationship with the lender, the level of the applicant's checking balance, the loan duration, the applicant's credit history, the type of loan (private versus professional), and an indicator of the applicant's employment status. The predictive success table for the fitted ML logit model appears in the upper right corner of table 1, and the predictive fit is relatively good for good-risk applicants (i.e., $Y_i = 1$) but is quite poor for the poor-risk cases. The predictive success table for the optimal penalized ML logit estimator appears in the lower right corner of table 1, and the predictive fit in both categories is improved relative to ML logit. The results for other values
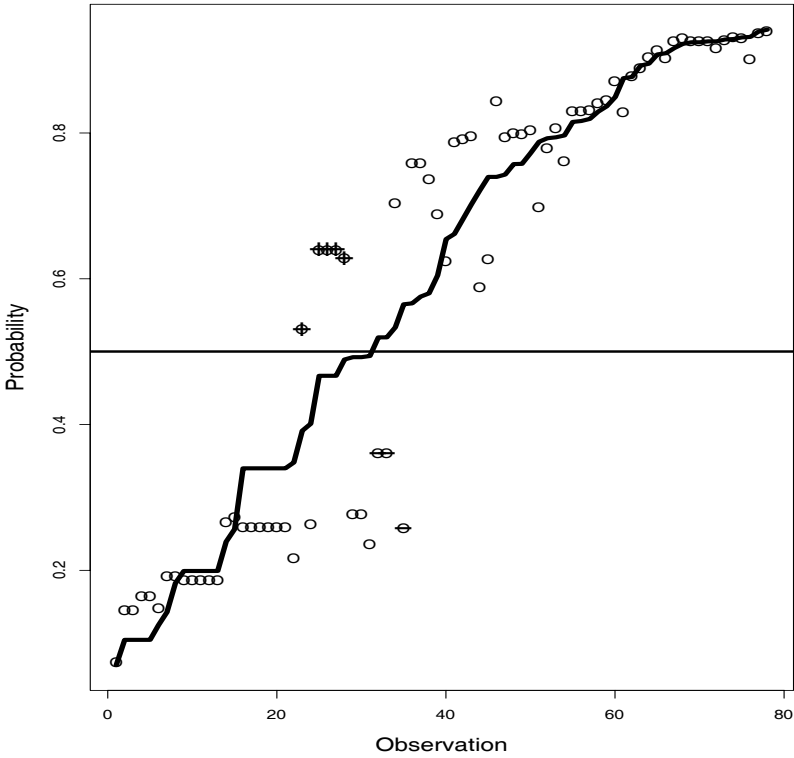
Figure 1: ML and optimal penalized ML logit predictions, Example 1.

of η are also reported in table 1, and the penalized ML logit estimator achieves perfect predictive fit for $\eta \geq 3,223$.

3.3 Out-of-sample predictive performance

Although Henley and Hand [11] show that the ML logit estimator is among the most accurate methods for predicting poor credit risks, lenders may achieve additional gains if they can further reduce the potentially large costs of making poor loans. To examine the predictive performance of the ML logit and penalized ML logit estimators, a bootstrap procedure is used to estimate the expected in-sample and out-of-sample predictive success tables given the data for Example 2. For each of $m = 5,000$ replications, $\bar{n} < n$ elements are drawn at random from the $n = 1,000$ observations, and the ML logit and optimal penalized ML logit parameter estimates are computed from the remaining $n - \bar{n}$ observations. The specified levels of the out-of-sample observation counts, $\bar{n} \in \{100, 150, 200, 250\}$, repre-

Table 2: In-sample and out-of-sample predictive success for Example 2.

	In-Sample		Out-of-Sample		
Optimal Penalized ML Logit Estimator					
	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	$\bar{n}$
$Y_i = 1$	0.9825	0.0175	0.6457	0.3543	100
$Y_i = 0$	0.3038	0.6962	0.2444	0.7556	
$Y_i = 1$	0.9837	0.0163	0.7119	0.2881	200
$Y_i = 0$	0.2996	0.7004	0.3182	0.6808	
$Y_i = 1$	0.9848	0.0152	0.7972	0.2028	400
$Y_i = 0$	0.2939	0.7061	0.4228	0.5772	
$Y_i = 1$	0.9861	0.0139	0.8749	0.1251	600
$Y_i = 0$	0.2876	0.7124	0.6023	0.3977	
Maximum Likelihood Logit Estimator					
	In-Sample		Out-of-Sample		
	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	$\bar{n}$
$Y_i = 1$	0.9097	0.0903	0.9057	0.0943	100
$Y_i = 0$	0.6410	0.3590	0.6488	0.3512	
$Y_i = 1$	0.9100	0.0900	0.9048	0.0952	200
$Y_i = 0$	0.6380	0.3620	0.6450	0.3550	
$Y_i = 1$	0.9098	0.0902	0.9052	0.0948	400
$Y_i = 0$	0.6356	0.3644	0.6387	0.3613	
$Y_i = 1$	0.9099	0.0901	0.8988	0.1012	600
$Y_i = 0$	0.6331	0.3669	0.6323	0.3677	

sent 10%, 20%, 40%, and 60% of the total observations in the data set. For each $\bar{n}$ and simulation trial $j = 1, \ldots, m$, the fitted ML logit and penalized ML logit models are used to predict the $n - \bar{n}$ in-sample and $\bar{n}$ out-of-sample bootstrap observations. The in-sample and out-of-sample prediction success tables are computed for each bootstrap trial, and the expected values of the tables are estimated by the sample averages of the replicated predictive success tables.

The bootstrap simulation results are reported in table 2. The in-sample and out-of-sample results for the ML logit estimator are quite close to the prediction success tables reported in table 1. For the optimal penalized ML logit estimator, the in-sample predictive success results are also quite comparable to the outcomes reported in table 1. As expected, the out-of-sample predictive fit is not as good for the good-risk category ($Y_i = 1$), and the ML logit estimator has better predic-

tive success. However, as noted above, the key decision error to avoid is offering a loan to a poor credit risk. For the poor-risk case ($Y_i = 0$), the optimal penalized ML logit estimator exhibits uniformly better predictive success, especially as the amount of in-sample information used to form the out-of-sample predictions increases relative to $\bar{n}$. In particular, the prediction success rate for poor credit risks is more than double the rate achieved by ML logit when $\bar{n}/n$ is only 10%. Given that the credit databases available for in-sample model estimation may be very large relative to the number of credit applications, the bootstrap evidence suggests that penalized ML logit may have significant advantages relative to ML logit in reducing the costs of extending credit to risky borrowers.

References

[1] Train, K., *Discrete Choice Methods with Simulation*. Cambridge University Press: New York, 2003.

[2] Miller, D., Penalized ML estimators of binary choice models with improved predictive fit. working paper, University of Missouri, 2006.

[3] Maddala, G.S., *Limited-Dependent and Qualitative Variables in Economics*. Cambridge University Press: New York, 1991.

[4] Kullback, S. & Leibler, R., On information and sufficiency. *Annals of Mathematical Statistics*, **22**, pp. 79–86, 1951.

[5] Manski, C., Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, **3**, pp. 205–28, 1975.

[6] Manski, C., Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics*, **27**, pp. 313–34, 1985.

[7] Horowitz, J., A smoothed maximum score estimator for the binary response model. *Econometrica*, **60**, pp. 505–31, 1992.

[8] Dhillon, U., Shilling, J. & Sirmans, C., Choosing between fixed and adjustable rage mortgages: a note. *Journal of Money, Credit, and Banking*, **19**, pp. 260–7, 1987.

[9] Hand, D. & Henley, W., Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, **160**, pp. 523–41, 1997.

[10] Fahrmeir, L. & Tutz, G., *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag: New York, 1994.

[11] Henley, W. & Hand, D., A k-nearest-neighbor classifier for assessing consumer credit risk. *Statistician*, **45**, pp. 77–95, 1996.

The use of quadratic filter for the estimation of time-varying β

M. Gastaldi[1], A. Germani[1, 2] & A. Nardecchia[1]
[1]*Department of Electrical and Information Engineering, University of L'Aquila, Monteluco di Roio, L'Aquila, Italy*
[2]*Istituto di Analisi dei Sistemi ed Informatica, CNR, Roma, Italy*

Abstract

The beta parameter is used in finance to estimate systematic risk and usually it is assumed to be time invariant. The literature shows that there is now considerable evidence that beta risk is not constant over time. The aim of this paper is the estimation of time-varying Italian industry parameter betas using a new approach based on the Kalman filter technique and on polynomial estimates. This approach is applied to returns of the Italian market over the period 1991-2001.
Keywords: time-varying beta, additive non-Gaussian noise, Kalman filter.

1   Introduction

The market effect on the returns of single assets is one of the most investigated arguments in finance. The *Capital Asset Pricing Model* (*CAPM*) suggests that the market effect is due to the relationship between the asset returns and the *market portfolio* returns. Moreover, the asset sensibility to the variations of the market portfolio returns produces the single asset expected returns. Parameter β measures the asset sensibility to the variations on the market returns [1].

In the classical financial analysis, parameter β is assumed to be time invariant and returns have a Gaussian distribution [2], but there is considerable general evidence that these assumptions are invalid in several financial markets as US markets [3] and Australia [4].

During the first 1970's researchers saw the first applications of the Kalman filter to the estimation of the systematic risk [5,6]. The proposed model for β was the *Random Walk Model* [7] requiring the estimation of the unknown variances. Many researchers investigated the validity of the *CAPM* in presence of

higher moments and their effects on asset prices. In [8] the *CAPM* was extended to incorporate the effect of skewness on the asset evaluation, while in [9] the effect of co-curtosis on the asset prices was examined.

In this work we suppose that the asset systematic risk β is time-variant non-Gaussian and we study the Italian financial market describing the relation between the assets return and the market index return by means of the market model. We assume that β follows a *Random Walk Model*. Starting from [10], where we supposed that random variables were Gaussian, we develop a new approach removing such hypothesis and we analyse a more realistic model where the random variables involved are non-Gaussian; since the knowledge of the asset return components is not complete, we assume that the moments of the random variables are unknown. Before starting with the estimation of β we need to estimate such moments, by means of a Markov estimate [11].

As already mentioned, β is non-Gaussian, therefore only the mean value and the variance of returns are not sufficient for the statistical characterization of the return distribution. In fact, it is known that in the Gaussian case the conditional expectation, which gives the minimum variance estimate, is a linear function of the observations and can be easily computed. In the non-Gaussian case this is not true, so that it is necessary to look for suboptimal estimates.

Following a state-space approach and adopting the minimum variance criterion [12], our aim is to find a more accurate estimate than the simple recursive linear one, that, as well known, admits the geometrical representation as the projection of the random β in the Hilbert space of the linear transformation of the output, namely $\text{L}(y)$. To improve such estimate our idea is to project it on the larger Hilbert space generated by the 2-nd order polynomial transformations of the output measurements, $\text{P}(y)$. Because $\text{P}(y)$ contains $\text{L}(y)$ the estimation error will decrease. Our approach requires the definition of an "extended system", in which the output is defined as the aggregate of the original output and of its second order Kronecker powers.

This paper is organised as follows. In section 2 the standard market model regression able to define an unconditional beta for any asset is presented whereas in section 3 Kalman methodology, applied to the "extended system" by which conditional time dependent betas may be estimated, is analysed. Section 4 is devoted to present time-varying betas generated for Italian data and finally section 5 presents some conclusions based on the empirical evidence obtained in this study.

2   The model

The relation between the asset return and the market index return can be expressed as follows:

$$R_{i,t} = \alpha_{i,t} + \beta_{i,t} R_{M,t} + \varepsilon_{i,t} \qquad\qquad t = 1,...,T \qquad (1)$$

where:
- $R_{i,t}$ is the return for the asset i during the period t;

- $R_{M,t}$ is the return for the market index during the period t;
- $\alpha_{i,t}$ is a random variable that describe the component of the return for the asset i which is independent from the market return;
- $\varepsilon_{i,t}$ is the random disturbance vector such that:
    - $E(\varepsilon_{i,t}) = 0;$ $\quad\quad \forall i, \forall t$
    - $E(\varepsilon_{i,t}\varepsilon_{j,t}^T) = 0;$ $\quad\quad \forall i, \forall j, \forall t, i \neq j$
    - $E(\varepsilon_{i,t}\varepsilon_{i,\tau}^T) = 0;$ $\quad\quad \forall i, \forall t, \forall \tau, \ t \neq \tau$
    - $E(\varepsilon_{i,t}R_{M,t}^T) = 0.$ $\quad\quad \forall i, \forall t$

Equation (1) shows that the return for the asset i during the period t, $R_{i,t}$, depends on the return for the market index $R_{M,t}$ on the same time. Moreover, the relation between these two variables is linear.

Coefficient β is the most important parameter; it shows how asset returns vary with the market returns and is used to measure the asset systematic risk, or market risk.

2.1 Random Walk model: hypothesis for our work

In literature there are many models able to describe systematic risk. All of them can be represented by a simple two equation model. There are numerous studies assuming that asset prices follow the *Random Walk* model (RW) [7]. The Random Walk model can be expressed as follows

$$R_{i,t} = \alpha_{i,t} + \beta_{i,t}R_{M,t} + \varepsilon_{i,t} \tag{2}$$
$$\alpha_{i,t} = \alpha_{i,t-1} + u_{i,t} \tag{3}$$
$$\beta_{i,t} = \beta_{i,t-1} + \eta_{i,t} \tag{4}$$

We assume that the random variable β_0 (initial condition) and the random sequences $\{\varepsilon_{i,t}\}$, $\{u_{i,t}\}$ and $\{\eta_{i,t}\}$ satisfy the following conditions for $t \geq 0$:

- $E\{\varepsilon_{i,t}\} = 0, E\{u_{i,t}\} = 0, E\{\eta_{i,t}\} = 0, E\{\beta_0\} = 0;$ $\tag{5}$
- all the noises moments up to the 4th order are finite;
- the noises $\{\varepsilon_{i,t}\}$, $\{u_{i,t}\}$ and $\{\eta_{i,t}\}$ are the sequences of independent non-Gaussian random variables.

We remark that no knowledge is assumed on the noises moments values; before proceeding is helpful to represent the Random Walk Model in the state space.

2.2 System equations

It is possible to define observation and state equations:
- ***observation equation*:**

$$R_{i,t} = y(t) = C(t)x(t) + \psi(t) \tag{6}$$

This equation represents the market model with time-varying coefficients. Matrix $C(t)$ has dimensions $T \times 2$ so that each row will represent the observations at certain point in time; this matrix has the following structure

$$C(t) = [1 \mid R_{M,t}] \tag{7}$$

and is assumed to be known.

The *state vector* $x(t)$ has dimensions 2×1 and represents the α and β

coefficients at time t:

$$x(t) = [\alpha_t \mid \beta_t]^T. \qquad (8)$$

$\psi(t)$ is the part of the asset return $y(t)$ which is not modelled and represents the random sequence $\{\varepsilon_{t,t}\}$. The first four moments of the output noise are unknown, assumed finite and indicated by $E\left(\psi^{[h]}(t)\right)$, $h = 1,\dots,4$.

- *state equation* assumes this general form

$$x(t) = Ax(t-1) + \zeta(t) \qquad (9)$$

The first four moments of the state noise $\zeta(t)$ are assumed finite (for hypothesis), indicated by $E\left(\zeta^{[h']}(t)\right)$, $h' = 1,\dots,4$ and its values are unknown.

In the model adopted in the present work (RW), matrix A is the 2×2 identity matrix while vector $\zeta(t)$ models the random part of the state vector:

$$\zeta(t) = [u_t \mid \eta_t]^T. \qquad (10)$$

Note that the values of the state noise moments depend on the moments of the random sequences $\{\eta_{i,t}\}$ and $\{u_{i,t}\}$, so that it is necessary to estimate six parameters – second, third and fourth moments of the sequences $\{\eta_{i,t}\}$, $\{u_{i,t}\}$ (for hypothesis all the random sequences are zero mean).

Moreover we must estimate second, third and fourth moments for the three noise considered sequences. We represent these unknown parameters as a vector represented by $\vartheta = \left(\sigma_u^2, \sigma_u^3, \sigma_u^4, \sigma_\eta^2, \sigma_\eta^3, \sigma_\eta^4, \sigma_\varepsilon^2, \sigma_\varepsilon^3, \sigma_\varepsilon^4\right)$.

3 The quadratic rilter and β estimation

As we have already seen in section 2, our aim is to find the minimum variance estimate of the state with respect to the output that coincides with its conditional expectation. While in the Gaussian case we obtain exactly a linear optimal solution, in our case the problem does not have an immediately recursive solution, so that we look for suboptimal estimates that are more accurate than the linear one.

To develop our approach, we need to use Kronecker algebra. Definitions and theorems that are necessary can be found in [13].

3.1 The extended system

To obtain the desired recursive estimates of (6) and (9) we define the 2-degree polynomial observation $Y \in \Re^\mu$, $\mu = m + m^2$, where m is the output dimension (in our case $m=1$)

$$Y(t) = \begin{bmatrix} y(t) \\ y^{[2]}(t) \end{bmatrix} \qquad (11)$$

and the extended state $X \in \Re^\chi$, $\chi = n + n^2$, where n is the state dimension (in our case $n=2$)

$$\mathsf{X}(t) = \begin{bmatrix} x(t) \\ x^{[2]}(t) \end{bmatrix} \tag{12}$$

where with $y^{[2]}(t)$ and $x^{[2]}(t)$ we denote, respectively, the 2nd Kronecker power of the vectors y and x.

We can now calculate the second Kronecker power of the state and the output equations

$$x^{[2]}(t) = A^{[2]}(t)x^{[2]}(t-1) + \zeta^{[2]}(t) + A(t)x(t-1) \otimes \zeta(t) + \zeta(t) \otimes A(t)x(t-1) \tag{13}$$

$$y^{[2]}(t) = C^{[2]}(t)x^{[2]}(t) + \psi^{[2]}(t) + C(t)x(t) \otimes \psi(t) + \psi(t) \otimes C(t)x(t) \tag{14}$$

where with the symbol $\otimes$ we denote the Kronecker product.

By using some properties of the Kronecker algebra, it is possible to rewrite previous equations in a compact form and give the equations of the extended system

$$\mathsf{X}(t) = \mathsf{A}\mathsf{X}(t-1) + \mathsf{N}'(t) + \mathsf{U}(t;\vartheta) \tag{15}$$

$$\mathsf{Y}(t) = \mathsf{C}(t)\mathsf{X}(t) + \mathsf{N}''(t) + \mathsf{V}(t;\vartheta)$$

where:

$$\mathsf{X}(t) = \begin{bmatrix} x(t) \\ x^{[2]}(t) \end{bmatrix} \quad \mathsf{Y}(t) = \begin{bmatrix} y(t) \\ y^{[2]}(t) \end{bmatrix} \quad \mathsf{A} = \begin{bmatrix} A & 0 \\ 0 & A^{[2]} \end{bmatrix}$$

$$\mathsf{C}(t) = \begin{bmatrix} C(t) & 0 \\ 0 & C^{[2]}(t) \end{bmatrix} \quad \mathsf{U}(t;\vartheta) = \begin{bmatrix} 0 \\ E\left(\zeta^{[2]}(t)\right) \end{bmatrix} \quad \mathsf{V}(t;\vartheta) = \begin{bmatrix} 0 \\ E\left(\psi^{[2]}(t)\right) \end{bmatrix} \tag{16}$$

$$\mathsf{N}'(t) = \begin{bmatrix} \zeta(t) \\ \zeta(t) - E\left(\zeta^{[2]}(t)\right) + \left(I_2 + C_{2,2}^T\right)\left[Ax(t-1) \otimes \zeta(t)\right] \end{bmatrix}$$

$$\mathsf{N}''(t) = \begin{bmatrix} \psi(t) \\ \psi(t) - E\left(\psi^{[2]}(t)\right) + 2\left[C(t)x(t) \otimes \psi(t)\right] \end{bmatrix}$$

indicating the dependence of vectors U and V on θ. Matrix I_n is the identity matrix of dimension $n \times n$ and matrix $C_{\cdot,\cdot}^T$ is a commutation matrix [14].

We call system (15) *augmented system*. Its state and observation noises ($\mathsf{N}'(t)$ and $\mathsf{N}''(t)$ respectively) are zero mean uncorrelated sequences and are also mutually uncorrelated at different times. For these noises we are able to calculate their autocovariances (for the initial hypothesis their cross covariance is null). Interested reader can found their expressions in [14]. Hence, for the augmented system the optimal linear state estimate can be calculated by means of the Kalman filter equations.

3.2 Quadratic filter

In economic systems, the covariance matrices for the various noise processes in the model are assumed to be known and assigned *a priori*. In this paper we estimate the covariance matrices by means of the observations of the returns to individual assets and the market portfolio.

We can define the following cost index to be minimized in order to obtain the desired estimation

$$J(\vartheta) = \sum_{t=1}^{T}\left(\left[\mathsf{Y}(t)-\mathsf{V}(t;\vartheta)-\mathsf{C}(t)\big(\mathsf{A}\hat{\mathsf{X}}(t-1)+\mathsf{U}(t;\vartheta)\big)\right]^{T}\left[\mathsf{C}(t)P(t\,|\,t-1;\vartheta)\cdot\right.\right.$$
$$\left.\left.\cdot\mathsf{C}^{T}(t)+R(t;\vartheta)\right]^{T}\left[\mathsf{Y}(t)-\mathsf{V}(t;\vartheta)-\mathsf{C}(t)\big(\mathsf{A}\hat{\mathsf{X}}(t-1)+\mathsf{U}(t;\vartheta)\big)\right]\right), \tag{17}$$

where $Pp(t|t\text{-}1;\theta)$ is the prediction covariance and $R(t;\theta)$ is the covariance of the output equivalent noise (16). The above function has been minimized by means of the Markov estimate [15]. When the estimation $\hat{\vartheta}$ of the parameter vector is calculated, the optimum estimation of the extended state vector is obtained by means of the Kalman filter, by using the system matrices evaluated for $\hat{\vartheta}$.

Using the obtained results and taking into account the deterministic and the stochastic input we can use the Kalman filter for the extended system.

The filter need to be initialised; initial conditions for the state vector and for the prediction covariance matrix are:

$$\hat{\mathsf{X}}(0\,|-1) = E\{\mathsf{X}(0)\} = 0,\ P(0\,|-1) = E\{\mathsf{X}(0)\mathsf{X}(0)^{T}\} = \Psi_{\mathsf{X}(0)}$$

Afterwards, it is possible to proceed with the estimation algorithm. At each time t, following steps are reiterated:

$$P_{p}(t) = \mathsf{A}\,P(t-1)\mathsf{A}^{T} + \mathsf{Q}(t;\hat{\vartheta}) \tag{18}$$

$$K(t) = P(t\,|\,t-1)\mathsf{C}^{T}(t)\big(\mathsf{C}(t)P(t\,|\,t-1)\mathsf{C}^{T}(t)+\mathsf{R}(t;\hat{\vartheta})\big)^{-1} \tag{19}$$

$$P(t) = [I - K(t)\mathsf{C}\ (t)]P(t\,|\,t-1) \tag{20}$$

$$\hat{\mathsf{X}}(t\,|\,t-1) = \mathsf{A}\,\hat{\mathsf{X}}(t-1)+\mathsf{U}\,(t;\hat{\vartheta}) \tag{21}$$

$$\hat{\mathsf{X}}(t) = \hat{\mathsf{X}}(t\,|\,t-1)+K(t)\big(\mathsf{Y}\,(t)-\mathsf{C}(t)\hat{\mathsf{X}}(t\,|\,t-1)\big) \tag{22}$$

where $K(t)$ is the filter gain, $P(t)$ and $P(t|t\text{-}1)$ are respectively the filter and prediction covariances.

The optimal linear estimate of the augmented state process $\mathsf{X}\ (k)$ with respect to the augmented observations $\mathsf{Y}\ (k)$ agrees with its optimal quadratic estimate with respect to the original observations $y(k)$, in the sense of taking into account the second power of $y(k)$. We obtain in this way the optimal quadratic estimate of the system (6) and (9). The optimal linear estimate of the original state $x(k)$ with respect to the same sets of augmented observations is easily determined by extracting the first n components in the vector $\hat{\mathsf{X}}\ (k)$ (recall that in our case $n=2$).

The optimal estimate of parameter at each time t is then determined by extracting the second component in the vector $\hat{x}(t)$.

We stress that the proposed algorithm, if we do not calculate the second power of the observations, produces the best linear filter, which coincides, as is well known, with the optimal filter when the noises are Gaussian. Consequently, it becomes necessary to consider higher order filters when the noises have distribution far from the Gaussian. By observing formulas that define the augmented system parameters, it becomes evident that the computational effort of the polynomial filter quickly grows with increasing filter order. However, we point out that even low-order polynomial filters (the quadratic filter considered in our case) which do not require a particular sophisticated implementation, show very high performances with respect to the linear filter.

3.3 Goodness of the proposed method

We assess the accuracy of the forecast using the MAE (Mean Absolute Forecasting Error indices) and MSE (Mean Square Forecasting Error) indices [16]:

1. Mean Absolute Forecasting Error: once we forecast $\hat{R}_{it}$ it is possible to measure estimation accuracy using a measure of forecast error which compares the forecast to actual values by

$$MAE_i = \sum_{t=1}^{T} \frac{\left| \hat{R}_{it} - R_{it} \right|}{T} \qquad (23)$$

A potential problem with the use of MAE measure is that all errors have the same weight. An alternative approach is to give an heavier penalty on outliers then the MAE measure with the use of squared term by the following index:

2. Mean Square Forecasting Error (MSE):

$$MSE_i = \sum_{t=1}^{T} \frac{\left(\hat{R}_{it} - R_{it} \right)^2}{T} \qquad (24)$$

Table 1:   Statistics for weekly returns data.

ISX Industry	Mean	Standard Deviation	Skewness	Kurtosis
Food (7)	0.0973	3.9622	7.0161	108.0010
Insurance (19)	0.1936	3.4726	0.6320	5.4050
Transport (13)	0.2253	2.8966	0.3902	5.0322
Banks (53)	0.2647	2.6466	0.8518	7.2513
Paper (2)	-0.0041	4.3221	0.9565	6.3610
Chemicals (21)	0.2173	2.5915	0.7134	4.9909
Building materials (13)	0.1973	3.2434	0.5810	4.2747
Distribution (6)	0.3348	3.5637	0.5201	4.7679
Publishing (11)	0.3351	3.8691	1.5298	11.6685
Electronics (29)	0.1712	2.5269	0.6609	5.4241
Diversified financials (4)	0.3863	4.7617	4.0280	34.2872
Financial holdings (29)	0.1610	3.2995	0.5838	4.5906
Real estate (21)	0.2525	3.3365	1.2264	7.2478
Equipments (9)	0.3017	3.0579	0.6262	5.4661
Miscellaneous industries (2)	0.1313	5.6184	0.1467	11.4537
Minerals (7)	0.2209	3.0296	0.7040	6.1690
Public utility (18)	0.3482	2.5856	0.4932	3.5527
Financial services (3)	0.0577	3.6974	0.6715	4.8213
Textile (27)	0.2305	2.9455	5.5834	77.8821
Tourism and leisure (14)	0.3072	3.0166	1.0332	6.0254
Market Index	*0.2198*	*2.1500*	*0.5214*	*4.7846*

4   Empirical results

The concept of beta is well known in the financial community and its values are estimated by various technical service organizations.

Generally speaking, we expect that aggressive companies or highly leveraged companies have high betas, whereas companies whose performance is unrelated

to the general market behaviour have low betas. In this paper the data used are weekly price relative information for 20 Italian Stock Exchange industries provided by TraderLink s.r.l.. Our full sample period is extended from May 1991 to June 2001. The data were expressed in Italian lyres and percentage returns were created for the analysis.

In table 1 are reported information about the distributional properties of the industry sector returns used in our study; in the first column the name of the industry and the number of considered firms in each sector (in parenthesis) are reported. Note that there is a correlation between the risk of each industry (the standard deviation) and the number of firms in each sector. In fact, the standard deviation for the industry with the largest number of firms (Banks – 53 companies) has a smaller value than the Paper industry (2 firms). Distribution of the industry return is leptokurtic. Moreover Diversified financials, Food and Textile exhibit high level of skewness.

Table 2:   MAE and MSE forecast error results.

ISX industry	MAE			MSE		
	Linear Filter	Quadratic Filter	Improvement ($|MAE_Q-MAE_L|$)	Linear Filter	Quadratic Filter	Improvement ($|MAE_Q-MAE_L|$)
Food	0.8061	1.7020e-2	0.7891	1.3679	6.1831e-4	1.3673
Insurance	0.7073	1.4641e-2	0.6926	0.9412	4.1161e-4	0.9408
Transport	0.5907	1.2154e-2	0.5785	0.6453	2.7302e-4	0.6450
Banks	0.4515	9.1051e-3	0.4424	0.3894	1.6032e-4	0.3892
Paper	1.2874	2.5247e-2	1.2622	3.2974	1.2571e-3	3.2961
Chemicals	0.4741	9.9739e-3	0.4641	0.4217	1.8257e-4	0.4215
Building materials	0.6840	1.4281e-2	0.6697	0.9303	4.2506e-4	0.9299
Distribution	0.8611	1.8579e-2	0.8425	1.3955	6.6215e-4	1.3948
Publishing	0.9296	1.8035e-2	0.9116	1.8858	6.9256e-4	1.8851
Electronics	0.4771	9.3423e-3	0.4677	0.4118	1.6036e-4	0.4116
Diversified financials	1.1486	2.1778e-2	1.1268	3.4293	1.2267e-3	3.4281
Financial holdings	0.5255	1.0353e-2	0.5151	0.4930	1.9494e-4	0.4928
Real estate	0.7240	1.3892e-2	0.7101	1.1918	4.4448e-4	1.1914
Equipments	0.7728	1.5111e-2	0.7577	1.1851	4.5278e-4	1.1846
Misc. industries	1.6283	3.2820e-2	1.5955	6.2683	2.4329e-3	6.2659
Mineral	0.7874	1.5627e-2	0.7718	1.2061	4.9693e-4	1.2056
Public utilities	0.6433	7.2584e-3	0.6360	0.7298	9.7258e-5	0.7297
Financial services	1.0708	2.1248e-2	1.0495	2.1797	8.8923e-4	2.1788
Textiles	0.5191	9.9448e-3	0.5091	0.4816	1.7282e-4	0.4814
Tourism and leisure	0.6969	1.3613e-2	0.6833	1.0309	3.8915e-4	1.0305

The standard market model was estimated for every Italian industry, using the domestic market index. To evaluate the performance of beta estimates we calculate the MAE and MSE metrics presented above ((23)-(24)). The MSE and MAE measures are presented in table 2.

Notice that the proposed method (the quadratic filter) produced in all 20 industries low level of forecast error demonstrating the effectiveness of the chosen estimation approach.

It is important to emphasize that quadratic filter follows variations of β parameter better than the linear one, so that the output restored by means of the estimated parameters in the case of quadratic filter is more similar to the true output than the output obtained by means of the linear filter, as shown in the

following figures 1 and 2.

In these figures is represented a comparison between a portion of the true output (returns for the Public utilities sector) and the restored output so that it is possible to better appreciate the performances of the two filters. It is evident that in the quadratic case the restored output practically coincides with the true output.
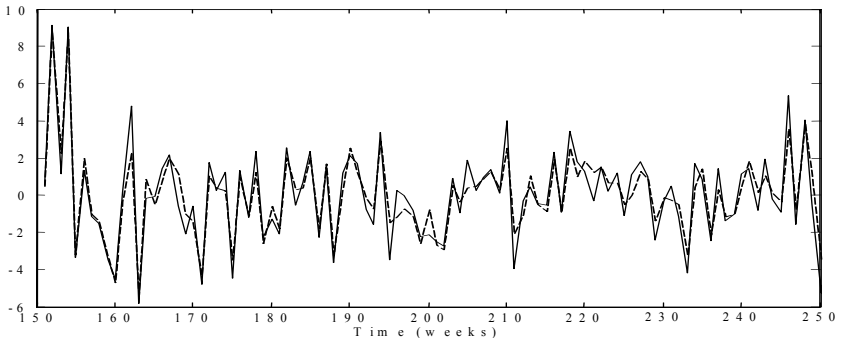

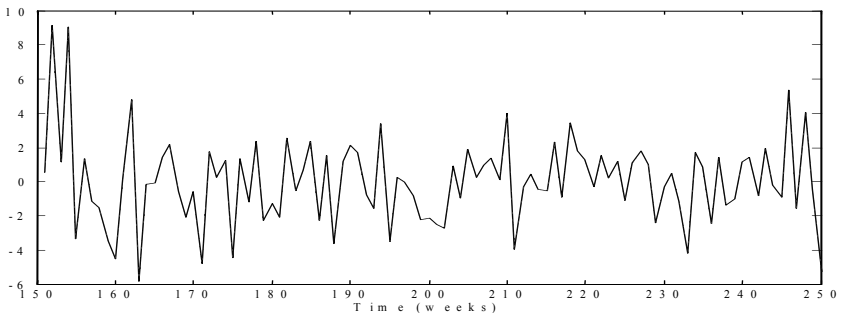Figure 1:   Matching between true and restored output (Linear filter).

Figure 2:   Matching between true and restored output (Quadratic filter).

5   Conclusions

In this paper we face the problem of systematic risk beta estimation. The presented results show that it is possible to estimate conditional time-dependent betas applying the quadratic filter to a sample of returns on Italian industry portfolios over the period 1991-2001. The obtained results by the proposed method are indeed much more accurate than those obtained by the classical linear filtering.

References

[1]   R.J. Fuller and J.L. Farrell, *Analisi degli investimenti finanziari*, McGraw-Hill: Milano; 1993.

[2]   E.F. Fama, "Risk, return and equilibrium: some clarifying comments", *Journal of Finance*, vol.23 n.1, 1968, pp 29-40.

[3]   F.J. Fabozzi and J.C. Francis, "Beta as a random coefficient", *Journal of Financial and Quantitative Analysis*, vol. 13, 1978, pp 101-115.

[4]   R.W Faff, J.H.H. Lee and T.R.L. Fry, "Time stationarity of systematic risk: Some Australian evidence", *Journal of Business Finance and Accounting*, vol. 19, 1992, pp 253-270.

[5]   M. Kantor, "Market Sensitivities", *Financial Analysts Journal*, vol.27 n.1, 1971, pp 64-68.

[6]   K. Garbade and J. Rentzler, "Testing the hypothesis of beta stationarity", *International Economic Review*, vol. 22 n.3, 1981, pp 577-587.

[7]   C. Wells, *The Kalman Filter in Finance*, Kluwert Academic Publishers: Dordrecht; 1996.

[8]   R.S. Sears and K.C.J. Wei, "Asset Pricing, Higher Moments, and the Market Risk Premium: a note", *Journal of Finance*, vol. 40, 1985, pp 1251-1253.

[9]   H. Fang and T-Y. Lai, "C-Kurtosis and Capital Asset Pricing", *The Financial Review*, vol. 32 n.2, 1997, pp 293-307.

[10]   M. Gastaldi and A. Nardecchia, "The Kalman filter approach for time-varying β estimation", *System Analysis Modelling Simulation*, vol.43 n.8, 2003, pp 1033-1042.

[11]   L. Lyung, System identification – theory for the user, New York: Prentice Hall; 1987.

[12]   F. Carravetta, A. Germani and M. Raimondi, "Polynomial Filtering for Linear discrete time non-Gaussian systems", *SIAM J.Control Optim.*, vol.34 n.5, 1996, pp 1666-1690.

[13]   R. Bellman, *Introduction to Matrix Analysis*. New-York: McGraw-Hill; 1970.

[14]   M. Dalla Mora, A. Germani and A. Nardecchia, "Restoration of Images Corrupted by Additive non-Gaussian Noise", *IEEE Trans. on Circuits and Systems I: Fundamental Theory and Applications*, vol.48 n.7, 2001, pp 859-875.

[15]   A.V. Balakrishnan, *Kalman Filtering Theory*. New York: Optimization Software, Inc., Publication Division; 1984.

[16]   M.D. McKenzie, R.D. Brooks and R.W. Faff, "The use of domestic and world market indexes in the estimation of the time-varying betas", *J. of Multinational Financial Management*, vol.10, 2000, pp 91-106.

Forecast of the regional EC development through an ANN model with a feedback controller

G. Jianquan[1,3], Fankun[2], T. Bingyong[1], B. Shi[3] & Y. Jianzheng[3]
[1]Donghua University, Shanghai, People's Republic of China
[2]Shanghai Maritime University, Shanghai, People's Republic of China
[3]University of Shanghai Science and Technology,
People's Republic of China

Abstract

This paper is to have a deep understanding of the way to forecast the economic development with the help of an Artificial Neural Network (ANN), putting forward a brand-new ANN forecast model, that is, the Back Propagation Networking Learning Algorithm (BP Networking Algorithm) with a feedback controller. The model has been used to overcome the deficiencies of the traditional BP Algorithm, as it is more accurate for forecasting, less dependable on initial data, and easier to select the needed number of hidden layers and hidden-layer neurons. In order to measure regional electronic commerce development we have set an evaluation system, which seems to be comparatively perfect and manipulative. With the model and the system, we carried out a regional EC forecast in Huai'nan, a medium-sized city in Anhui Province, China. The result of the case study has indicated that the model has an ideal extension, the number of its hidden-layer neurons can easily be decided, and we are to have a long-term forecast of the development without much initial data. With this model in hand, it is possible to cope with the problems of sparse, dispersed and hard-to-forecast statistical information in the development of electronic commerce.
Keywords: feedback controller, BP model, EC development, forecast.

1   Introduction

There has appeared all over the world a new business model—electronic commerce with the rapid development and application of information technology and communication technology represented by Internet and mobile technology. And based on this, a brand new economic formation has come into being—the Internet economy. As far as the nature of the Internet economy is concerned, it is global, but we may easily find that it has some regional characteristics in its development [1]. If we put it in the scale of the globe, it appears to be "North American" [2]. If we narrow our insight to the Mainland China, we also have the same phenomenon in this field. There is much faster EC development in the Yangtzi River Delta than those in the inner part of the mainland. Therefore, some economists advocate "the Ribbon Development Strategy"—focusing our attention of the development of the electronic commerce along the coastal regions, and "the Centralized Development Strategy" [3]—initiating the development of the E-business in Beijing, Shanghai, and Guangdong Province where there are adequate web users.

It is of great importance to have a study of the different level of the EC development in different regions. First of all, EC stands for the new economy or the Internet economy. The EC development represents to a great extent the development of the Internet economy.

Secondly, the world seems to run out of natural resources, and there are more and more countries and regions showing solicitudes for this. The Internet economy has become a platform for the growth in many economies as it has its inherent attributes of low energy costing, and many economic entities have been pursuing a sustainable development with as little consumption of natural resources as possible.

Thirdly, with the help of electronic commerce—a new business model, some less developed economies have got a short cut to catch up with the development of other countries and to have a close connection with the rest of the world. Anyway, national and regional competitiveness in the age of the Internet will require "being in the loop" more than ever before [4].

John C. Scott put forward a model called Internet Maturity in 2000, which highlighted the 4 stages of the development of the Electronic Commerce in businesses. It also explained thoroughly the way businesses stepped onto the highly developed stage of EC with such techniques as integrated skills and reengineering. This model was developed in somewhat the same way as the three stages of EC development presented by Yang Jianzheng in his *Principles and Applications of the Electronic Commerce* and the Tri-level Model of EC development and the Bi-level Model of EC development in 2003 China E-business Almanac. Unfortunately, these theories or models do not touch upon the study of the EC development in different regions.

It is considered difficult to implement the study of EC development in different regions because of the three handicaps: the construction or selection of models, the construction of measurement systems, and the collection of initial

data for statistics. This paper is to deploy studies in these three spheres respectively.

2   Construction of the model

The recently developed ANN Model is an active branch in artificial intelligence. ANN is a newly developed information processing system on the bases of the study of modern neurology, which simulates the biologic nerve system and seems to be able to process an array of information simultaneously. It can be used to process information by association, generalization, analogy, and reasoning. It has an advantage of self-learning, the capability of distilling features, summing up knowledge, and forecasting futures on the gained experiences. It is also full of adaptability, systematization, and an ability of learning, associating, infrastructure problem solving, and noise eliminating [5]. Therefore, ANN has its bright future in the economic forecast. A few Chinese specialists have set foot in this field. But if we use the traditional BP ANN model, it will be very difficult to ascertain the number of its hidden-layer neurons or the units in each layer, and will prolong the time for study [5]. On the bases of study of the economic forecast with ANN, we try to put forward a new ANN forecast model—the BP Model with Feedback Controller. Ours, we think, is more accurate, easier for the selection of the number of hidden-layer neuron and the units in each layer with fewer initial data needed. It has overcome the shortcomings of the traditional BP Models and become more applicable.

2.1 ANN with feedback controller

Our model is an amelioration of the Error Back Propagation Network. The BP Model is a multi-layer feed forward artificial neural network, which is composed of input layers, hidden layers, and output layers, and each layer has one or more neurons (Figures 1 and 2). There is no connection in the same layer, but there exists among the neighbouring layers.

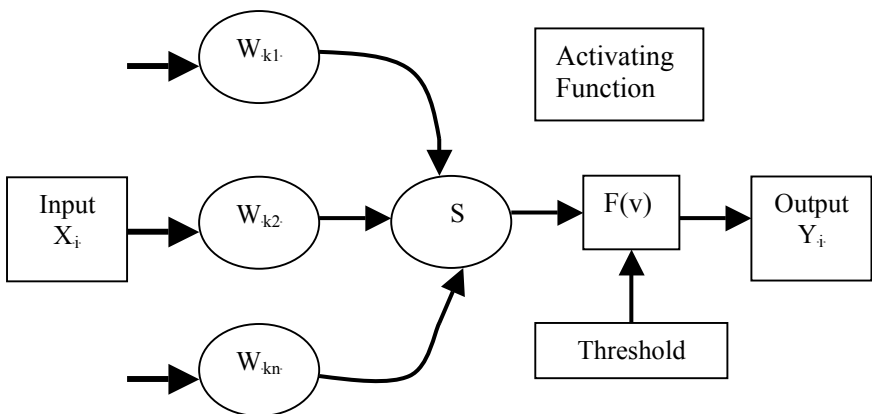
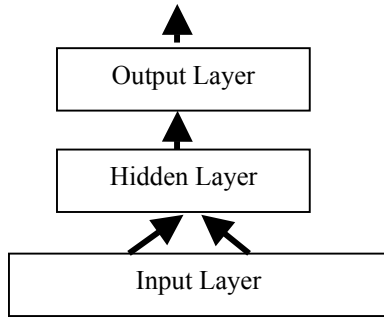Figure 1:   Neuron.

Figure 2:   Neural network constructions.
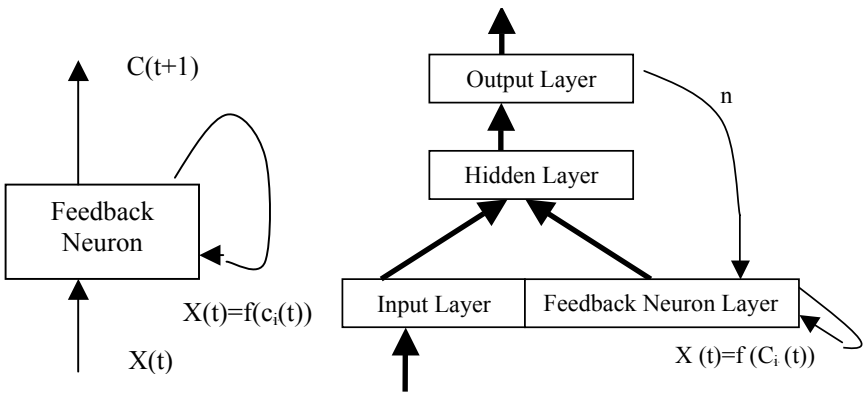

Figure 3:   Feedback neuron.

Figure 4:   Feedback   neural   network
              structure.

But the ordinary BP Networks could only achieve an ideal result of forecast with adequate samples and enough time for measurement when forecasting economic development. Practically, however, we always need to do some forecast on condition that there is not very much statistics. This is the reason why we try to improve the BP Networks with a feedback controller added. (Figures 3 and 4).

There could be one or more units for feedbacks accordingly to different questions. There may exist various kinds of feedback controlling functions, but usually simple function is enough to solve the ordinary problems. Our renovated neural network has developed from a static state to a dynamic one. Especially when $f(x) = x$, it will degenerate into a BP Network with some co-connected neurons. The net-learning arithmetic usually adopts Error Back Propagation Algorithm.

2.2 Network-learning algorithm

BP (Error Back Propagation) is a multi-layer artificial neural network, comprising input, hidden and output layers. There is full inter-layer connection but no intra-layer connection of neurons. Figure 5 demonstrates a three-layer BP neural network with nine neurons.
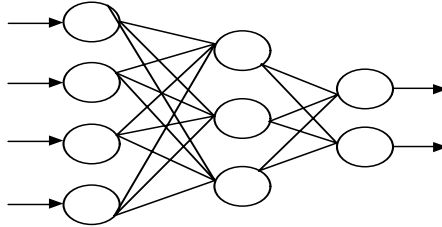

Figure 4:   A three-layer BP neural network.

2.2.1 Rationale of BP network

Suppose the input mode vector $A_k=(a_1, a_2, a_3,…a_n)$, k=1, 2, 3…m. Here, m is the number of learning modes; n is the number of neurons in the input layer. Correspondingly, the expected output vector $Y_k=(y_1, y_2, y_3,…,y_q)$, and q is the number of the output neurons.

The calculation process of the input of the neuron in each hidden layer is follows:

$$s_j = \sum_{i=1}^{n} w_{ij}a_i - \theta_j \; , j=1, 2,…,p \tag{1}$$

In this formula, w_{ij} is the connection weight ranging from the input layers to hidden layers; θ_j is the threshold value of neuron in the hidden layer; p is the number of the neurons in the hidden layer.

To simulate the non-linear features of the biologic neurons, make s_j the independent variable of the sigmoid function, so as to calculate the output of each neuron in the hidden layer. The Sigmoid function is as follows:

$$f(x) = 1/(1+e^{-x/x_0}) \tag{2}$$

Here $f(x)$ is activation function, and the activation value of the neurons in the hidden layer is:

$$b_j = f(s_j) \quad j=1, 2,…,p \tag{3}$$

While information is flowing from the input layer to the output layer, if we provide the input information, we can get an output as follows:

$$L_t = \sum_{j=1}^{n} v_{jt}b_j - \gamma_t \tag{4a}$$

$$c_t = f(L_t) \, , t=1,2,…,q \tag{4b}$$

It has been theoretically proved that there exists a three-layer network that can achieve the mapping action of any consecutive function with whatever accuracy required [6].

To carry out the mapping action, the network needs to be trained through the following steps:

1. Initialization of the weight value and threshold value. Choose at random an initialized weight value and threshold value from the interval (0,1);

2. Set input vector A and output vector Y;

3. Calculate the actual output vector C;

4. Revise the weight value, starting from the output layer, propagate the error signal backward, and try to minimize the error by revising different weight values;

5. Adopt $Y_k = (y_1^k, y_2^k, ..., y_n^k)$, the desirable output mode, and $\{C_t\}$, the actual network output, to calculate $\{d_j^k\}$, the error of different neurons in the hidden layer; its formula is as follows:

$$d_j^k = (y_t^k - c_t) \cdot c_t (1 - c_t) , \; t=1,2,...,q \tag{5}$$

6. Use $\{v_{jt}\}$, the connection weight, $\{d_t\}$, the error, and $\{b_j\}$, output of the hidden layer, to calculate the error of different neurons in hidden layers, namely $\{e_j^k\}$.

$$e_j^k = (\sum_{i=1}^{q} d_t \cdot v_{jt}) \cdot b_j (1 - b_j) , \; j=1,2,...,p \tag{6}$$

7. To revise v_{jt}, the connection weight, and γ_t, the threshold value by using $\{d_j^k\}$, the error of different neurons in the output layer and $\{b_j\}$, the output of different neurons in the hidden layers:

$$v_{jt}(N+1) = v_{jt}(N) + \alpha \cdot d_t^k \cdot b_j , \; j=1,2,...,p; \; t=1,2,...,q \tag{7}$$

$$\gamma_t(N+1) = \gamma_t(N) + \alpha \cdot d_t, (0 < \alpha < 1) \tag{8}$$

8. To revise $\{w_{ij}\}$, the connection weight, and $\{\theta_j\}$, the threshold value, by using $\{e_j^k\}$, the error of different neurons in the hidden layers and A_k, the input of different neurons in the input layers.

$$w_{ij}(N+1) = w_{ij}(N) + \beta \cdot e_j^k \cdot d_i^k , \; i=1,2,...,n; \; j=1,2,...,p \tag{9}$$

$$\theta_j(N+1) = \theta_j(N) + \beta \cdot e_j^k , \; j=1,2,...,p \tag{10}$$

9. Choose the next learning mode for the network, return to step 3, until all (m) modes are finished with the training.

10. Once again, choose at random a mode from m, return to step 3, if global error E is smaller than a preset small value, then the neural network is

convergent. Or else, if learning time is bigger than a preset value, which means the network cannot converge any more. The formula is as follows:

$$E = \sum_{k=1}^{m} \sum_{t=1}^{q} (y_t^k - c_t)^2 / 2 \tag{11}$$

BP algorithm is actually a kind of gradient algorithm, namely:

$$w(t+1) = w(t) + \eta(-\frac{\partial E}{\partial w})\Big|_{w=w(t)} \tag{12}$$

3   Construction of index system

The level of regional EC development can be used to reflect the integrated situation of the development of Electronic Commerce in that region. Therefore, it is necessary to select all the indexes from various spheres for the assessment.

With the consideration of the function of different sub-systems and the logical relationship between different levels of sub-systems, this paper will, in measuring the development level of the regional E-business Y, break the measurement system down into four first-grade sub-systems, which are: trading capability X_1, supporting trading capability X_2, development potential X_3 and governmental support X_4. Each first-grade sub-system is composed of several minor indexes. The particular index system is shown in the following table 1.

4   Case studies

4.1   Background information and initial data

This study is based on the practice in Huai'nan, Anhui Province. As a major city for coal and power generation, the medium-sized city has many big energy enterprises spread in several districts. Those businesses are generally advanced in information processing and hoist the EC development in the city. In order to promote the electronic commerce, the city started in 2004 a project called Digital Huai'nan. The project will be unfolded in all the 7 districts of the city, that is, tianjia'an, Panji, Maoji, Bagongshan, Xiejiaji, Datong and Fengtai.

4.2   Analysis of the model construction and calculation

This paper is to forecast the development of the EC transactions in the districts in Huai'nan with BP Model. The analysis has its foundation of assessments, and the logic of the assessment of the EC development is as follows:

The index X_1 is achieved by calculation of the 4 items: X_{11}, X_{12}, X_{13} and X_{14}. X_2 is achieved by calculation of the 3 items: X_{21}, X_{22}, and X_{23}. Of the 3 indexes X_{21}, X_{22}, and X_{23}, X_{21} is calculated through the following 5 items: X_{211}, X_{212}, X_{213}, X_{214}, and X_{215}. X_{22} is calculated through the 3 items: X_{221}, X_{222}, and X_{223}. X_{23} is calculated through the 4 items: X_{231}, X_{232}, X_{233}, and X_{234}. X3 is calculated through the 5 items: X_{31}, X_{32}, X_{33}, X_{34} and X_{35}. X4 is calculated

through the following 3 items: X_{41}, X_{42}, and X_{43}. The index Y is calculated through X_1, X_2, X_3, and X_4.

Table 1:   The index system.

Overall capability Y	Trading capability X_1	The percentage of e-business turnover in GDP X_{11}	
		The percentage of e-business dealers X_{12}	
		The extent to which the dealing cost has been reduced X_{13}	
		The extent to which the dealing time has been reduced X_{14}	
	Supporting trading capability X_2	Supporting trading capability of infrastructure X_{21}	Degree of popularity of computer X_{211}
			Degree of popularity among net-user X_{212}
			The percentage of enterprises net-users X_{213}
			Credit card per head X_{214}
			The proportion of investment on e-business in total investment X_{215}
		Supporting trading capability of labor resource X_{22}	The proportion of e-business personnel in the overall employed X_{221}
			The proportion of e-business personnel with bachelor degree or above in the overall employed X_{222}
			The proportion of e-business teaching program participators in the overall teaching program participators X_{223}
		Supporting trading capability of management and safety X_{23}	Available or unavailable of e-business safety center X_{231}
			The proportion of installation of anti-virus software in computers X_{232}
			The proportion of updating anti-virus software in computers X_{233}
			The proportion of virus-related damages in the overall business turnover X_{234}
	Potential of development X_3	The average ADR of net shares X_{31}	
		The average price-to-earnings ratio of net shares X_{32}	
		Degree of popularity among net-user X_{33}	
		The percentage of enterprises net-users X_{34}	
		The market accessibility of e-business X_{35}	
	Government support X_4	The availability of special fund to support e-business X_{41}	
		The availability of special project arrangements to support e-business X_{42}	
		The availability of government measures to support e-business X_{43}	

We calculate the level of EC development as follows:

The 3-layer BP Model is used for the calculation. We set different numbers of the input and output neurons according to the different requirements for indexes. At the same time, some adjustment was also made to the number of the hidden-layer neurons. For example, X_1, the capability of EC transactions, adopts 4 input neurons (X_{11}, X_{12}, X_{13} and X_{14}) and one output neuron (X_1), while 8 neurons were chosen for the number of neurons in hidden-layer. All the other indexes were processed more or less the same way as X_1.

By these ways the overall capability Y of each district is calculated.

4.3 Result of the calculations

4.3.1 The calculation for the forecast of EC development

The calculation is accomplished with the 3-layer BP model. Because the problem comes across as a non-linear time series problem, it is not proper to use an ordinary BP neuron network. Therefore, we use a BP neuron network with some controlling functions. Among which, there is one input neuron, one output neuron, and 10 neurons in the hidden-layer. Based on the initial data, we tried to forecast the 2005 EC development in various districts of Huai'nan. Listed below is only the result of the forecast of EC development in Tianjia'an District. (Table 2 and Figure 6.)

Table 2:   The forecast of EC development in Tianjia'an District.

Year	Tianjia'an District
2005	0.686299
2004	0.5815752
2003	0.4895732
2002	0.4097342
2001	0.3412378
2000	0.2830835
1999	0.23417

4.3.2 Analysis of the forecast of the EC development

We have got several unique characteristics from the result of forecast. First, there have been evident developments of the EC transactions in all the districts, which is relevant to the domestic and international economic environment. Second, as far as the EC development in the past few years is concerned, Tiania'an, Fengtai, and Xiejiaji Districts are the first three in transaction amounts and the growth rate. This reflects the global reality that in the launching stage of the EC development, the regions, which have solid economic foundations usually, take the lead. Third, we are once again assured from the assessment that the major driver of the EC development, that is, the government support, plays a very important role in this field. The draft of the EC development from 2002 to 2003 in various regions shows us that China Electronic Administration Year promoted greatly the EC development in these areas. Fourth, B2B transactions are the main force in the EC development in all the districts.
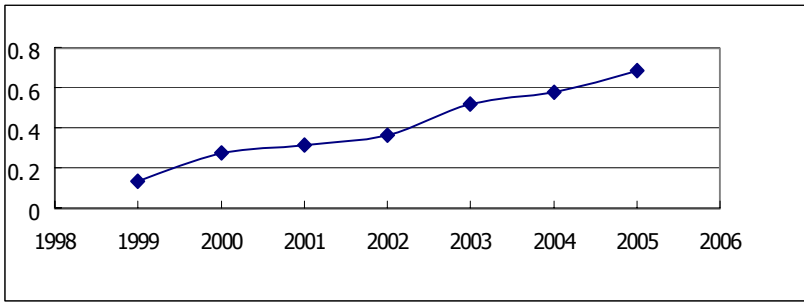
Figure 6:   The forecast of EC development in Tianjia'an District

5   Conclusion

This paper has introduced a new scientific means to assess and forecast the EC development in a region. The ANN with a feedback controller adopted in our study has solved the problem of sparse, dispersed and hard-to-forecast statistical information in the development of the electronic commerce. We have constructed a model for assessment and forecast, and implement some calculation with initial data from a sample region. The ANN is a data-oriented method of analysis. We took the model of this kind because the regional EC development is new area for study, and we have not had much systematic arithmetic analysis. One the other hand, the problem we have is systematically sophisticated, non-linear, multi-indexed, and non-adequate, so we are not able to deal with it with the traditional arithmetic models. The ANN is also full of the abilities of self-learning, self-organizing, self-adapting, and problem solving, and is a proper choice for the study of new and sophisticated systems.

References

[1] Guojianquan, Analysis of Network Economy. *Journal of East China Normal University*. **No.1**, pp.56-61, 2004.
[2] Joanne E Oxley, Bernard Yeung, E-commerce Readiness: Institutional environment and international competitiveness. *Journal of International Business Studies*, **(4)**, pp.705-706, 2001.
[3] Yang Jianzheng, *Principles and Applications of the Electronic Commerce*. Publication of University of Xi'an Electronic and Technology: Xi'an pp. 144~145, 2001.
[4] Edward E. Leamer, Michael Stroper, The Economic Geography of the Internet Age. *Journal of International Business Studies*, **(4)**, pp. 660~661, 2001.
[5] Jiao Licheng, *Algorithm of Neural Network*, Publication of University of Xi'an Electronic and Technology, Xi'an, pp. 249-294, 1995.
[6] Jiao Licheng, *System Theory of Neural Network*. Publication of University of Xi'an Electronic and Technology, Xi'an, 1995.

Section 5
Market analysis,
dynamics and simulation

This page intentionally left blank

The impact of the futures market on spot volatility: an analysis in Turkish derivatives markets

H. Baklaci & H. Tutek
Izmir University of Economics, Turkey

Abstract

The derivatives market in Turkey has been in operation since February 2005. This paper examines the impact of future trading on spot volatility by using Istanbul Stock Exchange 30 (ISE 30) Index future contracts which represent the most frequently traded future contracts in Turkish derivatives market.

The main objective of this paper is to investigate whether the existence of future markets in Turkey has improved the rate at which new information is impounded into spot prices and have any persistence effect.

The results gathered from the study indicate that even though it has been in operation for a short period of time, the futures market in Turkey has significantly increased the rate at which new information is transmitted into spot prices and that it has reduced the persistence of information and volatility in underlying spot market resulting in improved efficiency.

The results of this study have also some important implications for policy makers discussed in the final section of this paper.

Keywords: derivatives market, volatility, spot market, GARCH.

1   Introduction

There has been an ongoing debate on the impact of derivative markets on spot markets in terms of volatility, information flow, destabilizing spot markets and their speculative effects. Majority of the studies exploring the above impacts have been conducted on the developed markets, and particularly on U.S. (see for example, Board et al. [2]; Edwards [12]). On the other hand, there are only a few researches on emerging markets such as South Korea, India and Taiwan. (see for example Ryoo and Smith [18]; and Nath [16]).

However, this issue is even more important for the developing countries for the following reasons:

- Previous studies conducted on the developed markets have documented that futures markets have contributed to the efficiency of spot markets because of their impact on rapid impounding of information into prices. In previous studies it is observed that financial markets in the developing countries have been less efficient. Accordingly, it is crucial to examine whether futures trading has any effect on increasing the efficiency of spot markets in developing countries and whether the initiation of futures trading has a significant impact on price discovery in these markets.

- On the other hand, some studies revealed that (see for example Butterworth [16]), in case the futures market exerts a destabilizing influence on spot markets through speculative trading, then there should be some policy-making implications for governmental authorities.

- Turkey has been and will be one of the most appealing emerging markets in the near future for the institutional investors particularly for foreign investors. As solid evidence, the share of foreign investors in ISE (Istanbul Stock Exchange) has increased to 67% in 2005 (Istanbul Stock Exchange Statistics). Foreign direct investment has also increased in recent years, particularly in 2005 exceeding $9 billion (Turkish Central Bank, Balance of Payments 2005). In addition, being a candidate state to join EU with its rapid economic growth in the past few years, Turkey is considered to be one of the 'rising stars' for foreign investors in the near future. In this respect, as being one of the latest derivative market initiated in February 2005, the role of futures trading in Turkey and its impact on spot markets are crucial issues to be investigated.

Therefore, the objective of this study is to investigate whether the existence of future trading improve the rate at which new information is impounded into spot prices and have a persistence effect and also to determine whether the introduction of future trading has a significant impact on price discovery in the Turkish spot market.

The methodology used attempts to determine whether spot price volatility changes after the initiation of future trading.

2 Literature review

The previous literature includes various studies debating on how the introduction of derivatives market, particularly the futures market, has affected the volatility of associated spot markets. The majority of these studies has examined this effect by using stock indices and has reached mixed results.

One set of results gathered from these studies has concluded that the introduction of derivatives market had no effect or sometimes even decreased the spot market volatility. This result has been mainly attributed to the fact that derivatives market has increased the speed at which the information or news is impounded into spot prices. Thus, the proponents of this argument further claimed that the initiation of derivatives market has contributed to the efficiency of spot markets.

On the other hand, some studies have reached completely opposite results signifying that the derivatives market led to an increased volatility in underlying spot markets. These studies have associated this result to the existence of large speculative trading and activity, which in turn was claimed to destabilize and amplify the volatility in spot markets.

In the rest of this section, some of the selected studies including the controversial findings mentioned above will be discussed.

Holmes [15] has studied the impact of future trading on spot volatility using FTSE Index and Generalized Autoregressive Conditional Heteroskedasticity (GARCH from now on) methodology. He has proposed that the post futures volatility is less than pre futures in FTSE Index suggesting that the future trading increases the rate at which information is impounded into prices. He has also argued that the future trading has reduced the persistence of information flowing to underlying spot market.

Bologna and Cavallo [4] has reached similar results using GARCH modelling in Italian markets. Like Holmes, they have argued that the futures market has decreased spot market volatility by augmenting the speed at which the news is impounded into spot prices leading to increased market efficiency.

Two studies investigating the impact of futures trading on spot volatility in Indian market have come up with similar results. Nath [16] and Gupta and Kumar [14] have examined the impact of futures trading in Nifty and Nifty Junior indices and both have found that stock market volatility has declined after the introduction of futures markets in India.

Edwards [12] using a larger dataset including S&P 500 Index, Value Line Index, T-Bills and Eurodollar Time Deposits has investigated the change in asset price volatility following the derivatives markets. Likewise, he has also claimed that the introduction of futures has improved the speed and quality of information flowing to the spot market contributing to the spot market efficiency.

As another advocate of the same argument, Shenbagaraman [19] has deduced the fact that derivatives had no significant impact on spot market volatility, and that the persistence of information has diminished after derivatives resulting in more efficient spot markets.

In contrast to the findings of above studies, various researches have alleged that the introduction of derivatives market has augmented the volatility in underlying spot markets. Strikingly, some of the researchers have allied this outcome with the existence and volume of speculative activity in derivatives markets and thus have asserted the destabilizing impact of derivatives market on spot markets. On the contrary, some researchers have related the volatility

increase in spot market to the increased efficiency arising from the faster transmission of information from derivatives market to underlying spot market.

As a recent study, Ryoo and Smith [18] for instance have examined the impact of futures trading on spot Market in Korea. Their results signified that the future market in Korea has increased spot market volatility but also has increased the speed at which new info is impounded into spot prices leading to similar deductions as the supporters of counter arguments.

Using Mid250 future contracts, Butterworth [6] has gathered similar results arguing that the increase and persistence in volatility after futures trading could be adhered to the illiquidity of Mid250 contract. Antoniu and Holmes [1] and Chiang and Wang [10] have observed the same patterns for FTSE-100 Index and Taiwanese markets, respectively. Antoniu and Holmes have also acknowledged that the nature of volatility has not changed post-futures for FTSE-100 index following the futures trading.

Unlike many other researches utilizing GARCH model and daily closing prices Chiang and Wang [10], have tested the volatility impact by utilizing GJR model and by using high-low prices to proxy for the intraday volatility. Their results, have also displayed an increased volatility in Taiwanese market subsequent to futures trading.

Employing a larger sample, Yu [21] has detected volatility transmission between futures and spot markets for USA, France, Japan, Australia, UK, Hong Kong and has pointed out that the spot market volatility increases after stock futures in all countries except UK and Hong Kong.

3 Empirical analysis

The empirical analysis consist of three parts: First, the model used for testing the impact of futures market on spot volatility will be discussed followed by the explanation of data specifications. Finally, the results obtained from the analysis will be discussed along with their implications.

3.1 Methodology

The impact of futures trading on the underlying spot market can be examined by isolating price volatility peculiar to the underlying spot market by removing the impact of general market wide volatility. In order to capture the market wide volatility and isolate the market specific volatility on which futures contract is written, the spot price changes (returns) are regressed on a proxy variable for which there is no related futures contract by utilizing the following model [22]:

$$SPC_t = a_0 + a_1 EMIC_t + \varepsilon_t$$
$$\varepsilon_t = N(0, h_t)$$

$$(1)$$

where

$$SPC_t = \text{spot price change in period t (ISE 30)},$$

$EMIC_t$ = Price change in market proxy variable in period t (MSCI Emerging Market Index),

$$\varepsilon_t = \text{error term representing unexplained price changes}$$

The above model is used to isolate price volatility peculiar to the spot market underlying futures by removing the impact of global market wide volatility in which MSCI Emerging Market Index is used to proxy global market wide volatility. Thus, the error term captures the impact of factors specific to the futures market and variance of ε_t proxies price volatility specific to the futures market.

There are two major reasons for selecting MSCI Emerging Market Index as a proxy:

  a) There is no futures contract written on MSCI Emerging Market Index and it also includes Turkish Stock Market. Besides, by the increased effect of globalisation, the capital and information flow has amplified between emerging markets reflecting a higher correlation **.**

  b) A diagnostic test was made by regressing ISE30 on MSCI Emerging Market Index and the results of the regression are provided in Table 1. The results of the regression further support the argument that MSCI Emerging Market Index can be postulated as a good proxy since the coefficient parameter for MSCI Emerging Market Index (0.904) is close to unity and the R-squared as well as F-statistics for the model are quite high.

The error terms from Equation 1 representing market specific volatility for ISE30 are further analysed by the following GARCH representation :

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-i}^2 + \beta_1 h_{t-i} \qquad (2)$$

In Equation (2), α_1 represents the impact of new information and β_1 represents the persistence effect of information. Thus, the parameters in Equation (2) in the pre and post futures trading allows us to discover how futures trading has impacted the underlying spot market volatility and to what extent. Thus, an increase in α_1 in post-futures period proposes that news is impounded into prices more rapidly following the futures trading. Accordingly, a decrease in α_1 in post-futures period implies a slower information transmission into prices throughout the post-futures period. Similarly, a decline in β_1 specifies that information have a less persistent effect on price changes whereas an increase in β_1 signifies higher persistence. Thus, α_1 and β_1 parameters in Equation 2 for the pre and post-futures period would not only allow determining whether there

is a marked change in spot price volatility following the futures market but also allow determining whether the changes in volatility are due to more rapid impounding of information or by the destabilizing speculation effect which increases persistency of volatility and information transmission.

Table 1:   Diagnostic regression results. (ISE30 = Dep. Variable, MSCI Emerging Market Index = Independent Variable. t-statistics are provided in parentheses.)

	Coefficient
Intercept	0.126
	(1.778)
MSCI Em. Market Index	0.904 ***
	(10.982)
F statistics	120.62
R-squared	0.197
Observations	493

*** Significant at 1% level.

3.2 Data

The daily closing price indexes of ISE30 and MSCI Emerging Market Index for the period February 2004 to February 2006 are used to examine the impact of futures trading. In estimating Equation 1, the daily price changes are used to achieve stationarity. The data for ISE30 are gathered from www.analiz.com , an online financial data site and the data for MSCI Index are obtained from MSCI website. After excluding non-trading days for both indices and matching dates for both datasets, the final sample includes 493 observations. The whole sample is further segregated into two sub samples: The pre-futures period and post-futures spanning from February 2005 to February 2006, which includes 243 observations. (Due to the limited number of observations and data for the post-futures period, the pre-futures period observations were limited to one-year data to achieve consistency in the number of observations.)

3.3 Results

The descriptive statistics for the daily changes in ISE30 and MSCI Emerging Market Index for the pre and post-futures periods are provided in Table 2. As observed from Table 2, the mean and standard deviation of daily price changes exhibit similar changes for both indices. Particularly, while the mean of daily returns have increased for post-futures period for both indices, the standard

deviation of both indices have declined in the same period indicating that post-futures volatility is lower for both indices for the post-futures period. The skewness parameters for both indices, particularly for MSCI Index reveal that daily price changes do not conform to a normal distribution.

Table 2:   Descriptive statistics of return changes in ISE 30 and MSCI Emerging Market Index.

| Period | N | ISE 30 | | | MSCI | | |
		Mean	Std.Dev.	Skewness	Mean	Std.Dev.	Skewness
Pre-futures (Feb. 2004-Feb.2005)	250	0.1958	1.8009	-0.0851	0.0495	0.9175	-0.9621
Post-futures (Feb. 2005-Feb. 2006)	243	0.21	1.6788	-0.2605	0.1225	0.7862	-0.4441
Whole sample	493	0.2146	0.0772	-0.0752	0.1133	0.8571	-0.4975

The volatility impact of futures can be further analysed by examining the GARCH parameters in Table 3 obtained by estimating Equation 1 and 2 for both sub sample periods.

Table 3:   GARCH estimations.

Period	a_0	a_1	α_0	α_1	β_1
Pre-futures (Feb. 2004-Feb.2005)	0.1624 (1.51)	0.6743 (5.76)***	0.2665 (0.60)	0.0586 (1.17)	0.8471 (4.35)***
Post-futures (Feb. 2005-Feb. 2006)	0.0595 (0.67)	1.2283 (10.92)***	1.573 (2.54)**	0.1615 (1.65)*	0.0032 (0.01)

*** Significant at 1% level.
** Significant at 5% level.
* Significant at 10% level.

The results from the regression equation (Equation 1) and GARCH estimations (Equation 2) for each sub-period are provided in Table 3. The results are mixed in the sense that, even though α_0 and α_1 are statistically insignificant for the pre futures period, the same parameters turn out to be significant for the

post-futures period. Conversely, persistence parameter (β_1) is significant for the pre-futures period and insignificant for the post-futures period.

Likewise, there is a marked increase in the news coefficient (α_1) and a marked decrease in persistence parameter (β_1) after the futures trading. These results imply that the existence of futures market has increased the rate at which new information is incorporated into underlying spot prices and a fall in the persistence of information. These results are consistent with the findings from most of the other studies on this topic suggesting that the futures market improves the efficiency of spot markets by a faster transmission from futures to spot market and that the futures market has a stabilizing impact on the Turkish stock market. These results also suggest that the price discovery occurs first in futures market for the Turkish stock market. However, these findings have to be further analyzed but since the futures exchange in Turkey was established in February 2005, data for the post-futures period is limited to only one year.

These results also have some vital implications for policymakers in Turkey. Firstly, commencing from 2006, government has imposed 15% capital gains tax on the majority of marketable securities traded in Turkish financial markets. However, the capital gains from futures trading has been excluded from this tax burden to encourage trading since the volume of trading was considered to be thin for the derivatives market. In this respect, the results of this study also assert that policymakers should provide similar incentives such as reducing the minimum trading size for Turkish derivatives market because of its major contribution to the efficiency of underlying spot markets.

Secondly, controversial to some of the findings in other emerging markets, the results of this particular study show no destabilizing effect of futures market on spot market in Turkey arising from speculative trading. However, because of the limited data for the post-futures period at the time, the results might be subject to a sampling bias. Thus, the authorities should still monitor the speculative movements in Turkish derivatives market for their possible destabilizing effect on underlying spot markets for future periods. In this regard, failure to inspect the causes of any possible changes in derivatives market might lead to inapt policy recommendations for the regulation of futures trading.

4   Conclusion

Since its inception in February 2005, the trading volume and interest of investors in Turkish derivatives exchange has been steadily increasing. This paper examines the impact of futures trading on the underlying spot market volatility in Turkish stock market by using ISE30, a stock index comprised of 30 large size firms in Turkey, on which future contracts are written and traded. The impact of futures markets is investigated by separating the whole sample into two sub periods that contain pre and post-futures trading periods.

As of this date, this is the first study that examines the impact of futures market on spot market in Turkey. Thus, the results obtained from this study are

considered to have some important inferences for further study on this topic in Turkish financial markets.

The evidence gathered from this study demonstrate that despite its short history, the existence of futures market has significantly improved the rate at which new information is impounded into spot prices and has reduced the persistence of information and volatility in underlying spot market resulting in improved efficiency.

The results of this study have also some important implications for policy makers highlighting the fact that the incentives for the futures market should be strengthened because of its constructive effect on the underlying spot markets. However, these results must also be analyzed very cautiously. Since the sample for the post-futures period for this study cover only one year span, a possible rise in the speculative trading in the derivatives market for the future periods might have a detrimental influence on the underlying spot markets by their potential destabilizing effect. Thus, policy-making authorities should closely monitor the existence of speculative trading activity.

References

[1]   Antoniou, A., & Holmes, P. "Futures trading, information and spot price volatility: evidence for the FTSE-100 Stock Index Futures contract using GARCH.", *Journal* of *Banking* & *Finance*, 19 (1), p117-129, 1995.

[2]   Board, John, Sandmann G., & Sutcliffe C. "The Effect of Futures Market Volume on Spot Market Volatility", *Journal of Business Finance and Accounting*, 28(7) and (8), pp.799-819, 2001.

[3]   Bollerslev T., "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, 31, pp.307-327, 1986.

[4]   Bologna, P., & Cavallo, L. "Does the introduction of futures effectively reduce spot market volatility? Is the futures effect immediate? Evidence from the Italian stock exchange using GARCH", *Applied Financial Economics*, 12, pp.183-192, 2002.

[5]   Brailsford TJ. , Frino A., Hodgson A.,& West A. "Stock market automation and the transmission of information between spot and futures markets", *Journal Of Multinational Financial Management*, 9(3-4), pp.247-264,1999 .

[6]   Butterworth, D., " The Impact of futures trading on underlying stock index volatility: the case of the FTSE Mid250 contract", *Applied Economics Letters*, 7,pp.439-442, 2000.

[7]   Chan, K. "A Further Analysis of the Lead-lag Relationships between the Cash Market and Stock Index Futures Market", *The Review of Financial Studies*, 5, pp.123-152, 1992.

[8]   Chan, K., Chan, KC., & Karolyi G.A. "Intraday volatility in the stock index and stock index futures markets", *The Review of Financial Studies*, 4, pp.657-684, 1991.

[9]   Chatrath, A., Kamtah, R., Chakornpipat, R., & Ramchander, S. "Lead-lag associations between option trading and cash market volatility", *Applied Financial Economics*, 5, pp.373-381, 1995.

[10]   Chiang , Min-Hsien, Wang, Cheng-Yu. "The impact of futures trading on spot index volatility: evidence for Taiwan index futures", *Applied Economics Letters*, 9, pp.381-385, 2002

[11]   Darrat, A., Rahman, S., & Zhong, M. "On the Role of Futures Trading in Spot Market Fluctuations: Perpetrator or Volatility or Victim or Regret?", *The Journal of Financial Research*, 25 (3), pp.431-444, 2002.

[12]   Edwards, F. R. "Futures Trading and Cash Market Volatility: Stock Index and Interest Rate Futures", *Journal of Futures Markets*, , 8(4), pp.421-439, 1988.

[13]   Frino A., Walter T., and West A. "The Lead–Lag Relationship between Equities and Stock Index Futures Markets Around Information Releases", *Journal of Futures Markets*, , 20(5), pp.467-487, 2000.

[14]   Gupta O. P., Kumar M. "Impact of Introduction of Index Futures on Stock Market   Volatility: Indian   Experience",   2002, http://www.pbfea2002.ntu.edu.sg/papers/2070.pdf.

[15]   Holmes, Phil. "Spot Price Volatility, Information And Futures Trading: Evidence From A Thinly Traded Market", *Applied Economics Letters*, 3, pp.63-66,1996.

[16]   Nath, G. C., "Behavior of Stock Market Volatility after Derivatives", 2003, http://www.nse-india.com/content/press/nov2003a.pdf.

[17]   Racine MD. , Ackert LF. "Time-Varying Volatility in Canadian and US Stock Index and Index Futures Markets: A Multivariate Analysis, *Journal of Financial Research*, 23(2), pp.129-144, 2000.

[18]   Ryoo, Hyun-Jung, Smith, G. "The Impact of stock index futures on the Korean stock market", *Applied Financial Economics*, 14, pp.243-251 , 2004.

[19]   Shenbagaraman P. " Do Futures and Options Trading Increase Stock Market Volatility?" *NSE Working Papers,* Paper No: 60, 2003.

[20]   Soydemir G., Petrie G. " Intraday information transmission between DJIA spot and futures markets", *Applied Financial Economics,* 13, pp.817-827, 2003.

[21]   Yu, Shang-Wu. "Index futures trading and spot price volatility", *Applied Economics Letters*, 8, pp.183-186, 2001.

[22]   Holmes, Applied Economics Letters, July 1995.

A valuation model of credit-rating linked coupon bond based on a structural model

K. Yahagi & K. Miyazaki
The University of Electro-Communications, Japan

Abstract

A credit-linked coupon bond pays a coupon associated with its credit rating at the time of the coupon payment date, rather than an amount equal to the initially fixed coupon. The only existing corporate bond valuation model for credit-rating-triggered products was formulated by Jarrow et al. However, this model does not incorporate the fact that increases in the coupon payment resulting from downgrades may cause a further deterioration of credit ratings and of the likelihood that the company will be able to make future coupon payments. In this paper, we present a credit-linked coupon bond valuation model that considers this issue. Using a structural approach, we extend the classical model of Merton by introducing a threshold value corresponding to each credit rating, and a volatility of the company value process that depends on its credit rating. Given these extensions, our model is more flexible than the JLT model, and we are clearly able to capture the above effect via numerical simulations. Furthermore, from the perspective of practical implications, the JLT model tends to value credit-linked coupon bonds more cheaply than does our model when the initial credit rating is high, while the reverse is true for a low initial credit rating.
Keywords: risk management, derivative pricing, credit risk.

1   Introduction

The formulation and use of corporate bond valuation models dates from the work of Merton [5]. In the Merton model, the default of a bond is defined as a state in which the corporate value falls below the face amount of the bond, and in which the corporate value process follows a geometric Brownian motion. As a result of these assumptions, the Merton model may easily be used in conjunction with the Black-Scholes formula to value corporate bonds. Using valuation frameworks of this kind is typically characterised as following a "structural approach," and

many extensions of the Merton model have been derived. Another avenue for corporate bond valuation is relatively new and is known as the "reduced form approach." The latter approach assumes that the time to default may be modelled as a hazard rate. Famous and representative reduced form models include those of Jarrow and Turnbull [4] (the JT model), Jarrow et al. [3] (the JLT model), and Duffie and Singleton [2]. Among these structural and reduced form models, only the JLT model explicitly uses a rating transition matrix in modelling the time to default.

Given such preceding research on the valuation of the corporate bond, the JLT model at first glance appears the most suitable for the valuation of credit-rating-triggered bonds, such as the credit-rating-linked coupon bond. However, in order to incorporate the idea that the increased coupon payment due to downgrading deteriorates the potential for future coupon and notional payments, the impact of increased coupon payments on the balance sheet of the company must be considered, in addition to the credit-rating transition itself. In this paper, for the purpose of valuing credit-rating-linked coupon bonds, we further develop the ideas presented by Bhanot [1] by considering an analogue of the JLT model in a structural context.

The remainder of the paper is organised as follows. The next section briefly reviews the Merton and JLT models, and presents the motivation for our research. Section 3 proposes our valuation model and its means of calibration. Section 4 examines various features of the model using numerical examples. The final section summarises and concludes.

2 Prior research and the motivation for our model

2.1 Merton model

The Merton model assumes that the value of the company follows a next geometric Brownian motion:

$$\frac{dV_t}{V_t} = \mu dt + \sigma dW_t, \tag{1}$$

where μ, σ, and W_i are, respectively, the drift and volatility of the corporate value process and a standard Brownian motion under the usual statistical measure.

In order to value a corporate bond, the Merton model first transforms process (1) into one under a risk-neutral probability measure, such as process (2) below:

$$\frac{dV_t}{V_t} = rdt + \sigma d\widetilde{W}_t, \tag{2}$$

where r, σ, and W_t are, respectively, the risk-free short rate, the volatility of the corporate value process, and a standard Brownian motion under the usual risk-neutral measure.

The model then computes the risk-neutral expectation of the payoff expressing the corporate bond value min (V_r, B), where B denotes the face amount of the bond. Finally, the model discounts this expectation back to its

present value. Therefore, the model makes convenient use of the Black-Scholes formula.

2.2 The JT and JLT models

2.2.1 The JT model
Under an appropriate probability space and the assumption that the risk-free interest rate process and the default time process are independent, the JT model provides the value ($F(t,T)$) of the T-maturity discount corporate bond at time t as given by equation (3):

$$F(t,T) = p(t,T)\big(\delta + (1-\delta)\widetilde{Q}_t(\tau^* > T)\big), \tag{3}$$

where δ is the recovery rate, $p(t,T)$ is the price of the T-maturity risk-free discount bond at time t, and $\widetilde{Q}_t(\tau^* > T)$ is the probability under the risk-neutral probability measure that the default happens after the maturity of the bond.

2.2.2 The JLT model
The JLT model first describes the credit rating of a company using the state space $S = \{1, \ldots, k\}$. The first state indicates the highest credit rating (AAA), while the second state corresponds to the second-highest credit rating (AA), and so on. The final state k indicates default. The model initially adopts matrix (4) as the credit-rating transition probability matrix for a given point in time. In particular, the empirical credit-rating transition probability matrix is given by

$$Q = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,k} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k-1,1} & q_{k-1,2} & \cdots & q_{k-1,k} \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{4}$$

where $q_{i,j}$ is the probability that the credit rating of the company changes from i to j, and where, for all i, j, $q_{i,j} \geq 0$ and $q_{i,i}(t,t+1) \equiv 1 - \sum_{\substack{i=1 \\ j \neq i}}^{k} q_{i,j}(t,t+1)$. Moreover, the n-period transition probability matrix is then computed as $Q_{0,n} = Q^n$.

Under the usual assumptions that the market is complete and that the arbitrage-free condition is satisfied, the JLT model then introduces the transition probability matrix from time t to time $t+1$ under a risk-neutral measure:

$$\widetilde{Q}_{t,t+1} = \big[\widetilde{q}_{i,j}(t,t+1)\big]. \tag{5}$$

To retain its Markov character, the JLT model restricts the risk-neutral probability $\widetilde{q}_{i,j}(t,t+1)$ to

$$\widetilde{q}_{i,j}(t,t+1) = \pi_i(t)q_{i,j} \tag{6}$$

for all $i, j, i \neq j$, where $\pi_i(t)$ is the risk premium. The matrix form of equation (6) may be written as

$$\widetilde{Q}_{t,t+1} - I = \Pi(t)[Q - I], \tag{7}$$

where I is a $k \times k$ unit matrix $\Pi(t) = diag\left(\pi_1(t),\ldots,\pi_{k-1}(t),1\right)$, for all $i, j,$ $\pi_i(t) > 0$. Furthermore, $\tilde{q}_{i,j}(0,n)$ is defined as the probability that the credit rating of the company jumps from credit rating i to credit rating j over n periods, and this probability is expressed as the (i,j)-th entry on the left side of equation (8).

$$\tilde{Q}_{0,n} = \tilde{Q}_{0,1}\tilde{Q}_{1,2}\cdots\tilde{Q}_{n-1,n}. \tag{8}$$

Under the risk-neutral probability measure, the JLT model provides the probability $\tilde{Q}_t^i\left(\tau^* > T\right)$ that the a company with the i-th credit rating at time t does not default until the maturity T of the bond as

$$\tilde{Q}_t^i\left(\tau^* > T\right) = \sum_{j \neq K} \tilde{q}_{i,j}(t,T) = 1 - \tilde{q}_{i,k}(t,T), \tag{9}$$

where $\tau^* = \inf\{s \geq t : \eta_s = k\}$.

Using equation (10), the JLT model then evaluates the T-maturity, i-th credit rating discount corporate bond at time t, $F^i(t,T)$, simply by substituting $\tilde{Q}_t^i\left(\tau^* > T\right)$ in place of $\tilde{Q}_t\left(\tau^* > T\right)$ in valuation formula (3) of the JT model.

$$F^i(t,T) = p(t,T)\left(\delta + (1-\delta)\tilde{Q}_t^i\left(\tau^* > T\right)\right). \tag{10}$$

2.3 Characteristic features of the Merton and JLT models, and the motivation for our model

2.3.1 The Merton model

Strength:
Since it integrates a default based on the structure of the balance sheet of the company, the model easily incorporates the financial impact of credit-rating changes on the balance sheet.

Weaknesses:
1. The model does not explicitly describe credit ratings and, therefore, is not suitable for valuing credit-rating-triggered products.
2. With the exceptions of the risk-free interest rate r and the maturity T of the bond, the model has only three fundamental parameters, namely the volatility of the corporate value process σ, the initial corporate value V_0, and the face amount of the corporate bond B. Therefore, the model has too few parameters to fit the market credit spreads of all maturities flexibly.
3. In this regard, the volatility σ of the company value process does not depend on its credit rating and is constant across all credit states.
4. In the course of valuing a coupon bond, the model must determine whether the bond was in default at any coupon payment date, and this procedure is very time-consuming.
5. The model cannot incorporate the term structure of risk-free interest rates.

2.3.2 The JLT model

Strengths:
1. The model is based on credit ratings and is therefore suitable for valuing credit-rating-triggered products.

2. The model incorporates a credit risk premium $\pi_i(t)$ that depends both on the time t and the credit rating i provided in the risk-neutral credit-rating transition probability matrix $\widetilde{Q}$. Therefore, the model is flexible enough to fit market credit spreads for all maturities.

3. In this regard, not only the risk premium $\pi_i(t)$, but also the empirical credit-rating transition probability $q_{i,j}$ in the matrix Q, depend by definition on the credit rating.

4. The model easily values coupon bonds.

5. The model is able to incorporate the term structure of risk-free interest rates.

Weakness:

Since it models a default using a credit-rating transition probability matrix, the model does not incorporate the structure of the balance sheet of the company. For this reason, it does not consider the financial impact of the credit rating on the balance sheet.

In light of these characteristics, we propose a valuation model for the credit-rating-linked coupon bond that incorporates the impact of increased coupon payments on the potential of the firm to pay future coupons and to make face value payments. Our modelling approach is structural, although we recognise that structural models are in several respects weak in comparison to the JLT model. In short, we attempt to incorporate the benefits of the JLT model into an analogous structural model.

3   Our model and its calibration

3.1   Our model

Before introducing our model, we describe the correction of several weaknesses of the Merton model:

Weakness 1

As an analogue of the credit-rating state space $S = (1,...,k)$ in the JLT model, we introduced $k-1$ threshold values, $V^{*(i)}$, $i = 1, \cdots, k-1$. The $k-1$-th threshold value $V^{*(k-1)}$ is simply the coupon value $c^{(k-1)}$ of the bond at the coupon payment date and the face amount $B + c^{(k-1)}$ of the bond at Maturity.

Weaknesses 2 and 3

Instead of the common volatility of the corporate value process σ, we introduced the credit-rating-dependent volatilities $\sigma^{*(i)}$, for $i = 1, \cdots k-1$. In the case of $i = k$, no volatility exists, because the company defaults in that state. The volatility $\sigma^{*(i)}$ essentially corresponds to the empirical credit-rating transition probability matrix Q in the JLT model. We also introduced a credit-rating-dependent initial corporate value V_0^i, for $i = 1, \cdots k-1$, to increase the flexibility of the model.

Weakness 4

Since we adopted a Monte Carlo simulation method for the purpose of valuation, the analysis required very little time.

Our model:

Based on these revisions, the risk-neutral company value process in our model may be described as in equations (11) and (12) below.

At any time except that of the coupon payment,

$$dV_t^i = rV_t^i dt + \sigma^{*(1)}V_t^i dW_t, \qquad : V_t^i > V^{*(1)}$$
$$dV_t^i = rV_t^i dt + \sigma^{*(j)}V_t^i dW_t. \qquad : V^{*(j-1)} > V_t^i > V^{*(j)} \qquad (11)$$

In addition, at the coupon payment time t_l,

$$V_{t_l}^i = V_{t_l-}^i - c^{(j)}. \qquad : V^{*(j-1)} > V_{t_l-}^i > V^{*(j)} \qquad (12)$$

where, $V_{t_l-}^i$ is the just-before-t_l value of the corporate bond with initial credit rating i, and where $c^{(j)}$ is the coupon of a bond with the j-th credit rating at the date of issue.

Valuation procedure based on a Monte Carlo simulation:

Step1 : Simulate the sample path of the corporate value process given by equations (11) and (12), starting with the initial corporate value.

Step2 : Compute the cash flow (coupon + face amount) for each sample path.

Step3 : Invest the cash flow calculated in Step 2 in the risk-free asset for the maturity T of the corporate bond. Take the risk-neutral expectation of the invested cash flow at time T, and discount it backwards to its present value.

3.2 Calibration of our model

3.2.1 Parameters in our model

Exogenous parameters:

The exogenous parameters include the credit-rating-dependent company value volatilities $\sigma^{*(i)}$, for $i = 1, \cdots k-1$, as well as the coupon and face amounts of the bond, $c^{(j)}$ and B. As mentioned above, these values correspond to the empirical credit-rating transitional probability matrix Q in the JLT model.

Parameters to be estimated:

The parameters to be estimated included the credit-rating-dependent initial corporate values V_0^i, for $i = 1, \cdots k-1$, and $k-2$ threshold values, such as state $V^{*(i)}$, for $i = 1, \cdots k-2$, except the default state $V^{*(k-1)}$ and the total number of parameters was $2k-3$. To facilitate the calibration of the model, we restricted the $k-1$ threshold values $V^{*(i)}$, for $i = 1, \cdots k-2$, by $V^{*(i)} = (V_0^i + V_0^{i+1})/2$, for $i = 1, \cdots, k-2$, by $V^{*(k-1)} = c^{(k-1)}$ at the coupon payment date, and by $V^{*(k-1)} = B + c^{(k-1)}$ at maturity. Therefore, the total number of parameters to be estimated was simply $k-1$.

The $k-1$ initial company values V_0^i, for $i = 1, \cdots k-1$, in our model correspond to the risk premium $\pi_i(t)$ in the JLT model. We allowed the initial company values V_0^i, for $i = 1, \cdots k-1$, to depend on the maturity T of the

corporate bond. Under this allowance, the number of parameters $\pi_i(t)$ in the JLT model (discrete version) matches that of the parameters V_0^i in our model.

3.2.2 Calibration

Three remarks regarding the model calibration are in order. First, we allowed the initial company values V_0^i to depend on the maturity T of the corporate bond. Therefore, the estimated values of V_0^i could differ by maturity. Second, for each maturity T, we tried to estimate the $k-1$ initial company values by fitting the $k-1$ model credit spreads to the market credit spreads by numerically solving $k-1$ equations. Finally, we assumed that the coupon bonds observed in the market were par bonds, and that their coupons were the same as their yields.

4   Numerical experiments

Specification of the credit-rating-linked coupon bond, and valuation methods in numerical experiments:

Each credit-rating linked coupon bond was assumed to behave as follows. If the bond bore the same credit rating that it had on issuance, then it paid at each coupon date the amount of the corresponding coupon initially specified. If the bond was in default at the coupon payment date, the corporate value at that time was paid at the maturity T of the bond.

In several numerical experiments, we compared the various bond values derived from the three different valuation models: (1) the JLT model, in which, at the coupon payment date, the coupon corresponding to the credit rating was paid, as mentioned above; (2) Model A (our model); and (3) Model B, which was essentially the same as our model, except that the fall in company value resulting from coupon payments remained at the initial coupon amount, although the company paid the coupon corresponding its credit rating at the coupon payment date. In other words, we adopted a model that was economically incorrect as a reference point from which to evaluate the other models.

Data and the setting of external parameters:

We adopted six possible credit ratings: AAA, AA, A, BBB, BB, and D. Therefore, $k = 6$. The bond maturity was five years, and the term structure of the risk-free interest rate was flat. The face amount of each bond was 70 yen, and the coupon of the bond with each credit rating was the same as its yield.

Table 1:   The credit spreads.

Rating	AAA	AA	A	BBB	BB
Steep	5%	10%	20%	25%	35%
Flat	20%	20%	20%	20%	20%

Table 2:   The volatilities.

Rating	AAA	AA	A	BBB	BB
Steep	0.18%	0.44%	0.92%	1.85%	4.69%
Flat	0.16%	0.26%	0.46%	1.12%	2.05%

We adopted the average empirical credit-rating transition probability matrix Q in the JLT model that was announced by R&I (a Japanese rating agency) between 1994 and 2004. In this derivation, we lumped together all of the transition probabilities for credit ratings below BB, with the exception of the default state; these were given the corresponding credit-rating label "BB."

Moreover, in estimating the risk premium $\Pi(t)$, we used the estimation technique adopted by JLT (1997).

Table 3:   The cases of numerical experiments.

The Cases of Credit Spread(Flat):	Volatilities(Flat)	Volatilities(Steep)
Risk-free interest rate(1.21%)	Case1	Case2
Risk-free interest rate(3.21%)	Case3	Case4
The Cases of Credit Spread(Steep):	Volatilities(Flat)	Volatilities(Steep)
Risk-free interest rate(1.21%)	Case5	Case6
Risk-free interest rate(3.21%)	Case7	Case8

For both the volatility of the company value process and the credit spread of the bond corresponding to each credit rating, we allowed two different settings, and these are listed in Tables 1 and 2, respectively. In addition, we set the risk-free interest rate alternatively at 1.21% and 3.21%. Therefore, in total, we performed eight numerical experiments (Cases 1 through 8), the results of which are summarised in Table 3.

The results of the numerical experiments, and their implications:
The eight valuations, corresponding to Cases 1 through 8, of the credit-rating-linked coupon bond for each of the three valuation models are provided in Figures 1 through 8, respectively.
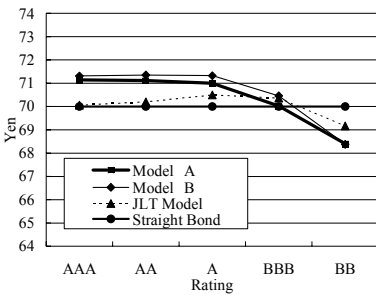
Figure 1:   The results of Case 1.
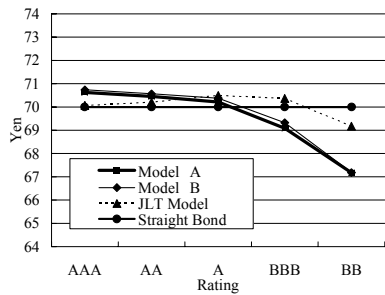
Figure 2:   The results of Case 2.
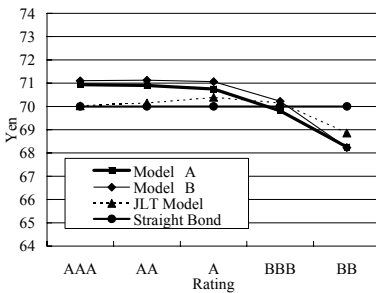
Figure 3:   The results of Case 3.
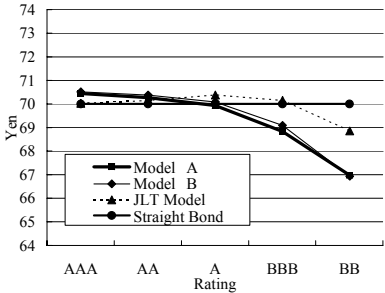
Figure 4:   The results of Case 4.
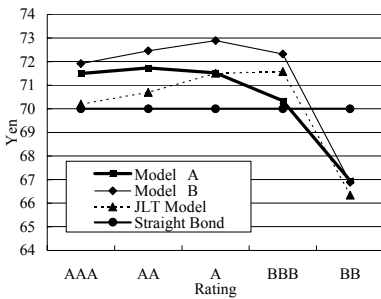
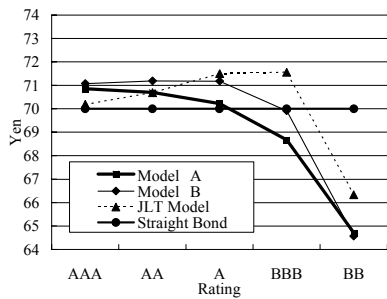Figure 5:   The results of Case 5.

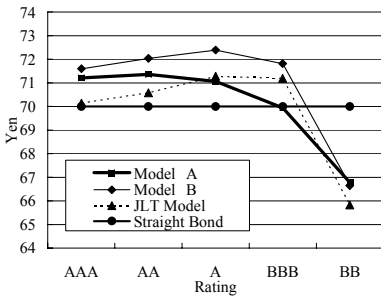Figure 6:   The results of Case 6.
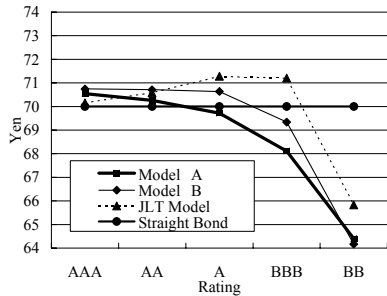

Figure 7:   The results of Case 7.

Figure 8:   The results of Case 8.

(1) Overview of the results

(a) All three models valued the credit-rating-linked coupon bond above the straight bond when the credit rating of the bond was relatively high (AAA, AA, A), while the opposite was true when the credit rating of the bond was relatively low (BBB, BB).

(b) The value of the credit-rating-linked coupon bond derived from the JLT model tended to be lower than those derived from Model A and Model B under a relatively high initial credit rating (AAA, AA, A); the reverse was true under a relatively low initial credit rating.

The first result was obtained because, under a higher initial credit rating, the effect of the coupon increase resulting from a downgrade swamped the resulting decrease in the potential of the company to make future coupon payments. Under a low initial credit rating, the situation was reversed. The second result was obtained because the coupon payment amount did not affect the credit-rating transition probability in the JLT model, while the increasing coupon amount increased the default probability, and the magnitude of this effect was larger under a low credit rating than under a high credit rating.

(2) The influence of the credit spread (comparison of Case 1 & Case 4 and Case 5 & Case 8).

The first result (1) appeared more salient for a large, steep credit-spread curve than for one that was small and flat. The reason underlying the first result in (1) also explains this observation.

(3) The influence of the volatility of the company value process (comparison of Cases 5 and 6)

(a) In both Models A and B, and for all credit ratings, the value of the credit-rating-linked coupon bond tended to be higher under a flat volatility structure (20% in all cases) than under a steep volatility structure (5, 10, 20, 25, and 35%, respectively, from the highest credit rating to the lowest).

(b) The valuation derived using Model B deviated from that of Model A to a greater extent under the flat volatility structure than under the steep one.

The first result may be explained as stemming from reason (1) above. The deviation of the value derived from Model B from that derived from Model A resulted from both the credit-rating probability and the difference between the initially set constant coupon and the credit-rating-linked coupon. For Cases 5 and 6, the latter impact was the same, but the former was larger under flat volatility than under steep volatility.

(4) The influence of the risk-free interest rate

For all of the initial credit ratings, the value of the credit-rating-linked coupon bond was higher when the risk-free interest rate was low. The difference between the initially set constant coupon and the credit-rating-linked coupon derived not from the risk-free interest rate itself, but rather from the credit spread. The risk-free interest rate only affected the value of the credit-rating-linked coupon bond through its impact on the discount rate of its cash flow.

5   Summary and concluding remarks

In this paper, we presented a structural valuation model for credit-rating-linked coupon bonds that incorporates the fact that an increased coupon payment resulting from a downgrade may deteriorate the potential of the issuing company to make future coupon and notional payments. Through numerical experiments, we demonstrated that our model reasonably captures this effect. A practical implication of our model is that the valuation of a credit-rating-linked coupon bond based on the JLT model tends to underestimate the value of the bond when its initial credit rating is high. However, the reverse is true when the initial credit rating is low.

References

[1] Bhanot K., Pricing Corporate Bonds with Rating-Based Covenants. *The Journal of Fixed Income*, March, pp. 57-64, 2003.
[2] Duffie, D. & Singleton, K., Modeling Term Structures of Defaultable Bonds. *Review of Financial Studies*, 12, pp. 687-720, 1999.
[3] Jarrow, R.A. David L. & Turnbull, S.M., A Markov Chain Model for the Term Structure of Credit Risk Spreads. *Review of Financial Studies*, 10(2), pp. 481-523, 1997.
[4] Jarrow, R. & Turnbull S.M., Pricing Derivatives on Financial Securities Subject to Credit Risk. *Journal of Finance*, 50, pp. 53-85, 1995.
[5] Merton, R.C., On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *Journal of Finance*, 29, pp. 449-470, 1974.

Dynamics of the top of the order book in a global FX spot market

E. Howorka & A. B. Schmidt
EBS Dealing Resources, Parsippany NJ, USA

Abstract

The order lifetime at the top of the order book is defined as the time between the order arrival at the top of the order book and its removal from the top of the order book. In this work, the average order lifetime in the EBS FX spot market is analyzed for two corresponding four-week periods in 2004 and 2005. The following currency pairs, EUR/USD, USD/JPY, USD/CHF, EUR/JPY, and EUR/CHF, are considered during the most liquid period of the working day, 7:00 – 17:00 GMT. Generally, the distribution of orders with a given lifetime at the top of the order book decays exponentially at short times. However, this decay follows a power law at longer time periods. The crossover times between the two decay forms are estimated. It is shown that the decays have steepened and the order lifetime has become shorter in 2005. In particular, 47.9% of the EUR/USD orders and 34.7% of the USD/JPY orders live less than one second on the top of the order book. Two possible causes of the power-law asymptote are indicated: orders with amounts significantly higher than the average value and the specifics of credit relations among the EBS customers. The only exclusion from the described pattern is the order dynamics of EUR/CHF in 2005 that does not have an exponential decay.
Keywords: high-frequency FX market, order lifetime.

1   Introduction

The global inter-bank FX spot market has dramatically changed since early 1990s when the electronic broking systems were introduced. Before that, a trader could either contact another trader directly (using telephone or a Reuters electronic system 2000) or trade via "voice brokers" who were collecting and matching the bid and offer orders over dedicated telephone lines. The electronic

broking systems do this matching at greatly increased speed and reduced cost. At present, the global inter-bank spot foreign exchange is overwhelmingly conducted via two electronic broking systems, EBS and Reuters 3000. While Reuters has significant market share in GBP-based currency pairs, EBS overwhelmingly dominates the electronic inter-bank spot EUR/USD and USD/JPY exchange. The daily transacted volume in the EBS market is approximately 120 billion USD. As a result, the EUR/USD and USD/JPY rates posted at any time on the EBS trading screens have become the reference prices quoted by dealers worldwide to their customers [1].

Yet, current empirical research of the high-frequency FX markets is overwhelmingly based on the Reuters indicative rates. The disadvantages of the indicative rates in comparison with the "firm" rates at which the inter-bank currency exchange is conducted are well documented (see e.g. [2, 3] for a review). In recent years, several studies of the high-frequency FX market based on the consolidated EBS proprietary data have been reported [1, 4, 5]. However, analysis of many intriguing properties of the high-frequency market requires an access to the customer-sensitive data that currently cannot be made publicly available. Therefore we feel that disclosing some of the EBS "in-house" findings based on analysis of these intimate data will benefit both the EBS customers and the academic community.

We define the order lifetime at the top of the order book as the time between the order arrival at the top of the order book and its removal from the top of the order book.

This report describes the average lifetime of the orders at the top of the EBS order book for two four-week periods starting on Mondays, 13 Sep 2004 and 12 Sep 2005, respectively. The following currency pairs, EUR/USD, USD/JPY, USD/CHF, EUR/JPY, and EUR/CHF, are considered during the most liquid time of the working day, 7:00 – 17:00 GMT.   We show that the distribution of orders with a given lifetime at the top of the order book generally decays exponentially at short times. However this decay follows a power law at longer time periods. The only exclusion from the described pattern is the order book dynamics of EUR/CHF in 2005 that does not have an exponential decay. The crossover times between the two decay forms are estimated and it is shown that the decays have steepened and the order lifetime has become shorter in 2005. In particular, 47.9% of the EUR/USD quotes and 34.7% of the USD/JPY quotes live less than one sec on the top of the order book. The report is organized as follows. The specifics of the EBS FX spot market pertinent to this work are listed in the next Section. The results and their discussion are presented in Section 3.

2   The EBS FX spot market

The EBS system has several specifics that are important for this work. First, only the limit orders are accepted (no market orders may be submitted). The EBS system has two types of orders: quotes and hits. Quotes stay in the order book until they are filled or are interrupted; hits are automatically cancelled if they have no matching counterpart when they reach the market. Hence, a hit is always

a taker order while a quote may be either a maker order or a taker order. If a quote matches an order arriving the market later than this quote, the quote is a maker order. If a quote matches another quote that was present in the market before this quote arrived, this quote is a taker order. In the EBS market, only takers pay the transaction fees. Here we consider only the quote dynamics.

Orders in the EBS market are submitted in units of millions (M) of the base currency (the first currency in the name of the currency pair, e.g., USD for USD/JPY and EUR for EUR/USD). This may be worth to remember while considering the triangle arbitrage opportunities. Indeed, if one buys an amount of USD for 1M of EUR (say 1208300 USD according to the exchange rate on 20 Jan 2006), then transforming this entire amount of USD into e.g. CHF is tricky as only an integer number of millions of USD can be submitted in the EBS market for exchange with CHF.

Trading in the EBS market is allowed only between the counterparts that have bilateral credit. Every EBS customer establishes credit with all other EBS customers and can change its credit to other customers at any time. This implies that the EBS best prices (highest bid and lowest offer) may or may not be available to an EBS customer, depending on whether this customer has bilateral credit with the makers of the best prices. In fact, entire market depth available to an EBS customer is determined by its credit relations with all other EBS customers. Four types of prices on the both bid and offer sides are shown on the EBS trading screen. Besides the EBS best prices and the best available (due to the credit restrictions) prices, there are also credit-screened regular prices. The regular amount is a notional volume specific for each currency pair. In particular, it currently equals 15M of EUR for EUR/USD and 15M of USD for USD/JPY. If the currently available volume is less than the regular amount, it is also displayed on the EBS trading screen. For example, current EUR/USD best available offer and the regular offer are 1.2083 and 1.2086, respectively. Also, current best available volume is 9M. Then while trading the regular amount, 9M can be bought at 1.2083 and 6M can be bought at a rate higher than 1.2083 but not higher than 1.2086.

The EBS global market has three regional order matching processes, so-called arbitrators. These arbitrators are located in London (LN), New York (NY), and Tokyo (TY). Since the order arrival time is smaller for intra-regional networks than for inter-regional networks, the regional order books may somewhat vary. Indeed, consider a case when two bids with the same (new) best price are submitted at the same second by a London customer and a Tokyo customer. One may expect that the London quote will arrive at the top of the London order book while the Tokyo quote will land on the top of the Tokyo order book. Then if the London best quote is filled, the top of the London order book is changed while the top of the Tokyo order book remains the same (the Tokyo quote is now at the top of both the London and Tokyo order books).

The EBS primary historical data base contains chronologically ordered records of all events ever occurred in the EBS market. Restoring an order book at a given time from the historical data base requires sophisticated software that replicates important arbitrator functions.

3   Results and discussion

Three types of events remove a quote from the top of the order book:
- Quote is completely filled
- Quote is interrupted
- Quote is replaced with another one that has a better price.

As it was indicated in the former Section, the regional order books can somewhat vary. Table 1 illustrates these differences for the period 3 Oct 2005 – 7 Oct 2005, 7:00 – 17:00 GMT. For a given currency pair, the percentages of filled, interrupted, and replaced quotes are very close in all three regions. Further, the data for NY are discussed. The estimates of the lifetime were done in seconds.

Table 1:   Changes at the top of the EBS regional order books for the period 3 Oct 2005 – 7 Oct 2005 between 7:00 – 17:00 GMT.

		Total changes at the top	Filled, %	Replaced, %	Interrupted, %
EUR/USD	NY	181259	66.1	27.1	6.8
	LN	184697	66.8	26.6	6.6
	TY	173445	64.6	28.4	7.0
USD/JPY	NY	88999	53.2	34.3	12.5
	LN	90237	53.9	33.8	12.3
	TY	87969	52.6	34.8	12.6
USD/CHF	NY	78062	34.0	38.7	27.3
	LN	78031	34.1	38.6	27.3
	TY	77026	33.2	39.2	27.6
EUR/JPY	NY	58546	27.8	41.2	31.0
	LN	58807	28.2	41.0	30.8
	TY	58334	27.6	41.4	31.0
EUR/CHF	NY	34838	45.9	39.7	14.4
	LN	34965	46.0	39.6	14.4
	TY	34470	45.3	40.1	14.6

If all quotes "were created equal", one might expect an exponentially decaying lifetime on the top of the order book, similarly to radioactive atom decay. Indeed, if some factors lead to removing N percent of quotes in the first second, the same N percent of the remaining quotes will be removed in the next

second, and so on. However, our results show that the distribution of the quote lifetime follows an exponential decay only at short periods of time. With exclusion of EUR/CHF in 2005 (see discussion below), these periods span from two seconds for EUR/USD to five seconds for EUR/JPY. At longer times, the quote lifetime follows a power law. This can be understood as the quote lifetime depends not only on the market activity but also on the quote size and on the creditability of its owner. Indeed, a quote with an amount notably exceeding an average value can stay at the top of the order book for some time until it is completely filled with smaller counterpart orders. Also, a quote submitted by a customer with a smaller credit may stay at the top of the order book for some time until someone with available bilateral credit is willing to match it.

We defined the crossover time between the exponential and power-law approximations of decay as the time at which the sum of the coefficients of determination, R^2, for the exponential fit and the power-law fit has a maximum. The analytical fits were estimated using the Microsoft Excel 2003 software.

The results of our analysis are summarized in Tables 2 and 3. The two most liquid currency pairs in the EBS market, EUR/USD and USD/JPY, have the same crossover time in 2004 and 2005 (2 sec for EUR/USD and 4 sec for USD/JPY). However decays for both these currency pairs have steepened in 2005. Namely, the percentage of EUR/USD quotes lived on the top of the order book for less than one second has increased from 44.8% to 47.9%. Similarly for USD/JPY, this percentage has changed from 30.9% to 34.7%. The decay of the quote lifetime at the top of the order book in 2005 is illustrated in Fig.1 and Fig.2 for EUR/USD and USD/JPY, respectively.

The most dramatic changes have occurred for less liquid currency pairs. In particular, the crossover times decreased from 7 sec to 3 sec for USD/CHF and from 8 sec to 5 sec for EUR/JPY. Moreover, the percentage of quotes that lived at the top of the order book less than one second almost doubled: from 22.3% to 32.3% for USD/CHF and from 18.6% to 26.9% for EUR/JPY. It should be noted also that these two currency pairs have significantly higher percentage of interrupted quotes at the top of the order book, particularly in 2005 (cf. 30.5% for USD/CHF and 32.0% for EUR/JPY versus 6.9% for EUR/USD and 12.7% for USD/JPY).

For the least liquid currency pair among those we considered, EUR/CHF, the percentage of quotes that lived at the top of the order book less than one second has also increased in 2005: from 21.4% to 25.1%. However, its decay did not follow the general pattern in 2005. Namely, while the exponential decay existed in 2004 at times greater than 6 sec, it was not found in 2005. As it can be seen in Fig. 3, the empirical curve has a small hump in the region from 2 to 4 sec, which complicates its simple analytical fit. It should be noted also that the exponential decay may still exist at times lower than one second. In future we are planning to make similar estimates on a grid finer than the one-second grid.

Table 2:   Quote lifetime at the top of the EBS order book for the period 13 Sep 2004 – 8 Oct 2004, working days between 7:00 – 17:00 GMT.

2004	EUR/USD	USD/JPY	USD/CHF	EUR/JPY	EUR/CHF
Crossover time (T), sec	2	4	7	8	6
Exponential law (x <= T)	$0.9140e^{-0.7054x}$	$0.4711e^{-0.4331x}$	$0.3195e^{-0.3179x}$	$0.2695e^{-0.2703x}$	$0.2643e^{-0.3015x}$
Exponent's R^2 (x <= T)	0.9996	0.9988	0.9945	0.9918	0.9914
Power law (x > T)	$0.7939x^{-2.1315}$	$0.8599x^{-1.8797}$	$0.9693x^{-1.7780}$	$1.2574x^{-1.8333}$	$0.4521x^{-1.4094}$
Power law's R^2 (x > T)	0.9931	0.9949	0.9954	0.9943	0.9969
Lifetime < 1 sec, %	44.8 ± 0.2	30.9 ± 0.3	22.3 ± 0.3	18.6 ± 0.4	21.4 ± 0.1
Filled, %	65.5 ± 0.3	52.1 ± 0.5	33.5 ± 1.0	26.6 ± 1.0	41.1 ± 0.1
Replaced, %	28.1 ± 0.2	36.3 ± 0.2	42.9 ± 0.4	44.4 ± 0.3	42.1 ± 0.1
Interrupted, %	6.4 ± 0.1	11.5 ± 0.3	23.6 ± 0.7	29.0 ± 0.7	16.8 ± 0.2

Table 3:   Quote lifetime at the top of the EBS order book for the period 12 Sep 2005 – 7 Oct 2005, working days between 7:00 – 17:00 GMT.

| 2005	EUR/USD	USD/JPY	USD/CHF	EUR/JPY	EUR/CHF
Crossover time (T), sec	2	4	3	5	-
Exponential law (x <= T)	$1.041e^{-0.7763x}$	$0.5149e^{-0.4746x}$	$0.5422e^{-0.5107x}$	$0.3867e^{-0.3818x}$	-
Exponent's R^2 (x <= T)	1.000	0.9927	0.9915	0.9912	-
Power law (x > T)	$0.7796x^{-2.1594}$	$0.8865x^{-1.9038}$	$0.6702x^{-1.7684}$	$0.8289x^{-1.7585}$	$0.3029x^{-1.255}$
Power law's R^2 (x > T)	0.9896	0.9913	0.9897	0.9910	0.9897
Lifetime < 1 sec,%	47.9 ± 0.6	34.7 ± 0.4	32.3 ± 1.1	26.9± 1.3	25.1 ± 0.4
Filled, %	65.8 ± 0.2	53.1 ± 0.3	31.0 ± 1.1	27.3 ± 1.4	44.8 ± 0.5
Replaced, %	27.3 ± 0.2	34.2 ± 0.1	38.5 ± 0.4	40.8 ± 0.4	
Interrupted, %	6.9 ± 0.1	12.7± 0.4	30.5 ± 1.2	32.0 ± 1.7	
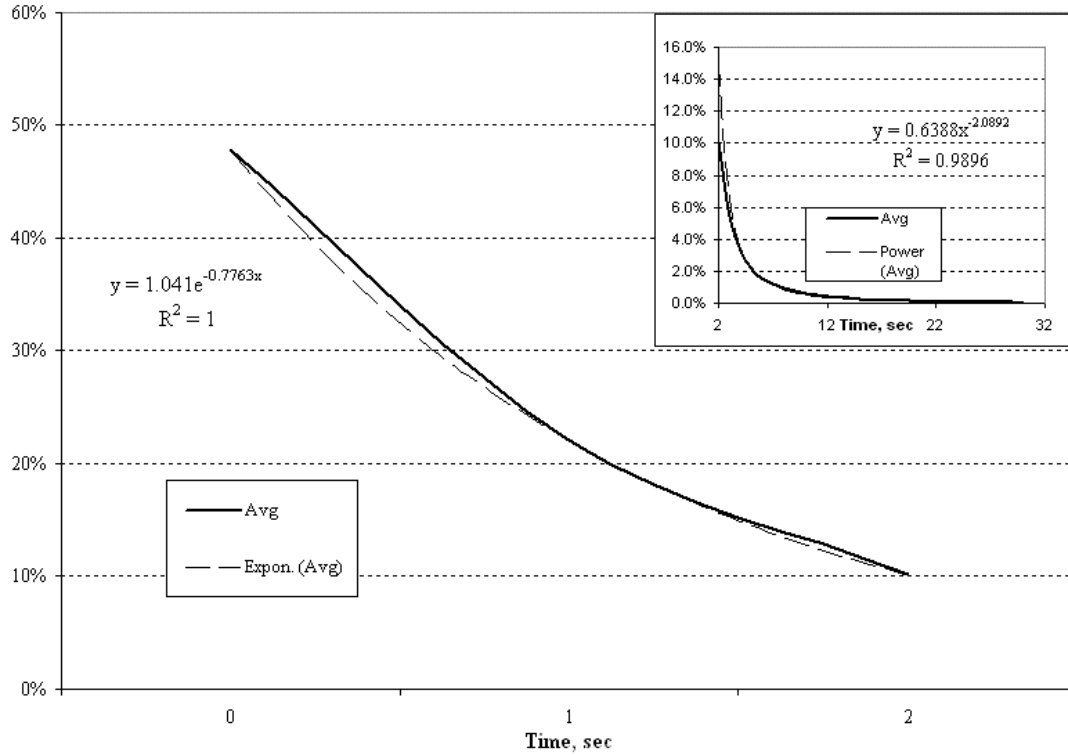
Figure 1:   Distribution of the EUR/USD quote lifetime at the top of the EBS order book (12 Sep 05 – 7 Oct 05, working days, 7:00 – 17:00 GMT).
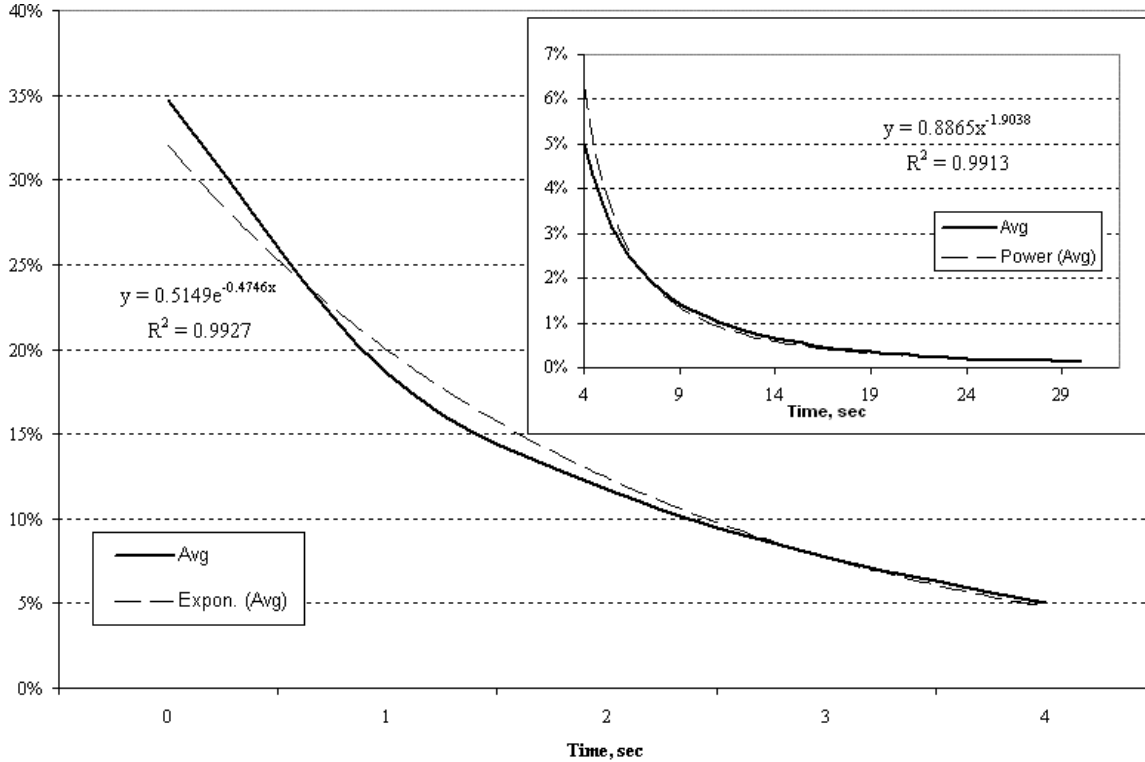
Figure 2:   Distribution of the USD/JPY quote lifetime at the top of the EBS order book (12 Sep 2005 – 7 Oct 2005, working days, 7:00 – 17:00 GMT).
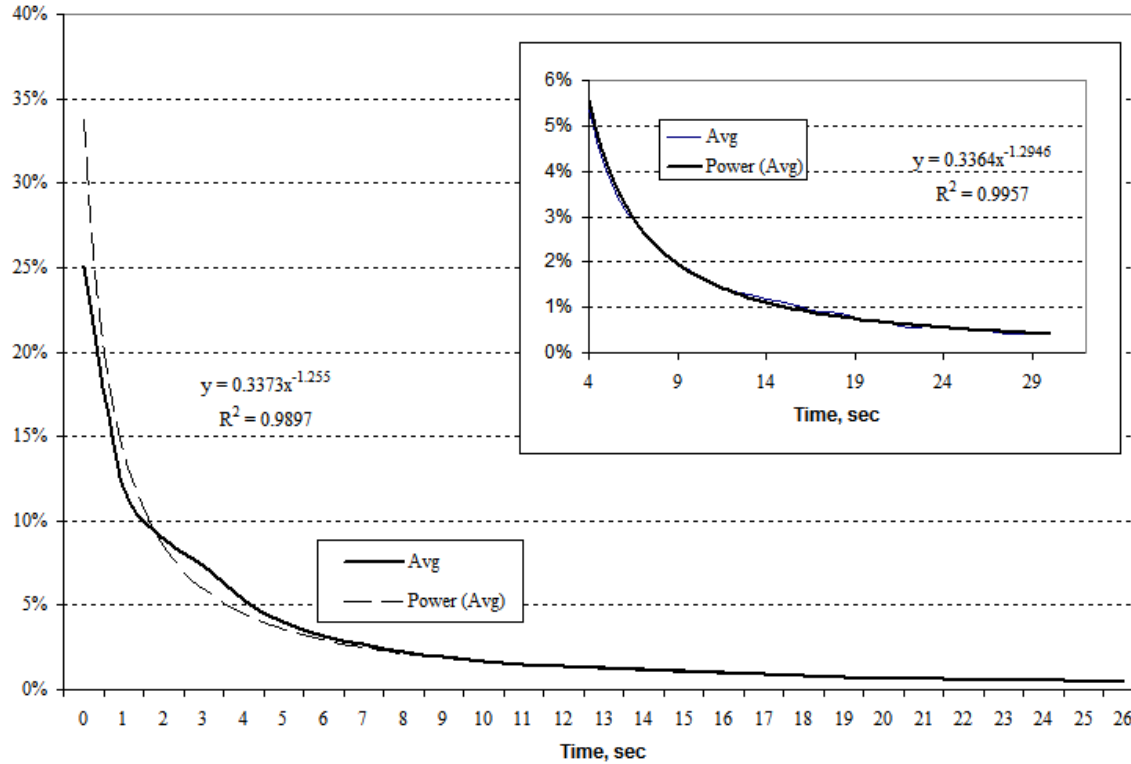
Figure 3:   Distribution of the EUR/CHF quote lifetime at the top of the EBS order book (12 Sep 2005 – 7 Oct 2005, working days, 7:00 – 17:00 GMT).

References

[1] A. P. Chaboud, S. V. Chernenko, E. Howorka, R. S. Krishnasami, D. Liu, and J. H. Wright, The High-Frequency Effects of US Macroeconomic Data Releases on Prices and Trading Activity in the Global Interdealer Foreign Exchange Market, *International Finance Discussion Papers*, 823 (2004).

[2] M. M. Dacorogna, R. Gencay, U. Muller, R.B. Olsen, and O.V. Pictet, *An Introduction to High-Frequency Finance*. Academic Press, 2001.

[3] C.A.O. Goodhart and M. O'Hara, High frequency data in financial markets: Issues and applications, *Journal of Empirical Finance*, 4, 73-114 (1997).

[4] W. P. Killeen, R. K. Lyons, and M. J. Moore, Fixed versus flexible: Lessons from EMS Order Flow, *NBER Working Paper* N8491, 2001.

[5] D. W. Berger, A. P. Chaboud, S. V. Chernenko, E. Howorka, R. S. Krishnasami, D. Liu, and J. H. Wright, Order Flow and Exchange Rate Dynamics in Electronic Brokerage System data, *International Finance Discussion Papers*, 830 (2005).

[6] T. Ito and Y. Hashimoto, Intra-day Seasonality in Activities of the Foreign Exchange Markets: Evidence from the Electronic Broking System, *Faculty of Economics, The University of Tokyo* (2006).

Seasonal behaviour of the volatility on European stock markets

L. Jordán Sales[1], R. Mª. Cáceres Apolinario[1], O. Maroto Santana[1]
& A. Rodríguez Caro[2]
[1]*Department of Financial Economics and Accounting,
Las Palmas de Gran Canaria University, Spain*
[2]*Departament of Quantitative Methods,
Las Palmas de Gran Canaria University, Spain*

Abstract

The existence of seasonal behaviour in return and volatility of different international stock exchanges may be considered as an indication of non-integrated financial markets. A type of this abnormal behaviour is the day of the week effect, which implies investment opportunities. This type of opportunity is studied in this paper, focused on the analysis of the day of the week effect on the major European stock markets using GARCH and T-ARCH models. Results show evidence in favour of day of the week effect in the volatility in the most of the studied countries.
Keywords: day of the week effect, volatility, GARCH, T-ARCH.

1   Introduction

The increasing internationalisation of the main economies from developed nations has given the investor additional choices when considering his portfolio. He is no longer obliged to focus his attention on the financial markets where the assets of his own country are listed in the stock market but instead may look towards other investment horizons whose markets offer opportunities to obtain greater results with respect to profit and risk. This scenery is characterised by significant relaxation of national barriers, thus allowing for the entrance of foreign capital, and its repercussions are seen in the considerable increase in international capital flows.

Nevertheless, it is necessary to remember that investment opportunities in international markets depend on the degree of integration or segmentation that said markets possess. The presence of anomalies in international financial markets can be a clear sign that a lack of integration among these markets exists, thus investment opportunities derived from different behaviours in the generation of returns are available. Several studies have centred on relative anomalies in the seasonality of distinct financial markets of developed countries as an explanation to why there is an absence of integration between international financial markets.

The objective of this paper is to empirically contrast the day of the week effect in the major European stock markets from July 1997 to March 2004. We will study not only return but volatility as well. The day of the week effect under a volatility context has not received much attention in the literature. The motivation for this paper comes from the growing process of integration of the distinct world economies and European economies in particular, resulting in an increasing correlation and synchronization among financial markets from different countries.

The paper is divided into the following sections. Section 2 presents a description of the database as well as the methodology employed in the paper. The estimations from the GARCH and T-ARCH models and the results are presented in Section 3. The paper ends with a summary of the main conclusions.

2   Data and methodology

The present paper used series of daily returns from the corresponding stock indices of the following European markets: Germany, Austria, Belgium, Denmark, Spain, France, The Netherlands, Italy, Portugal, The United Kingdom, The Czech Republic, Sweden and Switzerland.
The sampling dates begin with July 2, 1997 and end on March 22, 2004. The returns for each market are expressed in local currency and have been calculated as first differences in natural logarithms.

The analysis of the day of the week effect was carried out in the following manner. First we used five observation per week in order to avoid possible bias from the loss of information due to bank holidays. A total of 1754 yields were collected for each of the analysed markets. The indices used for each country market in our sample are DAX (Germany), ATX (Austria), BEL-20 (Belgium), KFX (Denmark), IBEX-35 (Spain), CAC-40 (France), AEX (Holland), MIB-30 (Italy), PSI-20 (Portugal), FTSE-100 (U. Kingdom), PX-50 (Czech Rep.), Stockholm General (Sweden), Swiss Market (Switzerland).

One of the most common seasonality anomalies is the day of the week effect. This analysis is based on the hypothesis that the yields produced by each security are not independent of the day of the week. An initial approximation that could contrast the day of the week effect can be carried out with a regression model. They included five dummy variables, one for each day of the week.

$$r_{it} = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \beta_5 D_{5t} + \varepsilon_t$$

where:

rit : is the daily yield of the financial asset

Djt : are dummy variables which take on the value 1 if the corresponding return for day t is a Monday, Tuesday, Wednesday, Thursday or Friday, respectively and 0 otherwise.

β*j* : are coefficients which represent the average return for each day of the week.

ε*t* : is the error term.

It is worth noting that even though the corresponding return on a specific day of the week is significantly different than zero, this does not imply seasonality. Thus it is necessary to perform a means test. This test verifies if the returns are independent of the day of the week that they are produced in, or on the contrary, they are characterised by statistically similar average returns. The rejection of the null hypothesis would imply that a day of the week effect is indeed present.

Nevertheless two serious problem arise with this approach. First, the residuals obtained from the regression model can be autocorrelated, thus creating errors in the inference. The second problem is that the variance of the residuals is not constant and possibly time-dependent.

A solution to the first type of problem was to introduce the returns with a one-week delay into the regression model, as used in the works by Easton and Faff [6], Corredor and Santamaría [5] and Kyimaz and Berument [11], among others.

$$r_{it} = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \beta_5 D_{5t} + \sum_{j=1}^{4} \beta_{j+5} \cdot r_{t-j} + \varepsilon_t$$

ARCH models are proposed in order to correct the variability in the variance of the residuals. Engle [7] used this approach and it has the advantage that the conditional variance can be expressed as a function of past errors. These models assume that the variance of the residual term is not constant through time and is distributed as $\varepsilon_t \sim iid\left(0, \sigma_t^2\right)$. The generalized version of these models was proposed by Bollerslev (1986) and is expressed by the sum of a moving-average polynomial of order q plus an autoregressive polynomial of order p:

Others works by Baillie and Bollerslev [2], Hsieh [9], Copeland and Wang [4] and Kyimaz and Berument [11] also include dummy variables which account for the possible stationary effects within the equation of variance. The result of this approach is that joint estimates of the day of the week effects are obtained, not only in the mean but also in the conditional variance.

$$r_{it} = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \beta_5 D_{5t} + \sum_{j=1}^{4} \beta_{j+5} r_{t-j} + \varepsilon_t$$

$$\varepsilon_t \sim iid\left(0, \sigma_t^2\right)$$

$$\sigma_t^2 = \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4 + \alpha_5 D_5 + \sum_{i=1}^{q} \alpha_{5+i} \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \gamma_i \sigma_{t-i}^2$$

This model is characterised by its symmetric behaviour since the volatility is invariant during gains and losses of the stock quotations. Nevertheless, it is well known that the impacts in the volatility in positive and negative yields need not

have the same effect. Kiymaz and Berumet [11] have argued that on many occasions the obtained volatility from a negative return is usually greater than the corresponding one during a gain in the stock quotation that is being analysed. The asymmetric T-ARCH model is used in this case to confirm the existence or absence of any asymmetric behaviour, which is known as the leverage effect.

The T-ARCH model introduced by Zakoian [14] and Glosten et al. [8] contains a structure which is similar to the symmetric GARCH model with one exception. They include a term where the λ parameter is used to indicate the existence of differentiated behaviour in the volatility against positive and negative shocks. The generalised structure of the T-ARCH model follows:

$$r_{it} = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \beta_5 D_{5t} + \sum_{j=1}^{4} \beta_{j+5} \, r_{t-j} + \varepsilon_t$$

$$\varepsilon_t \sim iid\left(0, \sigma_t^2\right)$$

$$\sigma_t^2 = \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4 + \alpha_5 D_5 + \sum_{i=1}^{q} \alpha_{5+i} \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \gamma_i \sigma_{t-i}^2 + \lambda \, \varepsilon_{t-1}^2 d_{t-1}$$

where dt-1 is a dicotomic variable which takes on value 1 when the stock quote falls in a period and 0 for increments of the stock quotation.

3 Estimation of the models and empirical results

The study of seasonality in the returns and volatility for the European stock markets that are included in our sample is carried out based on obtained estimates from the daily returns of each one of the stock markets considered.

3.1 The study of day of the week effect on returns

Four dummy variables have been used to account for seasonality in each of the stock exchanges for each workday except Wednesday. The regression model follows:

$$r_{it} = \alpha + \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_4 D_{4t} + \beta_5 D_{5t} + \sum_{j=1}^{4} \beta_{j+5} \cdot r_{t-j} + \varepsilon_t$$

The individual meaning for each one of the dicotomic variables could reveal the presence of an atypical yield during a day of the week with respect to that of Wednesday. Not only is the statistical significance of each dummy variable studied but also possible structure in the autoregressive portion and in the moving average which includes the regression model.

The obtained results are summarised in Table 1 and indicate that the day of the week effect is not evident in most European stock markets since the yield for each day of the week is not especially different than that of other days. This fact tells us that the return for the most important representative European markets is independent of the day of the week. Nonetheless, a stationary effect can be observed on Mondays for the representative indexes of France and Sweden since the yields on this day are greater than the rest of the week. This result does not coincide with those obtained in most empirical studies where average Monday

returns are usually significantly less than the average returns for the other days of the week. A similar finding is observed in Sweden where Friday yields are much greater than those for the other days of the week, thus recalling the Friday effect for this specific market.

Table 1:   Day of the week effect on returns.

Country	Significant variables	Country	Significant variables
Germany	--	Italy	MA(4)
Austria	MA(1), MA(3), MA(4)	Portugal	AR(1), AR(3)
Belgium	AR(1)	U. Kingdom	MA(3)
Denmark	AR(1)	Czech Rep.	AR(1)
Spain	--	Sweden	D1, D5, AR(1)
France	D1	Switzerland	AR(1)
Holland	--		

3.2 Day of the week effect on volatility

The importance of an analysis for the anomalies for distinct stock markets with respect to yields encountered for the day of the week cannot be ignored. The aim of each investor is to maximize the binomial yield-risk from his investment. Thus it is especially important to analyse fluctuations which are produced in the same markets. That is why both symmetric and asymmetric models have also been used to study their variance. We have included the earlier dummy variables to the equation of variance, similarly to Kyimaz and Berument [11] in order to collect possible stationary effects which may arise.

3.2.1 GARCH model

The structure for the equation of estimated variance follows:

$$\sigma_t^2 = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_4 D_4 + \alpha_5 D_5 + \sum_{i=1}^{q} \alpha_{5+i} \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \gamma_i \sigma_{t-i}^2$$

Table 2 presents the results derived from the day of the week effect on volatility for each stock market index, as well as the GARCH structure for each series.

Table 2:   Day of the week effect on variance: GARCH model.

Country	GARCH structure	Significant variables	Country	GARCH structure	Significant variables
Germany	(1,2)	D2, D5	Italy	(1,1)	D1, D4
Austria	(1,1)	D2, D5	Portugal	(1,1)	--
Belgium	(1,1)	D4, D5	U. K	(1,1)	D2
Denmark	(1,1)	D1, D5	Czech Rep.	(1,1)	--
Spain	(1,1)	D1, D4	Sweden	(1,1)	D2, D5
France	(1,1)	D4	Switzerland	(1,1)	D1, D4
Holland	(1,1)	D1, D4			

The table shows that the resultant structure for all markets except Germany is GARCH (1,1). This structure is the most appropriate for studying financial time series according to Lamoreux and Lastrapes (1990). The case of Germany is characterised by a GARCH (1,2) structure.

With regards to volatility during each day of the week, we did not find common behaviour in the day of the week effect in the equation of conditional variance. This finding is in agreement with Kyimaz and Berument [11]. There is, however, presence of abnormal volatility on Mondays and Fridays in Denmark. Other observations include significantly distinct volatility on Mondays and Thursdays, with respect to Wednesday, in Spain, Holland, Italy and Switzerland. The case is different for abnormal volatilities for the United Kingdom and France, where the days are Tuesdays and Thursdays, respectively. Seasonal behaviour is also apparent on Tuesdays and Fridays for the cases of Germany, Austria and Sweden. Abnormal volatility occurs on Thursdays and Fridays in Belgium. Finally, Portugal and the Czech Republic show no changes with regards to the day of the week.

A general statement can be made for all of the markets that exhibit seasonal behaviour in the volatility. Mondays and Thursday are always greater than Wednesdays, while the opposite is true for Tuesdays and Fridays, that is, the yields are lesser than those experienced on Wednesday, except Friday in the Belgian market. The results derived from the ARCH-LM test and the Q statistic of the standardised residuals reveal that an ARCH effect is not present in the corresponding residuals of the estimates for these financial markets. Thus, there is no problem of specification in these models.

Consequently the day of the week effect in volatility in distinct European financial markets is present even though no common behaviour is noted among the respective countries.

3.2.2 T-ARCH model

As pointed out earlier, volatility can differ significantly, depending upon the sign of the obtained yield for each period. For this reason we estimate volatility using a T-ARCH model which incorporates possible asymmetric behaviour. The structure for the equation of variance follows:

$$\sigma_t^2 = \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4 + \alpha_5 D_5 + \sum_{i=1}^{q} \alpha_{5+i} \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \gamma_i \sigma_{t-i}^2 + \lambda \, \varepsilon_{t-1}^2 d_{t-1}$$

Table 3 presents the obtained results from the analysis of the volatility in the day of the week for each stock market index in addition to the T-ARCH structure for each series.

The inclusion of a parameter which accounts for asymmetric behaviour produces clear results in this table. The most common structure in all of the markets is a GARCH (1,1), whereas Spain, France, Holland and Sweden follow a GARCH (0,1). Finally it should be noted that Germany resembles a GARCH (2,1) structure.

The asymmetric behaviour in all markets except the Czech Republic needs to be pointed out. Thus the gains and losses in each one of the stock markets in our

sample affect in volatility in a different way. The use of an additional parameter in the T-ARCH model for asymmetric behaviour leads to different results than those from the symmetric GARCH model, with the expected exception in the Czech Republic, whose results were the same for both models. The day of the week effect reveals a similar behaviour pattern in the equation of variance as in the earlier model, that is, greater volatility on Mondays and Thursdays with respect to Wednesdays, and lesser volatility on Tuesdays and Fridays, except on Mondays in the United Kingdom.

Table 3:   Day of the week effect on variance: T-ARCH model.

Country	GARCH structure	Significant Variables	Country	GARCH structure	Significant variables
Germany	(2,1)	D2	Italy	(1,1)	D1, D4
Austria	(1,1)	D2, D5	Portugal	(1,1)	D1
Belgium	(1,1)	D2	U.Kingdom	(1,1)	D1
Denmark	(1,1)	D1, D5	Czech Rep.	(1,1)	--
Spain	(0,1)	D1, D4	Sweden	(0,1)	D1, D2, D4, D5
France	(0,1)	--	Switzerland	(1,1)	D1, D4
Holland	(0,1)	D1, D4			

The results from the ARCH-LM test and the Q statistic from the standardized residuals indicate that no effect is present in the corresponding remainders of the estimates of the financial markets. Thus, we do not encounter specification problems in this model.

The following observations can be made regarding the day of the week effect based on the estimation of variance with an asymmetric model. First, a Monday effect takes place in Portugal and the United Kingdom, while a Tuesday effect occurs in Germany and Belgium. Secondly, all other countries except Sweden present seasonal behaviour in two days of the week. Thirdly, this behaviour is seen on Mondays and Thursday in Spain, Holland, Italy and Switzerland. On the other hand, Tuesdays and Fridays are statistically significant in Austria, as opposed to Mondays and Fridays in Denmark. Finally, the Swedish market demonstrates volatility each day of the week with respect to Wednesday.

4   Conclusions

Investors that are interested in including international markets in their portfolio need to know if these markets are integrated or not. We pursued the answer to this question by studying possible seasonality in international markets. Our analysis focused on an empirical comparison of the day of the week effect in the major European markets from July 1977 to March 2004, and included not only returns but volatility as well.

To begin with, we should note that most European markets do not reflect a day of the week effect since the results for each day do not differ significantly from the other days of the week. The returns in these markets are based on

representative indexes and reveal independence concerning which day of the week the return is calculated on. Nevertheless a seasonal effect can be observed on Mondays for the French and Swedish markets. The Swedish markets also reflects a significantly higher return on Fridays as opposed to the remaining days of the week.

With respect to the existence of abnormal volatility in the equation of conditional variance in the European markets, the following can be observed. A day of the week effect is present in all of the financial markets except in Portugal and the Czech Republic, where a symmetric model is applied. Exceptions are found in France and the Czech Republic, using an asymmetric T-ARCH model. Nevertheless, this effect does not agree with other analysed financial markets. However if we introduce a parameter which accounts for different behaviour in the volatility of the stock market indexes, then continuity in the day of the week effect becomes evident, differentiating the rise and fall of prices. Its presence is unlike that of the GARCH model because the statistical significance of the day of the week in the symmetric model in some cases could have been affected by asymmetric effects that were considered in the structure of the variance in the model.

Seasonality in conditional volatility in specific markets follow a similar behaviour pattern independent of the type of model that is being used. Mondays and Thursdays are more uncertain than on Wednesdays, while the Wednesday measure is lower than that of Tuesdays and Fridays.

Even though initially there does not seem to be a day of the week effect in yields from different European markets, an analysis of the conditional variance verifies that the extreme shifts observed in the major stock markets of each country indicate the absence of complete integration among all markets. This finding can be useful for an investor who is looking for investment instrument opportunities based on the change in volatility of these financial markets during specific days of the week.

References

[1] Aggarwal R. & P. Rivoli (1989): "Seasonal and day of the week effect in four emerging stock markets", Financial Review, 24, pp. 541-550.

[2] Baillie, R. T. & T. Bollerslev (1989): "The Message in Daily Exchange Rates: A Conditional-Variance Tale", Journal of Business and Economic Statistics, 7, 3, pp. 297-305.

[3] Climent, F. & V. Meneu (1999): "La Globalización de los mercados internacionales", Actualidad Financiera, noviembre, pp. 3-15.

[4] Copeland, L. & P. Wang (1994): "Estimating Daily Seasonality in Foreign Exchange Rate Changes", Journal of Forecasting, 13 , pp. 519-528.

[5] Corredor, P. & R. Santamaría (1996): "El efecto día de la semana: resultados sobre algunos mercados de valores europeos", Revista española de Financiación y Contabilidad, XXV, 86, pp. 235-252.

[6] Easton, S. & R. Faff (1994): "An Examination of the Robustness of the Day of the week Effect in Australia", Applied Financial Economics, 4, pp. 99-110.

[7] Engle, R.F. (1982): "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation", Econometrica, 50, pp. 987-1007.

[8] Glosten, L. R., R. Jagannathan & D. E. Runkle (1993): "On the relation between the expected value and the volatility of the nominal excess return on stocks", Journal of Finance, 48, pp. 1779-1801.

[9] Hsieh, D. A. (1988): "The statistical properties of daily foreign exchange rates: 1974-1983", Journal of International Economics, 24, pp. 129-145.

[10] Jacquillat, B. & B. Solnik (1978): "Multinational are Poor Tools for Diversification", Journal of Portfolio Management, 4, 2, Winter.

[11] Kyimaz, H. & H. Berument (2001): "The day of the week effect on Stock Market Volatility", Journal of Economics and Finance, 25, 2, pp. 181-193.

[12] Lamoreux C. & W. Lastrapes (1990): "Persistence in variance, structural change, and the GARCH model", Journal of Business and Economic Statistics, 2, pp. 225-234.

[13] Torrero, A. (1999): "La Importancia de las Bolsas en la internacionalización de las finanzas", Análisis Financiero, 79, pp. 62-77.

[14] Zakoian, J. M. (1990): Threshold Heteroskedasticity Models, manuscript, CREST, INSEE.

This page intentionally left blank

Simulating a digital business ecosystem

M. Petrou, S. Gautam & K. N. Giannoutakis
Electrical and Electronic Engineering, Imperial College,
London, UK

Abstract

A digital business ecosystem (DBE) is a closed or semi-closed system of small
and medium enterprises (SMEs), which will come together in cyberspace in the
same way that companies gather in a business park in the physical world. These
companies will interact with each other through buyer–seller relationships. The
purpose of this work is to develop a methodology that will allow one to study the
ecosystem under various conditions and we present here a model for the mutual
interactions between companies in a DBE and a methodology that can allow one
to study the dynamics of a digital business ecosystem. Furthermore we present a
quantitative model for studying the dynamics of such a system, inspired by human
physiology and attempting to capture many aspects of the way companies interact
with each other, including the quantitative modelling of trust and mistrust.

1 Introduction

A digital business ecosystem (DBE) is a closed or semi-closed system of small
and medium enterprises (SMEs), which will come together in cyberspace in the
same way that companies gather in a business park in the physical world. These
companies will interact with each other through buyer–seller relationships. The
purpose of this work is to develop methodology that will allow one to study the
ecosystem under various conditions. In particular, we would like to answer the
following question: "Under the assumption that the ecosystem is closed, and static,
ie no external influences, which of the companies in it are most likely to prosper
and survive, and which are most likely to be suppressed?". This is a situation
of competitive co-existence, where each unit receives excitatory and inhibitory
signals from all other units in the system. Such systems exist in biology, and their
states are known to oscillate between extremes, rather than to converge to a single

steady state. The reason of these oscillations is the asymmetry with which they influence each other: A influences B in a different way by which B influences A. So Markov random fields are not appropriate for studying such a system because the interactions between the units are asymmetric. On the other hand, biological systems consist of neurons which interact with each other in a non-symmetric way [11, 12]. Inspired from the work in [11, 12], we present here a model for the mutual interactions between companies in a DBE. This model yields a set of non-linear coupled differential equations which in the case of [4, 11, 12] govern the potential of each neuron in the visual cortex V1 and produce a saliency map of the viewed scene. In a DBE, instead of saliency of pixels we may have a fitness value of each company, or each product on sale. Following [4,11,12] we solve the system of non-linear coupled differential equations which govern these fitness values as a neural network of nodes that exchange messages. In a biological system, the membrane potential of a neuron, which corresponds to its activation level, changes with time: the stronger it is, the faster it decays. So this membrane potential obeys a differential equation. For example, we know that whatever the potential of a neuron is, in lack of any external stimulus, it will decay with time exponentially:

$$\frac{dy}{dt} = -\tau y \Rightarrow y = y_0 e^{-\tau t} \tag{1}$$

where τ is the time constant of the system, and y_0 is the value of y for the boundary condition $t = 0$.

In order to study the dynamics of an ecosystem, we must have first an instantiation of such a system. In Section 3 we show how we create a simulated DBE consisting of 100 companies which trade 20 products. The methodology we propose can be used to create realistic instantiations of a DBE, provided some statistical information is known from the real DBE we wish to study. In Section 4 we present the self-organising neural network we shall use for studying the dynamics of the simulated DBE. In Section 5 we present our experiments and results and we conclude in Section 6. We start, however, with a literature survey presented in Section 2.

2 Literature survey

There have not been many quantitative attempts to study DBEs. Most papers published follow the procedure of hypothesis generation, data collection by a survey or a questionnaire and finally hypothesis testing using statistical methods (e.g. [5,6,13,16]). There are various reasons for that: The complexity of the system, the multiplicity of the issues involved, and of course the lack of uniformity in the description of products and services, necessary to study the dynamics of a complex system [19]. The lack of such studies and the lack of quantitative measures that they could yield has consequences in the formation of economic policies for the internet [15]. We address the problem of lack of uniformity in this paper by created a realistic simulated DBE that shares its statistical properties with a real DBE.

In spite of the above mentioned difficulties, some attempts to quantify at least some of the relevant quantities in an e-commerse system have already been made. For example, Manchala in [14] makes a serious attempt to quantify trust by counting the number of transactions a vendor feels the need to verify before they proceed with the actual transaction. Manchala stresses the need for quantitative measures of trust in order to make quantitative studies of such systems. In this paper we quantify trust by invoking the psychophysical law of Weber Fechner (see Section 3). Our approach is not incompatible with the approach of Manchala: he starts from some objective measure; we start from qualitative categories of trust and try to infer from them some objective rankings. In a sense, if people were asked to use the quantitative measure of Manchala to create categories of trust, we believe that they would create categories that could be modelled by the psychophysical law of Weber Fechner. We believe that this law can bridge the gap between models like the one presented in [8], which uses qualitative terms like "low risk", "high risk" etc, and more quantitative studies like the one in [20]. Another attempt to use a quantitative model is that of Cheung and Liao [2] who produce a quantitative measure of shoppers' willingness to buy. The model is a simple regression formula, where the independent variables are the statistical scores of acceptance or rejection of certain hypotheses tested by surveys.

The importance of trust on web based transactions has been stressed by many researchers [10], to the point that there are even papers on how to build web-based systems that inspire trust to the customer [1, 16, 17]. Other people have studied the effect of trust by looking at the way web-sites evolve over time, their structure and of course by conducting surveys [18].

Most of the studies, qualitative or quantitative, concentrate on the binary interaction between supplier and buyer. One of the first attempts to try to model higher order interactions in a business environment is the one presented in [3]. This model, however, is still qualitative.

Our methodology of producing simulated DBEs may also allow the testing under controlled realistic conditions, of algorithms designed to work with real data, like for example the algorithm of Sung et al. [20] designed to cluster products according to their attributes in order to create product catalogues. Simulation experiments for studying on line stores are not novel. For example Gefen et al. used the model presented in [5] in a simulated environment to study the effect of trust using simulated scenaria.

3 A simulated DBE

Here we present methodology on how to construct a realistic simulated DBE, based on observations of a real DBE. We developed a software package which creates a database of companies and products that have the same statistical properties as in the observed DBE. This program has the flexibility to create a database of any number of companies and products. Each company created is assigned a SELL and a WANT list. The SELL list of a company is the list of the products the company wants to sell and the WANT list of a company is the list of products the company

wants to buy. Either of these lists might be empty, but not both. All products which appear in the SELL list of all companies make up the database of real products. All products which appear in the WANT lists of all companies make up the database of virtual products, because these products exist in the customers' minds. A product may appear in both databases, but it most likely will have different attributes in the two databases. A product has the same name, type and number of attributes no matter in which of the two databases it appears. What changes from one database to the other is the statistics of the attributes which characterise each product. Two types of attribute are catered for, numerical and symbolic. The statistics of the numeric attributes are characterised by their mean and standard deviation, which are assumed to be extracted by observing a real DBE. Each symbolic attribute takes values from a list of symbols, with probabilities according to the frequency with which each such value is encountered in the real DBE.

In addition, each company is assigned two other lists: the TRUST list which contains the names of the other companies in the ecosystem that it trusts, and the MISTRUST list which contains the names of the companies that it mistrusts. Any company that does not appear in either of the two lists is unknown to the company in terms of trustworthiness. We also model the effect of the spreading reputation of each company for these lists. When populating the TRUST or MISTRUST list of a company, we gave an extra weight to those companies which had already appeared in already created corresponding lists of other companies. To model the fact that good reputation propagates faster than bad reputation, the weights used for the TRUST lists are higher than the weights used for the MISTRUST lists.

Finally, we propose to use the psychophysical law of Weber-Fechner in order to convert the qualitative concepts of trust, indifference and mistrust to numerical weights for the case one wishes to construct numerical models to study these factors. The idea is to use this law to go from subjective classes of trust to relative numerical measurements that somehow reflect objectivity. According to this law, the degree of subjective judgement is proportional to the logarithm of an objective measure that measures the same effect. For example, if in your mind you create categories of untrustworthiness and you call them 1, 2 and 3, the people whom you classify in these categories have to lie to you twice, four times or eight times, respectively, for you to put them in the respective categories. So, we argue that categories of untrusted, indifferent and trusty correspond to some arbitrary objective numerical values proportional to 2, 4 and 8, respectively. As these values have to be used to weigh relatively the various possible transaction partners, their exact values do not matter. To make them into relative weights, these values are normalised to sum to 1, so in the model we shall present in the next section we shall use weights $\frac{2}{14} = 0.14$, $\frac{4}{14} = 0.29$ and $\frac{8}{14} = 0.57$ for undesirable, indifferent and desirable partner respectively.

4 Modelling the competitive co-existence of companies

Let us assume that each company C_i has with it associated a positive variable, y_i, which measures how well the company does and how strong it is, and let us call

it the *fitness* variable. This is an abstract quantity with no units associated with it, and it should not be confused with economic indicators like cash flow, volume of transactions etc. If the company is left on its own, in isolation, the value of y_i will decay according to equation (1) because a company of course cannot exist in isolation and with no interactions with other companies/customers. For simplicity, let us assume that for all companies, the decaying constant τ is the same. First we shall produce a simple model for variable y_i. The differential equation obeyed by y_i will have to model the following effects, yielding extra terms that have to be included on the right-hand-side of equation (1):

- The stronger a company is, the more strong it is likely to become. This is a self-excitation term, of the form $J_0 g_y(y_i)$. Self-excitation constant J_0 again is assumed to be the same for all companies. Function $g_y(y_i)$ is a sigmoid function: effects in real life are only linear over a certain scale. They saturate and the benefit we receive by changing the independent variable y_i levels off. On the other hand, before this positive feedback in the strength is triggered, a so called "critical mass" of strength y_i has to be reached. So, function $g_y(y_i)$ may be modelled as:

$$
g_y(y_i) = \begin{cases} 0 & \text{if } y_i < \Gamma_1, \\ \frac{(y_i - \Gamma_1)}{(\Gamma_2 - \Gamma_1)} & \text{if } \Gamma_1 \leq y_i \leq \Gamma_2 \\ 1 & \text{if } y_i > \Gamma_2 \end{cases} \tag{2}
$$

where $[\Gamma_1, \Gamma_2]$ is the range of linearity of the positive gain function.

- A term that models all excitatory signals the company receives from all other companies. First we have to quantify the excitatory signal a company C_i receives from another company C_j. A company will stimulate another company if they demand products that match those the other company sells. Let us say that one of the products a company C_i wishes to buy is product P, with attributes x_l^P for $l = 1, 2, ..., L_P$, with L_P being the number of attributes that characterise product P. There may be several companies C_j in the ecosystem that provide product P with attributes similar to those requested. The mismatch value of product P between the attributes company C_i requires and those of the same product company C_j sells may be computed as

$$
V_{Pij} \equiv \sum_{l=1}^{L_P} w_{Pl} V_{lPij} \tag{3}
$$

where if attribute l is numeric

$$
V_{lPij} \equiv \frac{|x_l^{Pj} - x_l^{Pi}|}{x_l^{Pi}} \tag{4}
$$

and if attribute l is symbolic:

$$V_{lPij} \equiv \begin{cases} 0 & \text{if } x_l^{Pj} = x_l^{Pi} \\ 1 & \text{if } x_l^{Pj} \neq x_l^{Pi} \end{cases} \tag{5}$$

V_{Pij} is the mismatch value of attribute l, between the required product by company i and the corresponding product supplied by company j. The weights w_{Pl} express the relative importance for each attribute. They are normalised so that they sum up to 1. If V_{Pij} is below a certain threshold T_1, we may assume that the products match. The more such products match, the more likely it is that company C_i will receive positive stimulation from company C_j. Let us say, therefore, that we count all products P which appear in the WANT list of company C_i and in the SELL list of company C_j and for which $V_{lPij} \leq T_1$ and find them to be E_{ij}. We may define then the excitatory signal C_i may receive from C_j as

$$J_{ij} \equiv 1 - e^{-E_{ij}} \tag{6}$$

Note that the higher E_{ij} is, the more J_{ij} will tend to 1, while when $E_{ij} = 0$, ie when no products match, $J_{ij} = 0$ too. Also note that the excitatory signal company C_j sends to C_i is not the same as the excitatory signal C_i sends to C_j. In other words $E_{ij} \neq E_{ji}$, as E_{ji} will count the pairs of products that are less dissimilar than T_1 from the sets of the WANT list of company C_j and the SELL list of company C_i. In addition, we must also realise that a company C_j will stimulate company C_i only if C_j is healthy and strong itself. A company that is very weak will probably not create much volume of trading. So, the excitatory signal J_{ij} must be modulated by $g_y(y_j)$ to account for that. In addition, company C_i will trade with company C_j only if it trusts it. So, this excitatory signal should also be weighed by the trust company C_i has to company C_j. This appears as a factor W_{ij}, which takes values $\frac{4}{7}, \frac{2}{7}, \frac{1}{7}$ when company C_j is trusted, is indifferent or mistrusted by company C_i, respectively. Finally, we must sum up all such positive influences C_i receives from all other companies in the ecosystem. So, the term we must add on the right-hand-side of (1) should be:

$$\sum_{j \in C, j \neq i} W_{ij} J_{ij} g_y(y_j) \tag{7}$$

- A term that models all inhibitory signals the company receives from all other companies. First we have to quantify the inhibitory signal a company C_i receives from another company C_j. A company will inhibit another company if both companies sell similar products. So, first we need to quantify the dissimilarity between a product P both companies sell. To do that we use equation:

$$U_{Pij} \equiv \sum_{l=1}^{L_P} w_{Pl} U_{lPij} \tag{8}$$

where if attribute l is numeric

$$U_{lPij} \equiv \frac{|x_l^{Pj} - x_l^{Pi}|}{x_l^{Pi}} \tag{9}$$

and if attribute l is symbolic

$$U_{lPij} \equiv \begin{cases} 0 & \text{if } x_l^{Pj} \equiv x_l^{Pi} \\ 1 & \text{if } x_l^{Pj} \neq x_l^{Pi} \end{cases} \tag{10}$$

U_{Pij} measures the dissimilarity between product P companies C_i and C_j sell. If this number is below a certain threshold T_2, we may assume that the products match. The more such products match, the more likely it is that company C_i will receive inhibitory signals from company C_j. Let us say, therefore, that we count all products that appear in the SELL lists of both companies for which $U_{Pij} \leq T_2$ and find them to be F_{ij}. We may define then the inhibitory signal C_i receives from C_j as

$$K_{ij} \equiv 1 - e^{-F_{ij}} \tag{11}$$

Note that the higher F_{ij} is, the more K_{ij} will tend to 1, while as $F_{ij} \to 0$, $K_{ij} \to 0$ too. We note that $F_{ij} = F_{ji}$, as F_{ji} will count the pairs of products that are less dissimilar than T_2 sold by both companies. In addition, we must also realise that a company C_j will inhibit company C_i only if C_j is healthy and strong itself. So, the inhibitory signal K_{ij} must be modulated by $g_y(y_j)$ to account for that. Finally, we must sum up all such negative influences C_i receives from all other companies in the ecosystem. So, the term we must add on the right-hand-side of (1) should be:

$$-\sum_{j \in C, j \neq i} K_{ij} g_y(y_j) \tag{12}$$

- We may also include a term which may be external input to the company, like total volume of transactions originating outside the DBE, or something like that, properly scaled to be a dimensionless number. Let us call this I_i.
- Finally, we may add a term that expresses the background input, eg the general economic climate, and it is the same for all companies in the ecosystem. Let us call it I_0.

If we put all the above together, we come up with the following differential equation that has to be obeyed by the fitness variable of company C_i:

$$\frac{dy_i}{dt} = -\tau_y y_i + J_0 g_y(y_i) + \sum_{j \in C, j \neq i} W_{ij} J_{ij} g_y(y_j) - \sum_{j \in C, j \neq i} K_{ij} g_y(y_j) + I_i + I_0 \tag{13}$$

This is a set of coupled differential equations concerning all companies in the ecosystem. If we solve it, we may be able to see the combination of values of the

fitness variables that will tell us which companies will dominate the ecosystem. Equation (13) may be solved as difference equations applied to the nodes of a fully connected network, the values of which are updated in an iterative scheme.

$$y_{i;new} - y_{i;old} = -\tau_y y_{i;old} + J_0 g_y(y_{i;old}) + \qquad (14)$$

$$\left\{ \sum_{j \in C, j \neq i} W_{ij} J_{ij} g_y(y_j) - \sum_{j \in C, j \neq i} K_{ij} g_y(y_j) \right\}_{old} + I_i + I_0$$

$$(15)$$

The values of y_i are initialised to be all equal to 1 at the first step. After each update cycle, we may remove from the system the companies the fitness value of which is below a certain threshold T_3. At the same time, we may allow the introduction of new companies with a certain rate, giving them as starting fitness value the average fitness of all other companies. In the next section, this model is investigated for various values of its fixed parameters, in order to observe the behaviour of the system under different conditions.

5 Some experimental results

We have started a series of extensive experiments in order to study the effect of each one of the parameters of the system to the dynamics of the system. The input data are the simulated DBE we constructed in Section 3. Some preliminary results are presented here. Figure 1 shows the number of companies that survive as a function of the number of iterations the system is allowed to run, for certain parameter values. In all these experiments, the following parameter values were used: $J_0 = I - i = I_0 = T_1 = T_2 = T_3 = 0.2$, $\Gamma_1 = 0.5$ and $\Gamma_2 = 1.5$. Figure 2 shows the fitness values of the various companies after 7 and 12 iterations when a monopoly was created. The parameter values that resulted in the monopoly were $\tau = 2.0$, $\Gamma_1 = 0.5$, $\Gamma_2 = 1.5$, $J_0 = I_i = I_0 = 2.5$ and $T_1 = T_2 = T_3 = 0.2$.

6 Discussion and conclusions

We presented here methodology that can allow one to study the dynamics of a digital business ecosystem. Such systems tend to be distributed in cyberspace and it is not possible to have real data for them. However, one may relatively easily acquire statistical data by using for example, a web robot, or another program designed for the purpose. The idea then is to use the gathered statistical data to produce a simulated version of the DBE which shares the same statistical properties as the real DBE. Such methodology has been used for many years by scientists to study complex systems that cannot be modelled in a deterministic way. For example, astronomers have learnt a lot about the dynamics of galaxies by studying simulated models of them.

Further, we presented a quantitative model for studying the dynamics of such a system, inspired by human physiology and attempting to capture many aspects of
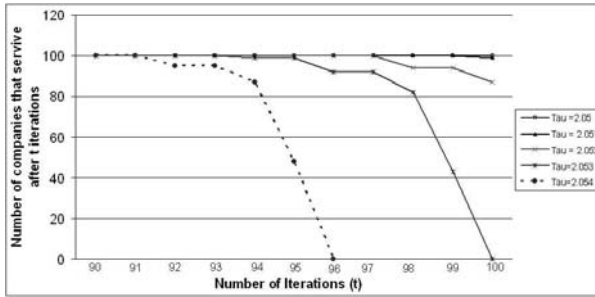
Figure 1: Number of companies that survive as a function of iterations, for various values of parameter τ and the remaining parameters fixed to $J_0 = I_i = I_0 = T_1 = T_2 = T_3 = 0.2$.
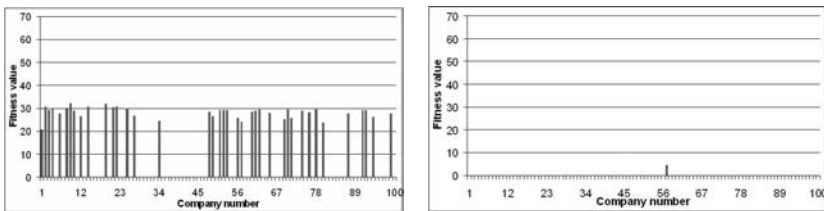

Figure 2: Monopoly created after 12 iterations. The fitness values of the companies after 7 and 12 iterations.

the way companies interact with each other, including the quantitative modelling of trust and mistrust. Several improvements to the model can be made. For example, one refinement one may make concerns the modelling of the mutual inhibition of two companies: At the moment we model this taking into consideration only the products both companies try to sell. This is fine in an open environment where the supply is infinite. However, in a closed environment when the supply is finite, two companies may exchange inhibitory signals even when they simply want to buy the same product or service. In this case we shall have to modify the calculation of term K_{ij} to rely also on the common products two companies seek to purchase. Other improvements will involve the injection of new companies into the system, in a random way.

Of course, the final step to make such a model really useful would be to be able to associate the values of its various parameters with real measurable values from the observed DBE. At this stage, only the values of the parameters that control the creation of the simulated DBE, according to Section 3, can be directly associated with measurable quantities. The values of the parameters used to study its dynamics, according to the model of Section 4, have also to be associated with real measurable quantities. This is an important big task on its own.

References

[1] Cook, D.P. & Luo, W., *The role of third-party seals in building trust on line,* e-Service Journal, Indiana University Press, 2003.

[2] Cheung, M.T. & Liao, Z., *Supply-side hurdles in internet B2C e-commerce: an empirical investigation,* IEEE Transactions on Engineering Management, **50(4)**, pp. 458–469.

[3] Choi, T.Y., Wu, Z., Ellram, L. & Koka, B.R., *Supplier-supplier relationships and their implications for buyer-supplier relationships,* IEEE Transactions on Engineering Management, **49(2)**, pp. 119–130, 2002.

[4] Dayan, P. & Abbott, L.F., *Theoretical neuroscience:computational and mathematical modelling of neural systems*, 2001.

[5] Gefen, D., *E-commerce: the role of familiarity and trust*, Omega, **28**, pp. 725–737, 2000.

[6] Gefen, D. & Straub, D., *Managing user-trust in B2C e-services,* e-Service Journal, Indiana University Press, 2003.

[7] Gefen, D., Karahanna, E. & Straub, D.W., *Inexperience and experience with online stores: the importance of TAM and trust,* IEEE Transactions on Engineering Management, **50(3)**, pp. 307–321, 2003.

[8] Jahng, J., Jain, H. & Ramamurthy, K., *Effective design of electronic commerce environments: a proposed theory of congruence and an illustration,* IEEE Transactions on System, Man and Cybernetics, Part A: Systems and Humans, **30(4)**, pp. 456–471, 2000.

[9] Jayaraman, V. & Baker, T., *The internet as an enabler for dynamic pricing of goods,* IEEE Transactions on Engineering Management, **50(4)**, pp. 470–477, 2003.

[10] Komiak, S.Y.X., Wang, W. & Benbasat, I., *Trust building in virtual salespersons versus in human salespersons; similarities and differences,* e-Service Journal, Indiana University Press, 2005.

[11] Li, Z., *Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex,* Networks:Computation in Neural Systems **10** pp. 187–212, 1999.

[12] Li, Z., *Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex,* Neural Computation, **13(8)**, pp. 1749–1780, 2001.

[13] Limayem, M., Khalifa, M. & Frini, A., *What makes consumers buy from internet? A longitudinal study of online shopping,* IEEE Transactions on System, Man and Cybernetics, Part A: Systems and Humans, **30(4)**, pp. 421–432, 2000.

[14] Manchala, D.W., *E-commerce trust metrics and models,* IEEE Internet Computing, pp. 36–44, March-April 2000.

[15] McKnight, L.W. & Bailey, J.P., *Internet Economics: when constituencies collide in cyberspace,* IEEE Internet Computing, pp. 30–37, November-December 1997.

[16] Noteberg, A., Christiaanse, E. & Wallage, P., *Consumer Trust in electronic channels,* e-Service Journal, Indiana University Press, 2003.

[17] Patrick, A.S., *Building trustworthy software agents,* IEEE Internet Computing, pp. 46–52, November-December 2002.

[18] Ruppel, C., Underwood-Queen, L. & Harrington, S.J., *e-Commerce: The roles of Trust, security and type of e-commerce involvement,* e-Service Journal, Indiana University Press, 2003.

[19] Singh, M.P., *The e-commerce inversion,* IEEE Internet Computing, pp. 4–5, September-October 1999.

[20] Sung, W.K., Yang, D., Yiu, S.M., Cheung, D.W., Ho, W.S. & Lam, T.W., *Automatic Construction of online catalog topologies,* IEEE Transactions on System, Man and Cybernetics, Part C: Applications and Reviews, **32(4)**, pp. 382–391, 2002.

This page intentionally left blank

Customer loyalty analysis of a commercial bank based on a structural equation model

H. Chi[1], Y. Zhang[1,2] & J.-J. Wang[1,2]
[1]Institute of Policy and Management,
Chinese Academy of Sciences, People's Republic of China
[2]Business School, University of Science and Technology of China,
People's Republic of China

Abstract

Customer Relationship Management (CRM) enjoys increasing attention since customers are known to be of pivotal importance to the long-term profitability and development of enterprises as well as commercial banks. With the competition among banks being more and more severe, customers' loyalty has become the decisive factor of a bank's profitability, as an increase in customer retention rate can be very profitable. In this paper, a structural equation model (SEM) is used to research into the measurement of customer loyalty and the factors that influence it. Based on an American Customer Satisfaction Index (ACSI) model, this model takes into consideration Chinese commercial banks' specific situations and improves the latent variables, manifest variables and the structure of the ACSI model. Then a partial least squares (PLS) approach is adopted to estimate the parameters in SEM. By using this model, further analysis can be conducted. A numerical example has been offered with the data deriving from a practical survey of a Chinese commercial bank. The results of the example have been analyzed and corresponding measures can be taken to improve services, thus increasing customer loyalty.
Keywords: customer loyalty, structural equation model (SEM), partial least squares (PLS) estimation procedure.

1   Introduction

Since China entered the World Trade Organization (WTO) in the year 2001, the opening pace of the Chinese financial industry has been much quicker. As a part

of the WTO commitments, China will completely open her bank sector to foreign banks from 2007. Consequently, Chinese commercial banks will face more and more severe competition in the very near future.

The customer is the source of bank profit. Studies indicate that 20% customers of retail bank yield more than 100% profit [1, 2], therefore international banking industry pays much attention to CRM. Now CRM enjoys increasing attention since customers are known to be of pivotal importance to the long-term profitability and development of commercial banks, and the customer-centric management idea has prevailed upon Chinese commercial banks. On the one hand, commercial banks must make great efforts to acquire new customers; on the other hand, they have to improve their service quality continuously in order to retain existing customers. M.T. Maloney and R.E. McCormick's study on customer value indicates that the cost of acquiring a new customer is 4~5 times that of retaining an existing customer [3]. Therefore, commercial banks may increase their profit by improving the customer retention rate and customer loyalty. Thus it can be seen that customer loyalty of commercial banks has become a decisive factor of their profitability as well as an important part of their core competence.

There are many factors that influence customer loyalty. How to find out the most important ones from all of the factors and take corresponding measures to improve banks' service quality and customer loyalty is a crucial issue for Chinese commercial banks.

Most researchers consider that customer loyalty is the measurement of customer behaviours [4–10]. Some researchers think that customer loyalty is the probability of customers' purchasing the products and services of an enterprise or the probability of repeated purchase behaviour [4–6], others think that customer loyalty is the measurement of customers' purchase frequency and purchase quantity [7–10]. Gerpott et al. [11] analyzed the relations among customer satisfaction, customer loyalty and customer retention in the German mobile communications market by using the LISREL method. David W. Wallace et al. [12] studied customer loyalty in the context of multiple channel retail strategies. Using a binomial logit model, Kim and Yoon [13] researched into the determinants of subscriber churn and customer loyalty in the Korean mobile telephony market, etc.

These studies analyzed the strategies to improve customer loyalty and the factors that influence customer loyalty. But how to measure customer loyalty still needs further research, and quantitative studies on how customer loyalty is influenced are even scarcer.

In this paper, an SEM, a multiple equation causality model, is established to study customer loyalty of a Chinese commercial bank. Using SEM, the relations among variables could be analyzed, and customer loyalty could be measured. We can also find out which variables influence customer loyalty most, and the degree of such influence could be quantified. Therefore, we are able to know in which aspects a commercial bank should make improvements to enhance customer loyalty.

The structure of this paper is as follows. In Section 2, a structural equation model is used to study how customer loyalty is influenced by other factors. A customer loyalty index (CLI) is put forward to measure customer loyalty. Section 3 explains why PLS is chosen to estimate the parameters. In Section 4, a numerical example has been offered. A questionnaire is designed for commercial banks' customers. Using this questionnaire, an investigation into a commercial bank's customers is conducted to collect the needed data. After testing the collected data, the PLS method is used to estimate the parameters of SEM. Furthermore, the factors that influence customer loyalty of the commercial bank are analyzed and the CLI is computed. Section 5 is our conclusions.

2   Structural equation model of customer loyalty

In some sense, customer loyalty is a kind of description of customers' psychological behaviour. Before choosing a corporation's products and services, customers always have certain anticipation, which affects their perception into the quality of products and services. Corporation's image and customer's perception into the quality of products and services jointly decide customers' satisfaction to this corporation. Ulteriorly, customer satisfaction will have some effects on customer loyalty. There are complicated causality among customer loyalty, perceived quality, perceived value and other variables. These variables are customers' psychological perception which could not be measured directly and accurately. Since SEM can be used to analyze the complicated relationship which involves latent variables, a structural equation model is constructed to study the measurement of commercial banks' customer loyalty and the factors that affect it.

SEM consists of two parts, the Measurement Model describing the relations between Latent Variables and their own measures, which are called Manifest Variables, and the Structure Equation Model describing the causal relations between Latent Variables. Variables like customer loyalty, perceived value and perceived quality describe customers' psychological perception. These variables can not be measured directly, so they are called Latent Variables. Every Latent Variable can be depicted by several variables which can be directly measured, and these variables are called Manifest Variables.

Many popular traditional methods (such as regression) allow dependent variables to have measurement errors, but they assume independent variables didn't have any errors. When neither dependent variables nor independent variables could be measured accurately, the traditional methods can not be applied to estimate the relationship among variables. In this circumstance, SEM can offer a better solution [14].

According to the customers' characteristics in China, the corresponding latent variables and manifest variables, the causal relations among them are designed by taking a new latent variable, corporation's image into consideration, as shown in Figure 1.

2.1 Structure equation model

$$\eta = B\eta + \Gamma\xi + \zeta$$

where $\eta' = (\eta_1, \eta_2, ..., \eta_m)$ and $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ are vectors of latent endogenous and exogenous variables, respectively. $B(m \times n)$ is a matrix of coefficient parameters for η, and $\Gamma(m \times n)$ is a matrix of coefficient parameters for ξ. This implies that $E[\eta\zeta'] = E[\xi\zeta'] = E[\zeta] = 0$.

2.2 Measurement model

$$x = \Lambda_x\xi + \delta$$
$$y = \Lambda_y\eta + \varepsilon$$

where $x' = (x_1, x_2, ..., x_q)$ and $y' = (y_1, y_2, ..., y_q)$ are the manifest exogenous and endogenous variables, respectively. $\Lambda_x(q \times n)$ and $\Lambda_y(p \times m)$ are the corresponding factor loading matrices. Here we have

$$E[\varepsilon] = E[\delta] = E[\eta\varepsilon'] = E[\xi\delta'] = 0.$$
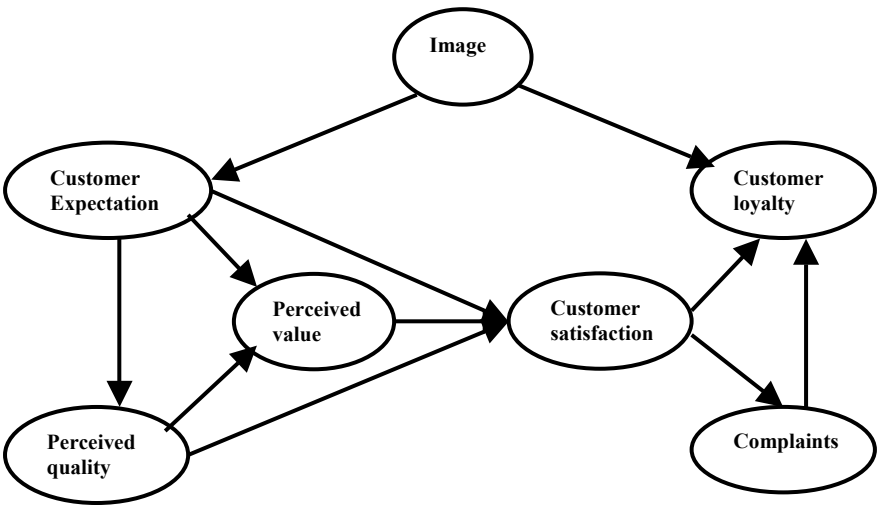
Figure 1:   Customer loyalty structural equation model.

2.3 Customer Loyalty Index (CLI)

In order to measure Customer Loyalty, Customer Loyalty Index is presented as follows:

$$CLI = \frac{10}{n}\left[\sum_{i=1}^{n}\pi_i\left(\frac{1}{m}\sum_{j=1}^{m}y_{ij}\right)\right]$$

where y_{ij} is the jth customer's opinion on the ith manifest variable of latent variable, Customer Loyalty. Suppose there are m customers whose questionnaires are valid, and there are n manifest variables. π_i denotes the weight of the ith manifest variable. In our questionnaire survey, all the manifest variables are scaled from 1 to 10. Scale 1 expresses a very negative point of view on the product and service, while scale 10 a very positive opinion. We use $10/n$ to normalize the index to ensure that the minimum possible value of CLI is 0 and its maximum possible value is equal to 100. Therefore, high CLI indicates a high level of customer loyalty.

3   Partial Least Squares (PLS) estimation procedure

There are two well-known estimation methods of SEM with Latent Variables: LISREL and PLS [15]. LISREL is a maximum likelihood method, while PLS is a least squares method. LISREL assumes a multivariate normal distribution of observed variables, and tries to fit the observed covariance matrix with the model covariance matrix estimated by model parameters. Its goal is, in a sense, to predict (which is another expression for "fit") a covariance matrix, rather than predict dependent variables such as customer's behaviour. Furthermore, LISREL requires large sample sizes.

   On the other hand, the PLS algorithm does not assume any distribution of the variables, and it specifies both the measurement and structural model as predictive relations. Once these predictive relations are specified, the estimation is a matter of minimizing the residual variances in the measurement models (measurement error variances) and structural equation models (prediction error variances). PLS weights are determined so that the least squares estimation of the model with given Latent Variables scores minimizes a combination of residual variances in the model. In addition, PLS only needs relatively small sample sizes.

   The basic PLS algorithm for SEM with Latent Variables is as follows [16]:

   We take Figure2 as an example to illustrate how PLS algorithm works.

   Stage 1: Giving arbitrary starting values to the weights $\pi_{\eta 1}$ and $\pi_{\eta 2}$

   Stage 2: Normalizing η to unit variance and regressing it on x1 and x2.

   Stage 3: Computing the predicted value of η in this regression.

   Stage 4: Normalizing η to unit variance and regressing it on y1 and y2. So we can get new weights $\pi_{\eta 1}$ and $\pi_{\eta 2}$. An iteration is completed. Repeating stages 1 to 4 until convergence is obtained, and the procedure is repeated k times.

   Stage 5: ξ is estimated by a similar procedure.

   Stage 6: At this point, the estimated $\hat{\eta}_k$ is regressed on the estimated $\hat{\xi}_k$ to obtain the structural relationship between the two. The regression of $\hat{\eta}_k$ on y1 and y2 gives the estimated criterion weights ($\pi_{\eta 1}$, $\pi_{\eta 2}$) and the regression of $\hat{\xi}_k$ on x1 and x2 gives the estimated predictor weights ($\pi_{\xi 1}$, $\pi_{\xi 2}$).
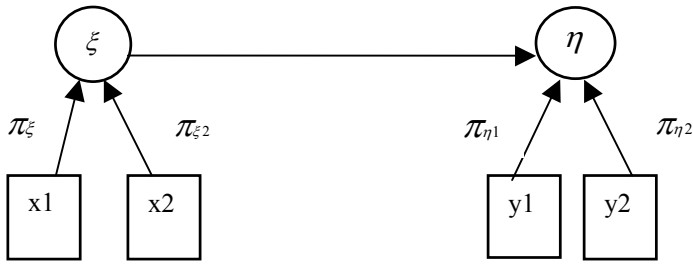
Figure 2:   A simple example of SEM.

Table 1:   The variables definition of customer loyalty SEM.

Latent Variables	Manifest Variables
Image ($\xi 1$)	a) It is a leader in banking industry (x1)
	b) Its services are customer-centric (x2)
	c) What do you think of the whole image of this bank (x3)
Customer Expectations of the overall quality ($\xi 2$)	a) Expectations for the overall quality of this bank at the moment you became customer of this bank (x4)
	b) Expectations for this bank to provide financial products and service to meet your personal need (x5)
	c) How often did you expect that things could go wrong at this bank (x6)
Perceived Quality ($\xi 3$)	a) Overall perceived quality (x7)
	b) Service attitude of this bank's clerks (x8)
	c) Professional efficiency of this bank's clerks (x9)
	d) Is the business hours convenient for you (x10)
	e) Accessibility of the bank's branches (x11)
	f) The bank's capacity to meet customers' individual demand (x12)
	g) Is the business procedure complicated (x13)
	h) Customers' waiting time before being served (x14)
Perceived Value ($\xi 4$)	a) Given the fees and prices that you pay for this bank how would you rate the quality of the products and services offered by the bank? (x15)
	b) Given the quality of the products and services offered by this bank how do you think the fees and prices that you pay for them? (x16)
Customer Satisfaction ($\xi 5$)	a) Overall satisfaction (x17)
	b) The gap between your practical perception and your expectation (x18)
	c) What do you think of this bank compared with your ideal bank? (x19)
Customer Complaints ($\xi 6$)	a) How many times have you complained about this bank last year? (x20)
	b) How well, or poorly, was your complaints handled? (x21)
Customer Loyalty ($\xi 7$)	a) If you would need to choose a new bank how likely is it that you would choose this bank again? (x22)
	b) If other banks decide to lower their fees and prices, but this bank stays at the same level as today. At which level of difference would you choose another bank? (x23)
	c) If your friends or colleagues ask you for advice, how likely is it that you would recommend this bank? (x24)

4   Numerical example

4.1 Data collection and variables definition

We sampled 65 customers of a Chinese commercial bank and conducted a questionnaire survey. After eliminating missing values and illogical responses, we got 61 valid questionnaires (93.8%). The Latent Variables and Manifest Variables of customer loyalty SEM are defined as in Table1.

All the items are scaled from 1 to 10. Scale 1 expresses a very negative point of view on the product and service, while scale 10 a very positive opinion.

4.2 Data analysis

Since LISREL requires that observed variables should meet normal distribution, we checked our sample data. Using One-Sample Kolmogorov-Smirnov Test, it can be found out that our data failed to meet normal distribution. Here is an example, the observed value of variable x12, to illustrate the case. Histogram of manifest variable x12 is as follows.
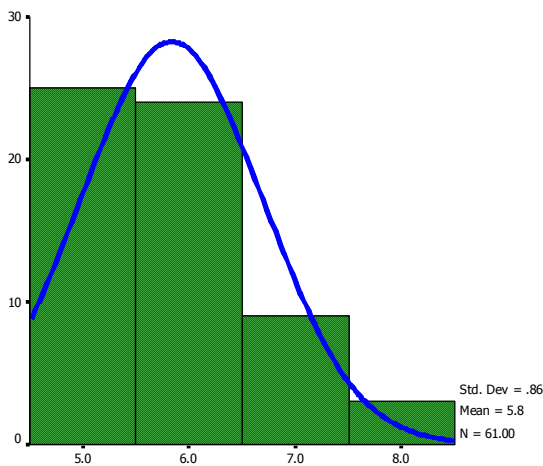
Figure 3:   Histogram of manifest variable x12.

From the histogram, it is not obvious whether x12 has normal distribution. Therefore we conduct One-Sample Kolmogorov-Smirnov Test. Table 2 shows the results.

It can be seen from Table 2 that asymptotic significance (two-tailed) of x12 is 0.001. It is less than 0.05 significance level and does not have normal distribution.

Table 2:   One-sample Kolmogorov-Smirnov test.

		X12
N		61
Normal Parameters(a,b)	Mean	5.8361
	Std. Deviation	.85985
Most Extreme Differences	Absolute	.244
	Positive	.244
	Negative	-.166
Kolmogorov-Smirnov Z		1.909
Asymp. Sig. (2-tailed)		.001

a. Test distribution is Normal. b. Calculated from data.

4.3 Parameters estimation

Utilizing the survey data, PLS estimation procedure is used to estimate the parameters of SEM. After five iterations, the parameters' convergence is obtained.

4.3.1 Measurement model

$$\xi1=0.7369*x1+0.0319*x2+0.0092*x3 \tag{1}$$

$$\xi2=0.7182*x4+0.1497*x5+0.0921*x6 \tag{2}$$

$$\xi3=0.0734*x7+0.3629*x8+0.3137*x9+0.0141*x10+0.0780*x11$$
$$+0.2679*x12+0.0226*x13+0.1183*x14 \tag{3}$$

$$\xi4=0.5589*x15+0.3828*x16 \tag{4}$$

$$\xi5=0.5609*x17+0.1984*x18+0.3728*x19 \tag{5}$$

$$\xi6=0.5603*x20+0.3364*x21 \tag{6}$$

$$\xi7=0.3654*x22+0.0714*x23+0.5082*x24 \tag{7}$$

4.3.2 Structure equation model

$$\xi2=0.6544*\xi1 \tag{8}$$

$$\xi3=0.3851*\xi2 \tag{9}$$

$$\xi4=0.2212*\xi2+0.4479*\xi3 \tag{10}$$

$$\xi5=0.0765*\xi2+0.7001*\xi3+0.1597*\xi4 \tag{11}$$

$$\xi6=-0.6632*\xi5 \tag{12}$$

$$\xi7=0.2472*\xi1+0.7483*\xi5-0.3044*\xi6 \tag{13}$$

4.4 Customer Loyalty Index (CLI)

$$CLI = \frac{10}{n}\left[\sum_{i=1}^{n}\pi_i\left(\frac{1}{m}\sum_{j=1}^{m}y_{ij}\right)\right] = 45.0894$$

4.5 The implications from the results of the example

1) Customer loyalty has a significant impact on Customer Satisfaction. Therefore to retain its customers, the commercial bank must make great efforts to enhance Customer Satisfaction.

2) Corporation image has some impact on Customer Loyalty, but it is not as important as Customer Satisfaction. It means that good image can help the commercial bank to attract customers, but what affect the commercial bank's customer retention most is Customer Satisfaction.

3) Customer complaints have certain negative impacts on customer loyalty. So the commercial bank should raise the service quality, deal with customer complains seriously and timely, and reduce the negative impacts to the minimum.

4) Perceived Quality has significant impact on Customer Satisfaction, so the commercial bank must improve their service level to enhance Customer Satisfaction.

5) Several manifest variables have significant impact on perceived quality, they are listed in descending order by their influence as follows: service attitude of bank clerks, technical level, bank's capacity to meet customers' personal business demand and customers' waiting time .

6) Perceived Value has some positive impact on Customer Satisfaction, but its influence degree is much less than Perceived Quality. It indicates that at present the fees and prices of this bank are relatively reasonable, what customers care most is the service quality offered by this bank.

7) The Customer Loyalty Index (CLI) of this bank is only 45.0894, which is much less than the maximum CLI (100). It means that this bank's customer retention is poor, and it must take powerful measures to keep its customers.

5   Conclusion

In this paper, according to the situation of the bank, the Structural Equation Model was constructed to study the relationships among customer loyalty and some factors. PLS procedure was employed to estimate the parameters of SEM based on a survey among 65 customers of a bank. From this bank's customer loyalty SEM we can find out that service quality and bank's capacity to meet customers' personal business demand are the most important factors influencing the commercial bank's customer retention. Only by sticking to be customer-centred and reinforcing customer relationship management, can the bank win the drastic competition in the Chinese finance market.

References

[1]   G. Hartfeil. Bank One Measures Profitability of Customers, Not Just Products. Journal of Retail Banking Services, 18(2) pp.24-31,1996.

[2]   Michael T. Maloney, Robert E. McCormick and Raymond D. Sauer. Customer Choice, Customer Value: An Analysis of Retail Competition in America's Electric Industry, Washington, DC: Citizens for a Sound Economy Foundation, pp.151-169,1996.

[3]   RL Desatnick, D Detzel. Managing to keep the customer, Boston, Massachusetts: Houghton Mifflin,1988.

[4]   Kuehn, A. A, An Analysis of the Dynamics of Consumer Behaviour and Its Implications for Marketing Management. Ph.D. dissertation, Graduate School of Industrial Administration, Carnegie Institute of Technology, 1958.

[5]   Lipstein, B. The dynamics of brand loyalty and brand switching. Proceedings of 5th Annual Conference of the Advertising Research Foundation, September 25th, pp.101-108,1959.

[6]   Massey, W.F., Montgomery, D.B. and Morrison, D.G. Stochastic models of buyer behaviour. Cambridge: MIT Press, 1970.

[7]   Cunningham, R.M. Brand loyalty-What, where, how much?. Harvard Business Review, 34(Jan-Feb), pp.116-128,1956.

[8]   Brody, RP and Cunningham. Personality Variables and the Consumer Decision Process. Journal of Marketing Research, 10, pp.50-57,1968.

[9]   A. Magi, C. Julander. Perceived service quality and customer satisfaction in a store performance framework-An empirical study of Swedish grocery retailers. Journal of Retailing and Consumer Services, 3 (1), pp.33-41,1996.

[10]   Hughes, C. E. Designing Products for Maximum Customer Satisfaction. IEEE, 1-5,1996.

[11]   Torsten J. Gerpott, Wolfgang Rams and Andreas Schindler. Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommunications Policy, 25(4), pp. 249-269, 2001.

[12]   David W. Wallace, Joan L. Gieseb, and Jean L. Johnson. Customer retailer loyalty in the context of multiple channel strategies. Journal of Retailing, 80(4), pp. 249-263, 2004.

[13]   Hee-Su Kim, and Choong-Han Yoon. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy, 28(9), pp. 751-765,2004.

[14]   C. Fornell, M.D. Johnson, E.W. Anderson. The American customer satisfaction index: nature, purse, and findings, J. Mark. 60, pp. 7-18,1996.

[15]   Jae Cha. CFI Group Methodology: PLS, Bridging Measurement and Management, Issue 2, 2000.

[16]   Wynne W. Chin. Partial Least Squares For Researchers: An overview and presentation of recent advances using the PLS approach, disc-nt.cba.uh.edu/chin/indx.html.

Do markets behave as expected?
Empirical test using both implied volatility and futures prices for the Taiwan Stock Market

A.-P. Chen, H.-Y. Chiu, C.-C. Sheng & Y.-H. Huang
Institute of Information Management, University of National Chiao Tung, Taiwan, Republic of China

Abstract

This study adopts derivative pricing as an indicator of market expectations, with those results suggesting that general investors can use market expectations to predict the final settlement value of underlying assets. Most investment textbooks note that one of the major functions of futures is price discovery. Similarly, the implied volatility associated with option prices can be used to discover the volatility of the underlying asset. This study combines futures price and implied volatility to establish a probability space of market expectations regarding the final settlement value of the underlying asset, and verifies this probability space using empirical data from the Taiwan stock market. The verification results suggest that market expectations closely reflect the actual behavior of the final settlement value of the underlying asset, and thus provide a practical perspective on future price behavior. According to this study, investors can easily estimate underlying asset behavior based on the behavior of the related futures and options and without incurring significant measurement error, which can be helpful in risk management and planning investment strategies.
Keywords: forecasting, market expectation, implied volatility, futures, probability space.

1   Introduction

Most financial texts confer that the futures markets aggregate diverse information and expectations regarding the future prices of underlying assets, and thus provide a common reference price which is known as the price

discovery function of futures. For example, three topics are commonly discussed in relation to the price discovery function of futures. The first topic deals with the lead-lag relationship and information transmission between the prices of national markets, or between different securities [6, 21]. The second topic involves the discussion of volatility spillover, since volatility is also a source of information [12, 19]. The third topic relates to the phenomenon of information transmission between stock index and index futures markets [7, 9, 14, 15, 18].

Similar to futures markets, options markets may also provide a common reference of subsequent real volatility (RV) by calculating the implied volatilities (IV). Early research on the predictive capability of IV found that IV explains variation in future volatilities better than that in historical volatilities (HV). For example, Lantane and Rendleman [17] found that actual option valuations were better explained by actual volatility over the life of the contract than by historical volatility. Chiras and Manaster [8] also tried to compare the predictive power of IV and HV using the CBOE data, and found that IV has superior forecasting power to HV. However, subsequent studies applying time serious methodologies to study the predictive power of implied volatility have yielded mixed results. Some studies have found that IV is a poor method of forecasting subsequent RV, while other studies have found that IV is a good method of forecasting RV. For example, Canina and Figlewski [5] found virtually no relation between the IV and subsequent RV throughout the remaining life of S&P 100 index options before maturity date. Moreover, Day and Lewis [11] and Lamoureux and Lastrapes [16] both found that GARCH associated with HV is better able to predict RV than IV. Meanwhile, other studies have found that IVs provide reasonably good information on the subsequent RVs of the underlying asset. For example, Harvey and Whaley [4] tested and rejected the hypothesis that volatility changes are unpredictable. Blair et al. [2] indicated that IV index provides good forecasts for S&P 100 index. Moreover, Fleming [13] examined the performance of the implied volatility of the S&P 100 for forecasting future stock market volatility, and found that although IV has an upward bias but it contains relevant information regarding future volatility.

Although the volatility discovery function of options yields mixed results, the mutual influences between options and futures cannot be ignored where futures and options contracts exist for the same underlying asset. Put-call parity is a strong arbitrage relation, first identified by Stoll [20], that exists between the prices of European put and call options that share same underlying asset with the same strike and expiry, and the combination of options can create positions that are equivalent to holding the underlying itself. Accordingly, the call price C, put price P, strike price K, and underlying stock price S must satisfy the following equation: $C - P = S - PV(Dividends) - K\,e^{(-r\,T)}$, where T denotes the option holding period, r represents a continuously compounded interest rate and $PV(Dividends)$ is the present value of the dividends received by the stock owner during the holding period. Although put-call parity identifies the relationship between options and underlying assets, in real world scenarios it is difficult to duplicate a stock index with limited stock positions but such an index can be easily duplicated using the related futures. Given the simultaneous existence of

both options and futures contracts for the same underlying securities and with the same strike and expiry, put-call parity becomes $C - P = FP - PV(Dividends) - K\,e^{(-r\,T)}$ where FP is the futures price of the relative underlying asset. Thus, it is necessary to verify that the combination of futures and options prices reflects market expectations regarding the final settlement value behaviours of the underlying asset. This study examines futures and options prices of the Taiwan Stock Exchange Capitalization Weighted Stock Index and finds that the market expectations can display the real final settlement value via a nearly normal distribution in which the mean equals the futures price and the standard deviation equals the implied volatility.

2   Determining market expectations using futures and options

Market expectations are difficult to observe directly. However, market expectations can be estimated using derivatives because derivatives prices reflect the expected behaviours of the final settlement value of the underlying asset on maturity. For example, if the spot price of a particular asset is $100 and its futures price is $101, the market expects the final settlement value of the underlying asset to be $1 higher than its current price. In this case, the no-arbitrage principle holds that investors will sell the futures and simultaneously buy the spot to perform risk-less arbitrage. Consequently, the spot and futures prices will be equal, thus weakening the price discovery function of futures. Although the no-arbitrage principle holds for most real assets, virtual assets such as stock indexes may violate this principle because of the difficulty of generating a reasonable position that precisely mirrors the behaviour of the index using limited stocks combinations in order to perform risk-less hedge. Thus, the prices of the stock index and its associated futures do not always equal to each other in the real world even after discounting the forward looking nature of futures. The price discovery function thus still holds for stock index futures.

Besides futures, options also provide information regarding market expectations, and Cox and Ross [10] established that the value of options is their expected return at maturity. Current markets widely adopt the Black–Scholes model [3] to calculate expected returns using the input variables of spot price S, exercise price K, risk-less interest rate r, time to maturity t and volatility σ of the underlying asset. The implied volatility (IV) can be calculated by including the option market price in the Black–Scholes model with fixed (S, K, t, r). Being derived from market price, IV can be interpreted as the market expectations of volatility during the time to maturity of the underlying asset. Combining the futures price (FP) and IV can calculate a range of market expectations regarding the value of the underlying asset on maturity, with the central point μ located at FP and standard deviation σ equalling the IV transformed into the scale of days to maturity V_t as shown in Figure 1.

Assuming that FP and IV form a market expectation probability space regarding the final settlement value, this study verifies the accuracy of such market expectations in the following section using four years of historical data from the Taiwanese stock market.
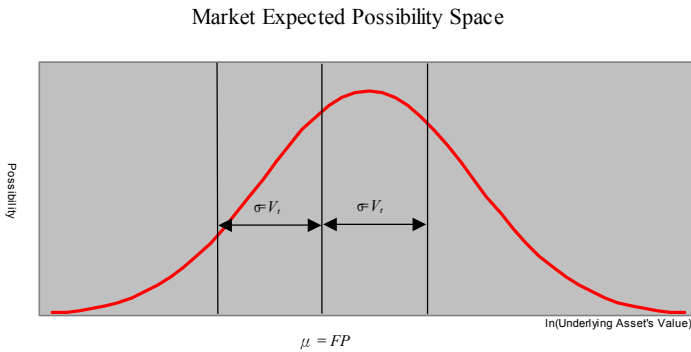
Market Expected Possibility Space

Figure 1: Market expected possibility space generated by (FV, V_t).

3 Empirical test of market expectations

This section outlines the verification procedure and results of testing the accuracy of market expectations formed by FP and IV. The notations used in this study are shown as Table1.

Table 1: Notations list.

C = Call Price
P = Put Price
σ = Volatility
r = Interest Rate
S = Spot Price
K = Exercise Price
T = Time to Maturity in years
t = Time to Maturity in days
$N()$ = Cumulative Standard Normal Distribution Function
$N_d(\mu,\sigma)$ = Normal Distribution Function with mean of μ and standard deviation of σ
IV = Implied Volatility, annual
V_t = Implied Volatility over t days.
FP = Futures Price
SP = Final Settlement value

The first step in generating a market expectation probability space is calculating the implied volatilities. Most studies on the observed market prices of various options based on a single underlying asset estimated the volatility using at-the-money or near-the-money options since these instruments are more sensitive to volatility changes and least susceptible to the influence of the bid-ask spread [22]. Notably, Beckers [1], Wiggins [23] and Canina and Figlewski [5] indicated that near-the-money options are better predictors of future real volatility than IVs of deep in or out of the money options. Thus, this study uses only the daily close price of near-the-money nearby options contracts with the same underlying asset of the same strike and expiry for verification.

According to the Black–Scholes model, the call price C and put price P of a European option can be calculated as follows:

$$C = SN(d_1) - Ke^{-rT}N(d_2) \tag{3.1}$$

$$P = Ke^{-rT}N(-d_2) - SN(-d_1) \tag{3.2}$$

$$d_1 = \frac{\ln(\frac{S}{K}) + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}} \tag{3.3}$$

$$d_2 = \frac{\ln(\frac{S}{K}) + (r - \frac{\sigma^2}{2})T}{\sigma\sqrt{T}} \tag{3.4}$$

This study modified the method used to calculate implied volatility to yield an unbiased probability space. Traditionally, calculating an implied volatility (IV) requires solving (3.1) or (3.2) repeatedly using different trial values for the volatility input. However, the derived IV of a call option (3.1) rarely equals the value obtained from a put option (3.2) with the same exercise price, and thus most IV related studies only consider call or put options. However, with the relationship indicated by put-call parity, the FP, C and P of the same underlying with the same strike and expiry are tightly coupled, and a slight change in the price of one will immediately affect the price of the other two. Thus, to verify the possibility space created by FP and IV, this study combined put and call IVs to yield a single value IV, i.e., the union IV (IV_u). Combining (3.1) and (3.2) can solve IV_u using the following formula:

$$C + P = SN(d_1) - Ke^{-rT}N(d_2) + Ke^{-rT}N(-d_2) - SN(-d_1) \tag{3.5}$$

Let $\mathcal{O}_c = \{C_1, C_2, ..., C_n\}$ denote the historical near-the-money call option prices in time interval $\mathcal{J}$, while $\mathcal{O}_p = \{P_1, P_2, ..., P_n\}$ represents the historical near-the-money put options prices in $\mathcal{J}$ and $\mathcal{F} = \{FP_1, FP_2, ..., FP_n\}$ is the futures prices in $\mathcal{J}$. For the ith historical data in $\mathcal{J}$, (C_i, P_i, FP_i) is used to generate the ith sampling distribution space and the actual final settlement value of (C_i, P_i) is SP_i.

Considering the ith historical data in $\mathcal{J}$, substituting C_i and P_i into (3.5), the annual IV_u of the ith day in $\mathcal{J}$ can be solved. Let $\mathcal{V} = \{V_1, V_2, ..., V_n\}$ denote the IV_u in $\mathcal{J}$. Furthermore, transfer the annual IV_u into the same scale as days to mature, and let v_t denote the union implied volatility of t days to mature:

$$v_t = \sqrt{\frac{t \times V_i^2}{365}} \tag{3.6}$$

Assume that SP_i and FP_i are the relative final settlement value and future price of the ith historical data in $\mathcal{J}$, then the actual SP is located at d_i standard deviations of an $N_d(\mu, \sigma) = N(FP_i, v_t)$ distribution space:

$$d_i = \frac{\ln(\frac{SP_i}{FP_i})}{v_t} \tag{3.7}$$

Thus, for each historical record in $\mathcal{I}$ a d_i can be calculated, and collectively these d_i form a sampling space of standard deviations $\mathcal{D} = \{d_1, d_2,..., d_n\}$. If the final settlement value follows the distribution of $N_d(\mu,\sigma) = N_d(FP_i, v_t)$ as expected by the market, the distribution of $\mathcal{D}$ will be a standard normal distribution, $N_d(\mu,\sigma) = N_d(0, 1)$.

To confirm the accuracy of the market expectations of (*FP, IV*), this study uses the option and futures prices of the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) from 24/12/2001 to 30/12/2005 to test the results, $\mathcal{I}$=[24/12/2001,30/12/2005]. The index values of TAIEX in $\mathcal{I}$ are demonstrated as Figure 2.
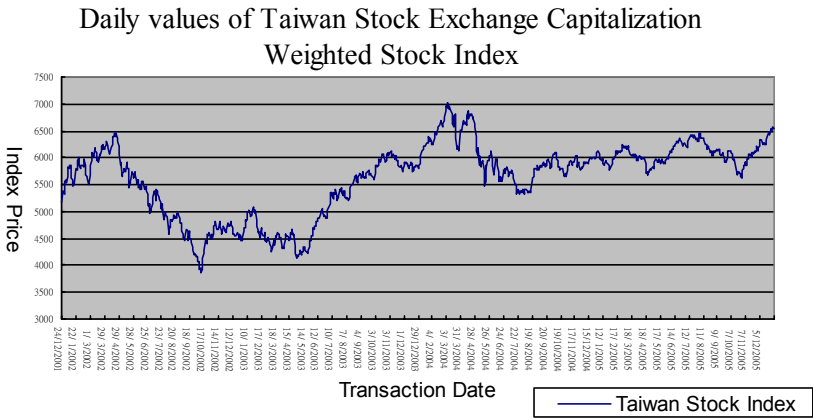
Figure 2:   Stock index values in I.

The risk-less interest rate *r* applied in this study is the monthly fixed deposit interest rate adopted by the Central Bank of Taiwan. The calculation results of $\mathcal{D}$ are collected in Figure 3.for comparison with the normal distribution $N_d(0, 1)$, and the mean, standard deviation, skewness and kurtosis of $\mathcal{D}$ are 0.0163, 1.1109, -0.2920 and 3.4500, respectively.

Figure 3 exhibits that the distribution of $\mathcal{D}$ approximates that of $N_d(0,1)$, and is skewed left and peaked compared to $N_d(0.0163, 1.1109)$ but slightly flatter and skewed right compared to $N_d(0, 1)$ also has thicker tails. That is, the actual final settlement values for Taiwan Stock Index futures and options roughly follow the market expectation formed by FP and IV, but have slightly larger volatilities. Another interesting observation is that two possibility peaks located at exactly +0.5 and -0.5 standard deviations.

Distribution Map of (FP, IV)
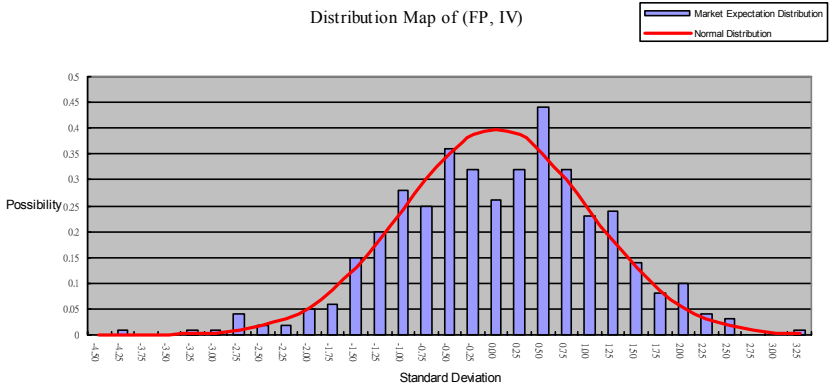
Figure 3:   Market expectation distribution map of (FP, IV) compared to the normal distribution, $D(\mu,\sigma,skewness,kurtosis) = D(0.0163, 1.1109, -0.2920, 3.4500)$.

Distribution Map of (S, IV)
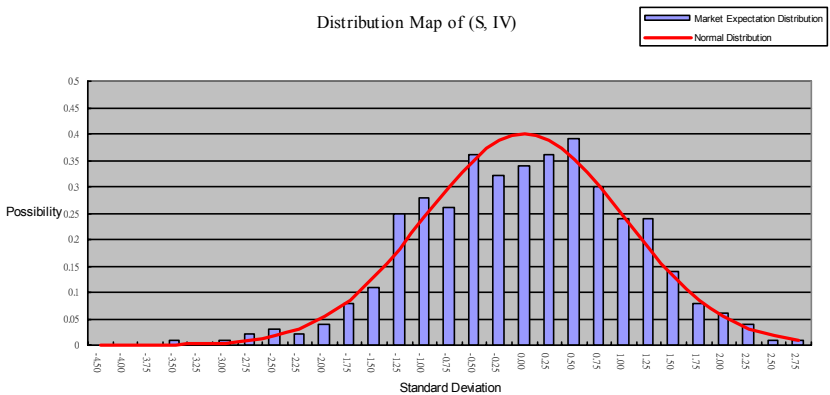
Figure 4:   The market expectation distribution map of (S, IV) compared to the normal distribution, $D'(\mu,\sigma,skewness,kurtosis) = D'(0.0050, 1.0609, -0.3102, 3.4146)$.

Using the closing price of the stock index rather than the FP in (3.7), another possibility space formed by IV only can be generated. Letting $\mathcal{D'} = \{d'_1, d'_2, \ldots, d'_n\}$, (3.7) becomes:

$$d'_i = \frac{\ln(\frac{SP_i}{S_i})}{v_t} \tag{3.8}$$

The computed results of $\mathcal{D'}$ in $\mathcal{J}$ are summarized in Figure 4. Compared to the normal distribution $N_d(0, 1)$, the mean, standard deviation, skewness and kurtosis of $\mathcal{D'}$ are 0.0050, 1.0609, -0.3102 and 3.4146, respectively.

Figure 4 displays that the distribution of $\mathcal{D}'$ resembles $N_d(0,1)$ more closely than does that of $\mathcal{D}$ and is skewed left and peaked compared to $N_d(0.0050,1.0609)$, but slightly flatter and skewed right compared to $N_d(0,1)$ and also has thicker tails. Comparing $\mathcal{D}$ and $\mathcal{D}'$, the actual final settlement values of Taiwan Stock Index futures and options roughly follow the market expectations formed by FP and IV but more closely follow the expectations of S and IV.

The cumulative possibility distributions from intervals $\pm\sigma$ to $\pm3\sigma$ of the normal distribution, $\mathcal{D}$ and $\mathcal{D}'$, are listed as Table 2.

Table 2: Cumulative possibility distributions.

		$\pm\sigma$	$\pm2\sigma$	$\pm3\sigma$
Normal Distribution	Value	0.6827	0.9545	0.9973
Market Expectation Distribution of $\mathcal{D}$(FP, IV)	Value	0.632	0.932	0.989
	Delta (%)	-7.4264	-2.3573	-0.8322
Market Expectation Distribution of $\mathcal{D}'$(S, IV)	Value	0.657	0.941	0.994
	Delta (%)	-3.7645	-1.4144	-0.3309

Table 2 reveals that the cumulative possibility of $\mathcal{D}$ is smaller than the normal distribution, being approximately 7.43% in the plus minus one standard deviation, and 0.83% smaller in $\pm3\sigma$. $\mathcal{D}'$ is even closer to the normal distribution in the testing example presented here, but both $\mathcal{D}$ and $\mathcal{D}'$ clearly demonstrate that the option price is slightly under estimated. However, after calculated the investment profits during the sampling period, the result indicates that the market maker only suffers 2% loss in four years. Observing Figure 3, Figure 4 and Table 2 reveals that the final settlement value of the stock index roughly meets market expectations as indicated by both the futures and options. This conclusion also indicates that investors can expect the final settlement value of a stock index to equal the futures price, with a standard deviation equalling the implied volatility derived from option price. Based on this conclusion, investors can expect the performance of their stock positions to roughly follow that of holding both futures and options combinations, but most likely with an expectation bias of ±0.5 implied volatilities. Another observation of this study also suggests that (S, IV) combinations more closely reflect stock index price behaviour than (FP, IV) combinations; that is, applying options only may achieve better risk management performance than using combinations of both options and futures.

4 Conclusions

This study used derivatives prices to indicate market expectations and found that the final settlement value of the stock index moves roughly in accordance with market expectations. Although market expectations cannot precisely forecast the final settlement value of the underlying assets, general investors can easily adopt

the futures price as the expected final settlement value, with standard deviation equaling the near-the-money implied volatility derived from option prices.

Another interesting finding was that market participants tend to slightly under estimate actual volatility in the Taiwanese stock market. Two possible explanations for this finding exist. The first explanation may be the pricing error of the naïve Black-Scholes model, which constantly underestimates the near-the-money option price. However, this study finds that option market makers suffer only a 2% loss during the four sample years, indicating that the Black-Scholes model only insignificantly underestimates the near-the-money option price. The second explanation is that general market participants tend to be conservative in anticipating stock index volatility and may slightly under-estimate option prices. The final and most interesting finding of this study was that overall market expectations exhibit two peaks, located at the +0.5 and -0.5 standard deviations. This phenomenon is particularly clear for the market expectations regarding FP and IV, and thus further research is required to explore this market behavior and its influence on general investors.

References

[1]   Beckers, S., Standard Deviations Implied in Option Prices as Predictors of Future Stock Price Variability, *Journal of Banking and Finance*, Vol. 5, No. 3, 1981, pp. 363-381, 1981.

[2]   Bevan J. Blair, Ser-Huang Poon, Stephen J., Taylor. Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. *Journal of Econometric*, 105, pp. 5-26, 2001.

[3]   Black, F., Scholes, M., The pricing of options and corporate liabilities. *Journal of Political Economy*, 81, pp. 637-654, 1973.

[4]   Campbell R. Harvey and Robert E. Whaley, Market volatility prediction and the efficiency of the S&P 100 index option market, *Journal of Financial Economics* 31, pp. 43-73, 1992.

[5]   Canina, L., Figlewski, S., The informational content of implied volatility. *Review of Financial Studies* 6, pp. 659-681, 1993.

[6]   Chan, K., A further analysis of the lead-lag relationship between the cash market and stock index futures market. *Review of Financial Studies*, 5, pp.123-152, 1992.

[7]   Chan, K., Chan, K. C., & Karolyi, G. A., Intraday volatility in the stock index and stock index futures markets. *Review of Financial Studies*, 4, pp. 657-684, 1991.

[8]   Chiras, D.P., Manaster, S., The information content of option prices and a test of market efficiency. *Journal of Financial Economics,* 6 (2/3), pp. 213-234, 1978.

[9]   Chu, Q. C., Hsieh, W. G., Tse, Y., Price discovery on the S&P 500 index markets: An analysis of spot index, index futures, and SPDRs. *International Review of Financial Analysis*, 8, pp.21-34, 1999.

[10]   Cox, John C. and Stephen A. Ross, The valuation of options for alternative stochastic processes, *Journal of Financial Economics,* 3, pp.145-166, 1976.

[11]   Day, T., Lewis, C., Stock market volatility and the information content of stock index options. *Journal of Econometrics*, 52, pp.267-287, 1992.

[12]   French, K. R., & Roll, R., Stock return variances: The arrival of information and the reaction of traders. *Journal of Financial Economics*, 17, pp.5-26, 1986.

[13]   Jeff Fleming, The quality of market volatility forecasts implied by S&P 100 index option prices. *Journal of Empirical Finance*, 5, pp. 317-345, 1998.

[14]   Kawaller, I.G. & Koch, P.D. & Koch T.W., The Temporal Price Relationship between S&P 500 Futures and the S&P 500 index. *Journal of Finance*, 42, pp.1309-1329, 1987.

[15]   Koutmos, G., & Tucker, M., Temporal relationships and dynamic interactions between spot and futures stock markets. *Journal of Futures Markets*, 16, pp.55-69, 1996.

[16]   Lamoureux, C.G., Lastrapes, W., Forecasting stock return variance: towards understanding stochastic implied volatility. *Review of Financial Studies* 6, pp.293-326, 1993.

[17]   Latane, H., Rendleman, R., Standard deviation of stock price ratios implied in option prices. *Journal of Finance,* 31, pp.369-381, 1976.

[18]   Raymond W So, Yiuman Tse, Price Discovery in the HANG SENG INDEX MARKETS: index, futures, and the tracker fund. *The Journal of Futures Markets*. Hoboken, Sep 2004.Vol.24, Iss. 9; pp. 887, 2004.

[19]   Ross, S., Information and volatility: The no-arbitrage martingale approach to timing and resolution irrelevancy. *Journal of Finance*, 44, pp.1-17, 1989.

[20]   Stoll, Hans R., The relationship between put and call option prices, *Journal of Finance*, 23, pp.801-824, 1969.

[21]   Stoll, H. R., Whaley, R. E., The dynamics of stock index and stock index futures returns. *Journal of Financial and Quantitative Analysis*, 25, pp.441-468, 1990.

[22]   Whaley, Robert, Valuation of American Call Options on Dividend-Paying Stocks: Empirical Tests. *Journal of Financial Economics*, 10, pp. 29-58, 1982.

[23]   Wiggins, J., Option values under stochastic volatility: Theory and empirical estimates, *Journal of Financial Economics*, 19, pp.351-372, 1987.

The simulation of news and insiders' influence on stock-market price dynamics in a non-linear model

V. Romanov, O. Naletova, E. Pantileeva & A. Federyakov
Plekhanov Russian Academy of Economics, Russia

Abstract

This work consists of two parts. The first is devoted to stock market simulation and is based on the model described by Li and Rosser. New factors were introduced in the mentioned model. "Bad" or "good" news arose at each moment. The model has a memory, which determines the rate of forgetting the news. In such a way, the news background is being formed by addition of decaying news intensities. The news background modifies the fundamental value of current market price according to the logistic law. The insiders at moment t-1 know the prices at the moment t. The market simulation by means of the proposed model shows that, in case of "good" news, the stock-market prices are rising, and in the case of "bad" news, the prices are falling. Moreover, the parameter that determines the news-forgetting rate changes the picture of rising and falling prices. The model also shows, that the effect of insiders activity depends on the return volume extracted, and when insiders' return approaches some crucial value, the fundamental value *v* abruptly falls down, and further with the increasing insiders' pressure the market explodes. In the second part of this work, we are considering the possibility for noise-trader to use market states recognition such as "bullish", "bearish", "cyclic" or "stable", instead of using MACD-strategy. With this aim in mind, we applied Kohonen neural net and non-supervised learning chart pattern recognition.
Keywords: market analysis, news influence, stock market, neural nets, market dynamics, simulation, insiders' role.

1 Introduction

In addition to the two types of traders in the asset market, described in [1]. *F-traders,* who could be referred to as "fundamentalists" or "information traders", and *N-traders*, in this article we are considering a small part of market participant, named "insiders". We also take into account another important factor determining the dynamics of the market: the news background. There exist several projects [2-3], the main idea of which is the assumption that news articles might give much better predictions of the stock market than analysis of price history. Thus, authors made an attempt to develop an NLP-system, which takes past news articles and stock prices data as inputs, and builds correlation links between certain features found in these articles and changes in stock prices. The new papers are classified on their influence on the stock market as "plunge" and "surge" or "bad" and "good" news. In Financial Information Grid project the "sentiment" concept is used, that is a qualitative, largely intuitive factor that may influence on the market agent decision, causing either bullish or bearish behavior. The results of this research nevertheless do not permit to analyze in dynamics cumulative interfering and impact of different factors, such as news strength, system memory, period of news accumulation and influence, and market agents strategies. Consequently, the simulation of stock market processes by mean nonlinear discrete models on computers presents powerful alternative way for studying the market, especially in proximity of bifurcation states boarder. In our paper we try to estimate factors that control asset prices in their interaction.

2 Implementing the news background in the model

The news influencing estimation is regarded as part of fundamental analysis. Fundamental analysis involves the in-depth analysis of a company's performance and profitability to determine its share price. As new information is released pertaining to the company's status, the expected return on the company's shares changes, affecting the stock price. The advantages of news analysis are its ability to predict changes before they show up on the charts [2-3]. Unfortunately, there is no models formalizing all this knowledge for purposes of automated decision making, and interpretation of news analysis results may be subjective at the high level. The most valuable feature of news analysis is that it makes possible to predict non- stationary points, as rising trend or falling trend, connected with time points, when essential news emerge.

To capture the significant contribution of news analysis to the decision making, we added the news background to our model, thus expanding the system described in article [1]. It replaced or added to the "market noise system". For each day, k random news events ξ_k are generated, whose value (or intensity) depends on the position of the "good-bad" slider, and the news sign ("bad" or "good") depends on the slider displacement from the neutral position. The news in this model has a property of being accumulated in time, but the strength of news is decaying with time. The accumulated news comprise the news

background, which may neutral, positive, or negative, depending on the sign of accumulated news. The weight of news emerged at the moment *k*, in background sum at the moment *j* is proportional to

$$Exp(-u*(j-k)),$$

where parameter *u* defines the rate of news decay. Then, the total news influence is calculated, based on the news data of all previous days being summed with their weights:

$$Q_j = \sum_{k=0}^{j} \xi_k \; Exp(-u*(j-k))$$

where ξ_k - intensity of news. Then the news background, computed in such way, is applied to modify the price fundamental value for the next day in form of logistic law:

$$v_{j+1} := v_j + h * (Exp \; Q_j - 1) / (Exp \; Q_j + 1),$$

where *h* is a constant defining the degree of the news background influence.
The noise factor reintroduced again later. Gaussian white noise with zero mean is generated, and the noise mean value is set by mean program interface window "noise". As a result of the two factors influencing the price fundamental value, the changes in the stock market variable *v* may be regarded as   random walk controlled by news background and noise:

$$v_{j+1} := v_j + (h * (Exp \; Q_j - 1) / (Exp \; Q_j + 1)) + \varepsilon_j,$$

Primarily, in building our model, we have been trying another system, in which the price fundamental value was affected noise as:

$$v_{i+1} = v_i + h * Q_i.$$

Though it showed similar results, it seemed to us that it was not economically adequate as long as traders rely on the news, and not on "noise".
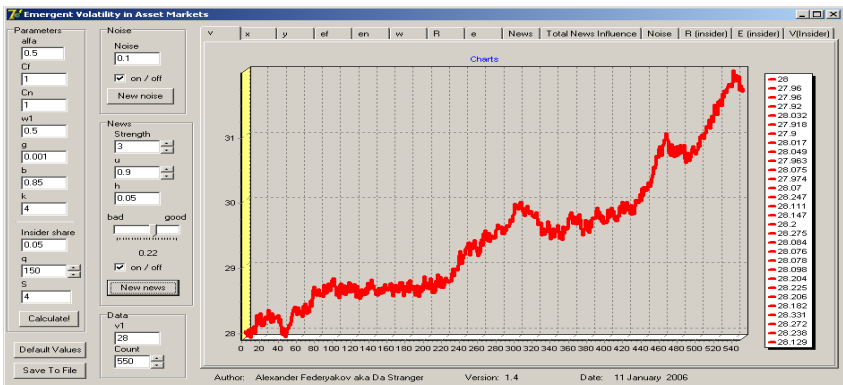
Figure 1:   Artificial price fundamental value rising with the influence of good news.

The simulation of the effect of "bad" and "good" news on the market volatility in the model, programmed on Delphi 7.0, showed that rising and falling prices take place in cases of "good" and "bad" news appearing on the market.

Moreover, the range of price changes depends on the nonnegative parameter b, which determines the market sensitivity to the aggregate excess demand e_t. Figure 1shows the typical picture of artificial course rising with the influence of good news. The lower the rate of news forgetting (long memory), the longer the range of price rising or falling, and the greater the difference between prices at start and end points of the range.

3   The insiders' model

The insiders are illegal or semi-legal participants of stock market who get information, which may have essential influence on stock market prices, before this information is officially published in mass media and, as a rule, against the existing laws that forbid its open circulation in the society. The term "insiders" reflects the fact that the specific information they possessed until the current moment was for internal use only inside a limited circle of users. As a result, these persons, getting the insiders information before its official publication, have an opportunity to make the certain moves at the stock market for obtaining superior profits, and thus destabilize the market. The "insiders" parameter was implemented into the model in the following way. Insiders have a certain share of the market in their hands (we designed it "l"). They always receive profit, the value of which can be calculated like this:

$eIns_t := q * (x_t - x_{t-l})^2$, q is a constant;

the insiders' past relative return of the two trade strategies during the past k periods is estimated as:

$$Rins = [x_t \sum_{j=t-k}^{t-1} eIns_j - \sum_{j=t-k}^{t-1} x_j eIns_j]/k .$$

If insiders receive profit, that means someone does not, so we suppose that insiders' past relative return decreases past relative return of other stock market agents:

IF $R_t > 0$ THEN,
$R_{t+1} = R_t - Rins_t$,
ELSE
$R_{t+1} := R_t + Rins_t$,

where R_t - the past relative return of the two trade strategies during the past k periods and $Rins$ – insiders' past relative return.

Now, in presence of three types of agents on the market, the aggregate excess demand of the whole market is:

$$e_t = w_t * ef_t + (1 - w_t - l) * en_t + eIns_t *l.$$

After that, we calculate the effect that the insiders might have on the price fundamental value. We will describe it here not exactly in all details, but giving a general idea of the process. We have tried 2 variants of insiders influence function on varying fundamental price value.

Variant insiders influence A:

IF($R_t + Rins_t) \neq 0$ THEN

*vins $_{t+1}$ = v$_t$ +(h * Exp(-s (1 − (Rins $_t$ /(R $_t$ + Rins $_t$)))))* (Exp(Q$_j$) − 1) / (Exp(Q$_j$) + 1)*
ELSE
*vins $_{t+1}$ = v$_t$ +(h * (Exp Q$_t$ − 1) / (Exp Q$_t$ + 1)) + ε$_t$.*

When we were applying this variant, we could check the combined influence of good news and insiders. Price fundamental value *v* varying under the influence of good news with the news intensity 0.22. By the period 540 time units, the artificial rate raises from 28 to 32. The same picture we can observe with the appearance of bad news, but with only one difference: the artificial course falls down.

Under the insiders impact, the relative rate of which between traders' community constitutes 0.05 and the multiplier value q=150, the common view of the plot remains the same, but the scale of course rising decreases and the maximum value achieved is only 28.07 in comparison with 32 without insiders.

Figure 2 shows the impact of the insiders on the fundamental price value in variant B insiders influence. It is interesting that the insiders get out the profit from the market in big steps as we can see from Figure 2.

In this case, *vins* represents the new price fundamental value, which includes the insiders' influence on the market. Using the graphical representation you can compare it with the case, what the price could have been without insiders (v$_t$).
Variant insiders influence B:

IF(R $_t$ + Rins $_t$) ≠ 0 THEN
*vins $_{t+1}$ = v$_t$ +(h * Exp(-s (Rins $_t$ /(R $_t$ + Rins $_t$))))* (Exp(Q$_j$) − 1) / (Exp(Q$_j$) + 1)*
ELSE
*vins $_{t+1}$ = v$_t$ +(h * (Exp Q$_t$ − 1) / (Exp Q$_t$ + 1)) + ε$_t$.*

In this variant, we explored the insiders' impact on the fundamental price value. In this case, it's interesting to note that, in spite of high level of good news (0.54), the fundamental price value after the little jump from 28 to 28.15 only, rises then very slowly (only to 28.24 at 250 days). The next interesting result we

Figure 2:   Insiders impact on the assets market price.

can see in the Figure 2. In spite of practically constant value of the fundamental price, the assets of market prices grow essentially from 28 to more than 30. Moreover, the plot of prices rising has a stepwise view. This is the consequence of insiders' super-profits (past relative return) which they get from the market. As it is shown in Figure 3, there are the half waves of decreasing heights of the insiders' past relative return strictly at the same moments, when prices plot (Figure 2) makes a jump to the next level.

Probably the time series chart pattern like the one represented in Figure 2 could indicate that the insiders are presenting on the market. It appears that inflation is increasing as insiders do not invest in industry development and rather spent money in consumption sector; and, as a result, the value of fundamental price does not increase while prices at stock market particularly are increasing.
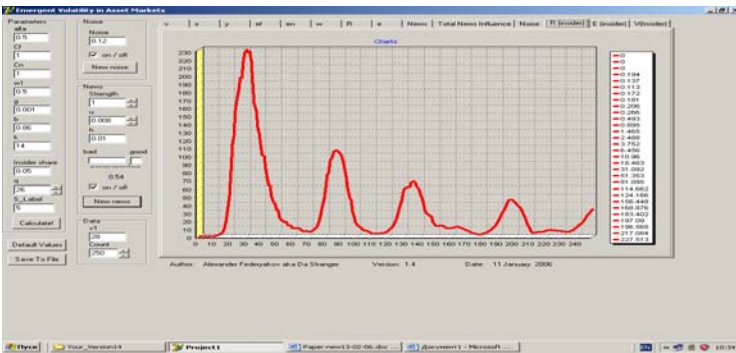
Figure 3:   Insiders past relative return.

4   Kohonen neural net exploring for possible changing N-trader strategy

In the model described, N-traders are traders of technical analysis or chartist's class, and they are sure that share prices move in trends as a consequence of the constantly changing attitudes of investors in response to different factors. As an alternative to the time series analysis, neural nets could be used. An advantage of neural nets is their capability of learning and adaptation. The neural nets work well in conditions of sharp prices jumps, noisy effects, and contradictory data; thus providing a good addition to the classical technical analysis.

In our case, a data base with variable values (mean weighted buying price, mean weighted selling price, maximum day price, minimum day price, volume) was imported to Excel media from the Central Bank of the Russian Federation site for EURO/ruble and USD/ruble rates for period from 03.10.2003 to 14.10.2005 (about 500 values). The forecast systems use successive values chosen as currency exchange rate:

$$x(t-n), x(t-n+1),... x(t-1), x(t) \rightarrow x(t+1).$$

We assume that there is a certain law that can define the value of the next observation in a time series as function of the past observation values. We suppose also that a finite number *n* of past observations is enough to determine the next one:

$$x(t+1) = f(x(t), x(t-1),... x(t-n-1),$$

The number n is fixed, and we can neglect the influence of other observations. Having made this assumption, we can speak about a "window" of a size n, which is moving along the time series. The Kohonen neural net must examine the set of values enclosed in the window as a pattern, and try to find pattern classes taxonomy for different window sizes and number of classes specified.

A matter of fact, there is no necessity to slice the input time series to parts according to the window size. It is sufficient to make a column from the input data, take a number of columns that equals the window size, and put these columns together, each of them moved upwards at one unit relatively preceding one. Then each row will represent parts of original time series sliced right at one unit. In this way, we can simulate the window, slicing upon time series and Kohonen net will "see" one after another pieces and try to classify it.
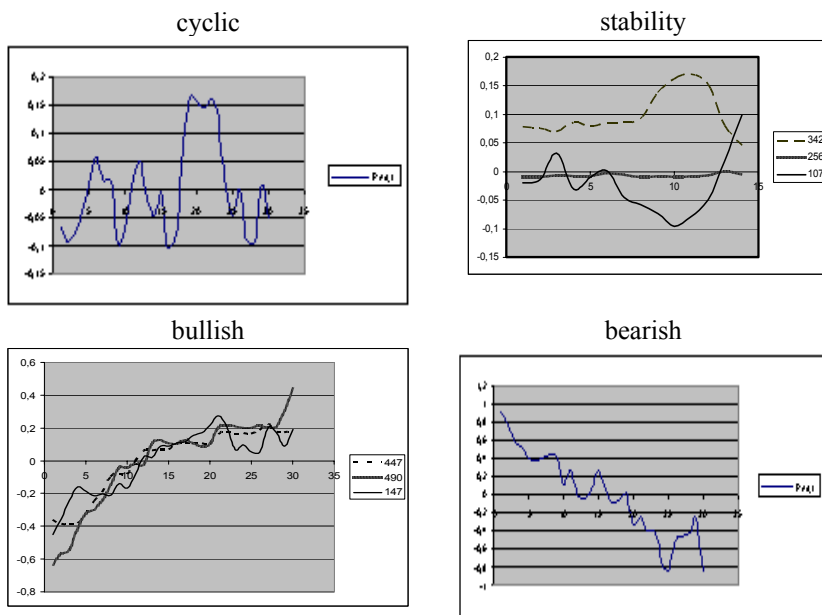
Figure 4:   Some pattern classes discovered by Kohonen network.

We use information on different currencies, and its quotations differ from each other very much. We create time series using the original data set (the set of closing prices), and then the average value of time series is calculated. As a

result, the new time series are obtained after the subtraction of the average from each of the series values. It is possible therefore to avoid the classification based on the level of the prices. Transformed data is imported to *ST Neural Networks.*

In our work, we have explored 6×6, 5×5, 4×4 and 3×3 Kohonen network, e.g. the network should recognize from 36 to 9 different classes and the window sizes 36, 28, 21, 14 and 7 days.

In Figure 4, the class discovered by Kohonen network is presented. The class pattern recognition is shown in Figure 5, where the members of the class recognized by the Kohonen net in the time series of EURO/ruble exchange rate are indicated by the arrows.
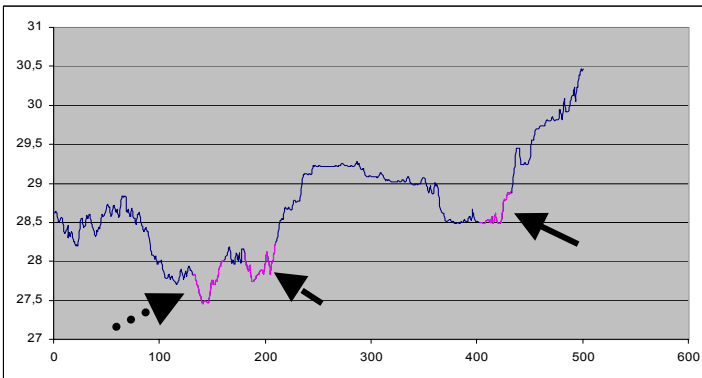
Figure 5: The parts of reversed time series USD/ruble where the arrows indicate that trough was discovered.

The Kohonen networks attempts to find the structure of the data and classes to which data belong. The Kohonen networks can learn to recognize clusters of data, and can also relate chart patterns to each other as well as to clusters. As classes of data are formed, they can be labeled, so that the network achieves capability of classification of new data.

Initially starting with randomly chosen centers of sets, the algorithm gradually adjusts them to reflect the clustering of the training data. The iterative training procedure also arranges the network so that units representing the chart pattern, which are similar in the input space, are situated close together on the topological map. We then say that the KNN has learned. After many epochs, when the KNN has been trained, the weights are frozen, and the test set data are presented to the KNN to examine how well it is able to classify it using the weight values frozen after the training of the KNN.

Our model in some conditions and parameters values has deterministic chaos properties. The R/S-analysis was used for evaluating of persistence features of time series [4]. Our program allows to lead the R/S-analysis of time series and to receive an estimation of Hurst exponent value H.

Figure 6 shows the results of the R/S-analysis and time series for euro at the rate of the Central Bank of the Russian Federation for the period of October 3rd, 2003 through October 14th, 2005. The value of variable H has turned out to be in

the range of 0, 75-0, 8. Calculations were carried out according to the equations published in the article by Skiena [7] As the obtained value of Hurst's exhibitors is concluded in an interval 0,5 <h ≤ 1, it means that the processed time series is persistent and is characterized by the effect of long-term memory.
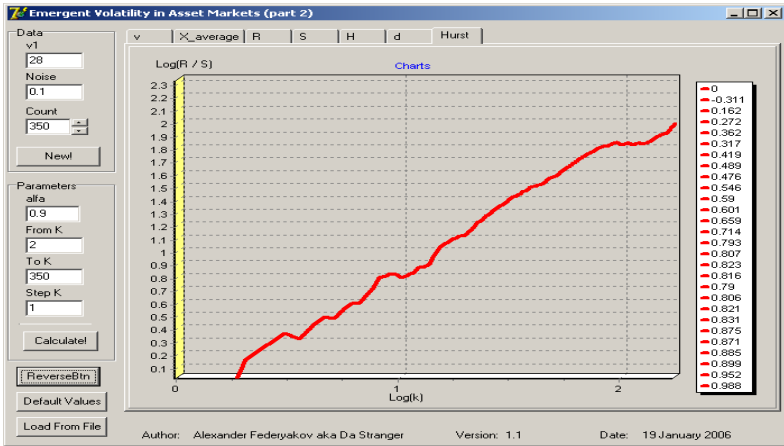

Figure 6:   R/S-analysis for euro at the rate of the Central Bank of the Russian Federation (3.10.2003–14.10.2005).

The learning algorithm uses a time-decaying learning rate, which determines increment of neuron weight adjustment and ensures that the weights alterations become more subtle as the epochs pass. Kohonen net uses topological ordering property, which means that each output neuron has its neighborhood. For getting the best training results, the training process as a rule is conducted in two distinct phases: a relatively short phase with high learning rates and a neighborhood, and a long phase with low learning rate. In our case, the training is conducted in 3 phases, in each phase the number of epochs (10000; 5000; 2000), the learning rate (0.1; 0.05; 0,01) and neighborhood (1;0;0) are different. After finishing the network learning, the network may be used for detecting chart pattern in time series.

When training cases were formed, we examine them to see which meaning can be assigned to the clusters in terms of "rise", "fall", "cycle", "stability" and other possible classes. Once the topological map has been built up in this way, new samples from another time series could be submitted to the network for classification.

5   Conclusions

The traders' decision must be based on more precise model that we have today, and the future prices forecast must be based on the results of using technical and fundamental analysis. The aim of this article was to research the conditions of traders' activity on the market and to evaluate the effect of different factors.

The simulation of "bad" and "good" news effect on the market volatility showed that rising and falling of prices takes place in cases "good" and "bad" news appearing on the market. Moreover, the range of price changing depends on the nonnegative parameter b, which determines the market sensitivity to the aggregate excess demand e_t.

The combined influence of good news and insiders in our model shows that, under the impact of insiders, the common profile of rising of the plot course caused by the "good" news remains the same, but the scale of the course rising decreases. It is interesting that the insiders take the profit from the market in big steps.

In one of the model variants, we show that the past insiders' relative return has the profile of periodic peaks corresponding to the stepwise increase in market prices. At the same time, the fundamental price value retains approximately constant.

The experiments with Kohonen network demonstrate that there is a possibility of Kohonen Network learning to recognize the chart pattern specific for time series at hand. The N-trader then can make the decision based on the results of program detecting one or another class pattern. The program recognizes the part of time series visible in the "window" as a member of one of the classes that were formed by Kohonen Network.

References

[1] Honggang Li, J. Barkley Rosser, Jr. *Emergent volatility in asset markets with heterogeneous agents. Discrete Dynamics in Nature and Society*, vol. 6, no. 3, pp. 171-180, 2001, (www.hindawi.com/journals/ ddns/volume-6/S1026022601000188.html)

[2] K. Ahmad, T. Taskaya-Temizel, D. Cheng, L. Gillam, S. Ahmad, H. Traboulsi, J. Nankervis, *Financial Information Grid –an ESRC e-Social Science Pilot.* Dept. of Computing, University of Surrey, Dept. of Economics, University of Essex.

[3] http://www.ncess.ac.uk/events/past/2004-09-ahm/2004-09-FinGrid-paper.pdf

[4] http://www.allhands.org.uk/2004/proceedings/papers/144.pdf, 2004

[5] Hugo van Bunningen, *Master of Science Thesis. Augmented Trading: From news articles to stock price predictions using syntactic analysis,* University of Twente, (http://www.cyc.com/doc/white_papers /Arthur_Thesis.pdf, 2004)

[6] Edgar E. Peters, *Chaos and Order in the Capital Markets: A New View of Cycles, Prices, and Market Volatility, 2nd Edition,* Wiley finance editions, pp. 288, 1996.

[7] Steven S. Skiena "*CSE 691 - Lecture 12 - Fractal Analysis*" http://www.cs.sunysb.edu/~skiena/691/lectures/lecture12/lecture12.html

T-outlier and a novel dimensionality reduction framework for high dimensional financial time series

D. Wang[1], P. J. Fortier[2], H. E. Michel[2] & T. Mitsa[2]
[1]*State Street Corporation, U.S.A.*
[2]*Department of Electrical and Computer Engineering,*
University of Massachusetts Dartmouth, U.S.A.

Abstract

Outlier trading strategy is one of the more popular strategies in the financial markets. In reality, the outlier trading strategy tends to provide a higher return and is accompanied by a respective higher risk. Even so, more fund managers and investment companies turn to outlier trading strategies to maximize their return. The key is to minimize the risk by finding the most accurate and relevant outliers.

The framework presents a T-outlier concept in high dimensional time series data space. The framework generalizes the characteristics of time series outliers and proposes a novel dimensionality reduction framework in high dimensional time series data space of financial markets to support further mining. The framework consists of horizontal dimensionality reduction and vertical dimensionality reduction algorithms. On the vertical dimension, an Attribute Selection algorithm reduces the vertical dimension by identifying significant dimensions among a given set of attributes that closely represent the data characteristics. On the horizontal dimension, a Segmentation and Pruning algorithm reduces the selected attributes' horizontal data points to concisely represent the data object in piecewise linear representation form. This paper discloses a series of experiment results that illustrate the effectiveness of the framework.

Keywords: T-outlier, outlier, data mining, computational finance, dimensionality reduction, piecewise linear representation, high dimension.

1 Introduction

An outlier is "…an observation that deviates so much from other observations as to arouse suspicion that it was generated from a different mechanism."[1]. Outlier analysis strives to detect data objects whose characteristics are dissimilar from the rest of the objects in a given data collection. Outlier analysis can be categorized into distribution based [2], depth [3] based, and distance based approaches [4, 5]. Recently, additional outlier detection algorithms have been proposed for high dimensional data [6], spatial outlier detection [7], and OLAP based outlier data mining [8].

In this paper, we propose a definition of T-outliers (outliers in time-series data) and a framework to detect outliers efficiently in high dimensional space. The framework is composed of an Attributes Selection (AS) algorithm and a Piecewise Linear Representation (PLR) based a Segmentation and Pruning (SP) algorithms. The Attributes Selection algorithm utilizes the correlation coefficient and other statistics techniques to find significant attributes by applying a unique attribute weighting procedure. The identified significant attributes represent most of the data characteristics of a given data set. The Attribute Selection algorithm processes the vertical dimension reduction technique. Horizontally, we use the PLR based Segmentation and Pruning algorithms to concisely and precisely identify significant attributes' time dimensional data points – changing points.

The rest of the paper is organized as follows: Section 2 presents the problem statement to be addressed and summarizes the major contribution of this work. The background and related work is reviewed in Section 3. Section 4 describes the generalized T-outlier concept, the business similarity from an industry perspective using two definitions: changing point and object definition. Section 5 presents the data transformation algorithms including the Attribute Selection and PLR based Segmentation and Pruning algorithm. The experiments and result analysis are presented in Section 6, followed by Section 7 containing the conclusion and future work discussions.

2 Problem statement and contributions

Outlier trading strategy is a popular strategy in financial markets and tends to provide a higher return though with a respective higher risk. The key is to minimize the risk by finding the most accurate and relevant outliers.

Sophisticated tools are needed to help investors filter out outliers from hundreds or even thousands of stocks, based on their criteria. The goal is to develop a method that not only improves the efficiency in identifying outliers, but also provides a very high level of accuracy, thereby reducing the risk component in investment decision making processes.

The major contributions of the research presented are as following.

❖ We propose a general dimensionality reduction framework for time series data based on a data object transformation from high dimensional space to low dimensional space.

❖ Secondly, we propose novel horizontal and vertical dimension reduction algorithms that maintain significant data characteristics and concisely represent the data in the time axis.
❖ The resultant dimensionality reduction framework in high dimensional space can also be applied to other industries' data such as medical data, weather data, credit card usage data, and so on.

3   Background and related work

3.1   General outlier research

The outlier notion is formalized in a classic distance based approach by Knorr and Ng [4, 5]: In general, outlier detection research are classified into two categories. The first is distribution based, where a standard distribution is used to find the best fit with outliers defined thereafter. A major drawback of this category is that except for a handful of tests, most of the distributions used are single variant which does not match the reality of most problems.

The second category of outlier studies is referred to as depth based. Each data object is represented as a point in a k-dimension space, and is assigned a depth. Outliers are more likely to be data objects with smaller depths. In theory, depth-based approaches work for large values of k. However, depth-based approaches generally become inefficient for large datasets with $k \geq 4$ in Breunig [3], though there exist efficient algorithms for k = 2 or 3 in practice. Their notion generalizes many concepts from the distribution-based approaches, and enjoys better computational complexity than the depth-based approaches for large values of k. Recently, additional outlier concepts have been proposed such as high dimensional outlier [6], spatial outlier [7], and OLAP based outlier [8].

3.2   Dimensionality reduction research

The most promising data mining solution techniques involve dimensionality reduction with multidimensional index structures. Many dimensionality reduction techniques have been developed, including Singular Value Decomposition (SVD), Discrete Fourier transform (DFT), Discrete Wavelet Transform (DWT), and Adaptive Piecewise Constant Approximation (APCA). These are good horizontal dimensionality reduction techniques but can not solely be used for high dimensional data objects such as stock equities. The framework we presented is based on a Piecewise Linear Representation (PLR) of horizontal data with a vertical dimensionality selection algorithm to represent data objects. It produces satisfactory results in our experiments as is shown in section 7.

4   T-outlier concept

4.1   PLR based changing point

In this research, objects are assumed to move in a piecewise linear manner. Specifically, we assume that time is the horizontal continuous dimension and is

one dimension in the total data space. We extend the definition presented in Wu et al [9] to define the changing points including upper changing point and lower changing point which are symmetric. The lower changing point is defined here.

Definition 4.1: **Changing Point Definition**

Given the user specified time window h and current time t_c, a lower change point in Price dimension $P_j(\overrightarrow{D_{ij}}, t_j)$ at time j within user specified time period h, where i is the ith attribute, should satisfy the following:

❖ $D_{price-j}$ = MIN(D_{price} value of current sliding window), where the sliding window can be various sizes which specify m as the maximum number of points contained after the last identified end point and m is given as a user specified parameter.

❖ $P_j(\overrightarrow{D_{ij}}, t_j)$ is the last point satisfying the above two properties.

❖ $\overrightarrow{D_{ij}}$ = (D_{1j}, D_{2j}, $D_{3j,}$ $_{...}D_{nj}$), where n is the total number of selected dimensions.

4.2 T-outlier concept

T-outliers consist of a subset of data that differs from a given super data set and exhibits four major features: periodicity, trend, predictability, and continuity. Continuity is the major differentiating factor when compared with other data mining fields, and also differentiate from streaming data.

We illustrate the T-outliers concept within the following scenarios:

First, consider only one object (stock object) in a two dimensional space (Time is the first dimension, the other dimension is Price). There are generally four types of unusual movements: isolated outliers, level shifts, slope change, and changes in frequency. An isolated outlier could be due to a sudden temporally localized change. A level shift represents a change in value that endures for at least a while and represents a new local mean value for the temporal data. A slope change can be unusual if the previous slope such as a high slope either decreases or increases. Changes in frequency means the frequency of change in the time series data. We believe all the above types apply to T-outliers in two dimensional spaces. We call these types of outliers: horizontal outliers because we only need consider the historical data of a particular object. Much research exist for horizontal T-outliers: such as burst in Zhu and Shasha [10], event in Guralnik and Srivastava [11], shocks in Jong and Penzer [12], or changing point in Yamanishi and Takeuchi [13].

Second, consider many objects (many stock objects) within two dimensional spaces. We define the outlier object as the object whose properties are dissimilar from remaining objects by comparing the recent period. The length of period is a user defined parameter h. The object property is denoted as $((t, \overrightarrow{D_1}, \overrightarrow{T_1}), t')$. In Figure 1:(c) we present only one segment to simplify the illustration. There are

two data points at t and t'. $\vec{D_1}$ and $\vec{D_2}$ are the normalized Price attribute value of the current point at time t' and the last end point of the price dimension at time t. $\vec{T_1}$ refers to the normalized degree of change associate with the point at time t. We call this type of time series outliers a two dimensional vertical T-outlier.

Third, consider many data objects within high dimensional spaces. The object property is denoted as $((t_1, \vec{D_1}, \vec{T_1}), (t_2, \vec{D_2}, \vec{T_2})\ldots(t_n, \vec{D_n}, \vec{T_n}), t_{0,} t_c)$, where $\vec{D_1}$ is a vector which includes the normalized attribute value in the Price dimension and normalized values for other attributes. $\vec{T_1}$ is the normalized degree of change for the Price dimension. (T`-T) is the time difference which equals to the last end point time and current time. As shown in Figure 1:(d), there are multiple additional attributes associated with time t and t'. We should take all of these attributes into account when performing a similarity search to find the outliers. Figure 1, only shows one segment with two changing data points for simplicity with each changing point having three different attributes. Only one attribute has trend and the other two attributes only take the moving average value at a time stamp. There could be many segments and changing data points. We call this type of outlier as high dimensional vertical T-outliers. In this study, we explore this high dimensional vertical T-outlier.

(a) and (b) two dimensional and one object space T-outlier; (c) two dimensional and many objects space T-Oulier; (d) high dimensional and many objects space T-Outlier
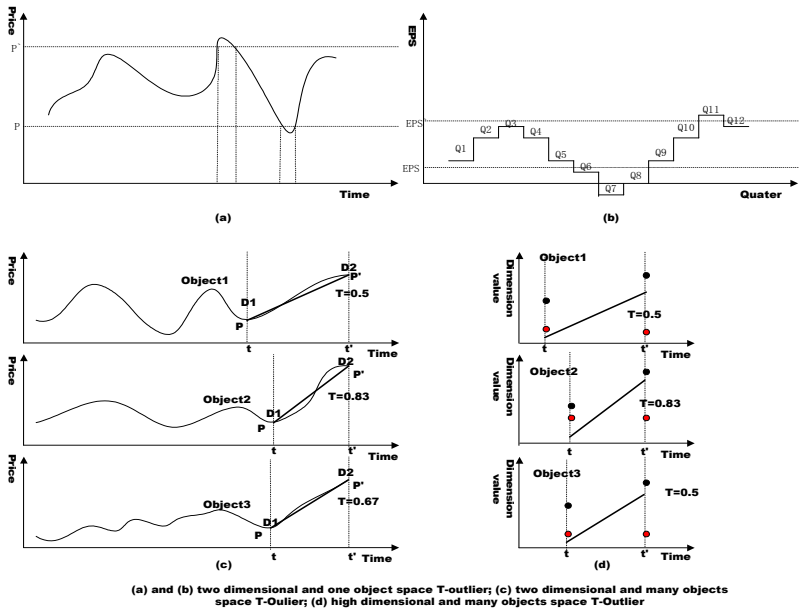
Figure 1:   Sample T-outliers.

5 Dimensionality reduction framework

In this section, we will illustrate our vertical and horizontal dimensionality reduction algorithms.

5.1 Attribute Selection algorithm

It is important to know that the rational behind the algorithm is based on the correlation coefficient theory. Higher correlation coefficient value means higher co-dependence. Higher co-dependence means higher contribution to the related attribute. A high number of set co-dependence attributes however is problematic. We need to eliminate some attributes by performing further analysis.

Based on the rational above, we select the significant attributes in this study first and then filter out attributes by examining the co-dependence. Typically, the number of attributes ranges from 4 to 8 in financial market analysis for most of the research problems. We will try different numbers of attribute candidates in our experiments.

In our Attribute Selection algorithm we require the maintenance of a correlation coefficient matrix for the computation. For each stock object, we calculate the correlation coefficient value of all attributes against Price attribute (plast) and record the value in the matrix. While computing this, we perform pruning according to given conditions. Some stock objects are pruned because of its attribute(s) have 5 consecutive business days values equal to Null, or empty, or ∞. After completion of all computations for every stock object, we sort the values in ascending order. We then read the top 20 fields across all the stock objects, which are a matrix composed of only 20 rows and 500 columns. Finally, we select the attributes that represent 80% of occurrences in the new matrix. In addition, we use an order weighted concept to distinguish the contribution of the attributes having a high correlation coefficient value across different stock objects.

5.2 Segmentation and Pruning algorithm

Our PLR based Segmentation and Pruning algorithm with the selected attributes values is based on a user specified time period *h*. The algorithm can be separated into two parts. The first performs segmentation for each valid stock. The segmentation is based on solely b%. A b% time series is calculated before this algorithm in the Attribute Selection algorithm. Most of the b% values range from 0 to 1. Only in an occasional scenario the data varies slightly away from 0 or 1. The algorithm starts with the assumption that the first segment is upward. As differentiate from Wu et al [9], our algorithm can handle continuously up or down scenarios. Once the segments are identified as b% series, the time stamps of the changing points are mapped to the Price (plast) series to produce the segments in the Price series. The pruning process is performed in the second part. There are multiple steps of pruning.

First we prune on b% based on the given threshold. We proved in our experiments that a b% pruning threshold of 10~12.5% works for most of the

stock objects and yields adequate segments. The b% pruning process removes the changing points that do not qualify for the threshold and the time stamps associate with the changing points. The result is then mapped according to the new time stamps in Price series. If there is no need to perform b% pruning, the algorithm will prune by Price based on the Price pruning threshold. We proved that a Price (plast) pruning threshold of 10–15% works well for most of the stock objects. The result of the Segmentation and Pruning algorithm produce a PLR form of object which concisely and precisely represents the original Price series.

6   Business implications

6.1 Bollinger Bands and Bollinger Band Percent b%

Bollinger Bands are widely used indicators for American financial markets. The Bollinger bands measure relative definitions of the high and low values for financial time series for the Price dimension. The Bollinger bands are three curves drawn above and below a moving average using a measure of standard derivation. Bollinger Band Percent (b%) is another popular indicator which indicates the current state within the Bollinger bands. We perform segmentation using Bollinger bands in our Segmentation and Pruning algorithm. The reasons are: the b% has a smoothed moving trend similar to the Price movement; it represents a normalized value of the real price (between -1 and 2); and is very sensitive to the price change.

6.2 Attribute selection and construction

The selection of the initial attribute set was conducted with inputs from American financial domain experts and scholars from the financial markets. We selected three categories of attributes: market attributes, financial attributes, and environmental attributes. Besides the attributes we can directly obtain from the market, others were created to better measure stock performance. It is considered best practice in the financial domain to construct composite attributes to better measure stock from different perspectives. The market attributes are the variables gained in the stock market such as daily closing price, daily highest price, daily lowest price, and daily trading volume. In addition, we believe some variable manipulation is important to reflect characteristics of the market, i.e. 5 days average price which measures the relative price stability over a short period of time. The financial variables refer to the variables found in the financial balance sheet, income statement, and cash statements such as sales, revenue, earnings per quarter etc. The environment variables include federal bank interest rates, government behaviors, consumer market indicator, GDP, etc. We believe the Price attribute movement is highly correlated to the selected attributes.

6.3 Business aspect of the framework

It is a common practice to find a set of significant independent attributes for a given dependent attribute in the American markets. For instance, find a set of important attributes for implied volatility to support the option pricing model. The common ways are linear regression, multi-linear regression, and correlation coefficient analysis but is usually limited to one object with many attributes. The object here can be equity, option, and bond etc. It is a special challenge to find common significant attributes across objects. In this research, we present a unique attribute selection algorithm to identify common significant attributes across a pool of equities. In big financial institutions, there are terabytes of data sets for financial instruments with hundreds of attributes with over 20 years of historical daily (even intraday) data. The computation of such instruments becomes a nightmare for financial analysis which further supports other financial models in many cases in reality. We observed that there is no need to compute each data point of the object in some cases like outlier-ness calculation between data objects. We presented a novel PLR based Segmentation and Pruning algorithm to reduce the data points in time axis without loosing major data characteristics but with significant performance gain.

In addition, the novel framework presented is not specific to the U.S. equity market but represents a general framework to perform dimensionality reduction in high dimensional time series data. The framework can be applied to U.S. non-equity markets data, international financial markets data, weather time series data, and medical time series data etc. as long as the following properties are satisfied. First, the data object of the problem has a high number of attributes with long historical data in time series formats; second, part of the problem is to find the common significant attributes across objects; third, computational performance is important while not sacrificing result accuracy.

7 Experiment and result analysis

7.1 Raw data and attribute set

The Standard & Poors 500 Composite (S&P 500) stocks from BLOOMBERG© were selected as our raw data. We obtained 15 years of historical data with 95 attributes ranging from the pricing attributes to fundamental financial attributes. In order to adapt the environmental effect, also collected were 11 key economic indicators for the United States over those 15 years from DATASTREAM.

7.2 Experiment environment

The Windows XP platform is used on a machine with 1 GB memory, and a 1.4 GB Intel® Pentium processor. The program is implemented in the Java language with a set of property files.

7.3 Perform Attribute Selection and result analysis

First, cleaned data was generated by changing different values of moving average parameters whose value include 5 days, 10 days, 20 days, 30 days, 40 days, 50 days and 60 days respectively.

Second, for each different number of days scenario, we ran our Attribute Selection algorithm using the correlation coefficient calculation. Then selected the top fields (5~20 user specified) with the highest correlation value for each stock in each scenario. This produces a new matrix with 500 columns and 5~20 rows. The column represents the individual stock and the row represents the attributes for all stocks. We count the occurrence of each attribute in the new matrix and then sort the attributes in ascending order in terms of occurrence value. The algorithm then calculates the top attributes representing 80% of the total occurrences.

In order to analyze parameter significance, we used a PARETO chart. From analysis of the charts, we observed that the slope of the increasing part of the chart is very steep implying that most of the data characteristics are represented by the beginning attributes in the NAME axis. In testing, most of data characteristics could be represented by the top 5–10 attributes occurring in the matrix based on the observation.

Third, we calculate the percentage of occurrence for each attribute in the matrix by taking the occurrence of the attribute as nominator and the total cells of the matrix as denominator. We select the top attributes that represent user specified significant value of the data characteristics. We perform a business attribute analysis afterwards since the data object represents individual stocks in real financial markets and is part of standard practice within the financial industry. Based on both the scientific and business attribute analysis, the following attributes are selected: P2bk, P2sl, and P2ebita are the market data to financial data ratio attributes which reflect some fundamental financial data of the company. P2bk is the price to book value. It presents the current trading price to the company's equity value in the book. P2sl is price to company quarterly sale revenue which reflects company revenue ability. P2ebita is the price to earning before income tax amortization which reflects the profitability ability. pxcncel is related to number of trading get cancelled which is indirectly related to trading volume. plow, phigh, popen, and lsttrd are market trading related attributes which represent the stock market characteristics of the stock. ewapx is the monthly equity weighted average trading prices and dpavg is a constructed attribute which represents the past 20 business days moving average of the price.

7.4 Piecewise Linear Representation via Segmentation and Pruning algorithm

Cleaned data was then processed by our PLR segmentation and pruning algorithm using different periods of days (5–60) to find out which values best simulate the original data. In our experiment, we found that a 20 day period moving average yields good results. The segmentation pruning threshold (segbpcntThreshold) over a b% value of 10% works best for all scenarios. Most

of the b% values fall between 0 and 1. Therefore we can use standard values such as 10% as the threshold. We varied segmentation and pruning thresholds (segplastThreshold) from 0%, 2%, 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, to 22.5%. The experimental results show that 12.5% and15% yields better results.

The example of STT (State Street Corp.)for the period from January 3, 2000 to December 31, 2003, Figures 2–5 shows a detailed comparison. As we can see from the figures intuitively, Figure 3: precisely represents the raw data object from Figure 2 without losing any major data characteristics. Figure 3: is the result of performing segmentation on b% series and mapping the time stamps that identified b% back to the Price dimension. The logic is explained in the Segmentation and Pruning algorithm. Also, the data in Figure 3: produced an appropriate zigzag shape according to our theory.

Since the b% values fall between 0 and 1, we use a uniform threshold for pruning our object based on the b%. A b% of 10% is proven to be suitable for most of the stock objects [9]. We prove here through our experiments that this is true. As we can see in Figure 4 and 5, the data gets further reduced concisely, without losing any major characteristics.

Based on the observations from experiments we can conclude two important rules. First rule, the degree of simulation accuracy depends on the plast pruning threshold. The smaller the threshold the closer simulation though it yields more segments. More segments adds more complexity to the HAC clustering and yields fewer similar objects but with higher similarity. The second rule, the user specified time period *h* is balanced with the plast pruning threshold. The shorter period of h requires a smaller plast pruning threshold to produce a reasonable amount of segments. The longer period of h requires a larger plast pruning threshold to produce segments. The algorithm itself is flexible to the user.
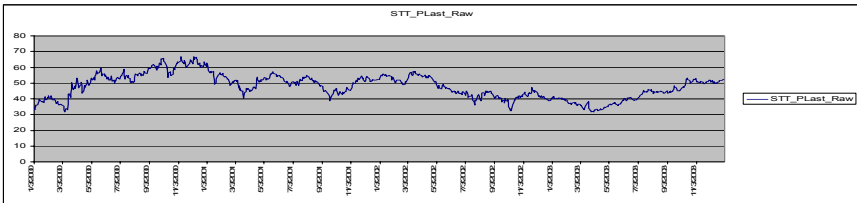
Figure 2:   The raw plast data stream of STT.
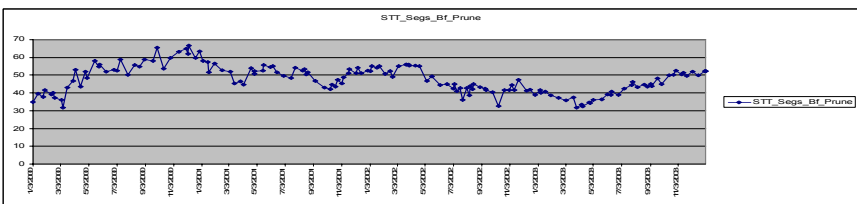
Figure 3:   Segmentation results of raw data stream without any pruning based on sliding window size 10 and average moving period 20 days.
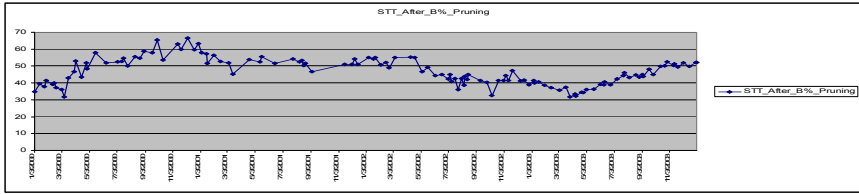
Figure 4:   The result after perform b% pruning before performing plast pruning(b% = 10%).
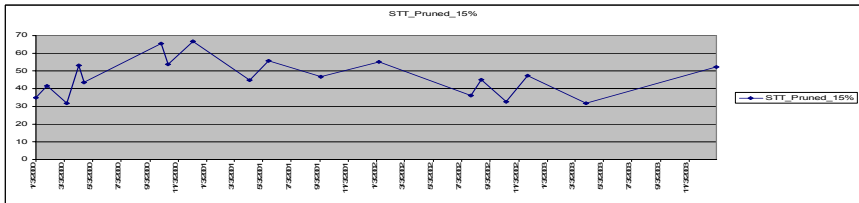

Figure 5:   The pruning threshold of plast is 15%.

8   Conclusion and future work

In this study, we introduced a new T-outlier concept and a novel dimensionality reduction framework based on a Segmentation and Pruning algorithm over high dimensional financial data along with an Attribute Selection algorithm. Experiments show that vertical and horizontal dimension reduction algorithms are able to present the objects in concise and accurate form for further mining in the future.

The experiments illustrate that our algorithms are able to reduce the dimensionality efficiently within a large set of high dimensional data and could be used as foundation of developing more sophisticated tools.

This research focused on static time series data. Extending this research to include moving data, may improve predictive capabilities. The second extension looks to perform clustering and classification to find out regularities among large set of data objects based on data representation transformed by this framework. Additionally, we can perform prediction analysis based data object representation by this framework. Lastly, the framework itself can also be extended to other time series applications such as weather outlier detection, POS transaction outlier detection, network intrusion outlier detection and so on.

References

[1]   Hawkins, D.M., Identification of outliers. Chapman and Hall, London, 1980.

[2]   Barnett, V., and Lewis, T., Outliers in Statistical Data (Wiley Series in Probability & Mathematical Statistics). John Wiley and Sons Ltd. February, 1994.

[3]   Breunig, M.M., Kriegel, H.P., Ng, R.T., and Sander, J., LOF: identifying density-based local outliers, The 2000 ACM SIGMOD international conference on Management of data, Pages: 93 - 104, 2000.

[4]   Knorr, E.M., and Ng, R.T., Algorithms for Mining Distance-Based Outliers in Large Datasets, The 24th International Conference on Very Large Data Bases, Pages: 392 - 403, 1998.

[5]   Knorr, E.M., and Ng, R.T., Finding Intensional Knowledge of Distance-Based Outliers. The VLDB Journal, Pages: 211-222, 1999.

[6]   Aggarwal, C.C., and Yu, P.S., Outlier Detection from High Dimensional Data, The 2001 ACM SIGMOD international conference on Management of data, Pages: 37 – 46, 2001.

[7]   Shekhar, S., Lu, C. T., and Zhang, P., A Unified Approach to Spatial Outliers Detection. University of Minnesota - Computer Science and Engineering Technical Report Abstract, December 10, 2001.

[8]   Lin, S., and Brown, D.E., Outlier-based Data Association: Combining OLAP and Data Mining, Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, http://www.sys.virginia.edu/techreps/2002/sie-020011.pdf.

[9]   Wu, H., Salzberg, B., and Zhang, D., Online Event-driven Subsequence Matching over Financial Data Streams, ACM SIGMOD international conference on Management of data, Pages: 23 - 34, 2004.

[10]   Zhu, Y., and Shasha, D., Efficient elastic burst detection in data streams, The ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages: 336 - 345, 2003.

[11]   Guralnik, V., and Srivastava, J., Event Detection from time series data, The fifth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages: 33 - 42, 1999.

[12]   Jong, P.D., and Penzer, J., Diagnosing Shocks in Time Series. Journal of the American Statistical Association, 93, Pages: 796 - 806, 1998

[13]   Yamanishi, K., and Takeuchi, J., A unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time series Data. The eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages: 676 - 681, 2002.

Section 6
Portfolio management
and asset allocation

This page intentionally left blank

Integrating elements in an *i*-DSS for portfolio management in the Mexican market

M. A. Osorio[1], A. Sánchez[2] & M. A. Gómez[3]
*[1,2]School of Computer Sciences, Universidad Autónoma de Puebla,
Ciudad Universitaria, Puebla, México*
*[3]Actuarial Sciences and Statistics Department,
Universidad de las Américas Puebla, Cholula, México*

Abstract

We present the main elements to design an intelligent decision support system (*i*-DSS) for portfolio management in the Mexican market, including the financial investment considerations for a knowledge database and the requirements for a portfolio optimization model. We take into account the tax impact and the Mexican market conditions and volatility into the models; and modify the classic Markowitz model, augmenting constraints to propose stochastic linear and stochastic quadratic models. The extended Markowitz model is solved with an efficient adapted Evolutionary Simulated Annealing algorithm. The efficient frontier is calculated using the augmented model. We present a proposal for the integration of an *i*-DSS (intelligent Decision Support Systems) with the knowledge and preference system, the optimization model and the Mexican market characteristics.
Keywords: intelligent decision support systems, portfolio optimization.

1   Introduction

Intelligent Decision Support Systems (*i*-DSS), as successors of Management Information Systems (MIS), traditionally follow the decision logic line of thinking and include in MIS algorithmic tools to improve the choice activity of decision makers. This includes optimization methods, mathematical programming, multi criteria models etc. They are "structure related", normally assume that the decision problem can be formulated mathematically and do not stress information processing and display. "Knowledge" is represented in many

different ways, such as, frames, semantic nets, rules etc. Knowledge is processed in inference machines, which normally perform symbol processing, i.e., truth values of antecedents, conclusions etc. Some experts consider ES as part of *i*-DSS [1]; others see *i*-DSS and ES as basically different systems and others combine the two approaches into "Knowledge-based DSS" [2].

In this paper, we approach the design of decision support systems for portfolio management in Mexican market. In section 2, we describe the capabilities and potentialities of intelligent decision support systems. The Mexican market is described in section 3, including definitions, history, etc. In section 4 we describe the elements for a portfolio optimization model, taking into account the Mexican market conditions and volatility into the constraints in the model. Section 5 describes in detail an evolutionary simulated annealing algorithm that proved to be efficient in solving the Markowitz's models. The conclusions in section 6, present an integration of the proposed stochastic linear and stochastic quadratic models in an *i*-DSS framework.
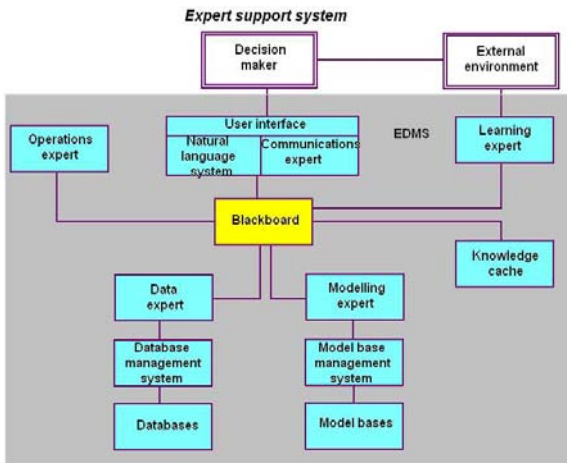
Figure 1:   Expert support system [3].

2   Intelligent decision support systems

The idea of making "intelligent" DSS (*i*-DSS) is supported by most major researchers in the field. The resulting system has been referred by different authors. It has been called an *i*-DSS, and intelligent support system, and expert *i*-DSS (EDSS), and a know-ledge-based *i*-DSS (*i*-KBDSS). It has been suggested that an *i*-DSS could support higher levels of decision making if each of its three subsystems is enhanced by AI. For example, knowledge representation techniques could be used to enhance the data subsystem; a natural language interface could serve as a front-end to the dialogue subsystem; and reasoning rules could strengthen the model subsystem. The main forms of integration are:

- ES integration intro various *i*-DSS components.
- ES integration as a separate component of the *i*-DSS.
- *i*-DSS models and a data access as a component of an ES.

In some cases, there are other problems with *i*-DSS-ES integration; integrating a general problem support system with a specific narrow problem can be a formidable task. It is not clear that the expertise in ES is always transferable across problem domains as an *i*-DSS should support. Furthermore, the time and cost involved in the design and development of such an interesting system may be difficult to justify. A new form of integration is necessary to deal with the preceding problems. The new model of *i*-DSS-Es integration is called an expert support system (ESS) [3] (see Fig. 1).

3 Mexican market

In the early 1990's the Mexican authorities (officially) lifted capital controls on foreign exchange transactions. Other financial market reforms came about at the same time. However, institutional participants remained subject to tight investment regimes precluding them from adding foreign assets to their portfolios. Since 2003 mutual funds can invest foreign stocks, bonds and other mutual funds. Since mid 2003, the Mexican's Securities Market: BMV (*Bolsa Mexicana de Valores*) established a UTP platform (Unlisted Trading Privileges for foreign stocks, equity index ETFs (Exchange Traded Funds) and fixed income ETFs. Since early 2004, pension funds can operate equity, foreign exchange and fixed income derivatives. In March 2004, Mexico started with an options market and in November 2004, pension funds started investing in domestic equities and foreign stocks and bonds. Savings through SIEFORES reaches 55% of *traditional* savings.

Table 1:   Capital controls and financial regulation.

Period	Controls on Public	Bank	Institutional Investors	Mutual Funds
1976-1982	No	No	No	N.A.
1982-1991	Yes	Yes	Yes	Yes
1991-2002	No	Yes	Yes	Yes
2002-2003	No	Yes	Yes	No
2004	No	Yes	No	No

Even if the impact on Mexico's economic growth has not been properly measured, according to other experiences, these changes can contribute to the development of Mexico's capital market, even if some people think that this liberalization may make Mexico more vulnerable to changes in domestic and external macroeconomic conditions. Based on other experiences, adding equities and foreign instruments to the set of eligible assets will induce better rates on Mexican Treasures and corporate bond yields will tend to go up (Reynoso [7]).

The trend to lift capital controls has the purpose to develop the local intermediation industry, according to experiences that say that cross listings in domestic and foreign markets adds to liquidity and market development, lifting

controls on local investors can positively influence national savings, productivity and growth and may be good for the economy. The global market for the BMV has already 200 stocks and ETFs and includes: 31 stocks of the DJI index, 48 stocks of the DJ STOXX50 Index, 20 NASDAQ listed stocks, 78 NYSE and AMEX listed stocks, 5 TSX listed stocks, 9 Exchange Traded Funds.

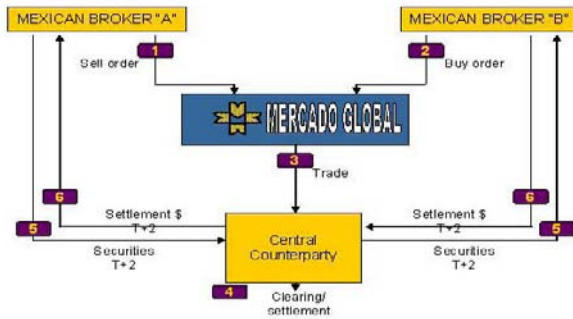The global market components for the BMV for onshore and offshore (cross border) stocks can be seen in Figs. 2 and 3.

Figure 2:   Global market BMV: onshore stocks.
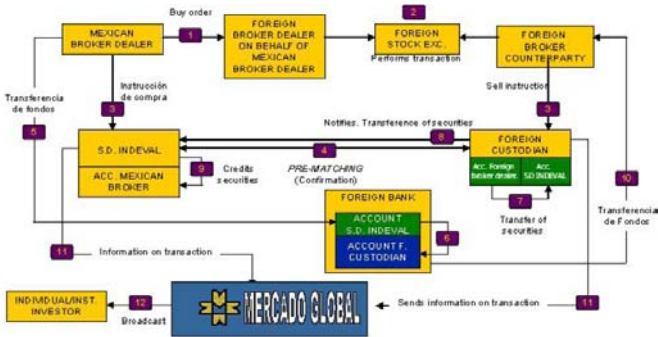
Figure 3:   Global market BMV: offshore stocks.

Has it been a good strategy to invest in the BMV? To answer this question we will explore BMV's history [7]. Fig. 4 shows the investment's index evolution taking in consideration the period: 1980-2004. Trends show that the index's growth has been very high in the long term even if the growth is lower in the short term.

The BMV's risk return frontier shows in Fig. 5 that diversification plays an important role in the expected returns. As more stocks are included in the investment, as higher is the expected return for the same risk level.

The BMV individual stocks' growth for the period 1996-2004 is shown in Table 2. It can be seen that most of the stocks have had a stable behavior and a positive growth.
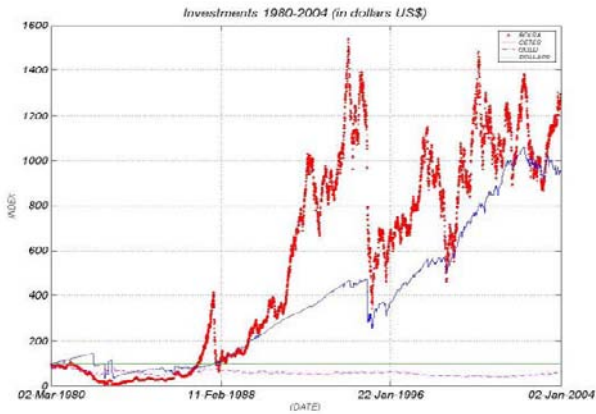
Figure 4:   Index evolution: 1980-2004.
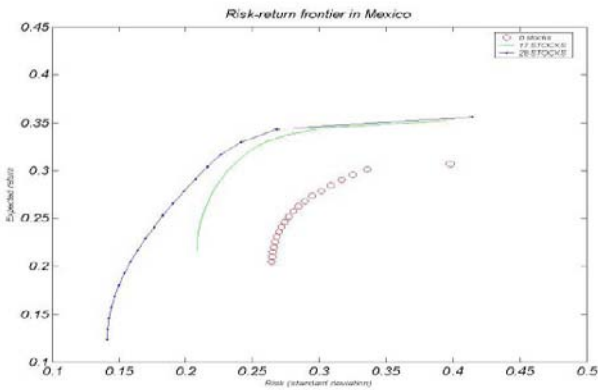
Figure 5:   BMV's risk-return frontier.

Conditions for the BMV are changing. The new regulation will likely achieve the objective of significantly improving the risk-return mix available to domestic investors. Allowing institutional investors to include domestic and foreign equities in their portfolios will translate into a much larger demand for foreign securities than local stocks. The regulation will effectively impede the amplification of exogenous shocks via the impact of the portfolio composition of pension funds on the balance of payments.

4   Financial investment elements

The definitions and elements that should be present in the knowledge database, considering taxes and the investor's scenario can be reviewed to detail in [4]. They were first introduced in the model by Osorio and Sanchez [5] and Osorio et al. [6].

Table 2:   Individual stocks in the BMV (1996-2004).

Portfolio num.	1	3	5	7	9
BIMBO	0.0357	0.0322	0.0287	0.0251	0.0244
CEMEX	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
COMERCI	-0.0000	-0.0000	-0.0000	0.0000	0.0000
CONTAL	0.2119	0.2118	0.2117	0.2116	0.2115
DESC	0.0000	0.0000	0.0000	0.0000	0.0000
BANCOMER	0.0215	0.0252	0.0289	0.0327	0.0335
BANORTE	0.0907	0.0935	0.0962	0.0990	0.0996
GISSA	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
ICA	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
KIMBER	0.0000	0.0000	0.0000	0.0000	0.0000
KOF	0.2200	0.2222	0.2244	0.2268	0.2272
MASECA	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
PENOLES	0.0151	0.0136	0.0121	0.0106	0.0103
SAVIA	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
SORIANA	0.0068	0.0062	0.0055	0.0048	0.0047
TELMEX	0.2550	0.2558	0.2567	0.2575	0.2577
TELEVISA	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
GFINBUR	0.0000	0.0000	-0.0000	-0.0000	-0.0000
GMODELO	0.0471	0.0425	0.0379	0.0331	0.0322
ELEKTRA	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
GMEXICO	0.0000	0.0000	0.0000	0.0000	0.0000
CIE	0.0848	0.0868	0.0887	0.0908	0.0912
TELECOM	0.0114	0.0103	0.0092	0.0080	0.0078
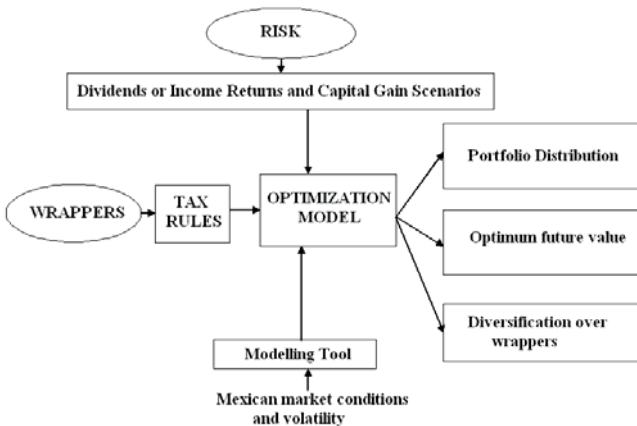
Figure 6:   Elements for a portfolio optimization model.

From a global point of view, we should consider the elements shown in Fig. 6 in a portfolio optimization model. The elements for a portfolio optimization model are constraints and uncertainty representation in scenario trees. The constraints in the model are:

- Initial Allocation
- Cash Balance Equations
- Wealth Balance

- Cumulative Taxes
- Cumulative Returns
- Diversification Constraints

These models require a coherent representation of uncertainty. This is expressed in terms of multivariate continuous distributions. Hence, a decision model is generated with internal sampling or a discrete approximation of the underlying continuous distribution. The random variables are the uncertain return values of each asset on an investment. The discretization of the random values and the probability space leads to a framework in which a random variable takes finitely many values. At each time period, new scenarios branch from the old, creating a scenario tree. Scenario trees can be generated based on different probabilistic approaches as simulation or optimization as presented in Osorio et al. [6].

5   An evolutive simulated annealing algorithm for the extended Markowitz model

The problem of optimally selecting a portfolio among n assets was formulated by Markowitz in 1952 as a constrained quadratic minimization problem [10]. The main assumption on Markowitz's model is that the aim of the investor is to design a portfolio which minimizes risk while achieving a predetermined expected return. An investor should be compensated with an increase in the portfolio's return if she accepts an increased level of risk.

The model considers that each asset is characterized by a return varying randomly with time. The variance of its return measures the risk of each asset. If each component x_i of the vector $X=\{x_1, x_2, .., x_n\}$ represents the proportion of an investor's wealth allocated to asset i and r_i is the individual asset return contained in a vector $R=\{r_1, r_2, .., r_n\}$, then the total return of the portfolio is given by the scalar product of X by R. The set of optimal solutions of the Markowitz's model, parameterized over all possible values of the expected return constitutes the mean-variance frontier of the portfolio. The basic mean-variance model has been studied extensively but is often too simplistic to represent the complexity of real-world portfolio selection problems in an adequate fashion. In order to enrich the model, we introduced more realistic constraints mentioned above (see [4]). With these additional constraints the resulting problem is a mixed integer and combinatorial problem so the decision of which assets to include or not is a crucial one which turns out to be NP hard problem.

The use of a hybrid search algorithm that combines a simulated annealing (SA) procedure and some elements taken from evolutionary strategies to find a set of solutions for the mean-variance frontier, proved to be very efficient to approach this problem [11].

Unlike the standard version of SA where just one portfolio is considered this algorithm uses a whole "population" of portfolios which are permanently ranked. According to evolutionary principles, the best of the population's members are favored whereas the worst are eliminated and replaced either by a clone of one of

the best portfolios or some new individual which is endowed with properties of the best portfolios.

The algorithm starts with a random initialization of a solution each representing a portfolio. This includes selecting k of the N available assets and assigning them random positive weights w_i such that their sum adds up to 1; i.e., w_i is the weight of asset i in a portfolio during iteration t .The subsequent iterations consist of three stages: evaluation and ranking of the portfolios; modification of the portfolio structure; replacement of the poorest portfolios in the population.

Evaluation ranks the solutions according to their fitness. This evaluation and ranking procedure is important to decide which solutions to reinforce and which to eliminate. Modification process uses the set of assets contained in a portfolio and changes the weights originally assigned to each one according to investor's rules and, with random probabilities, it may incorporate new elements to the portfolio. Having changed the weights and standardized them such that the sum of w_i is equal to one, the fitness of the modified portfolio P is calculated according to the principles of SA, the modifications are accepted with probability p which comes from the Metropolis function. Replacement is used to reinforce promising tendencies on the one hand and eliminate rather disappointing ones on the other. The worst portfolios of the current population are replaced with new portfolios which are considered to have high potential. We distinguish two alternatives of high potential replacements: Clone and Average idol. In clone, the probabilities for selecting an existing portfolio are calculated such that the better a prodigy's fitness the higher the probability it is chosen. In average idol, an average weight for each asset is calculated based on the elitists and prodigies portfolios.

The algorithm finishes when a fixed number of iterations is reached and the best solution, which corresponds to the elitist portfolio is reported. The process is repeated for each required value of the return in order to create the mean-variance frontier.

6   *i*-DSS proposed

The elements in the portfolio optimization models and the knowledge databases can be integrated in an *i*-DSS scheme, showed in Fig. 7. The scheme presented is based in the scheme first introduced by Trippy and Lee [9], but instead of prioritized variables we generated additional constraints with enriched elements to manage a wide range of withdrawals in different wrappers for the portfolio elements considered, took into account the investors' scenario first introduced by Osorio et al. [5], and introduce the Mexican market conditions and volatility into the knowledge and preference system.

7   Examples and conclusions

We considered an example where an investor sold a factory, would like to invest $8,000,000 and to live from this money the next eleven years. He would like to

get an annual withdrawal of $500,000. How would he have to invest his money in order to maximize the money he will receive back after eleven years?
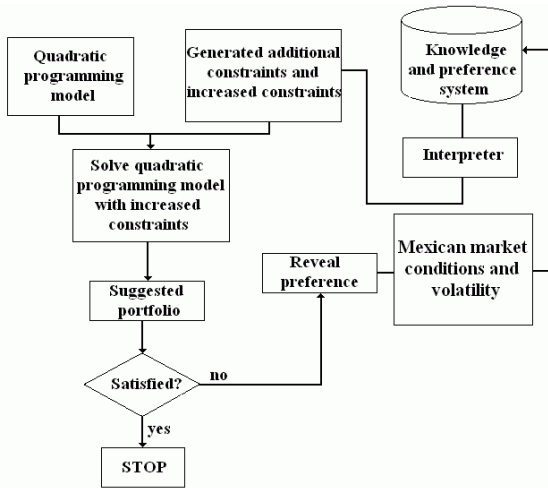

Figure 7:   *i*-DSS that integrates knowledge and portfolio optimization models.
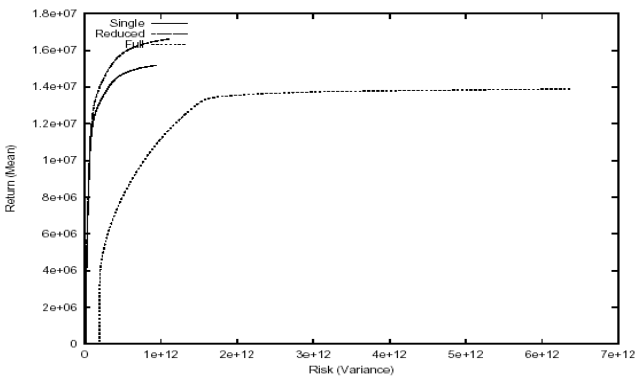

Figure 8:   Impact of scenario trees.

Results show that scenario tree structure plays an important role in the efficient frontier obtained by the Markowitz models, as can be seen in Fig. 8. In the example, we compared a one branch scenario tree with a tree with four branches for the first period and one branch for the rest, and with a full binary tree.

The impact of real life situations in the efficient frontier can be seen in Fig. 9. We introduced the options of gift, inheritance and emigration. Finally, we can conclude that the integration of knowledge in portfolio optimization models can help to develop flexible and efficient *i*-DSSs for investment management.
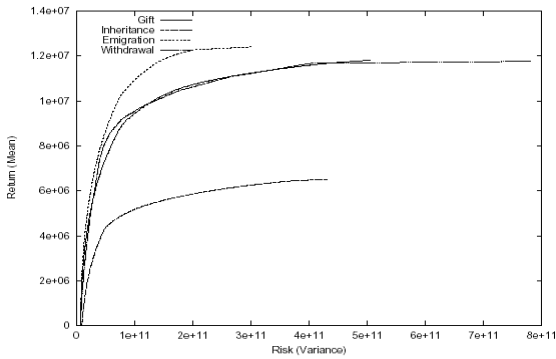
Figure 9:   Impact of real life situations.

References

[1]   Turban, E.: "Decision Support and Expert Systems", 2nd Ed., Macmillan, New York (1988)

[2]   Klein, R. L., Methlie, L. B.: "Knowledge-Based Decision Support Systems", 2nd Ed., Wiley (1995)

[3]   El-Najdawi, M. K., Stylianou, A. C.: "Expert Support Systems: Integrating AI Technologies", Communications of the ACM, Vol. 36, No. 12, (1993) 55-103

[4]   Osorio L. M., Sánchez L. A.: "Designing Decision Support Systems for Portfolio Management", Technical Report, School of Computer Sciences, Universidad Autónoma de Puebla (2005)

[5]   Osorio M. A., Settergren R., Rustem B., Gulpinar N.: "Post Tax Optimal Investments, Financial Engineering, e-Commerce and Supply Chain", Kluwer Academic Publishers, (2002) 153-173

[6]   Osorio M. A., Gulpinar N., Settergren R., Rustem B.: "Post-Tax Optimization with Stochastic Programming", European Journal of Operational Research, Vol. 157, (2004) 152-168

[7]   Reynoso A.: "Opening up a Securities Market: Mexico's New Push for Liberalization 2003-2004", Mexican Stock Exchange, SCID Latin America Conference (2004)

[8]   Reynoso A.: "Can Subsidiaries of Foreign Banks Contribute to the Stability of the Forex Market in Emerging Economies? A Look at some Evidences from the Mexican Financial System", Working paper 8864, National Bureau of Economic Research, Cambridge, MA (2002)

[9]   Trippy, R., Lee, J.: "Artificial Intelligence in Finance and Investing", IRWIN Professional Publishing (1996)

[10]   Markowitz, H. M.: "Portfolio selection", Cambridge, MA, Blackwell. (1959)

[11]   Mora, X., Gomez, M. A.: "An evolutive simulated annealing algorithm for the enriched Markowitz's Model", Technical Report, Actuarial and Statistics department, Universidad de las Américas Puebla (2004).

Timing inconsistencies in the calculation of funds of funds net asset value

C. Louargant[1], L. Neuberg[2] & V. Terraza[3]
[1]*University of Metz, Grefige-Ceremo, France*
[2]*Fortis Investments, Luxembourg*
[3]*LSF, University of Luxembourg*

Abstract

The purpose of this paper is to analyze the impact upon tracking errors of timing inconsistencies in the calculation of Funds of Funds (FoF) net asset value (NAV). We examine how these timing inconsistencies produce noise in the NAV of FoF and therefore noise in the tracking error. We construct Funds of Funds and calculate NAVs of these FoF using underlying NAVs at different dates. We then compare series of tracking errors to analyze the impact of the timing inconsistencies and formalize a relation adjusting the tracking error including the error term generated by these timing inconsistencies.
Keywords: Funds of Funds, NAV calculation, tracking error.

1   Introduction

A fund of funds is a mutual fund, which invests in other mutual funds. These funds were designed to achieve even greater diversification than traditional mutual funds. In the literature we have found many researches on funds but little on fund of funds. Nevertheless, the number of fund of funds increases each year and the proportion of this tool increases too since 2002. Identifying and selecting the most appropriate fund to use in an investor's portfolio are the major aspects of investment strategies. Two of the most important quantitative measures traditionally used are tracking error (TE) and the information ratio (IR). When used properly, these tools give interesting information to make decisions. However, one important problem recently advocated in the literature [1, 2], concerns the difficulty to estimate TE. We compute annualised tracking errors for 1700 Funds and 280 Funds of Funds. We compare the two non Gaussian

distributions of TE using Kolmogorov Smirnov and Wilcoxon-Mann Whitney tests. The results of tests reveal that the two sample of TE don't follow the same distribution. In this paper, we consider another source of bias to estimate TE relative to the analysis of fund of funds: the excess volatility of Tracking Errors due of time inconsistencies. Although the problem of non-synchronous data has been already shown in mutual funds by previous studies [3–7], it has found more pronounced in fund of funds due to specifics microstructure effects in fund of fund markets. We show that time inconsistencies in the NAV of FOF create autocorrelation in series and induced a biased indicator of risk. For that we introduce a correction term in the TE formula. Our analysis proceeds in the following steps: section 2 describes the construction of fund of funds. Section 3 presents the simulation results of impact upon tracking errors of timing inconsistencies in the calculation of FoFs NAV. We compare the return series of a FoF with the return series of its benchmark, and show the problem due to the timing inconsistencies in the calculation. We examine how these timing inconsistencies produce noise in the NAV and therefore noise in the tracking error. In section 4, we analyze the autocorrelation of simulated FoF and propose a correction term in TE formula. In section 5 we conclude.

Table 1:   The construction of funds of funds.

	Name	Initial Weight
1	Schroder ISF Japanese Equity C Acc	5%
2	Pictet F-Emerging Markets-P	5%
3	Pioneer Funds Top European Players A No Dis EUR	5%
4	Ofima Cible	5%
5	JPMF Europe Strategic Value A EUR	5%
6	MLIIF US Focused Value A2 USD	5%
7	Vanguard US Opportunities Institutional USD	5%
8	Templeton Euroland A Acc	5%
9	SGAM Fund Equities US Concentrated Core B	5%
10	SGAM Fund Equities US Relative Value A	5%
11	Fidelity Funds - European Aggressive Fund	5%
12	CA Funds Emerging Markets I Cap (USD)	5%
13	ACM Bernstein-European Value Portfolio A EUR	5%
14	AXA Rosenberg Eurobloc Equity Alpha A EUR	5%
15	Fidelity Funds - Japan Fund	5%
16	Henderson HF Pan European Equity A2	5%
17	INVESCO GT Pan European A	5%
18	Gartmore CS Eurobloc	5%
19	Franklin US Equity A Acc USD	5%
20	GAM Star European Equity EUR Accumulation Class	5%

2   Construction of funds of funds

The first stage in our FoF construction concerns the benchmark definition. As the objective of the study is to analyze the impact of timing inconsistencies on the FoF' TE, an interesting benchmark has to incorporate the larger possible range of international markets in order to maximize the overlapping effect induced by time zones. We choice the following benchmark with the proportion

in the brackets: MSCI EMU (20%), MSCI Europe ex-EMU (20%), MSCI USA (20%), MSCI Japan (20%), MSCI Emerging Markets Free (20%). The second stage is the construction of the FoF itself. We dispose of every NAV of 20 underlying funds for every market day during the period of May 2002 to March 2005. These funds are equally weighted, and we keep the number of parts of underlying funds unchanged during time without any buy and sell, avoiding by the way transactions costs. These data comes from the Lipper Database. Our simulated FoF is presented in table 1.

3 Impact upon tracking errors of timing inconsistencies in the calculation of FoFs NAV

3.1 Timing inconsistencies

When the NAV of a FoF is calculated, the NAVs of underlying funds are not always available at the same market date. This inconsistency is minimum (even null) when we dispose of all underlying NAVs at the same market day and maximum when we have only fifty percents of the available NAVs at one market day and fifty percents at another market day. At the optimal situation, the most recent underlying NAVs are available for the day before the calculation day. This is usually the case when the FoF has a benchmark defined on a single market. Nevertheless, every manager of FoF compare the return series of his FoF with the benchmark return series delayed by one day (the day before). This can be due to the time of publication of the underlying funds NAVs and also due to the fact that the funds relate to different markets with different closing times. The NAV of the FoF are therefore calculated using diverse market days. In practise, every manager of FoF compares the return series of his FoF at time t with the benchmark return series delayed by one day (at t-1). This strategy creates timing inconsistencies. A solution should be to use all NAVs available two days before (t-2) in place of some in t-1 and some in t-2. Nevertheless, managers don't use this solution because legal rules avoid arbitrage opportunity that can be generated by Late Trading or Market Timing.

3.2 The tracking error ratio measure

Although there are a great number of risk measurement frameworks, the focus of the paper is the market risk of fund of funds relative to its benchmark. Tracking error (TE) is a commonly used summary statistical measure of relative risk to provide an acceptable range of relative performance. TE was first defined by Tobe [8] as the percentage difference between the portfolio (in our application the FOF) and its benchmark index the fund was designed to replicate. TE is estimated as the annualized standard deviation of the difference in returns. For investment funds, it represents the percentage change in the Net Asset Value (NAV) for each day of the whole time period required:

$$\text{Percentage change in the NAV} = \frac{NAV\ on\ day(t) - NAV\ on\ day(t-1)}{NAV\ on\ day\ (t-1)} \qquad (1)$$

Mathematically, the tracking error (TE) is:

$$TE = \sqrt{\frac{\sum_{t=1}^{n}(R_{FOF} - R_B)^2}{N-1}} \tag{2}$$

where R_{FOF} is the return of FOF, R_B the return of the benchmark and N the number of return periods

The annualized TE for daily observations is:

$$TE * \sqrt{250} \tag{3}$$

Lower the TE, closer are the returns of the fund to that of the benchmark.

3.3 Impact upon relative risk measures of timing inconsistencies

Lets assume that we calculate the NAV of our FoF at time t. At this time, the underlying funds haven't yet published their own NAV. We create fictive situations starting from an optimal situation on which we dispose of every underlying NAV at time t-1 to the worst situation where the all underlying NAV are available only on time t-2. Between these two extreme situations, we have NAV available on both t-1 and t-2. As we have 20 underlying funds, we create 21 simulations. Table 2 describes the simulation procedure. For each simulation, we compute the returns series on a daily basis, the annualised tracking errors (from the benchmark return series delay by one day), the beta, the correlation and finally the total risk (the volatility). Table 2 shows the influence on timing inconsistencies on measures of risk portfolio.

Table 2:   Simulations results.

Simulation	Number of NAV in T-1 (in %)	Number of NAV in T-2 (in %)	Tracking Error	Beta	Correlation	Volatility
1	100	0	5.88%	0.90	0.93	15.27%
2	95	5	6.19%	0.87	0.92	14.80%
3	90	10	6.22%	0.83	0.92	14.18%
4	85	15	6.69%	0.78	0.91	13.62%
5	80	20	7.32%	0.74	0.89	13.18%
6	75	25	8.01%	0.70	0.86	12.81%
7	70	30	8.81%	0.67	0.83	12.64%
8	65	35	9.88%	0.63	0.78	12.78%
9	60	40	10.94%	0.60	0.73	13.01%
10	55	45	11.48%	0.55	0.69	12.60%
11	50	50	12.51%	0.53	0.64	13.00%
12	45	55	13.42%	0.48	0.58	13.11%
13	40	60	14.39%	0.45	0.52	13.42%
14	35	65	14.54%	0.42	0.50	13.08%
15	30	70	15.20%	0.38	0.46	13.11%
16	25	75	15.75%	0.34	0.41	13.07%
17	20	80	16.76%	0.31	0.36	13.76%
18	15	85	17.66%	0.26	0.30	13.98%
19	10	90	18.47%	0.23	0.25	14.44%
20	5	95	19.07%	0.20	0.21	14.64%
21	0	100	20.01%	0.16	0.17	15.28%

We can observe on figures 1 and 2, linear relationships between them: the evolution of the TE is positively linear while the evolution of the beta is negatively linear.
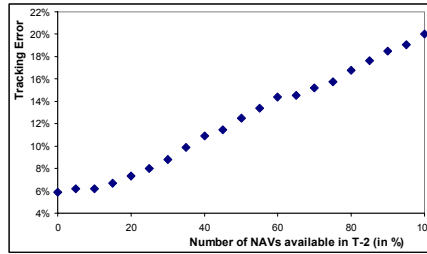

Figure 1:   Influence on timing inconsistencies on tracking error.
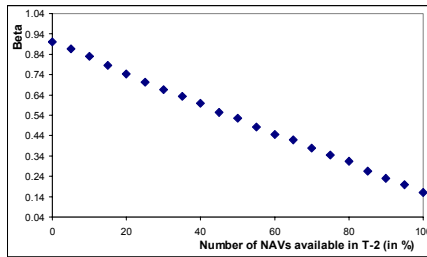

Figure 2:   Influence on timing inconsistencies on beta.

Indeed, the higher is the correlation between the benchmark and the FoF and the lower is the tracking error. As timing inconsistencies increase (100% of NAV available in t-2 for the worst case), the correlation between the benchmark and the FoF falls and the TE increases exponentially.

To verify if the outcome of simulations is correlative to the choice of the benchmark, we compute the same simulations using different benchmarks and we obtain the same kinds of behaviours in regards of time inconsistencies.

But due to our diversified benchmark, which maximises the overlapping effect induced by time zones, we observe the maximum timing inconsistency effect in computing TE, beta and correlation for our simulations.

4   Linear filtering and the "adjusted" tracking error

The next step in the analysis is the modeling of the linear dependencies within the funds of funds data sets. Indeed, time inconsistencies can create

autocorrelation in series and then can explain one part of excess volatility in tracking error.

This phenomenon has already been observed in mutual funds. However, the goals of this step are to determine the effects, if any, that such linear pre-filtering has on tracking error index.

We apply the Ljung Box statistics to test the autocorrelation in simulation, in 250 FoF and 250 funds. The results of the tests on real series reveal that statistics is more often significant for FoF than funds.

The Ljung Box statistics reveal the presence of autocorrelation both in our simulations and the benchmark.

To take into account the autocorrelation effect, we pre-filter the original returns. This series is in effect an integrated series. More specifically, it follows an AR(1) or an AR(2) process. Thus, the elements of the processes must next be filtered out before the final, proxy series of returns could be obtained.

For the benchmark, the I(1) elements is removed by taking the first differences within the series, leaving the following:

$$R_{B_t} - 0.1472 R_{B_{t-1}} = \eta_t$$
$$(0.036)$$

$$(4)$$

For our fund of funds series, the I(1) or I(2) elements is removed using the following regressions:

$$R_{FOF_t} - \rho_1 R_{FOF_{t-1}} - \rho_2 R_{FOF_{t-2}} = \varepsilon_t \qquad (5)$$

where the parameters estimation of equation 5 is given in table 3.

It is the series of residuals from the models (4) and (5), which are $\hat{\eta}_t = R_B^{Proxy}$ and $\hat{\varepsilon}_t = R_{FOF}^{Proxy}$ that finally serves as the proxy series for returns.

The next step is to define an "adjusted" Tracking Error from proxy returns of funds of funds:

$$\text{Adjusted TE} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{\varepsilon}_t - \hat{\eta}_t)^2}{N-1}} \qquad (6)$$

This indicator is based on a linear pre-filtering approach to estimate the tracking error ratio. In comparison with the traditional TE statistics, the adjusted tracking error pre-filter the original returns in order to avoid linear dependencies. The following picture compares the two indicators.

Table 3:   Regression estimations.

Simulations	$\hat{\rho}_1$		$\hat{\rho}_2$	
1	0.2488	(0.0358)	/	
2	0.2739	(0.0356)	/	
3	0.3223	(0.035)	/	
4	0.3496	(0.0346)	/	
5	0.3746	(0.0343)	/	
6	0.4898	(0.0366)	-0.1338	(0.0366)
7	0.4892	(0.0366)	-0.1416	(0.0366)
8	0.4709	(0.0366)	-0.1314	(0.0366)
9	0.4709	(0.0366)	-0.1314	(0.0366)
10	0.5132	(0.0365)	-0.159	(0.0365)
11	0.522	(0.0364)	-0.1651	(0.0364)
12	0.6245	(0.0359)	-0.2337	(0.0359)
13	0.5957	(0.0361)	-0.2099	(0.0361)
14	0.6346	(0.036)	-0.2258	(0.036)
15	0.6319	(0.0362)	-0.2039	(0.0362)
16	0.5374	(0.0367)	-0.1145	(0.0367)
17	0.4994	(0.0367)	-0.1013	(0.0367)
18	0.4021	(0.0367)		
19	0.4354	(0.0368)	-0.0833	(0.0368)
20	0.3303	(0.0348)	/	
21	0.299	(0.0352)	/	

(.) standard deviations
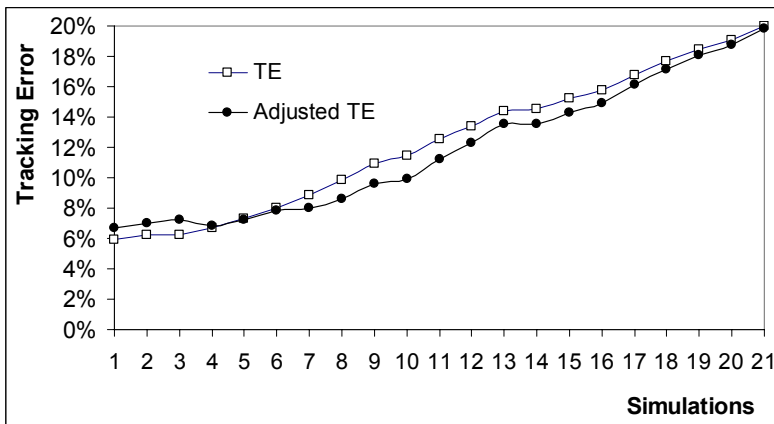
Figure 3:   Comparison between TE and Adjusted TE.

We can observe a substantial improve of our results from figure 3: the correction of autocorrelation gives a better estimation of the Tracking Error. However, linear filtering is not sufficient. Indeed, even if the effect of time inconsistencies can be reduced by taking account autocorrelation of the series it seems that it's cannot explain the excess of volatility observed in tracking error of funds of funds. Some others sources can be due to statistical properties like residual non-linear dependencies, or perhaps in the construction on funds of funds itself.

5   Conclusion

We analyzed another source of bias in the calculation of tracking error: time inconsistencies. We concluded that these inconsistencies create excess volatility in Fund of Funds tracking error. Moreover, we showed that time inconsistencies in the NAV of Fund of Funds create autocorrelation in series and so induced a biased indicator of risk. We construct an adjusted tracking error formula estimated from autoregressive processes of returns. Our principal result is that taking into account the autocorrelation effect in NAV of funds of funds improves our risk estimation. The next step will be to ameliorate the quality of the correction of Tracking error formula by improving the analysis of the residuals terms.

References

[1] Clarke R.G., Krase S. and Statman M., Tracking errors, Regret and tactical asset allocation, *The Journal of Portfolio Management,* Spring, pp.16-24, 1994.
[2] Rudolf M., Wolter H and Zimmermann H., A linear model for tracking error minimization, *The Journal of Banking and Finance*, 23, pp.85-103, 1999.
[3] French and Roll, Stock return variances: The arrival of information and the reaction of traders *Journal of Financial Economics* ,17 , pp.5-26, 1986.
[4] Kadlec G.B. and Patterson D.M., A transaction data analysis of nonsynchronous trading, *Review of Financial Studies*, fall, 12 (3), pp.609-630, 1999.
[5] Goetzmann W.N., Ivkovich Z. and Rouwenhorst K.G., Day trading international mutual funds: evidence and policy solutions, working paper ICF00-03, Yale School of Management, Yale University, 2000.
[6] Chalmers J.M.R., Edelen R.M. and Kadlec G.B., Predictable changes in NAV: the Wildcard option in transacting mutual-fund shares, working paper 08-00, Wharton School, University of Pennsylvania, 2000.
[7] Ammann M. and Zimmermann H., Tracking error and tactical asset allocation, *Financial Analysts Journal*, 57 (2) March/April, pp.32-43, 2001.
[8] Tobe C, Tracking error: An essential tool in evaluating index Funds, the *Journal of Investment Consulting*, November, pp.8-11, 1999.

Strategic asset allocation using quadratic programming with case based reasoning and intelligent agents

E. Falconer, A. Usoro, M. Stansfield & B. Lees
School of Computing, University of Paisley, Scotland, UK

Abstract

There is an abundance of intelligent agent decision support systems (DSS) that have been used to assist with portfolio management decisions. These systems have tended to focus on fundamental, technical and trader behaviour analyses whereas efficient markets hypothesis, as argued by researchers such as Clarke et al., suggest that effort should rather be spent on strategic asset allocation. All of the artificial intelligent studies that have been found have focused on analysis methods which are aimed at identifying investment opportunities with above average returns. According to Clarke *et al*, although giving support to these types of decision is possible with today's computing power, to be effective these systems must identify an opportunity before the rest of the marketplace. There is evidence from the efficient markets hypothesis that the stockmarket reflects new information so quickly that these types of analysis technique are ineffective. As an alternative to these methods Markowitz developed an approach known as mean-variance efficiency analysis which aimed at allowing an investor to gain a specific level of return for a corresponding degree of risk. Sharpe introduced an alternative way of implementing mean-variance efficiency theory by defining stocks and other investments in terms of asset classes. A literature search has revealed that there has as yet been no attempt to investigate this potential. This paper therefore aims to fill this gap by reporting on research that involves the incorporation of quadratic calculations in an intelligent agent and case based reasoning tool for strategic asset allocation. This paper provides the results of an evaluation into the effectiveness of intelligent agents and case based reasoning in dealing with this problem.
Keywords: intelligent agents, quadratic optimization, portfolio management, asset allocation, case based reasoning.

1   Introduction

The efficient markets hypothesis has led academics and practitioners to divide investment management approaches into two primary types: i) active investment management, where an investment manager uses available information and forecasting techniques to seek better performance than a diversified portfolio; and ii) passive investment management, where an investment manager has minimal expectations about the returns that can be achieved and relies instead on diversification to match the performance of a market index [1, 2]. Although each of these approaches has benefits, there has been strong evidence to support the claim that the profitability of active investing is minimal, especially after transaction costs, and therefore passive investing is more profitable [1, 3, 4]. One of the implications of choosing a passive investment approach is that portfolio managers must concentrate on other strategies for improving the return from their portfolio [1]. Examples of these strategies include cost management, diversification and asset allocation strategies. This research concentrates on asset allocation.

There exist many different approaches to asset allocation. These variations include strategic, tactical and dynamic asset allocation, amongst others. This research concentrates on strategic asset allocation. The optimal strategic asset allocation can be determined in one of two ways: by maximizing return for a given risk level or alternatively, minimizing risk for a particular return objective [5–7]. A strategic asset allocation decision is made up of a quadratic component and a utility component [5]. The quadratic component is concerned with constructing the portfolio under quadratic constraints whilst the utility component is concerned with determining the investor's attitude towards risk. An earlier stage of the work described in this paper was presented at the International conference on intelligent agents, web technologies and internet commerce – IAWTIC – in Vienna Austria [8].

This research considers the application of a combined intelligent agent and case based reasoning (CBR) approach to the strategic asset allocation problem. Broadly speaking, literature on intelligent agent and case based reasoning in finance has focused on three areas: i) portfolio monitoring, ii) stock selection, and iii) behavioural finance. Portfolio monitoring is defined as the ongoing, continuous, daily provision of an up-to-date financial picture of an existing portfolio [9]. Stock selection is defined as an analysis technique which uses valuation and forecasting techniques to identify mis-priced assets [10]. Behavioural finance is defined as the computational study of economies modelled as evolving systems of autonomous interacting agents. The determination of an investor's utility or support of quadratic programming does not appear to have been considered. This research therefore assesses the effectiveness of intelligent agents and CBR in dealing with the complexities associated with this problem. This paper concentrates on the quadratic program calculations under mean variance efficiency. Utility theory will be considered in a subsequent paper.

The paper reports preliminary results on the effectiveness of a prototype CBR and agent-based system towards the quadratic programming problem. The rest of this paper provides: (i) a background to the research problem and an overview of the data preparation steps; (ii) an overview of agent and CBR characteristics; (iii) an outline of the agent architecture; (iv) an evaluation of case based reasoning as it applies to the problem and preliminary results on the intelligent agent approach; (v) conclusions and future directions.

2 Agent and CBR characteristics

Strategic asset allocation is characterised by a variety of complexities that have to be considered in the design of a decision support system [5]. These complexities are related to the problem of matching investors with portfolios. Some of these complexities are listed below.

(i) The strategic asset allocation task is distributed over various parts; this means that different factors contribute to the matching of an investor with a portfolio.

(ii) Some tasks require input from other tasks to arrive at a conclusion; thereby necessitating the need to share data and control the flow of information between tasks.

(iii) There is a chance that the same set of data inputs could be used multiple times; this means there is a potential for task duplication.

(iv) There is not one solution that can be applied to every set of circumstances; this means that each portfolio or investor could be either unique or at best similar but often not exactly the same as those encountered previously.

(v) In order to match up an investor with a portfolio it is necessary to have input from a user; and this user could require a different level of assistance depending on their preferences.

In the light of these observations a review was conducted of available technologies and intelligent agent and case based reasoning identified as displaying the types of attribute that could deal with the complicated nature of the strategic asset allocation decision. Agent and CRB technologies have been used to successfully deal with the complexities of portfolio management [9, 11, 12].

2.1 Intelligent agents

Intelligent agents are software artefacts that have a number of characteristics that distinguish them from other technologies. In particular, an agent can be given a goal or task which it can complete autonomously [13]. Agents may also carry out tasks or goals in parallel with the existing computer operations and ask for advice from humans if they become "stuck" in a problem [14]. Agents can act proactively and communicate with human users, providing a personalised service to some users and communicating with other agents in a multi-agent environment [13]. Wooldridge [14] discusses how intelligent agents can interact with, and react to, a user's preferences. This means that software can be

designed to recognise a user's actions and modify the user/agent interaction to ensure the level of support is appropriate to that user.

Decker and Sycara [9] also discuss how intelligent agents can incorporate a separate control component to regulate the flow of information between distributed tasks. This control component can organise the flow of information that passes between the various parts of the problem and ensures that a conclusion is reached. This ability to control distributed tasks is a key attribute of the anticipated solution. Intelligent agents therefore demonstrate various abilities that suggest they may be appropriate for resolving this problem. In particular the ability of intelligent agents to interact with human experts, organise data flow, construct knowledge bases and learn from experience are perceived to be important for overcoming the complexities associated with matching portfolios with investors.

2.2 Case based reasoning

Case-based reasoning (CBR) is a methodology that can allow software to learn from experience. Case based reasoning subdivides a problem into a series of tasks and then combines the results of each task to form a case. Each case is then compared with previous cases to determine whether new experiences can be learned. Althoff [15] identifies four tasks that exist as a part of most general CBR programs: retrieve task – to retrieve similar cases; reuse task – to reuse information and knowledge from previous cases to solve a problem; revise task – where the problem is revised in line with previous experiences; retain task – where experience is retained that could be used for future cases. Althoff [15] also discusses how CBR systems gather case data from distributed tasks and join the outputs of these tasks together to formulate a combined state. In this way a distributed problem can be resolved, with each distributed solution contributing to all future solutions.

Althoff [15] also discusses how CBR systems can be used to subdivide each solution into different parts and then pass requests for analysis onto other components. By initiating this task division CBR software can perform distributed tasks at the same time as other tasks are performed and then combine the parts to arrive at a more complete solution. Althoff [15] identifies that CBR systems evaluate each set of results to determine whether the current case is new, or whether it matches a previous case. CBR was found to offer additional capabilities that would strengthen the intelligent agent approach. It is the ability of CBR to organise distributed tasks, subdivide solutions into categories, historically evaluate solutions and learn from experience which means it may be applicable for dealing with the complexities associated with strategic asset allocation.

3   Preliminary results

There are two sets of results to report, findings related to: (i) the quadratic component and the use of agents to overcome the limitations of infeasible solutions; (ii) the application of CBR to the quadratic programming problem.

3.1   The quadratic component and the quadratic program

In contrast to the standard approach the securities in this case are in the form of global asset prices. The global approach differs slightly from the standard asset allocation approach as the focus is upon prices that reflect a certain global market. In contrast, asset allocation approaches usually focus on much broader asset allocation classes, such as real estate, stocks, bonds and cash. Despite the differences, there are many advantages to adopting a global approach. This decision is supported by the globalisation of markets discussed in literature and the importance of global diversification [16–20]. The historical global asset prices were obtained from Morgan Stanley Corporate International (MSCI©). The calculation used global equity asset classes for 9 specific asset classes. However, although the assets are limited, the model could easily be extended to include additional asset classes, cash assets or bond assets, with minimal effort. The QP used a quadratic function to determine the optimum investment amount to place in each asset class. The following objective function was used to maximise return or minimise risk (from an article by Fylstra [21].

Maxmise $c^T x - x^T Q^o x$
Subject to: $Ax >= b$
                 $\Sigma x = 1$
                 $xi >= 0$ *for all i.*

Q^o is a matrix containing covariances for the assets held within the portfolio. c^T is a vector containing returns on each of the assets held within the portfolio. A is also a vector containing the returns on each of the assets held within the portfolio. b is a value that represents the minimum portfolio return. x is the amount to be invested in the ith asset class within the portfolio. The data tables are available from the author by request. A front end application was developed and is illustrated in Figure 1. The application allows a user to choose asset classes, choose a minimum portfolio return, view covariances and mean returns, and view the optimised result. Yet despite the initial simplicity of the quadratic program in calculating the optimum results, from close inspection, it was apparent that at times it was not possible to produce results that were useful. An extensive evaluation was therefore conducted of the quadratic program and the outputs that were produced under various constraints. This problem is discussed by Fylstra [21] who indicates that analysts take great care to ensure that their portfolio models remain calculable under quadratic or linear constraints. The following four categories identify the types of result that were returned.
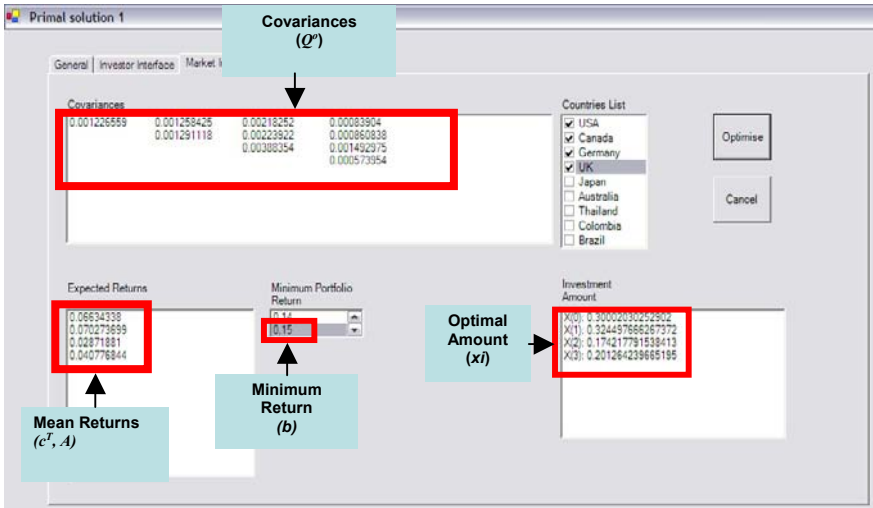
Figure 1:   Front end application for portfolio choice.

It was found from the evaluation that each quadratic result fell into one of these four categories. However, if category two, three or four occurred the user was required to change their choice. In order to resolve this problem intelligent agents are being used to search for alternative solutions. Multiple agents are incorporated in the solution to conduct parallel searches for more effective solutions. The aim of each intelligent agent is to determine alternatives which contain either different asset classes or a different minimum portfolio return.

The closeness of the solution is considered in terms of the distance between the optimum investment amounts returned by each of the amounts invested. The intelligent agent systems attempt to define alternative solutions by modifying the minimum portfolio return or adjusting the asset classes within the solution, reinitiating the quadratic calculations through the MOSEK tool and therefore resolving the problem.

The agent tasked with adjusting the minimum portfolio return keeps the asset classes the same and simply amends the portfolio return to achieve feasible results. On the other hand the agent system focuses on adjusting the asset classes held within the portfolio by using four factors to determine the similarity of each country with another: global influence, average volume, length established and emerging rating. These factors are used to determine the suitability of each country as an alternative to other countries chosen by the user. A numeric value was determined for each country factor which allowed a comparison of asset classes to occur.

The agent adjusting the minimum portfolio return and the asset classes used threshold values to determine the suitability of alternative choices. From an evaluation of the quadratic results it was determined that a threshold value of 0.02 was appropriate for the minimum portfolio return. In so doing it was possible in most cases to determine an alternative portfolio. The agent changing

the chosen asset classes also incorporated a threshold that was set for each of the similarity factors (mentioned above as global influence, length established, average volume and emerging rating).   Using the similarity factor and the threshold value the agent program attempts to determine an asset class or asset classes that are sufficiently similar. The agent program proceeds to optimise the result and determine whether the portfolio is more suitable.

Results of an evaluation of the agent program in achieving these aims shows that in most cases an alternative solution can be found. By incorporating multiple agents into the solution it is therefore possible to overcome the limitations of the quadratic program.

3.2 Case based reasoning search and retrieve method

This section outlines the development of the case-based-reasoning solution and the identification of the case features. The CBR program operates by first having the user provide the system with a set of responses in relation to risk and the user is asked to choose a portfolio for this type of investor. The system uses the quadratic program to provide the problem and solution variables to the CBR component.   The CBR program then evaluates and monitors cases as new problems and solutions are introduced. In general the CBR tool carries out the following steps:

1. Searches from its memory to find portfolios that have been previously chosen for certain investors and makes an assessment of similarity.
2. Infers an answer from the most similar matches.
3. Adjusts the portfolio to investor matching solution for changes in circumstances by inferring relationships between sets of historical data.
4. Modifies solutions from step three and records them for future use.

The search and retrieve stages are completed in the first and second steps and will use the nearest neighbour retrieval method. The nearest neighbour retrieval method identifies possible matches by defining values that can be used to index each case. The indexes are then weighted in terms of importance. Using the nearest neighbour retrieval method there are three states that the CBR program can find itself in when evaluating alternatives: (i) an exact match where the same problem has been used before; (ii) no match where the current problem does not match any previous case; or (iii) a partial match where the current case is similar but not exactly the same as a previous case. The features that relate to this particular problem have been identified and are shown in Table 1. Each feature has been allocated a weight based on the importance of that feature to the problem.   Therefore for instance the country name and number of countries (given a score of 10) are considered more important than global influence, average volume (given a score of 5). Each feature either relates to an aspect about either the portfolio or the investor.

The portfolios will be given a score for each portfolio and investor combination. This score is then used to determine how close this case may be to previous cases. The values will be interpreted by considering the potential risks associated with the individual portfolio as it compares with other portfolios.

Table 1:   Case Features as they apply to the quadratic problem and investor profiler.

Rule Description	Exact Match	Partial Match	No Match
CountryName	10	0	0
Number of Countries	10	5	0
Global Influence	5	2	0
Average Volume	5	2	0
Length Established	5	2	0
Emerging Rating	5	2	0
Covariances	10	5	0
Minimum Portfolio Return	10	5	0
Mean Return	10	5	0
Mosek Result	10	5	0
Risk Attitude	10	5	0
Risk Perception	10	5	0
Knowledge and Experience	5	2	0
Optimal Risk Level	5	2	0
Complexity	10	7	0
Investment Goal	10	5	0
Investment Objective	10	5	0
Investment Time Horizon	10	7	0
Investment Amount	10	7	0
Estimated Year Return	5	2	0
Investor Expected Return	5	2	0
Total	180	87	0

The following is an example of the CBR process. During the first step the user is presented with a set of responses and is asked to choose an asset class and a minimum portfolio return. In this example the user chooses the following portfolio for their investor: Asset classes: USA, UK, Germany and Canada; minimum portfolio return: 0.15. During the second stage all previous cases are compared with the portfolio chosen in step 1 for the type of investor identified by the user. This retrieval is based on the values identified in table 1. The CBR program initially finds three portfolios that are similar to the current investor. The first portfolio contains USA, UK, Thailand and Canada and has the same portfolio return as the current portfolio. The investor has some similarities to the current investor however the time horizon and risk profile are different. This portfolio/investor combination therefore achieves a 58% similarity based on the scores received. The second portfolio contains USA, UK, Canada and has a different portfolio return. However, the investor has different goals and investment amount. The second portfolio/investor combination achieves a 60% similarity. The third portfolio contains USA, UK, Canada, Germany and has a different portfolio return. The investor however is very similar to the current set of investor responses provided by the investor. The third portfolio/investor combination achieved an 87% similarity. The final stage is based on the results of the previous step where the alternatives are evaluated. In order for the current problem to be identified as a possible match the minimum-portfolio-return and the features relating to the investor are all required to be very close to that of the current case. All three portfolios are displayed as possible matches to the user with the score displayed for each portfolio. The final stages of general CBR cycles involves revising the proposed solution and updating the case base and retaining experience. As yet these have not been considered and will be reported at a later stage.

4   Conclusions and future direction

This paper outlines a decision-support system to assist strategic asset allocation decisions based on an intelligent agent. The system uses case-based reasoning to make portfolio optimisation calculations under quadratic constraints. These calculations are performed as part of an optimisation of the expected return and the minimisation of risk within a portfolio of broad asset classes.

The results attained from an evaluation of the quadratic program present sufficient evidence to support the decision to use intelligent agents for the problem of matching investors to portfolios. In addition to this the ease with which the variables associated with this problem fit with the general approach to CBR also support the decision to use CBR within the approach.

Combining the capabilities of intelligent agent and CBR based systems therefore present the opportunity to support strategic asset allocation decisions. By combining these technologies the system can learn from its interactions with the user and also its experiences when evaluating problems.

The next piece of work to be undertaken will involve extending out the CBR component and carrying out further validation on the agent program. Additional work is also being undertaken to identify more opportunities for using the CBR and agent programs to provide more assistance.

References

[1]   Clarke, J; Mendelker, G; Jandik, T (2001) *Expert Financial Planning : Investment Strategies from Industry Leaders*, (Edited by Robert C Arffa), 2001, chapter 10, Wiley and Sons.
[2]   Fama E.F (1991) "Efficient Capital Markets: II", Journal of Finance, December.
[3]   Bernstein, W (2000) *The Intelligent Asset Allocator: How to Build Your Portfolio to Maximize Returns and Minimize Risk* ; McGraw-Hill Education.
[4]   Bernstein, W (2002) The Four Pillars of Investing: Lessons for Building a Winning Portfolio, McGraw-Hill Education.
[5]   Nemati, H; Iyer, L.S (1999) "Intelligent decision support system for asset allocation", *Americas Conference on information Systems*.
[6]   Kaplan, P (1998) "Asset Allocation Models Using the Markowitz Approach", http://www.misp.it/doc/materiali_doc/Curti1-EN.pdf, accessed 17th February 2006.
[7]   Todd, G.P; Markowitz, H (2000) *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, Frank J. Fabozzi Associates.
[8]   Falconer, E; Usoro, A; Stansfield, M; Lees, B (2006); A Prototype Intelligent Agent Based DSS for Quadratic Portfolio Optimization Calculations with Case Based Reasoning, *Agent, Internet and Web Technologies International Conference*, Vienna Austria.

[9]   Decker, K.S; Sycara, K. (1997) "Intelligent Adaptive Information
       Agents", *Journal of intelligent Information Systems*, Kluwer Academic
       Publishers, volume 9, pp 239 – 260.
[10]   Luo, Y; Davis, D; Liu, K (2002) "A multi-agent system framework for
       decision making support in stock trading", *The IEEE Network Magazine
       Special Issue on Enterprise Networking Services*, Volume 16, No 1,
       Jan/Feb.
[11]   Davis, D; Luo, Y (2001) "Using KADS to Design a Multi-Agent
       Framework for Stock Trading", *Multi-Conference Event - Agent for E-
       Business on the Internet*, Las Vegas, June 2001.
[12]   Davis, D; Luo, Y; Liu, K (2000) "A Multi-Agent Framework for Stock
       Trading", *World Computing Conference, Beijing*, August.
[13]   Turban, E; Aronson, AE (2001) *Decision Support Systems and intelligent
       systems*, Sixth Edition, Prentice-Hall Incorporated.
[14]   Wooldridge. M (2002) *An Introduction to Multiagent Systems*, John Wiley
       and Sons Ltd.
[15]   Althoff, K (2001), "Case-Based Reasoning", *ICCBR-01 Workshop On
       Case-Based Reasoning Authoring Support Tools*, 31st July.
[16]   French, K (1991) "Investor Diversification and International Equity
       Markets", *American Economic Review*, MIT Press.
[17]   Lummer, S; Riepe, MW (1994a) "Taming Your Optimisation: A Guide
       through the pitfalls of Mean Variance Optimisation", *Global Asset:
       Techniques for Optimising Portfolio Management*, Edited by J.S.
       Lummer, M.W. Riepe, , John Wiley & Sons.
[18]   Lummer, S; Riepe, MW (1994b) "The Role of Asset Allocation in
       Portfolio Management", *Global Asset Allocation: Techniques for
       Optimising Portfolio Management*, Edited by J.S. Lummer, M.W. Riepe,
       John Wiley & Sons.
[19]   Kaplan, P (1998) "Asset Allocation Models Using the Markowitz
       Approach",   http://www.misp.it/doc/materiali_doc/Curti1-EN.pdf,
       accessed 17th February 2006.
[20]   Dempster, Germano, Villaverde, (2003) "Global Asset Liability
       Management", *British Actuarial Journal*, Volume 9, Number 1,
       pp. 137-195(59)
[21]   Fylstra, D (2005) "Introducing Convex and Conic Optimization for the
       Quantitative Finance Professional", *Wilmott Magazine*, pp. 18-22.

Heuristic approaches to realistic portfolio optimisation

F. Busetti
Andisa Securities, South Africa

Abstract

Previous research in portfolio optimisation has incorporated some real-world aspects but, to the best of our knowledge, none has incorporated all of them in a model applicable to actual real-world portfolios.

We therefore develop a realistic model, investigate the efficiency of its solution by two heuristic methods, genetic algorithms and tabu search, and then examine the insights provided by the optimisation of real portfolios.

Our model is based on the classical mean-variance approach, enhanced with floor and ceiling constraints, cardinality constraints and nonlinear transaction costs that include a substantial illiquidity premium, and is applied to a large 100-stock portfolio.

We find that for large portfolios the performance of genetic algorithms is three orders of magnitude better than that of tabu search.

The results confirm that both floor and ceiling constraints have a substantial negative impact on real portfolio performance. Optimal portfolios with nonlinear costs and cardinality constraints often contain a large number of stocks with very low weightings. Their function is to diversify risk, and floor constraints hamper this, damaging portfolio performance. In addition, nonlinear transaction costs that are comparable in magnitude to forecast returns tend to diversify portfolios materially.

Keywords: portfolio optimisation, efficient frontier, heuristic, genetic algorithm, tabu search.

1   Introduction

While the basis for portfolio optimisation was established by Markowitz [1] in a seminal paper over 50 years ago, it is often difficult to incorporate real-world constraints into the classical theory. In practical portfolio construction there are

floor and ceiling constraints. In addition, the impact of transaction costs on performance can be large. These costs are variable and there is also an additional "liquidity premium" that drives up the cost for large transactions in illiquid stocks. This premium is strongly nonlinear and can be up to two orders of magnitude larger than the base costs. Combined with cardinality constraints this results in a mixed integer nonlinear programming problem that classical algorithms are unable to optimise efficiently.

2   Related work

Mulvey [2] approximated a transaction cost function by a piecewise linear convex function. Zenios and Kang [3] applied a linear programming model, using mean-absolute deviation (MAD) as the risk function, to a mortgage-backed securities portfolio. Konno and Wijayanayake [4] use a branch-and-bound algorithm to solve the MAD optimisation model for a concave cost function. Loraschi *et al*. [5] presented a distributed genetic algorithm for the unconstrained portfolio optimisation problem, while Crama and Schyns [6] developed a model incorporating many types of constraints but costs were ignored. The algorithms of Bienstock [7] exploit the fact that the objective function is quadratic and that the covariance matrix is of low rank, while Borchers and Mitchell [8] use an interior point nonlinear method. Mansini and Speranza [9] used a MAD approach and considered floor constraints. Chang *et al*. [10] constructed a cardinality-constrained Markowitz model incorporating floor and ceiling constraints that was solved using genetic algorithms, simulated annealing and tabu search, but costs were not addressed. Tabu search was applied by Glover *et al*. [11] to a portfolio optimisation problem. Lobo *et al*. [12] described an approximate method incorporating ceiling constraints, risk constraints and costs. Only linear and fixed transaction costs were used, and cardinality constraints were not incorporated.

3   The real-world model

The unconstrained Markowitz model is as follows. If:
N = the number of assets in the investable universe
R_i = the expected return of asset i ($i = 1 \ldots N$) above the risk-free rate r_f
σ_{ij} = the covariance between assets i and j ($i = 1 \ldots N$, $j = 1 \ldots N$)
x_i = the fractional weight in the portfolio of asset i ($i = 1 \ldots N$)
w = the risk-aversion parameter ($0 \leq w \leq 1$)
then the problem becomes:

$$\text{maximise} \qquad (1\text{-}w) \sum_{i=1}^{N} R_i x_i - w \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} x_i x_j \qquad (1)$$

$$\text{subject to} \qquad \sum_{i=1}^{N} x_i = 1$$

$$0 \leq x_i \leq 1 \quad i = 1, \ldots N$$

Solving the quadratic programming equation (1) for various values of *w* results in combinations of portfolio return and variance that map the efficient frontier. This curve represents the set of Pareto-optimal or non-dominated portfolios. Our real-world extensions to the above model are as follows.

3.1 Floor and ceiling constraints

If l_i = the minimum weighting that can be held of asset *i* ($i = 1 \ldots N$)
  u_i = the maximum weighting that can be held of asset *i* ($i = 1 \ldots N$)
then the constraint is simply formulated as

$$l_i \leq x_i \leq u_i \tag{2}$$

where $\qquad\qquad\qquad 0 \leq l_i \leq u_i \leq 1 \ (i = 1 \ldots N)$

3.2 The cost function

The conceptual shape of the transaction cost function is shown in figure 1, where units can refer to either number of stocks or transaction size in monetary units.

Our model is restricted to the high end of the cost curve, as this is the region relevant to institutional investors. If:
m = marketable securities tax (MST) rate
f = a fixed charge component
v = value-added tax (VAT) rate
b = brokerage rate
s = transaction value
p = the illiquidity premium (a function)
c' = total unit transaction cost
C = total transaction cost
then the total transaction cost is given by

$$C = (1 + v)f + [(1 + v)(b + p) + m]s \tag{3}$$

and the total unit transaction cost is

$$c' = C/s = (1 + v)f/s + [(1 + v)(b + p) + m] \tag{4}$$
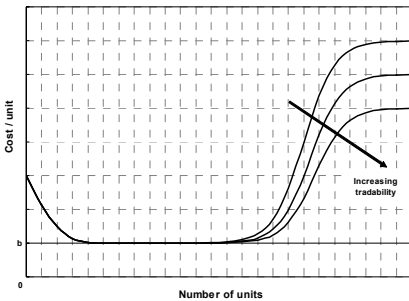
Figure 1:   Illustrative   transaction cost functions.
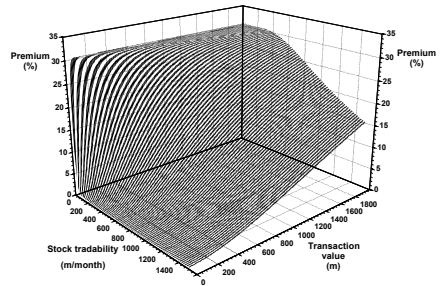
Figure 2:   Illiquidity   premium surface.

The illiquidity premium is mainly a function of transaction size relative to the stocks' tradability. The illiquidity premium is therefore given by a function of s/t, where t = stock tradability (average value traded on the market per time period). Our estimates of the illiquidity premium p are fitted empirically with a two-term exponential function of the form

$$p(s/t) = a[1 - ke^{-\{(k-1)/k\}\,d\,(s/t)} + (k-1)e^{-d\,(s/t)}] \tag{5}$$

using the three parameters a, k and d.

The resulting set of curves of the illiquidity premium for various values of t is shown in figure 2. As stock tradability increases the cost curve declines and also becomes more linear.

3.3 Cardinality constraint

Let $z_i = 1$ if any amount of asset i ($i = 1, \ldots N$) is held

               $z_i = 0$ otherwise

               K = the maximum number of assets allowed in the portfolio

Then the cardinality constraint becomes

$$\sum_{i=1}^{N} z_i = K \tag{6}$$

where $K \leq N$ and $z_i \in [0,1]$ is the integrality constraint.

4 Solving the real-world model

4.1 The heuristic approach

Heuristic solution methods are approximate algorithms that are applied to intractable problems and ensure ever-improving solutions that will result in "acceptable" answers. Examples of heuristic algorithms include simulated annealing, tabu search and genetic algorithms. Since it is difficult to find commercial software for simulated annealing, we test tabu and genetic algorithms on the problem.

4.2 Cardinality-unconstrained case

The cardinality-unconstrained case is first examined to establish the impact of the real-world features of this model. A small portfolio of 10 stocks is sufficient for this purpose and is easily optimised with a simple solver. The 10 largest stocks by market capitalisation are used in this analysis, with a floor constraint of 2% and ceiling constraint of 15%.

4.2.1 Effect of floor and ceiling constraints

The effect of floor and ceiling constraints on the efficient frontier is shown in figure 3. Risk is shown as the standard deviation returns. Floor constraints force an exposure to every stock, including those with poor returns, thus reducing the

portfolio's return. For low levels of risk aversion, the portfolio will normally tend to consist of only one or two stocks, i.e. those with the highest returns.
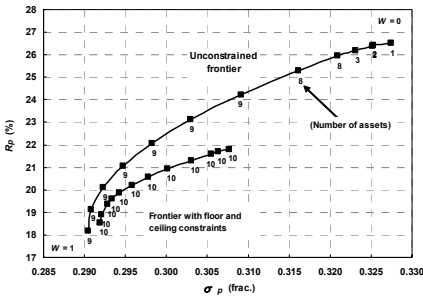
Figure 3:   Effect of floor and ceiling constraints (no costs).
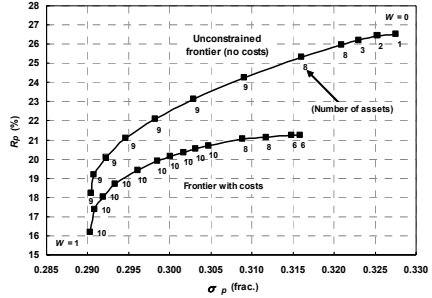
Figure 4:   Effect of costs (no floor and ceiling constraints).

The ceiling constraint, however, will make this high level of optimal exposure to these stocks impossible, and again force an exposure to lower-return stocks, which will reduce the portfolio's return.

The lowest-risk constrained portfolio has a higher volatility than that of the unconstrained portfolio, since the constraints also interfere with the optimal weights for risk reduction.

4.2.2 Effect of nonlinear costs

The impact of nonlinear costs on the portfolio, without any floor or ceiling constraints, is shown in figure 4.

Without costs, the highest-returning portfolio again consists only of one stock. However, the large size of the transaction results in a high transaction cost since costs increase exponentially with transaction size because of the illiquidity premium. If this cost is of a magnitude comparable to the forecast returns, the portfolio tends to diversify into more stocks in order to reduce total transaction costs and their adverse impact on returns.

Therefore, realistic nonlinear transaction costs tend to diversify portfolios.

4.2.3 Combined effect of floor and ceiling constraints and nonlinear costs

The impact on the portfolio of floor and ceiling constraints together with nonlinear costs is shown in figure 5.

The negative impact on the constrained and cost-laden portfolio is cumulative. The characteristics of the three realistically-constrained frontiers are summarised in table 1.

Risk for the highest-return portfolio is reduced by the introduction of constraints and nonlinear costs, since they diversify the portfolio. However, for the lowest-risk portfolio, floor and ceiling constraints will increase risk since they force exposure to higher-risk stocks. An interesting result is that this may in some cases also be accompanied by a corresponding increase in return. The

impact of nonlinear costs on risk is thus ambiguous, depending on the degree of diversification required for cost reduction. The number of stocks in the portfolio invariably increases as a result of constraints (trivially), nonlinear costs and their combined effect.
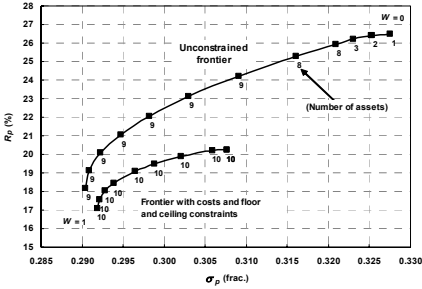

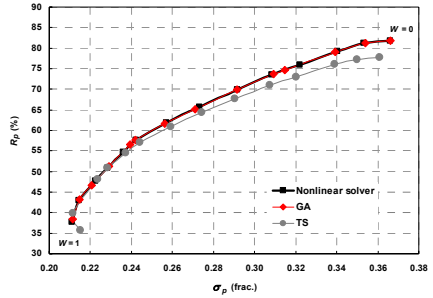Figure 5: Combined effect of constraints and costs.

Figure 6: Heuristic-derived 100-stock cardinality-unconstrained frontier.

Table 1: Summary of constraint and cost effects.

Model	Portfolio		Highest-return portfolio	Lowest-risk portfolio
No constraints	Return	(%)	26.50	18.18
	Risk	(frac)	0.328	0.290
	Stocks	no.	1	8
With floor and ceiling constraints	Return	(%)	21.78	18.53
	Risk	(frac)	0.308	0.292
	Stocks	no.	10	10
With nonlinear costs	Return	(%)	21.24	16.16
	Risk	(frac)	0.316	0.290
	Stocks	no.	6	9
With floor and ceiling constraints and nonlinear costs	Return	(%)	20.24	17.07
	Risk	(frac)	0.308	0.292
	Stocks	no.	10	10
Difference between all constraints and no constraints	Return	(%)	-6.26	-1.11
	Risk	(frac)	-0.020	0.002
	Stocks	no.	9	2

4.3 Cardinality-constrained case

4.3.1 Testing the heuristic methods

The incorporation of cardinality constraints is essential in any realistic portfolio optimisation.

For the real-world cardinality-constrained, large 100-stock portfolio with both floor and ceiling constraints and nonlinear costs which we now consider, there is

no classical method of calculating the exact efficient frontier because of the size and nonlinearity of the problem, and hence no way of benchmarking the heuristic methods against the exact solution. Therefore, to test initially the effectiveness of the heuristic methods and establish their suitability, they are first used to find the efficient frontier without the cardinality constraints.

4.3.2 Evaluating the heuristics on the cardinality-unconstrained case

The efficient frontiers generated by the two heuristic methods for the cardinality-unconstrained but large 100-stock portfolio with floor/ceiling constraints and nonlinear costs are shown in figure 6.

The comparative results of genetic algorithms (GA) and tabu search (TS) are summarised in table 2. The closeness of the calculated frontier to the exact frontier is measured by the arithmetic average of the absolute percentage differences in both return and risk. Solution times are for a 2.8 GHz computer.

Table 2:   Performance of heuristics.

Absolute difference in:		Return (%)	Risk (%)	Objective function (%)	Solution time (min)	Best trial (no.)	Total trials (no.)
TS	Median	3.30	0.54	8.56	30	371	723
	Standard deviation	2.06	0.46	20.62	3	271	175
	Mean	3.06	0.73	16.48	29	379	736
	Combined mean	1.89		-	-	-	-
GA	Median	0.36	0.24	0.07	4	14 819	14 819
	Standard deviation	0.95	0.62	0.09	4	10 656	10 656
	Mean	0.85	0.51	0.08	5	18 419	18 419
	Combined mean	0.68		-	-	-	-

While tabu search works well on small (around 20-stock) problems, its rate of convergence to the solution slows significantly when the problem size increases towards 100 stocks. A stopping rule of 30 minutes is therefore implemented.

A simple efficiency measure is provided by combining absolute accuracy and time by using their product as the criterion. Figure 7 shows both methods' performance across the frontier for various values of w.

Based on this measure, the performance of the genetic algorithm is better than that of tabu search by approximately three orders of magnitude and is able to find solutions arbitrarily close to the correct value, with calculation times of around 5 minutes.

Interestingly, both methods find the centre portion of the efficient frontier the most difficult to generate. A possible reason is that at the upper end (highest returns, low risk aversion, $w \Rightarrow 0$), the selection of the highest-return stocks is relatively straightforward, and at the lower end (lowest risk, high risk-aversion, $w \Rightarrow 1$) the optimisation strategy is also simple: select the lowest-risk stocks. However, in the central part of the frontier there is a much larger number of

combinations of stocks that will result in middle-of-the-road return and risk levels.
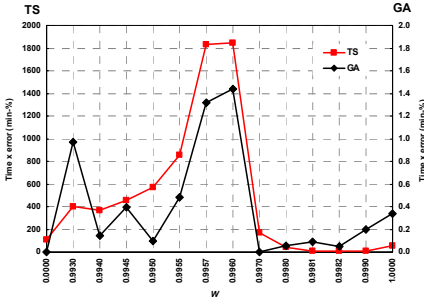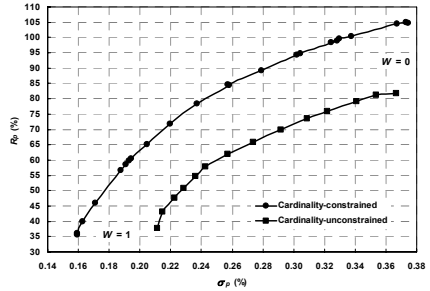

Figure 7: Efficiency of heuristic methods.

Figure 8: Cardinality-constrained efficient frontier.

Since the efficient frontier found by the heuristic methods consists of suboptimal points, the optimal portfolios generated for the cardinality-constrained case will always be conservative in terms of displaying returns per unit of risk that are lower than their true values.

4.3.3 Application to the cardinality-constrained case

In the cardinality-constrained case the floor constraint is subsumed into the cardinality count, i.e. if $x_i > l_i$ then $z_i = 1$, otherwise it is zero. The ceiling of 15% is retained. A cardinality constraint of 40 stocks within the 100-stock universe is selected.

The introduction of cardinality constraints may result in a discontinuous efficient frontier, as shown by Chang *et al* [10]. The efficient frontier for 40 stocks is shown in figure 8. There are no signs of any discontinuities in this particular cardinality-constrained efficient frontier.

The cardinality-constrained portfolio completely dominates the cardinality-unconstrained portfolio. On average, for the same level of return the cardinality-constrained frontier exhibits risk that is lower by between 5% and 12%. Conversely, for equal risk levels, the cardinality-constrained portfolio produces returns that are 24% to 30% higher across the efficient frontier.

The next step in optimising real-world portfolios is to determine the risk-aversion factor w. This can be estimated from the original Markowitz theory, i.e. by drawing the capital market line from the point representing risk-free T-bills to the efficient frontier and noting the value of w at the point of tangency.

Using this value of w, optimal 40-stock portfolios are then generated from the 100-stock universe. Their characteristics are compared with those of the universe in table 3, which also shows the effect of changing the floor constraint from 0.5% to 2%. Since the higher floor constraint shifts the frontier, a different value of w arises.

Table 3:   Optimal portfolio characteristics.

	Weight range x_i (frac)	Total costs $C(x)$ (%)	Excess return less costs Ri (%)	Risk (beta) (x)
100-stock universe				
Maximum return	0.150	1.72	146.5	1.81
Equally-weighted	0.010	0.29	40.2	1.11
Minimum risk	0.000	0.00	-10.4	0.26
Optimal portfolio	-	0.60	39.9	1.11
40-stock optimal portfolios				
Parameters:		Floor =	0.005	w = 0.996
Maximum return	0.150	1.72	146.5	1.61
Equally-weighted	0.025	0.73	59.5	1.00
Minimum risk	0.005	0.59	17.5	0.26
Optimal portfolio	-	1.28	83.5	0.89
Parameters:		Floor =	0.020	w = 0.998
Maximum return	0.150	1.72	146.8	1.61
Equally-weighted	0.025	0.65	56.5	0.99
Minimum risk	0.020	0.62	-4.6	0.26
Optimal portfolio	-	0.79	60.1	0.90

The substantially better performance of the 40-stock cardinality-constrained portfolios in terms of providing substantially higher returns at lower risk is readily apparent. The optimal portfolios generated usually consists of relatively few stocks with high weights that are at, or close to, the ceiling constraint, a larger but still relatively small number of medium weightings and a long tail consisting of many stocks at, or close to, the floor constraint. In terms of the number of stocks, this tail can be around 70%-80% of the portfolio. What happens is that the highest weightings are usually allocated to stocks with high forecast returns. However, these stocks normally also have above-average risk, which raises the portfolio's risk level. This risk is then diversified away by the large number of stocks with very low weightings.

5   Conclusions

We have shown that realistically large portfolios, incorporating floor and ceiling constraints, nonlinear transaction costs including a substantial illiquidity premium, together with cardinality constraints, can be optimised in reasonably short times using genetic algorithms.

It should be noted that the addition of embellishments to the model, such as setting individual floor and ceiling constraints for each stock; applying other constraints such as market capitalisation or tradability, as well as class, sector or style constraints; providing individual cost curves for individual stocks; setting the cardinality constraint as a range and constructing more complex objective functions and cost curves (which could incorporate step functions or even be discontinuous) will be equally easily handled by this approach.

References

[1] Markowitz, H., Portfolio selection. *Journal of Finance,* **7**, pp.77-91, 1952.
[2] Mulvey, J. M., Incorporating transaction in models for asset allocation. *Financial Optimization*, ed. W. Ziemba et al, Cambridge University Press, pp. 243-259, 1993.
[3] Zenios, S. A. and Kang, P., Mean-absolute deviation portfolio optimization for mortgage-backed securities. *Annals of Operations Research,* **45**, pp. 433-450, 1993.
[4] Konno, H. and Wijayanayake, A., Mean-absolute deviation portfolio optimization model under transaction costs. *Journal of the Operations Research Society of Japan,* **42**, pp. 422-435, 1999.
[5] Loraschi, A., Tettamanzi, A., Tomassini, M. and Verda, P., Distributed genetic algorithms with an application to portfolio selection problems. *Artificial Neural Nets And Genetic Algorithms*, ed. D. W. Pearson, N.C. Steele and R. F. Albrecht, Springer-Verlag, pp. 384-387, 1995.
[6] Crama, Y. and Schyns, M., Simulated annealing for complex portfolio selection problems. *European Journal of Operational Research*, **150**, pp. 546–571, 2003.
[7] Bienstock, D., Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, **74**, pp. 121-140, 1996.
[8] Borchers, B. and Mitchell, J. E., A computational comparison of branch and bound and outer approximation algorithms for 0-1 mixed integer nonlinear programs. *Computers & Operations Research*, **24**, pp. 699-701, 1997.
[9] Mansini, R. and Speranza, M. G. Heuristic algorithms for the portfolio selection problem with minimum transaction lots. *European Journal of Operational Research*, **114**, pp. 219–233, 1999.
[10] Chang, T.-J., Meade, N., Beasley, J. E. and Sharaiha, Y. M., Heuristics for cardinality constrained portfolio optimization. *Computers & Operations Research*, **27(13)**, pp. 1271–1302, 2000.
[11] Glover, F., Mulvey, J. M. and Hoyland, K., Solving dynamic stochastic control problems in finance using tabu search with variable scaling. *Meta-heuristics: theory & applications*, ed. I. H. Osman and J.P. Kelly, Kluwer Academic Publishers, pp. 429-448, 1996.
[12] Lobo, M. S., Fazel, M. and Boyd, S., Portfolio optimisation with linear and fixed transaction costs and bounds on risk. (Revise-and-resubmit decision from *Annals of Operations Research*, January 2005.)

Selection of an optimal portfolio with stochastic volatility and discrete observations

N. V. Batalova[1], V. Maroussov[2] & F. G. Viens[3]
[1]*Bank of America Securities, London, U.K.*
[2]*Department of Physics, Purdue University, West Lafayette, IN, USA*
[3]*Department of Statistics, Purdue University, West Lafayette, IN, USA*

Abstract

We give a numerical method to calculate the optimal self-financing portfolio of stock and risk-free asset to maximize the wealth's expected future utility, in the case of stochastic volatility and discrete observations: the portfolio stock allocation is only allowed to change discretely in time at fixed time intervals. We use a particle-filtering and Monte-Carlo-type algorithm, which we implement forward in time in the case of power utility.
Keywords: portfolio optimization, stochastic volatility, particle filtering, Monte-Carlo method, expected utility, diffusion processes, numerical implementation.

1 Introduction and summary

For many markets, the Black-Scholes (BS) model's basic assumption, that a stock's volatility is constant, is far from being satisfied. Empirical evidence for this inadequacy is known to include the so-called volatility smile for implied volatilities, and other phenomena not visible within the BS model. Many natural extensions posit the volatility itself is random; when it is a stochastic process, this is the *stochastic volatility* (SV) model, which we use in this article. We propose a systematic way to optimize a portfolio of continuous-time SV stock and risk-free asset using a discrete-time strategy, thereby offering a way to minimize transaction costs.

Section 2 presents a short overview of the optimal portfolio selection problem, and its position in the literature, including some references on stochastic volatility, and a short list of ways which have been suggested to tackle the issue of non-constant volatility. Section 3 describes the SV model, along with tools needed to

understand our portfolio optimization problem, including (i) optimal stochastic volatility filtering under discrete observations, and its interacting particle method; (ii) the solution, developed in Viens [12], of the discrete observation optimization problem using a Bellman principle and stochastic dynamic programming. Because the algorithm in Viens [12] is too complex to be implemented, we show, in the special case of *power utility* in Section 4 how to reduce the complexity tremendously by using a simple mathematical argument, and a reasonable additional approximation. In section 5, results of our numerical implementation are presented, showing that our method typically outperforms the standard Hamilton-Jacobi-Bellman solution based on the BS model (the classical case of R. Merton, see Bjork [1, Ch. 14]). Conclusions are drawn in Section 6.

2 Scientific context

In the SV class of models, volatility depend on a latent unobserved stochastic process, which can be interpreted as a rate at which new economic information is absorbed by the market. See Ghysels et al. [10] for a survey of SV, Fouque et al. [6] for a detailed study of SV option pricing under fast mean reversion, and Cont et al. [11] for a number of recent advances by various authors. The main question in much of these works is that of SV estimation, a challenging statistical problem. The methodology we have chosen is a Bayesian one, specifically that of optimal non-linear stochastic filtering, where the discretely observed stock prices contain the useful information: see Section 3. Other SV Bayesian techniques which do not intersect ours have been proposed. In Frey and Runggaldier [7], an entirely different type of filtering is performed using the information contained in random observation times. Another different kind of "filtering" is the statistical estimation, pioneered by D.B. Nelson (see Fornari and Mele [8]), of coefficients in time series models of ARCH/GARCH type, which approximate SV models for high-frequency observations. The "method of moments" can be considered as a further distinct type of filtering: see Gallant et al. [9].

The portfolio optimization problem, in its basic form, is to maximize the expected future wealth (as measured using a concave *utility* function) of a portfolio of stock and risk-free account, using a self-financing dynamic strategy. In the SV literature, little attention has been given to this question. The few solutions rely on the assumption of high-frequency or continuous trading, e.g. Fouque et al. [6, Chapter 10]. Only Viens [12] addresses the question with discrete-time observation and trading, but gives no hope for a practical implementation, because the algorithm is akin to a finite difference method for a PDE whose state space has a desperately high dimension. Our paper is the first to give a solid practical algorithm for portfolio utility optimization under continuous-time SV with dynamic discrete time updating; we are also among the first (see also Florescu and [5], and Desai et al. [4]) to show how optimal stochastic volatility filtering can be useful numerically.

3 Stochastic volatility, discrete observations, and optimization

In our partial information setting there is no means of obtaining an arbitrarily accurate estimate of stochastic volatility. This section provides the estimator given by the stochastic volatility filter and its particle method. We then show how this "smart" (interacting) Monte-Carlo-type algorithm, combined with a further Monte-Carlo algorithm for one-step portfolio maximization, approximates the optimization problem.

3.1 The model and the problem

Under the market's probability measure $\mathbf{P}$, the stochastic volatility model for the stock price X and the risk-free asset B is

$$dX_t = X_t\mu dt + X_t\sigma(Y_t)dW_t, \quad B_t = e^{rt} \tag{1}$$

for all $t \geq 0$, where W is a Brownian motion, eqn (1) is in the stochastic Itô sense (see Bjork [1]), μ is the constant mean rate of return, and $\sigma(Y_t)$ is a deterministic function of a stochastic process Y_t that satisfies a diffusion equation driven by another Brownian motion Z with $corr(W, Z) = \rho$ where $0 \leq |\rho| < 1$, i.e.

$$dY_t = \alpha(Y_t)dt + \gamma(Y_t)dZ_t. \tag{2}$$

The choice of the function σ is not fundamental: many choices will yield models which are difficult to distinguish empirically. One commonly assumes $\sigma(x) = \exp(x)$. The choice of the law of Y is crucial. A popular choice, esp. in the case of highly traded assets and indexes (see Fouque et al. [6]), is a *fast-mean-reverting* process such as the Ornstein-Uhlenbeck process with a large α, i.e.

$$dY_t = \alpha(m - Y_t)dt + \sqrt{\alpha}dZ_t. \tag{3}$$

For simplicity of notation, we assume now and throughout that our observation times are the non-negative integer $i = 0, 1, ...N$, with N our time horizon. For a fixed *scenario* $\bar{x} = (x_0, ..., x_N)$, let $\mathcal{F}_i^{\bar{x}}$ be the event $\mathcal{F}_i^{\bar{x}} := \{X_0 = x_0, ..., X_i = x_i\}$ (information from observations up to time i). We will often use the notation $\bar{x}_i := (x_0, ..., x_i)$. The stochastic volatility filtering problem is to estimate the conditional probability distribution $p_i^{\bar{x}}(dy) := \mathbf{P}[Y_i \in dy | \mathcal{F}_i^{\bar{x}}]$. This is called the *optimal* stochastic filter because it minimizes its distance to the actual value of Y in the sense of least squares, i.e. using L^2 norms under conditional expectation given the observations $\mathcal{F}_i^{\bar{x}}$.

A portfolio is defined by $(a, b) = (a_i, b_i)_{i=0}^N$, where a_i represents the number of units of stock X in our portfolio at time i, and b_i the number of units of the risk-free asset (one unit is worth one dollar at time 0). This portfolio's wealth process $\mathcal{W}$, with constant stock holdings in each time interval to ensure that no transaction costs are incurred, is thus for every $s \in [i, i+1]$: $\mathcal{W}_s = \mathcal{W}_s^{a_i, b_i} = a_i X_s + b_i B_s$.

Using the convenient substitution $b_i = (\mathcal{W}_i - x_i a_i)e^{-ri}$, the self-financing condition now simply reads

$$\mathcal{W}_s = a_i X_s + (\mathcal{W}_i - X_i a_i)e^{r(s-i)} \tag{4}$$

for all $i = 0, 1, 2, \cdots, N - 1$, and for all $s \in [i, i + 1]$. Since in addition the initial wealth $\mathcal{W}_0 = w_0$ is known, we see that the system is defined solely using the control variable a_i and the state variable x_i in each interval. Let U be an increasing and concave function on $\mathbf{R}_+$, the *utility function*. In this paper we will use the popular so called HARA case $U(w) = w^p/p$ for some $p \in (0, 1)$. The only condition we must impose on the set A_0 of admissible sequences $(a_i)_i$ is that a_i depend solely on what information is available, i.e. $a_i = a_i(w_0, \bar{x}_i)$. Our task is to find a dynamic portfolio $a^* = (a_i^*)_{i=1}^n$ that attains the supremum in the **P**-expectation

$$V(0, x_0, w_0) = \sup_{a \in A_0} \mathbf{E}\left[U\left(\mathcal{W}_N^{a,b}\right) \middle| X_0 = x_0, \mathcal{W}_0 = w_0 \right]. \tag{5}$$

3.2 Filtering with stochastic volatility

Viens [12] can be consulted for a recursion formula for the filter $p_i(dy)$. However, this formula cannot be evaluated explicitly. The *interacting particle* (or *smart Monte Carlo*, or *MCMC*) algorithm established in Del Moral et al. [3], yields a decent approximation (order $n^{-1/3}$) of $p_i^{\bar{x}}(dy)$ as the empirical distribution of a family of n particles $(Y_i^k)_{k=1}^n$:

$$\hat{p}_i^{\bar{x}}(dy) := \frac{1}{n} \sum_{k=1}^n \delta_{Y_i^k}(dy). \tag{6}$$

We refer to Florescu and Viens [5] for a complete description of the algorithm, and to Del Moral et al. [3] for a proof of the convergence result. Summarizing, let us note that the particles evolve according to the iteration of a two-step (selection/mutation) process. In the mutation process, they evolve independently according to the Euler approximations of appropriate diffusions (dynamics given by eqns (1) and (2)), with number of time steps $m = n^{1/3}$ in each interval of length 1. The proof presented in Del Moral et al. [3] assumes that $\rho = 0$, and shows the convergence in L^1 of $\hat{p}_i^{\bar{x}}(f)$ to $p_i^{\bar{x}}(f)$ for deterministic bounded test functions f. The extension to the case of $\rho \in (-1, 1)$ is stated in Viens [12]. Our simulations indicate that our particle filter is good at stabilizing quickly, following the actual mean-reverting signal Y, and converges relatively fast to the optimal filter; in fact, our empirical convergence speed seems to be on the order of $m^{-3/2} = n^{-1/2}$, which is considerably faster than the result in Del Moral et al. [3], and is consistent with the comments in that reference indicating that ergodic signals should see faster convergence.

4 Numerically implementable approximation

We embed our portfolio optimization problem (5) at time 0 into a dynamic one as follows: for all $w, x, \bar{x}$, for all $i = 1, 2, ..., N$, for all $s \in [i, i+1]$, find

$$V(s, x, w) = V(s, x, w; \bar{x}_i) = \sup_{a \in A_0} \mathbf{E}[U(\mathcal{W}_N^a)|X_s = x, \mathcal{W}_s^a = w, \mathcal{F}_i^{\bar{x}}]. \quad (7)$$

A discrete Bellman principle was established in Viens [12] for the dynamic optimization problem (7), showing that the entire optimization can be solved by a discrete iteration of continuous HJB equations in the individual successive time intervals. Unfortunately, implementing this iteration backwards in time starting from $V(N, \bar{x}_N, w) = U(w)$ is essentially hopeless because it is akin to a finite-difference method whose state variable $\bar{x}_i$ has a time-dependent dimension which is extremely high even for moderate i.

   In order to circumvent this problem, we take advantage of the HARA case $U(w) = w^p/p$ which features the possibility of preserving, approximately, the power form of V in the parameter w. With such an approximation, we will see that a time-forward algorithm can be developed in order to calculate the optimal strategy $a_i^*(w, \bar{x}_i)$ directly, without needing to know $V(i, \bar{x}_i, w)$. If one then wishes to find the initial maximum expected future utility, one then only needs to use the optimal strategies a^* computed for a number of scenarios, keeping track of the corresponding wealth processes, so that $V(0, w_0, x_0)$ can be obtained a posteriori by any Monte-Carlo method in a low-complexity way directly from eqn (5), replacing the supremum by the evaluation for a^*.

4.1 Mathematical analysis

The SV filtering algorithm results in a sequence indexed by time i of a set of n pairs of particles $\left(\hat{X}_i^k, Y_i^k\right)_{k=1}^n$ which approximate the distribution of $\left(X_i, p_i^{\bar{X}}\right)$ given the observations $\bar{x}_i$. Assume that we have constructed, as in Viens [12], an algorithm using these particles that outputs functions $\hat{V}(i, \bar{x}_i, w)$ as an approximation to $V(i, \bar{x}_i, w)$. In accordance with the Monte-Carlo method of [12], using eqn (4), and the notation

$$\hat{\beta}_k(\bar{x}_i) := \hat{X}_{i+1}^k(\bar{x}_i) - x_i e^r, \quad (8)$$

one can check we must have

$$\hat{V}(i, \bar{x}_i, w) = \max_{a \in \mathbf{R}} \frac{1}{n} \sum_{k=1}^n \hat{V}\left(i+1, \bar{x}_i, \hat{X}_{i+1}^k, a\hat{\beta}_k(\bar{x}_i) + we^r\right). \quad (9)$$

We now prove that $\hat{V}(i+1, \bar{x}_{i+1}, w) = (w^p/p) K(i+1, \bar{x}_{i+1})$ for some function K, using induction. For $i = N$, $\hat{V}(N, \bar{x}_N, w) = w^p/p$ obviously has the correct form with $K_N \equiv 1$. Plugging this form for $\hat{V}(i+1, \cdot)$ above, to evaluate

$\hat{V}(i, \cdot)$ by evaluating the extremum, we only need to consider the zeros of the derivative with respect to a for the resulting function, namely those $a = we^r\lambda$ where λ solves

$$\sum_{k=1}^{n} K(i+1, \bar{x}_i, \hat{X}_{i+1}^k)\hat{\beta}_k(\bar{x}_i)(\lambda\hat{\beta}_k(\bar{x}_i) + 1)^{p-1} = 0. \tag{10}$$

The derivative w.r.t. λ of the quantity above is always negative, which proves the sole extremum $\lambda = \lambda^*$ does correspond to a maximum $a^* = a_i^*(w, \bar{x}_i) = we^r\lambda^*$ in eqn (9). This allows us to conclude

$$\hat{V}(i, \bar{x}_i, w) = \frac{w^p}{p}e^{rp}\frac{1}{n}\sum_{k=1}^{n} K(i+1, \bar{x}_i, \hat{X}_{i+1}^k)(\lambda^*\hat{\beta}_k(\bar{x}_i) + 1)^p \tag{11}$$

$$=: \frac{w^p}{p}K(i, \bar{x}_i).$$

Since λ^* solving eqn (10) depends only on $\bar{x}_i$ and the particles, not on w, the same holds for K_i in the last expression above, which also serves as a backward induction formula to calculate K_i. Also note that since w_i depends only on w_0 and $\bar{x}_i$, the same holds for a^*.

4.2 Simplification: our algorithm

In order to make the above solution forward in time, assume for the moment that the quantity $K(i+1, \bar{x}_{i+1}, \hat{X}_{i+1}^k)$ does not in fact depend on k. Then λ^* is the unique solution of

$$0 = \sum_{k=1}^{n} \hat{\beta}_k(\bar{x}_i)(\lambda^*\hat{\beta}_k(\bar{x}_i) + 1)^{p-1} \tag{12}$$

computable forward in time with only the knowledge of $\hat{\beta}_k(\bar{x}_i)$ which can be calculated via their definition (8) at the same time as the filter. Equation (12) is the only equation the portfolio manager needs to solve to find the approximate optimal allocation $a^* = we^r\lambda^*$.

This assumption on $\hat{K}$ corresponds to approximating each $\hat{X}_{i+1}^k$ by their empirical average $\check{X}_{i+1} = \frac{1}{n}\sum_{k=1}^{n}\hat{X}_{i+1}^k$. The variance of the error of each such approximation (which, by the propagation-of-chaos results in Del Moral [2] are known to be approximately IID for reasonably large n) is on the order of the length of the time interval Δt used to simulate $\hat{X}_{i+1}^k$ starting from the observed x_i. In our presentation, $\Delta t = 1$, but this is only for convenience. The accumulation of these errors in eqn (11) yields a variance of order $n^{-2}n\Delta t = \Delta t/n$, showing that unless Δt is quite large, the error introduced by this new approximation will be of a smaller order than the particle filtering error $n^{-1/3}$. A proper mathematical justification of this approximation would take us beyond the scope of this article. We now give a detailed summary of our new forward-only algorithm. Since this algorithm runs using a single sequence $\bar{x}$ of observations, we have suppressed the notation $\bar{x}$ throughout.

4.2.1 Preliminary steps

Decide on a number of particles n such that the error $1/\sqrt{n}$ is satisfactorily small. Let $m = n^{1/3}$, the number of Euler time steps per each time step $[i, i+1]$. Let $\hat{V}(N, w) = w^p/p$. Let $X_0^k = x_0$, $Y_0^k = y_0^k$, for all $k = 1, \cdots, n$. A preliminary filter can be run prior to starting the optimization (time 0), in order to generate realistic initial particles y_0^k: it is a way to implement a "break-in" period for the filter, one after which the filter's stability property will eliminate any misspecification of its initialization.

4.2.2 On-line time loop

For all $i = 0$ to $N - 1$, repeat

Particle Filtering: use the method of Del Moral et al. [3, Section 5] (see Florescu and Viens [5]) to calculate the approximate filter $\hat{p}_i^{\bar{x}}$ from $\hat{p}_{i-1}^{\bar{x}}$, as the empirical measure of the n particles $\{Y_i^k\}_{k=1}^n$.

Euler/Monte-Carlo step for the optimization: For each $k = 1, \cdots, n$, let $\hat{X}_{i+1}^k$ be the endpoint $\hat{X}_i^k(m)$ of the m-step Euler method to simulate (X_{i+1}, Y_{i+1}) using the (X, Y) dynamics starting at time i from the starting point (x_i, Y_i^k), where Y_i^k is the k-th particle of the filter $\hat{p}_i^{\bar{x}}$. Specifically, with $(\chi_m^{i,k})$ and $(\zeta_m^{i,k})$ independent families of IID standard normals, for all $j = 0, \cdots, m - 1$

$$\hat{X}_i^k(j+1) = \hat{X}_i^k(j) + \hat{X}_i^k(j)\mu m^{-1} + \hat{X}_i^k(j)\sigma(\hat{Y}_i^k(j))m^{-1/2}\chi_m^{i,k},$$
$$\hat{Y}_i^k(j+1) = \hat{Y}_i^k(j) + \alpha(\hat{Y}_i^k(j))m^{-1} + \gamma(\hat{Y}_i^k(j))m^{-1/2}\zeta_m^{i,k}.$$

Maximization step: Evaluate the quantities $\hat{\beta}_k := \hat{X}_{i+1}^k - x_i e^r$ and find the solution $\lambda = \lambda_i^*$ of $\sum_{k=1}^n \hat{\beta}_k(\lambda\hat{\beta}_k + 1)^{p-1} = 0$, using any numerical procedure known for finding a root of nonlinear equation, such as simulated annealing.

Portfolio selection. The portfolio manager changes stock and risk-free account allocations in a self-financing way to obtain $a_i^* = w_i e^r \lambda_i^*$ where w_i is the current wealth before the allocation change.

4.2.3 Computation of initial maximal expected future utilities (optional)

In order to present possible objectives for the client, for each fixed initial wealth of interest w_0, the portfolio manager may run repeated simulations of Step 1 for a large number of varying stock scenarios $\bar{x}$, and then calculate the average of all terminal wealth utilities $(w_N)^p/p$ to yield a good approximation for $V(0, x_0, w_0)$.

5 Numerical results

We have coded our algorithm, implementing both steps 1 and 2 above. Our main code outputs the sequence of optimal strategy values $a_i^*(w_0, \bar{x}_i)$ and its corresponding dynamic wealth sequence $w_i(w_0, \bar{x}_i)$ for any given sequence of observed stock prices $\bar{x}$. In order to obtain the maximum expected future utility, as described in step 2 above, we have implemented a further Monte-Carlo method by simulating, independently of step 1, a large number of sequences of stock prices
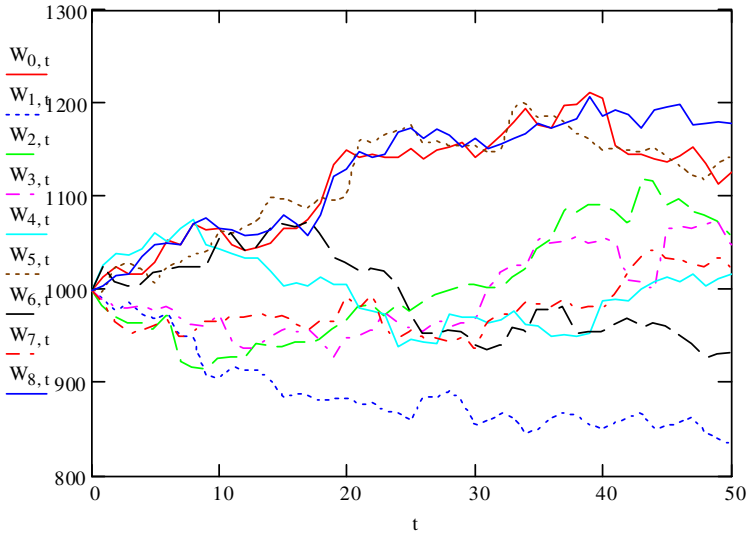
Figure 1: Optimal wealth $\mathcal{W}$ for eight typical scenarios. Params: $m = 230$, $N = 50$, $\sigma_0 = 0.25$, $\alpha = 5$, $\mu = 0.05$, $r = 0.02$, $p = 0.04$.

and volatilities following the true dynamics of (X, Y) as in eqns (1), (2); each of these gives rise to a sequence $\bar{x}$ which is fed into the main code, resulting in a wealth at each time $i \leq N$. For $i = N$, these wealths are then injected into $U(w)$ and averaged yielding a Monte-Carlo approximation of $V(0, w_0)$. Using this approximation, we have been able to observe that a misspecification of the model, by which one assumes that volatility is constant (equal to the level of mean reversion), yields a significant decrease in expected utility, as one would hope. The wealth for a fixed scenario $\bar{x}$ can also be plotted individually (fig. 1), and compared to other strategies for the same scenario, such as pure stock, pure risk-free account, Merton's constant volatility scheme, and arbitrary randomly chosen strategies (see fig. 2). In fact, our fig. 2, which is typical of many scenarios for our choice of parameters, represents an average of dynamic utility over several scenarios. The solid red line corresponds to our method, while the dashed red line is Merton's method. We have strived to show pictures of typical situations, rather than repeating our algorithm until a favorable picture was obtained. This can be seen from the fact that the all-in-bank method (blue) outperforms many other strategies, because of poor stock performance. Nevertheless, our strategy is nearly systematically outperforming the constant-volatility method. It also does better than most random strategies. All-in-stock (brown) is clearly the loser in our simulated bearish market. As a measure of prudence, we have made sure that, when volatility is nearly constant (case of very small α), our method yields, as it should, nearly the same portfolio as the Merton case (optimization for the BS model): $a^* = \sigma^{-2}(\mu - r)/(1 - p)$, where $\sigma = \exp Y_0$ which we chose $\equiv 0.25$.

Red - optimal strategy; Brown - all-money-in-stock; Blue - all-money-in-bank; Magenta/Green/Cyan/Black - Random strategies
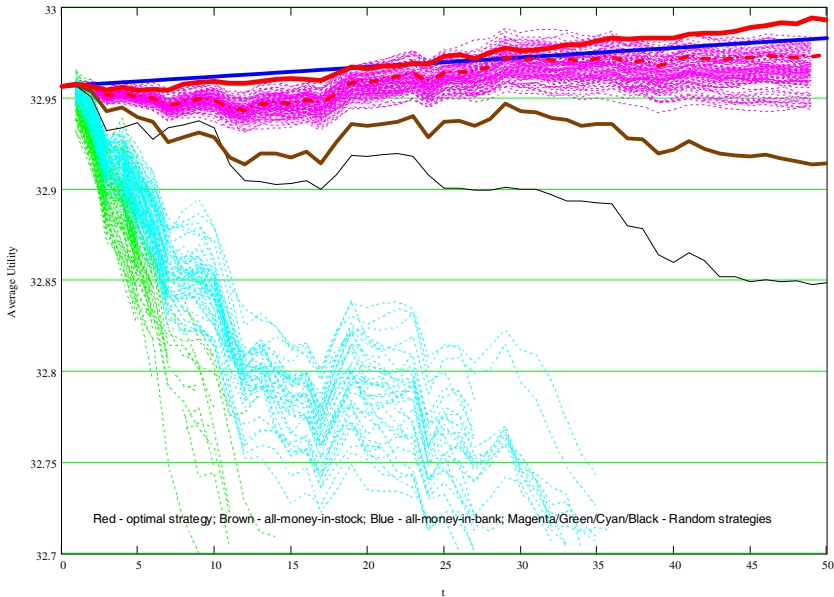
Figure 2: Dynamic average utility over 240 scenarios. Parameters are the same as in fig. 1.

6 Conclusions

Beyond solving the implementability issue for the algorithm of Viens [12] by way of an explicit calculation and an approximation tailored to the HARA case, our work has an important practical economic consequence. For investors who cannot observe and trade SV stock at high-frequency because of prohibitive transaction costs, rather than using an *ad-hoc* discrete adaptation of a continuously traded strategy, our algorithm provides the *optimal* portfolio selection method based solely, and dynamically, on moderate-frequency observations of a high-frequency asset price modeled by a continuous-time SV diffusion.

References

[1] Bjork, T. *Arbitrage theory in continuous time.* Oxford U.P., 1998.

[2] Del Moral, P. *Feynman-Kac Formulae; Genealogical abd Interacting Particle Systems with Applications.* Springer-V., 2004.

[3] Del Moral, P.; Jacod, J.; Protter Ph. The Monte Carlo method for filtering with discrete-time observations. *Prob. Th. Rel. Fields.* **120** (2001), no. 3, 346–368.

[4] Desai, R.; Lele, T.; Viens, F. A Monte-Carlo method for portfolio optimization under partially observed stochastic volatility. CIFEr 03 IEEE Int. Conf. on Comp. Intelligence for Fin. Eng. (2003), 257-263.

[5] Florescu, I; Viens, F. Stochastic volatility: option pricing using a multinomial recombining tree. Preprint, 2006, available at http://www.stat.purdue.edu/~viens/quatrtree7.pdf.

[6] Fouque, J.-P.; Papanicolaou, G.; Sircar, K. R. *Derivatives in financial markets with stochastic volatility*. Cambridge University Press, Cambridge, 2000.

[7] Frey R.; Runggaldier W.G. A nonlinear filtering approach to volatility estimation with a view towards high frequency data. *Int. J. Theor. Appl. Finance* **4** (2001), no. 2, 199–210.

[8] Fornari, F.; Mele, A. *Stochastic volatility in financial markets – Crossing the bridge to continuous time.* Kluwer A.P., 2000.

[9] Gallant, A. R.; Hsieh, D.; Tauchen, G. E. Estimation of stochastic volatility models with diagnostics, *Journal of Econometrics* **81** 159-192 (1997).

[10] Ghysels, E.; Harvey, A.C.; Renault, E. Stochastic volatility. *Statistical methods in finance*, 119–191, Handbook of Statist. **14**, North-Holland, 1996.

[11] Special issue on volatility modelling. M. Avellaneda and R. Cont, eds. *Quantitative Finance* **2**(1), 2002.

[12] Viens, F.G. Portfolio optimization under partially observed stochastic volatility. *COMCON 8. W. Wells, Ed*. 1-12. Optim. Soft., Inc, Pub. Div., 2002.

Section 7
Risk management

This page intentionally left blank

Monte Carlo risk management

M. Di Pierro & A. Nandy
CTI, DePaul University, Chicago, IL, USA

Abstract

In this paper we propose a Monte Carlo based approach to Risk Management. Our approach applies to any system subject to random uncorrelated losses under very general conditions. Our methodology consists of the following steps: model the distribution of losses and the distribution of time intervals between losses via an analysis of historical data; perform a Monte Carlo simulation of a finite period of time in the future; use the Monte Carlo data to estimate the 99.9% VaR.

1 Introduction

In this paper we present an approach to Risk Management based on the Monte Carlo technique. The proposed approach is very general and it can be applied to any system characterized by discrete losses. We have made the following broad assumptions: a) loss events are independent; b) number of loss events occurring in any time interval ΔT is independent of loss events occursing before the time interval considered; c) the probability of two events occurring at exactly same time is zero.

Discrete losses due to internal fraud (i.e. Operational Risk as defined by Basel II accord) for a particular department in a Bank provide a good example of application [1].

Our approach is based on the following steps: 1) model the time distribution and the severity distribution of losses; 2) simulate possible future scenarios compatible with the observed time and severity distributions; 3) compute the 99.9% VaR as the monetary amount that is greater than the total loss in 99.9% of the simulated scenarios.

Notice that our approach is free from bias or assumptions about the distribution of the total loss. The validity of the assumptions a), b) and c) can be verified directly from the data, as we show in the example of section 3.

In the next section of this paper we present the general approach and provide examples of distribution functions that can be used to model the time distribution and the severity distribution of losses.

In the third section we provide an example of application. We build a hypothetical portfolio of stocks and compute the 99.9% VaR due exclusively to mini-crashes. In the example we define a mini-crash as the event when all stocks considered move simultaneously downward. We stress again that our method is very general and can be applied seamlessly to any portfolio for any quantifiable definition of crashes [2] (as long as they can be identified as discrete events), as well as to the analysis of other types of losses (for example Operational Risk losses).

We want to stress that in the cases examined here, usual analytical formulas for computing the VaR without performing a simulation does not apply because the distribution of the total loss over a finite period of time is not Gaussian.

2 Model

Our approach is based on the following assumptions:

a) loss events are independent.
b) the number of loss events occurring in any time interval ΔT is independent of the number of loss events occurring before the time interval considered;
c) the probability of two events occurring at exactly same time is zero.

The consequence of assumptions a), b) and c) is that we can separately model the frequency distribution of losses and the severity distribution of losses. More specifically, assumptions b) and c) indicate that the process underlying the frequency distribution of losses is a Poisson process.

With the above assumptions, our approach requires the following steps that are described in detail in the following subsections:

1. Model the time distribution and severity distribution of losses separately;
2. Simulate a large set of possible future scenarios compatible with the observed time and severity distributions;
3. Compute the 99.9% VaR as the monetary amount that is greater than the total loss in 99.9% of the simulated scenarios.

2.1 Modeling time and severity distributions

Each loss event is characterized by the time when the event occurs, T_i, and by the severity of the loss, S_i. In the example below we assume T_i is measured in days and S_i is measured in dollars. We also assume that losses are sorted by time so that $T_i < T_{i+1}$. We model the time distribution of $\{T_i\}$ and $\{S_i\}$ separately.

As consequence of assumptions a), b) and c), the process underlying the frequency distribution of losses is a Poisson process. This means that the number of events occurring in any finite time frame T follows the Poisson distribution. It

also means that the distribution of the time intervals between subsequent losses, $t_i = T_{i+1} - T_i$ is exponential

$$p(t) = \lambda e^{-\lambda t} \tag{1}$$

where λ is one of the parameter of our model and it equals the average number of events per unit of time (per day). It can be measured from the data as

$$\lambda = (E[t])^{-1} = \left(\frac{1}{N-1} \sum_{i=0}^{i<N-1} (T_{i+1} - T_i) \right)^{-1} \tag{2}$$

where N is the number of historical loss events.

Modeling the severity distribution is a more difficult task and it depends on the data set at hand. Some general characteristics are that losses are always positive and can be arbitrarily large or small, therefore we expect the underlying distribution to be appreciatively log-normal. In the most general case, consistently with extreme value theory [3–5], we expect the presence of a fat tail for high losses, as observed in [1, 6]. A fat tail is characterized by an inverse polynomial behavior as opposed to an exponential behavior as for a normal or lognormal distribution. The natural choice for modeling this fat tail is therefore a generalized Pareto distribution (GPD).

Our approach consists of modeling the severity distribution using the following distribution function

$$p(x) = a\theta(x - x_0)\frac{1}{x}e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} + ab\theta(x_0 - x)(x + c)^{-\beta} \tag{3}$$

which is discussed in detail in Appendix A. This distribution is exactly log-normal for $x < x_0$, exactly GPD for $x > x_0$, and it is continuous and differentiable everywhere. It reduces to a log-normal for $x_0 \to \infty$ and to a GPD for $x_0 = 0$. It depends on only four parameters: μ, σ (mean and variance of the log-normal contribution), x_0 (the point where change of behavior is observed), β (the power of the polynomial behavior of the fat tail). a, b, and c are constants that depend on the above four parameters and their explicit analytical expressions are reported in Appendix A.

Fig. (1) shows various plots of $p(x)$ (left) and the corresponding cumulative distribution functions $F(x)$ (right) in logarithmic scale, for different values of x_0, and fixed values of μ, σ, and β. The greater the value of x_0, the smaller the fat tail.

In summary, the parameters of our model are λ, μ, σ, x_0, and β. While λ is measured via eq. (2), the other 4 parameters can be measure via a fit of the cumulative distribution function implied by historical severity data.

2.2 Simulation

Once the model parameters are extracted from historical data, we proceed by performing a simulation of multiple future scenarios. Each simulated scenario
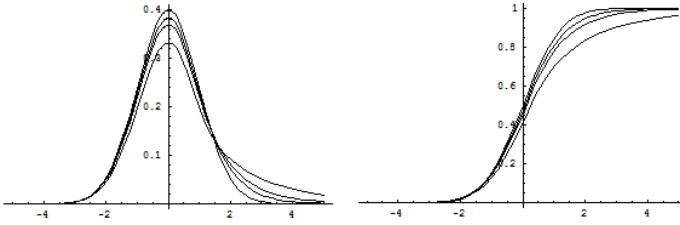
Figure 1: Log-plot of $p(x)$ (left) and $F(x) = \int_{-\infty}^{x} p(x)\mathrm{d}x$ (right) for different values of $\beta = 1.5, 2, 2.5$ at $\mu = 0$, $\sigma = 1$, and $z_0 = 1$ fixed, compared with a Gaussian distribution.

covers a finite time interval from 0 (today) until some future time T. This is the time over which we wish to compute the Value-at-Risk. Each simulated scenario is characterized by a set of simulated loss events occurring within the time interval T.

The simulation of each simulated scenario starts at time 0 and recursively proceeds to simulate the next loss event. The next loss event will occur in time t and will have a severity s. t and s are random numbers generated according to the frequency distribution and the severity distribution respectively. The simulation of each scenario proceeds until a loss event occurs beyond the time interval T being considered.

Multiple scenarios are simulated using the method described above. Then, for each simulated scenario k, we compute the total loss L_k by summing the discounted s losses occurring under scenario k within the considered time interval.

2.3 Computing the VaR

The 99.9% Value-at-Risk is defined as the value such that the probability of losing more than its value is 0.1%. Since our simulated scenarios are generated using the same procedure, they all occur with equal probability $1/m$, where m is the number of simulated scenarios. The 99.9% VaR can therefore be determined as that value that is greater than the total loss L_k in 99.9% of the simulated scenarios, and less the total loss in the remaining 0.1% of the simulated scenarios. This computation is done by sorting the scenarios according to their total loss L_k so that $L_{k+1} > L_k$ and choosing (the notation $\lceil x \rceil$ indicates excess rounding of x)

$$99.9\% \text{ VaR} = L_{\lceil 0.999m \rceil}$$

The same argument applies to the computation of the VaR for any other percentile. The larger is the number of simulated scenarios m, the more precise is the determination of the VaR.

Notice how, in general, the usual formula to compute the VaR in terms of the standard deviation of the losses does not apply. In particular that formula does not apply if the distribution of the total loss is not Gaussian. Fig. (2) shows that this is clearly not the case for the systems under consideration in this paper.
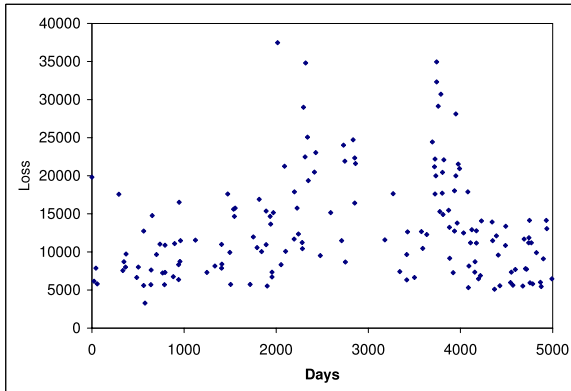
Figure 2: Loss events associated to the mini-crashes (all stocks moving downward) of our sample portfolio during the time frame March 16, 2001 - February 21, 2006.

3 Example

In this section, as an example application of our approach, we consider the computation of VaR due to mini-crashes in a portfolio of stocks. This number represents the economic capital that one would have to save to provide insurance against such events at 99.9% confidence level. We define a mini-crash as the event when all the stocks considered move simultaneously downward.

In order to provide a concrete case we consider a portfolio comprised of the following stocks:

MSFT, RTN, HCA, EP, AVP, HNZ, USB, GD

with a constant capital of $100,000 invested in each of them.

The above set of stocks have been chosen because of their relative low correlation. If the stocks where uncorrected, the probability of all of them moving downward (a mini-crash) in one day would be about 0.4%. Our analysis of historical market data indicates that this occurred 159 times in the time period starting March 16, 1992 and ending February 21, 2006, which averages to approximately once every 31 calendar days, or once every 21 trading days. This corresponds to a mini-crash probability of 5%, ten times more likely than the naive expectation.

Fig. 2 shows past losses associated with our portfolio in the time frame considered.

Fig. 3 shows the distribution of time intervals between two consecutive past losses, superimposed with the exponential distribution using the value $\lambda = 1/31$ extracted from the data. The plot indicates that the probability of two consecutive mini-crashes occurring in two consecutive days is more likely than predicted by the exponential distribution. This is a small deviation from the assumption and
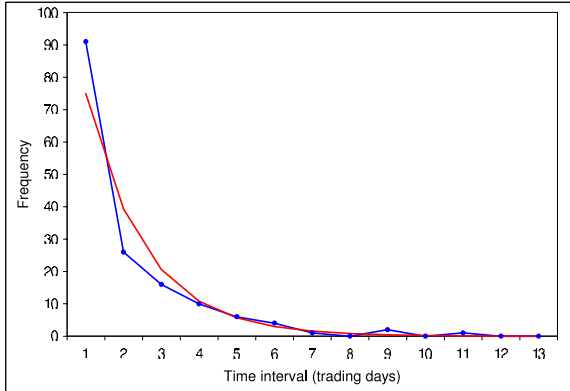
Figure 3: Distribution of time intervals between two consecutive past losses, superimposed to the exponential distribution at $\lambda = 1/31$.
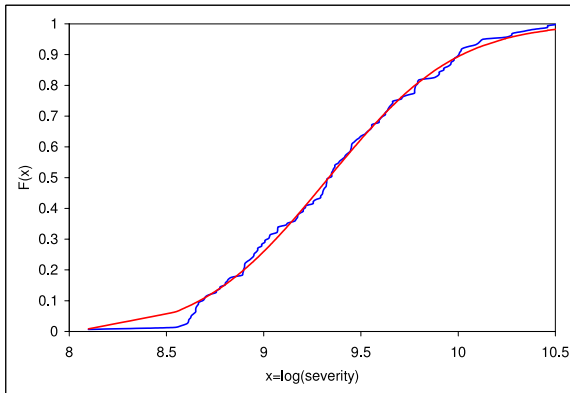
Figure 4: Log-plot of the cumulative distribution function $F(x)$ for the severity of historical losses, and best fit using the distribution in eq. (3).

therefore it will be ignored in the rest of the paper, although it could be taken into account by correcting the frequency distribution.

Fig. 4 shows, in logarithmic scale, the cumulative distribution function for the severity of historical losses and its best fit using the distribution in eq. (3).

Fig. 5 shows the total loss for each simulated scenario for $m = 10000$ simulations. The simulated scenarios are sorted according to their total loss. The 99.9% VaR reads directly from this plot as the y-axis value corresponding to the x-axis value $\lceil 0.999m \rceil = 9990$.

Other VaR percentiles can be calculated in a similar fashion.

Our result is 99.9% VaR = \$360,000. This is 45% of the total funds invested in the portfolio. (The number exceeds the regular 99.9% VaR associated with the portfolio because it only consider events when all stocks jump downward. In fact,
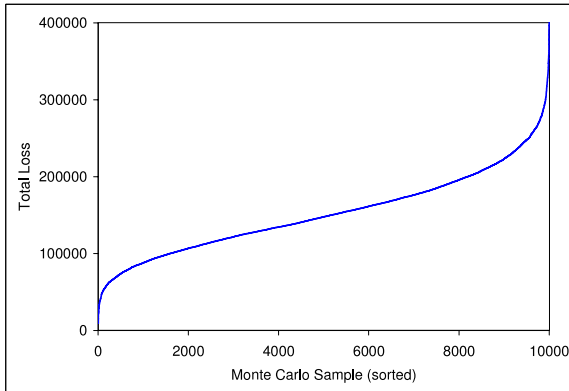
Figure 5: Total loss of each of the 10000 simulated scenarios. The 99.9% VaR reads directly from this plot as the y-axis value corresponding to the x-axis value $\lceil 0.999m \rceil = 9990$.

in the example considered, positive jumps offset most of the risk discussed in this paper. Our approach allows us to isolate the risk associate to this specific type of discrete event.)

4 Conclusion

In this paper we present a Monte Carlo approach to the computation of Value-at-Risk for systems subject to discrete losses. The main ingredients of our approach are that it model the frequency distribution of losses as a Poisson process and the severity distribution of losses using the novel fat-tail distribution discussed in this paper. The VaR is computed by performing a Monte Carlo simulation of future losses and sorting these scenarios according to their relative total loss. This procedure is very general and does not require any assumption about the distribution of the total loss.

In this paper, as an example, we have applied our technique to the computation of the VaR due to mini-crashes of a given stock portfolio.

Acknowledgements

Appendix A: Combined Log-normal $+$ Pareto distribution

The particular distribution utilized in this paper to model the severity of losses has the following form

$$p(x) = a\theta(x - x_0)\frac{1}{x}e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} + ab\theta(x_0 - x)(x + c)^{-\beta} \qquad (4)$$

This distribution is exactly log-normal for $x < x_0$, exactly Pareto for $x > x_0$, and it is continuous and differentiable everywhere, including $x = x_0$.

μ, σ, x_0 and β are free parameters, while a, b and c are coefficients that depend on those parameters via the following exact relations:

$$a = \left(\sqrt{\frac{\pi}{2}}\sigma\left(1 + \text{erf}(\frac{\log x_0 - \mu}{\sqrt{2}\sigma})\right) + \frac{b}{\beta - 1}(x_0 + c)^{1-\beta}\right)^{-1} \qquad (5)$$

$$b = \frac{(x_0 + c)^\beta}{x_0}e^{-\frac{(\log x_0 - \mu)^2}{2\sigma^2}} \qquad (6)$$

$$c = \frac{x_0\beta\sigma^2}{\sigma^2 + \log x_0 - \mu} - x_0 \qquad (7)$$

where we defined

$$\text{erf}z = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt \qquad (8)$$

Appendix B: Statistics about the sample portfolio

Average daily log-return and daily volatility in the time period starting March 16, 1992 and ending February 21, 2006.

	MSFT	RTN	HCA	EP	AVP	HNZ	USB	GD
$\mu/100$	7.2%	1.6%	3.8%	3.1%	6.3%	3.6%	5.6%	11.4%
σ	2.2%	2.1%	2.2%	2.0%	2.0%	1.4%	1.9%	1.8%

Average correlation of daily log-returns in the same period.

	MSFT	RTN	HCA	EP	AVP	HNZ	USB	GD
MSFT	-	8%	13%	13%	10%	12%	24%	16%
RTN	8%	-	13%	10%	14%	13%	17%	29%
HCA	13%	13%	-	9%	17%	18%	21%	14%
EP	13%	10%	9%	-	7%	12%	15%	12%
AVP	10%	14%	17%	7%	-	24%	20%	14%
HNZ	12%	13%	18%	12%	24%	-	22%	14%
USB	24%	17%	21%	15%	20%	22%	-	17%
GD	16%	29%	14%	12%	14%	14%	17%	-

References

[1] M. Di Pierro and A. Nandy, "Comprehensive Modeling of Operational Risk Data An Empirical Approach", (publication-pending)

[2] M Rubinstein, "Portfolio Insurance and the Market Crash", Financial Analysts Journal, (1988)

[3] J. Beirlant, J. Teugels, P. Vynckier, P. "Practical analysis of extreme values", Leuven University Press, Leuven (1996)

[4] P. Embrechts, C. Klüppelberg,T. Mikosch, "Modeling extremal events for insurance and Finance, Springer, Berlin (1997)

[5] A. Roehr, "Modelling Operational Losses", Algo Research Quanterly, Vol. 5, n.2 (2002)

[6] de Fontnouvelle, Patrick, De Jesus-Rueff, Virginia, Jordan, John S. and Rosengren, Eric S., "Using Loss Data to Quantify Operational Risk" (April 2003). http://ssrn.com/abstract=395083

[7] OpRisk Calculator, www.metacryption.com (a software tool developed by MetaCryption LLC that implements the methodology described in this paper).

This page intentionally left blank

Path dependent options: the case of high water mark provision for hedge funds

Z. Li[1] & S. S.-T. Yau[2]
[1] *Global Derivative Products, Bank of America New York, USA*
[2] *Department of Mathematics, Statistics and Computer Science,*
University of Illinois at Chicago, USA

Abstract

High Water Mark (HWM) provision is an important feature in the hedge fund industry. The framework of the option pricing with HWM provision for hedge funds is developed in this paper. The closed forms of HWM look-back put option, Russian option and stop-loss option are derived. We also obtain the internal relationship between HWM look-back put and the traditional look-back option. We show that HWM look-back put is cheaper than the traditional look-back put, and the higher the incentive fee, the lower the option price.
Keywords: high water mark options, stochastic processes, hedging.

1 Introduction

Hedge funds are pooled investment vehicles; most set up as private limited partnerships and investors buy an interest into the partnership. As such, they have more freedom and flexibility than mutual funds. In the past ten years, the number of hedge funds has risen about 20% per year. Currently, there are estimated to be 4000-5000 hedge funds managing $200-$300 billion. While the number and size of hedge funds are small relative to mutual funds, their growth reflects the importance of this alternative investment category. One important feature of the hedge funds industry is the structure of the fee paid to fund managers.

The fee in a hedge fund's account mainly comprises management fee and incentive fee. Management fee is charged on the account balance whether the account is profitable or not. Management fee normally ranges between 1%-2.5% annually,

and typically it is 2%. Hedge funds also share in profits generated in the account by charging an incentive fee based on the difference between each old and new profit high for the account. This is the so-called high water mark (HWM) provision. The up to date highest asset value in this account is the HWM. The incentive fee can only be earned by producing on-going new HWM, i.e. new profits for an account. It works this way: If the manager has an incentive fee of 20% and his/her current HWM is 10 million. Say the manager has a 50% return in one month which increases the asset value to 15 million. He/She then pockets a 1 million incentive fee and the HWM changes to 15 million. Suppose next month the asset value (15 - 1 = 14 million) shrinks to 10 million, the manager cannot collect any incentive fee. Additional incentive fees are due only to the extent the manager pushes the fund above 15 million. Moreover all commission charges and per trade transaction costs must be made up before an incentive fee is applied. Some funds have a hurdle rate provision as well, meaning that a certain level of return must be met in order to trigger the incentive allocation, or against an index such as S&P 500 or treasury rates. The incentive fee is normally between 5%-25%, with majority of 20%. These fees are usually paid from the account on a monthly or quarterly basis.

The fee structure of hedge funds has been studied intensively. Recently, Goetzmann et al. [3] use an option approach to calculate the present value of the fees charged by money managers. They show that incentive fee takes a large part. For example, for a money manager with volatility of 15%, the incentive fee can be as high as 13% of the total managed money. Fung and Hsieh [2] provide a rationale for how hedge funds are organized and they show that the incentive fee paid to successful managers can be significantly higher than the fixed management fee.

One natural question is how significant this HWM provision impacts on option pricing. Because every time a money manager charges the incentive fee, the asset value correspondingly is reduced. Option with HWM provision is clearly path dependent. And the path is changed every time a hedge fund reaches a new HWM, since a certain amount of incentive fee is charged. Surprisingly, there is not much study on this kind of option in literature. This paper sets up the framework of option pricing model with HWM provision. The closed forms of HWM look-back put option, Russian option and stop-loss option are derived. Moreover we obtain the internal relationship between HWM look-back put and the traditional look-back option. We show that HWM look-back put is cheaper than the traditional look-back put. The higher the incentive fee, the lower the option price.

2 HWM option pricing framework

We work in a continuous-time framework and assume that, in the absence of management fee, the net asset of the fund follows a lognormal diffusion process with expected rate of return μ_t and variance σ_t^2. Let g_t be the rate of the management fee. The evolution of the asset of the fund, S_t, is assumed to be the solution to a stochastic differential equation of the form

$$dS_t = (\mu_t + g_t)S_t dt + \sigma_t S_t d\omega_t, \quad S_t < H_t.$$

where ω_t is a standard Brownian motion.

In the simplest case the HWM is the highest level the asset value has reached in the past. For some incentive contracts, the HWM can be changed due to other conditions, such as some indices or treasury rates. Because they are not locally random in our model, the change of H_t is locally deterministic. So for $S_t < H_t$, the change of H_t is $dH_t = G_t H_t dt$, $S_t < H_t$ where G_t is a deterministic function defined in the HWM provision. When the asset value reaches a new high, the HWM is reset to this higher level.

Following the arguments in Black and Scholes [1] and Merton [5], by applying Ito's lemma (See Øksendal [6] and adjusting the hedge position, we can find the diffusion function of the option price. When the fund's assets are below the HWM, the option price V_t satisfies the following partial differential equation (PDE)

$$\frac{\partial V_t}{\partial t} + \frac{1}{2}\sigma_t^2 S_t^2 \frac{\partial^2 V_t}{\partial S_t^2} + (r_t + g_t)S_t\frac{\partial V_t}{\partial S_t} + G_t H_t\frac{\partial V_t}{\partial H_t} - r_t V_t = 0, \qquad S_t < H_t \quad (1)$$

The payoff function is

$$V(S_T, H_T, T) = \Lambda(S_T, H_T) \tag{2}$$

where $\Lambda(S_T, H_T)$ is defined in the contract.

Another condition applies along the boundary $S_t = H_t$. When the asset value rises above H_t to $H_t + \epsilon$, the HWM is reset to $H_t + \epsilon$, and a incentive fee of $k\epsilon$ is paid to the manager reducing the asset value to $H_t + \epsilon(1-k)$. Therefore, the option price before any adjustments of the incentive fee and HWM is $V(H_t + \epsilon, H_t, t + \Delta t)$, and the option price after the adjustments of the incentive fee and HWM is $V(H_t + (1-k)\epsilon, H_t + \epsilon, t + \Delta t)$. Let $\epsilon \to 0$ and $\Delta t \to 0$ this gives the boundary condition

$$k\frac{\partial V_t}{\partial S_t} = \frac{\partial V_t}{\partial H_t} \qquad on \ S_t = H_t \tag{3}$$

Hence PDE (1) together with conditions (2) and (3) give the solution of the option price with the HWM provision.

3 Special cases

It is not easy to get a closed form for the HWM option. The followings are some special cases, in which some closed forms are obtained. For simplicity, we drop the subscript t for convenience when it does not cause confusion.

3.1 Look-back put HWM option

A look-back put option is an option with payoff determined by the price of the asset value and the maximum value of the underlying asset within the life of the option. Look-back options can somehow capture investor's fantasy of buying low, selling high, and minimize regrets, as Goldman et al. [4] argues.

Assume G_t is zero, i.e. the HWM doesn't change when the asset value is less than the HWM. For simplicity, let $g_t = 0$. Also assume that $r = r_t$ and $\sigma = \sigma_t$ are constant, then the following PDE gives the HWM look-back put option

$$\frac{\partial V_{HWM}}{\partial t} + \sigma^2 S^2 \frac{\partial^2 V_{HWM}}{2\partial S^2} + rS\frac{\partial V_{HWM}}{\partial S} - rV_{HWM} = 0, \qquad 0 \leqslant S < H \quad (4)$$

$$V_{HWM}(S, H, T) = max(H_T - S_T, 0), \tag{5}$$

$$V_{HWM}(0, H, t) = H, \tag{6}$$

$$k\frac{\partial V_{HWM}}{\partial S} = \frac{\partial V_{HWM}}{\partial H} \qquad \text{on } S = H. \tag{7}$$

Notice the only difference between V_{HWM} and the traditional look-back put option V_{TRD} is the boundary condition at $S = H$. The traditional look-back put option is just a special case of the HWM look-back put option when the incentive fee is zero.

Proposition 1. *The HWM look-back option $V_{HWM}(S, H, t)$ has the following relationship with the traditional look-back option $V_\theta(S, H, t)$.*

$$V_{HWM}(S, H, t) = H^{\theta-1}V_\theta(S, H, t) \tag{8}$$

where $V_\theta(S, H, t)$ is a look-back option with payoff function $H^\theta(1 - S/H)$ but without HWM provision.

Proposition 1 shows deeper relationship between HWM look-back put and a traditional look-back option. Every time when the asset value reaches a new HWM, a certain amount of incentive fee is charged. Hence there is an adjustment of the asset value that changes the path with a small downside jump. While Proposition 1 shows that a HWM look-back put is nothing but a traditional look-back option with a different payoff function, with a justification of the HWM to the power of $\theta - 1$.

Proposition 2. *The HWM look-back put option price*
$V_{HWM}(S_0, H_0, 0) = H_0 W(S_0/H_0, 0)$, *where*

$$\begin{aligned}
W = &-\frac{\theta\sigma^2}{2\nu + \theta\sigma^2} e^{-rT} \left(\frac{H_0}{S_0}\right)^{\frac{2\nu}{\sigma^2}+1} N(d_1) \\
&-\frac{(1-\theta)\sigma^2}{2\nu + (1+\theta)\sigma^2} e^{(\nu-r+\frac{1}{2}\sigma^2)T} \left(\frac{H_0}{S_0}\right)^{\frac{2\nu}{\sigma^2}} N(d_2) \\
&+\left(\frac{2\nu + 2\theta\sigma^2}{2\nu + \theta\sigma^2} - \frac{2\nu + 2\theta\sigma^2}{2\nu + (1+\theta)\sigma^2}\right) e^{(\theta\nu-r+\frac{\theta^2}{2}\sigma^2)T} \left(\frac{S_0}{H_0}\right)^\theta N(d_3) \\
&-e^{(\nu-r+\frac{\sigma^2}{2})T} \left(\frac{S_0}{H_0}\right) N(d_4) \\
&+e^{-rT} N(d_5)
\end{aligned} \tag{9}$$

and

$$H_0 = \text{high water mark at time 0} \tag{10}$$

$$S_0 = \text{asset value at time 0} \tag{11}$$

$$d_1 = -\frac{ln(H_0/S_0) + \nu T}{\sigma\sqrt{T}} \tag{12}$$

$$d_2 = -\frac{ln(H_0/S_0) + (\nu + \sigma^2)T}{\sigma\sqrt{T}} \tag{13}$$

$$d_3 = \frac{ln(S_0/H_0) + (\nu + \theta\sigma^2)T}{\sigma\sqrt{T}} \tag{14}$$

$$d_4 = \frac{ln(H_0/S_0) - (\nu + \sigma^2)T}{\sigma\sqrt{T}} \tag{15}$$

$$d_5 = \frac{ln(H_0/S_0) - \nu T}{\sigma\sqrt{T}} \tag{16}$$

$$\nu = r + g - \frac{1}{2}\sigma^2 \tag{17}$$

$$r = \text{riskfree rate} \tag{18}$$

$$g = \text{management fee} \tag{19}$$

By replacing T with $T - t$, S_0 with S_t and H_0 with H_t, we verify that solution (9) satisfies PDE (4)-(7). This is exactly what we expect since both probability way and PDE way should yield the same option price.

It is neither surprising that HWM look-back put option formula (9) is more complicated than the traditional look-back put option formula . But which option is more expensive? The following proposition answers this question.

Proposition 3. *If $\theta = 1$, i.e. the incentive fee is zero, then the HWM look-back put has the same price as the traditional look-back put, i.e. $V_{HWM} = V_{TRD}$. If $\theta < 1$, i.e. the incentive fee is greater than zero, then $V_{HWM} < V_{TRD}$. Moreover, the higher the incentive fee, the lower the option price. The lower the ratio of S_0/H_0, the less impact of the HWM provision. And the price of HWM look-back put converges to the traditional look-back put option as S_0/H_0 tends to zero.*

It is a trivial case when the rate of incentive fee is zero, since it is reduced to the traditional look-back put option. Notice θ is between $\frac{1}{2}$ and 1, since the incentive fee k is between 0 and 1, and $\theta = \frac{1}{1+k}$. When the rate of incentive fee is greater than zero, every time when the asset value goes above the past HWM, a certain amount of incentive fee is charged. Hence the asset value is reduced. It is surely that it is more difficult for the asset value to reach a new high. HWM provision reduces both the asset value and the HWM. Proposition 3 reveals the different significance of the impact of HWM provision on the asset value and the corresponding HWM. It shows that the incentive fee reduces more HWM level than it does on the asset value itself. Since incentive fee only applies when S_t

reaches a new high, when S_0/H_0 is small, the probability that S_t goes above H_0 is small. Hence the less impact of the HWM provision and the price of HWM look-back put converges to the traditional look-back put option as S_0/H_0 tends to zero.

3.2 Russian HWM option

The term "Russian Option" was coined by Shepp and Shirayaev [7] to describe a perpetual American option, which, at any time chosen by the holder, pays out the maximum realised asset price up to that date. The solution of the Russion option with HWM is $V = c_1 H \xi^{\eta_1} + c_2 H \xi^{\eta_2}$. where

$$\eta_1 = \frac{-(r + g - \frac{1}{2}\sigma^2) + \sqrt{(r + g - \frac{1}{2}\sigma^2)^2 + 2\sigma^2 r}}{\sigma^2} \tag{20}$$

$$\eta_2 = \frac{-(r + g - \frac{1}{2}\sigma^2) - \sqrt{(r + g - \frac{1}{2}\sigma^2)^2 + 2\sigma^2 r}}{\sigma^2} \tag{21}$$

$$c_1 = \frac{1 - (1 + k)\eta_2}{(1 - (1 + k)\eta_2)\xi_0^{\eta_1} - (1 - (1 + k)\eta_1)\xi_0^{\eta_2}} \tag{22}$$

$$c_2 = \frac{1 - (1 + k)\eta_1}{(1 - (1 + k)\eta_1)\xi_0^{\eta_2} - (1 - (1 + k)\eta_2)\xi_0^{\eta_1}} \tag{23}$$

$$\xi_0 = \left(\frac{\eta_1 (1 - (1 + k)\eta_2)}{\eta_2 (1 - (1 + k)\eta_1)} \right)^{1/(\eta_2 - \eta_1)} \tag{24}$$

3.3 Stop-loss HWM option

A stop-loss option may be thought of as a perpetual barrier look-back with a rebate that is a fixed proportion of the maximum realised value of the asset price. If at any time t the asset price S falls to λH, where H is the HWM and $\lambda < 1$ is fixed, the option is triggered and pays off λH.

The solution to the Stop-loss HWM option is very similar to the Russian HWM option with c_1 and c_2 defined as

$$c_1 = \frac{(1 - (1 + k)\eta_2)\lambda}{(1 - (1 + k)\eta_2)\lambda^{\eta_1} - (1 - (1 + k)\eta_1)\lambda^{\eta_2}}, \tag{25}$$

$$c_2 = \frac{(1 - (1 + k)\eta_1)\lambda}{(1 - (1 + k)\eta_1)\lambda^{\eta_2} - (1 - (1 + k)\eta_2)\lambda^{\eta_1}}. \tag{26}$$

4 Numerical example

Some numerical examples on HWM look-back are given. We also compare the price of the HWM look-back put with that of the traditional look-back put.

   Table 1 shows 3 month horizon look-back put vs. incentive fee. In this example, risk-free rate $r = 5\%$, initial HWM $H_0 = 100$, initial asset value $S_0 = 100$ and volatility $\sigma = 20\%$. When the incentive fee $k = 20\%$, the HWM look-back option price is 6.49, while the traditional look-back is 6.58. The difference is about 9 basis points. It is an interesting observation that the difference is not significant. It shows that as incentive fee goes to zero, the HWM look-back put converges to the traditional look-back put. The higher the incentive fee, the lower the option price.

   Table 2 compares the HWM look-back put with the traditional look-back put with different initial asset value, assuming the HWM is 100. In this case, $r = 5\%$, $k = 20\%$, $\sigma = 20\%$. This table shows that the traditional look-back put is always more expensive than HWM look-back put. As the initial asset value tends to zero, the difference of the traditional look-back put and HWM look-back put also tends to zero.

5 Conclusion

In this paper we developed the framework of option pricing with HWM provision for hedge funds. The closed forms of HWM look-back put option, Russian option and stop-loss option are derived. We also show that HWM look-back put is cheaper than the traditional look-back put, and the higher the incentive fee, the lower the option price.

   As to our knowledge, option with HWM is not traded in the market. Partly because hedge funds industry is not well known to the public until recently. Another reason is hedge funds are free from most disclosure and regulation requirements that apply to mutual funds and banks. It is difficult to hedge an option on hedge funds.

   Hedge funds are not strangers to leverage and derivatives. It is quite natural to trade options on hedge funds. As hedge funds industry grows and more regulations on position reporting, we believe that it is a matter of time that derivative products on hedge funds themselves will also appear.

Acknowledgements

References

[1] Black, F. and Scholes M., 1973, The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, 81:637-654.
[2] Fung, William, and Hsieh, David, 1999, A Primer on Hedge Funds, *Journal of Empirical Finance*, 6:309-331
[3] Goetzmann, W.N., Ingersoll, J., and Ross, S.A., 2001, High-Water Marks and

Hedge Fund Management Contracts, Working paper, International Center for Finance, Yale School of Management

[4] Goldman, B., Sosin, H. and Gatto, A., 1979, Path Dependent Options: Buy at the Low, Sell at the High, *Journal of Finance*, 34:1111-1127.

[5] Merton, R., 1973, The Theory of Rational Option Pricing, *The Bell Journal of Economics and Management Science*, 4: 141-183.

[6] Øksendal, Bernt, 1998 (5th edition), *Stochastic Differential Equations: An introduction with applications*, Springer.

[7] Shepp, L. and Shiryaev, A.N., 1993, The Russian option: Reduced regret, *Annals of Applied Probability*, 3: 631-640.

Section 8
Time series analysis
and forecasting

This page intentionally left blank

Macroeconomic time series prediction using prediction networks and evolutionary algorithms

P. Forsberg & M. Wahde
Department of Applied Mechanics, Chalmers University of Technology, Sweden

Abstract

The prediction of macroeconomic time series by means of a form of fully recurrent neural networks, called discrete-time prediction networks (DTPNs), is considered. The DTPNs are generated using an evolutionary algorithm, allowing both structural and parametric modifications of the networks, as well as modifications in the squashing function of individual neurons.

The results show that the evolved DTPNs achieve better performance on both training and validation data compared to benchmark prediction methods. The importance of allowing structural modifications in the evolving networks is discussed. Finally, a brief investigation of predictability measures is presented.
Key words: time series prediction, recurrent neural networks, evolutionary algorithms.

1 Introduction

Prediction of time series is an important problem in many fields, including economics. Due to the high level of noise in macroeconomic time series, models involving two parts, one deterministic and one stochastic, are often used. One such method is ARIMA [1]. For one-step prediction, the results obtained by these simple predictive methods (such as exponential smoothing, which is a special case of ARIMA models), are difficult to improve much due to the high levels of noise present. However, even a small improvement can translate into considerable amounts of money for data sets that concern e.g. an entire national economy. The aims of this paper is (1) to introduce a class of generalized, recurrent neural

networks and an associated evolutionary optimization method and (2) to apply such networks to the problem of deterministic prediction of macroeconomic time series, with the aim of extracting as much information as possible, while keeping in mind that the noise in the data introduces limits on the achievable performance.

2 Macroeconomic data

Two different data sets were considered, namely US GDP (quarterly variation, from 1947, first quarter to 2005, second quarter), and the Fed Funds interest rate (monthly values, from July 1954 to July 2005). The raw GDP and interest data were first transformed to a relative difference series, using the transformation

$$Z_{\mathrm{RD}}(t) = \frac{Z_{\mathrm{raw}}(t) - Z_{\mathrm{raw}}(t-1)}{Z_{\mathrm{raw}}(t-1)}. \tag{1}$$

Next, this series was further transformed using a hyperbolic tangent transformation

$$Z(t) = \tanh(C_{\mathrm{TH}} Z_{\mathrm{RD}}(t)). \tag{2}$$

For the GDP and interest rate series transformations, the values $C_{\mathrm{TH}} = 25$ and $C_{\mathrm{TH}} = 5$ were used, respectively. The aim of the hyperbolic tangent transformation was to make the data points as evenly distributed as possible in the range $[-1, 1]$.

Both data sets were divided into a training part with M_{tr} data points, and a validation part with M_{val} data points. During training, only the results (i.e. the error) over the training data set were used as feedback to the optimization procedure (see below). The rescaled GDP data set contained 233 data points. For training, steps 16-115 were used ($M_{\mathrm{tr}} = 100$) and for validation, steps 126-225 were used ($M_{\mathrm{val}} = 100$). During training, the first 15 steps were used to initialize the short-term memory of the DTPN. A similar initialization procedure was applied during validation. For the Fed Funds data set, with 612 data points, steps 26-475 were used for training ($M_{\mathrm{tr}} = 450$) and steps 486-605 ($M_{\mathrm{val}} = 120$) were used for validation.

3 Methods for prediction

3.1 Discrete-time prediction networks

Neural networks constitute a commonly used blackbox prediction model. In most cases, feedforward neural networks (FFNNs) are used. In such networks, the computational elements (neurons) are placed in layers. The input signals (i.e. earlier, consecutive values of the time series) are distributed to the neurons in the first layer, and the output signals of those neurons are then computed and used as

input in the second layer etc. The output of a given neuron i is computed as

$$x_i(t+1) = \sigma \left(b_i + \sum_{j=1}^{N} w_{ij} y_j \right), \qquad (3)$$

where b_i is the bias term, w_{ij} are the weights connecting neuron j in the preceding layer to neuron i, N is the number of neurons in the preceding layer, and σ is the squashing function, usually taken as the logistic function

$$\sigma_1(z) = \frac{1}{1 + e^{-cz}}, \qquad (4)$$

where c is a positive constant, or the hyperbolic tangent

$$\sigma_2(z) = \tanh cz. \qquad (5)$$

Given a set of training data, i.e. a list of input vectors and their corresponding desired output, such networks can be trained using gradient-based methods, such as e.g. backpropagation.

However, there are fundamental limitations in the prediction that can be achieved using FFNNs, due to their lack of dynamic (short-term) memory. Stated differently, an FFNN will, for a given input, always give the same output, regardless of any earlier input signals [2, 3]. Thus, such networks are unable to deal with situations in which identical inputs to the network (at different times along the time series) require different outputs. Earlier work [2] has shown that dynamic short-term memory *does* make a difference in neural network-based time series prediction.

Furthermore, the requirement that it should be possible to obtain a gradient of the prediction error, in order to form the derivatives needed for updating the weights (during training), restricts the shape of the squashing functions. Without such restrictions, squashing functions such as e.g.

$$\sigma_3(z) = \text{sgn}(z), \qquad (6)$$

and

$$\sigma_4(z) = \begin{cases} \tanh(z+c) & \text{if} \quad z < -c \\ 0 & \text{if} \quad -c \leq z \leq c \\ \tanh(z-c) & \text{if} \quad z > c \end{cases} \qquad (7)$$

could be used.

To overcome the limitations of FFNNs, it is possible to introduce feedback couplings in the networks, transforming them into recurrent neural networks (RNNs). Such networks have been used in many financial and macroeconomic applications, see e.g. [3, 4]. A problem with many standard training techniques for neural networks is that they require that the user should set the structure of the network (i.e. the number of neurons and their position in the network), a

procedure for which one often has to rely on guesswork and rules-of-thumb [5]. An alternative training procedure is to use an evolutionary algorithm (EA) which, if properly designed, can handle both structural and parametric optimization [6].

In this paper, a new kind of network (and an associated evolutionary optimization method), well suited for the problem of time series prediction, will be used, with dynamical memory, arbitrary structure, and (in principle) arbitrary squashing functions. Each of the n neurons in these networks which, henceforth, will be called *discrete-time prediction networks* or DTPNs for short) contains arbitrary connections from the n_{in} input elements and from other neurons (including itself). In addition, each neuron has an *evaluation order tag* (EOT) such that, in each time step, the output of the neurons with the lowest EOT values is computed first, followed by the output of the neurons with the second lowest EOT values etc. The output neuron, i.e. the neuron with highest EOT (arbitrarily chosen as neuron 1) is evaluated last. Thus, the equations for neurons with the lowest EOT become

$$x_i(t+1) = \sigma\left(b_i + \sum_{j=1}^{n_{\text{in}}} w_{ij}^{\text{in}} I_j(t) + \sum_{j=1}^{n} w_{ij} x_j(t)\right), \tag{8}$$

where w_{ij}^{in} are the input weights, w_{ij} the interneuron weights, and b_i is the bias term. I_j are the inputs to the network which, in the case of time series prediction, consist of earlier values of the time series $Z(t)$, i.e. $I_j(t) = Z(t - j + 1)$. The number of inputs can thus be referred to as the *lookback* (L) of the DTPN. For neurons with the second lowest EOT, the equations look the same, except that $x(t)$ is changed to $x(t+1)$ for neurons with lowest EOT etc. Finally, the output neuron gives the following output

$$x_1(t+1) = \sigma\left(b_1 + \sum_{j=1}^{n_{\text{in}}} w_{1j}^{\text{in}} I_j(t) + w_{11} x_1(t) + \sum_{j=2}^{n} w_{1j} x_j(t+1)\right), \tag{9}$$

since, at this stage, all neurons except neuron 1 have been updated. It is evident that the EOTs introduce the equivalent of layers. Thus, while most DTPNs will contain many recurrent connections, an FFNN is a special case of a DTPN. More precisely, a DTPN is equivalent to an ordinary FFNN if and *only* if (1) all squashing functions are of the same type (either σ_1 or σ_2), (2) only neurons with the lowest EOT values receive external input, and (3) w_{ij} (i.e. the weight connecting neuron j to neuron i) is equal to zero if $\text{EOT}(j) \geq \text{EOT}(i)$.

3.2 Benchmark predictions

In order to evaluate the results obtained using DTPNs, a comparison will be made with two standard prediction techniques, namely *autoregressive moving average* (ARMA) and *exponential smoothing*. The general simple ARMA(p, q) model

$$\phi(\Lambda)Z(t) = \theta(\Lambda)\epsilon(t), \tag{10}$$
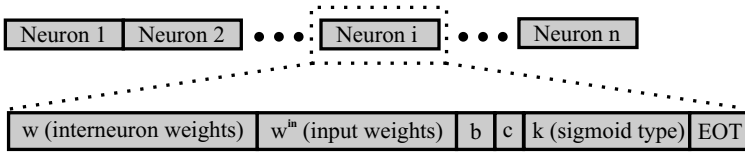
Figure 1: A chromosome encoding a DTPN.

where Λ is the lag operator, ϵ is the disturbance $Z - \hat{Z}$, and

$$\phi(\Lambda) = 1 - \phi_1\Lambda - \ldots - \phi_p\Lambda^p, \tag{11}$$

and

$$\theta(\Lambda) = 1 + \theta_1\Lambda + \ldots + \theta_q\Lambda^q, \tag{12}$$

gives the one-step prediction $\hat{Z}(t+1|t)$

$$\hat{Z}(t+1|t) = \sum_{i=0}^{p} \phi_i Z(t-i) + \sum_{i=0}^{q} \theta_i \epsilon(t-i). \tag{13}$$

ϕ_i and θ_i are parameters to be estimated in order to find the lowest error. The exponential smoothing technique (without trend) is described by the ARIMA(0,1,1) equation

$$(1 - \Lambda)Z(t) = (1 - \theta_1\Lambda)\epsilon(t). \tag{14}$$

This model gives the prediction

$$\hat{Z}(t+1|t) = \frac{1 - \theta_1}{1 - \theta_1\Lambda}Z(t) = \theta_1\hat{Z}(t|t-1) + (1-\theta_1)Z(t). \tag{15}$$

As a special case, if $\theta_1 = 0$, the *naive prediction* $\hat{Z}(t+1|t) = Z(t)$ is obtained.

4 Evolutionary algorithm

The DTPNs were generated using an evolutionary algorithm (EA) [7]. The EA used here employed a non-standard chromosomal representation, shown in Fig. 1, in which each gene represented a neuron in the network, encoding its interneuron weights (w_{ij}), input weights (w_{ij}^{in}), bias term (b_i), sigmoid parameter (c), sigmoid type, and EOT. During the formation of new individuals, crossover was only allowed between individuals containing the same number of neurons. Several different forms of mutations were used, both parametric mutations modifying the values of the parameters (including the EOT) listed above, and structural mutations which could either add or subtract a neuron from the DTPN. No upper limit was set on the number of neurons. A lower limit of 2 neurons was introduced, however. In addition to the mutations just listed, a sigmoid type mutation was introduced

as well, allowing a neuron to change its sigmoid type by randomly changing the index k of the sigmoid σ_k (see Sect. 3.1 and Eq. (17) below). Finally, in order to *allow* (not force) the EA to produce sparsely connected networks, some runs were carried out in which parametric mutations of interneuron weights, input weights, and biases not only could modify the value of the parameter in question, but also (with low probability) could set it exactly to zero. Thus, these mutations essentially functioned as on-off toggles, and were therefore called *zero-toggle mutations*. The number of input elements (and therefore the lookback L) was fixed in each run. The fitness measure F used by the EA was taken as the inverse of the RMS prediction error over the training set, i.e. $F = 1/e_{\mathrm{RMS}}$ where

$$e_{\mathrm{RMS}} = \sqrt{\frac{1}{M_{\mathrm{tr}}} \sum_{i=1}^{M_{\mathrm{tr}}} \left(Z(i) - \hat{Z}(i) \right)^2} \tag{16}$$

Note that the use of an EA implies that any form of sigmoid function can be used in the networks. In addition to the four functions $\sigma_1 - \sigma_4$, a fifth sigmoid, namely

$$\sigma_5(z) = \frac{cz}{1 + (cz)^2}, \tag{17}$$

was also allowed in the simulations reported below.

5 Prediction results

A large number of runs were carried out, using different number of inputs and different EA parameters in order to test the ability of the evolutionary algorithm to generate DTPNs with low prediction error for the two data sets under consideration.

The results are summarized in Table 1. The table shows the prediction error for the DTPN with lowest validation error. In addition, the prediction errors obtained using naive prediction, exponential smoothing, and ARMA (all with optimized parameter values), are shown.

As is evident from the table, the best DTPNs outperform the two other prediction methods. Table 2 gives a more detailed description of the best DTPNs, obtained with different values of n_{in}. For comparison, note that the best *training* errors obtained with exponential smoothing were $e_{\mathrm{ES}}^{\mathrm{tr}} = 0.2512$ for the GDP data and $e_{\mathrm{ES}}^{\mathrm{tr}} = 0.3477$ for the Fed funds data. Using the ARMA model, the best training errors were $e_{\mathrm{ARMA}}^{\mathrm{tr}} = 0.2108$ and $e_{\mathrm{ARMA}}^{\mathrm{tr}} = 0.3248$, respectively.

6 Predictability measures

The fact that the DTPNs outperform the benchmark prediction methods does not imply that these networks extract all the available information in the time series under study. One way of determining whether additional information can be extracted would be to devise a measure $P(t)$ of predictability such that, in addition

Table 1: Minimum errors over the *validation* part of the data set, obtained using naive prediction (e_N), exponential smoothing (e_{ES}), ARMA (e_{ARMA}), and DTPNs (e_{DTPN}). Only the results for the very best DTPN are shown.

Data set	e_N	e_{ES}	e_{ARMA}	e_{DTPN}
Fed funds interest rate	0.2018	0.1901	0.1887	0.1837
GDP	0.1771	0.1490	0.1473	0.1305

Table 2: Examples of the performance of evolved DTPNs. The second column shows the number of inputs to the network, and the third column shows the probability of a mutation being of the zero-toggle type, i.e. a mutation that sets the parameter in question to zero. The fourth column shows the (evolved) number of neurons, and the fifth column shows the (evolved) number of layers (n_L), i.e. the number of distinct EOT values in the evolved network. The two final columns show the errors over the training and validation parts of the data set.

Data set	n_{IN}	P_{zero}	n	n_L	e_{DTPN}^{tr}	e_{DTPN}^{val}
Fed funds, run 1	2	0.00	7	5	0.3072	0.1837
Fed funds, run 2	2	0.25	5	5	0.2968	0.1881
GDP, run 1	5	0.00	4	4	0.2095	0.1423
GDP, run 2	4	0.00	6	4	0.2173	0.1399
GDP, run 3	3	0.00	5	4	0.2131	0.1360
GDP, run 4	3	0.20	11	5	0.2094	0.1305

to the prediction $\hat{Z}(t+1)$ of the next value in the time series, one would obtain an estimate of the error $e(t+1) = Z(t+1) - \hat{Z}(t+1)$. Ideally, the measure should be such that $P(t) = f(e(t+1))$ where f is a known, monotonous function.

Several different predictability measure can be formed. The amount of (local) information in a time series can, for instance, be estimated analytically using random matrix theory, based on the correlation matrix formed from the delay matrix D [8]. In addition, various empirical measures can also be generated, based on the prediction errors obtained in previous time steps. An investigation was made involving both the analytical measure and a few different empirical measures, applied to the rescaled difference series $Z(t)$. However, in all cases, the results were negative, i.e. the proposed predictability measure showed near-zero correlation with the actual prediction error, and therefore these measures will not be described further here.
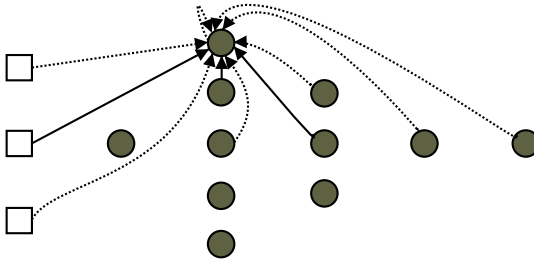
Figure 2: The best evolved network (run 4) for the prediction of the GDP series. Input elements are shown as squares and neurons as filled circles. The neurons are arranged in layers based on their EOT values. For clarity, only the inputs to one neuron are shown. Solid lines indicate positive weights and dotted lines negative ones.

7 Discussion and conclusion

This investigation has shown that it is possible to improve, albeit only slightly, the predictions obtained from standard prediction methods using a generalized version of neural networks (called discrete-time prediction networks, DTPNs) with the possibility of adding a short-term memory through feedback couplings.

In earlier work [2], continuous-time recurrent neural networks were considered for time series prediction. The DTPNs introduced here do not require continuous-time integration, i.e. the network output is obtained by discrete-time equations rather than differential equations, making the evaluation of the networks much faster, while still allowing a rich dynamical structure, including dynamic short-term memory.

The use of an EA for the optimization of the networks removes all restrictions regarding both the behavior of individual neurons as well as the structure of the network as a whole, while still allowing standard feedforward neural networks as a special case.

The importance of structural modifications in the network is illustrated by the fact that, in any given run, the structure of the current best network varied significantly during the run. The final networks often contained rather few neurons and used only a few input elements, illustrating another advantage of using recurrent networks: because of their ability to form a short-term dynamic memory, such networks need not use as many inputs as a feedforward network, thus also reducing the number of networks weights and hence the risk of overfitting.

The best network for prediction of the GDP series, shown in Fig. 2, had a slightly more complex structure. However, in the run generating that network, zero-toggle mutations were used, and indeed the resulting network was far from fully connected, and therefore had, in fact, a somewhat simpler structure than would have been suspected on the basis of the number of neurons involved.

The fact that the predictability measures all gave negative results was expected, and it indicates that the DTPNs really do extract all, or almost all, information available in the time series.

References

[1] Harvey, A., *The econometric analysis of time series*. London School of Economics handbooks in economic analysis, New York; London: Philip Allan, 2nd edition, 1990.

[2] Hulthén, E. & Wahde, M., Improving time series prediction using evolutionary algorithms for the generation of feedback connections in neural networks. *Proc. of Comp. Finance 2004*, 2004.

[3] Giles, C.L., Lawrence, S. & Tsoi, A.C., Noisy time series prediction using a recurrent neural network and grammatical inference. *Machine Learning*, **44(1/2)**, pp. 161-183, 2001.

[4] Tino, P., Schittenkopf, C. & Dorffner, G., Financial volatility trading using recurrent neural networks. *IEEE-NN*, **12**, pp. 865-874, 2001.

[5] Herbrich, R., Keilbach, M., Graepel, T., Bollmann-Sdorra, P. & Obermayer, K., Neural networks in economics: Background, applications and new developments. *Advances in Computational Economics*, **11**, pp. 169-196, 1999.

[6] Yao, X., Evolving artificial neural networks. *Proc of the IEEE*, **87**, pp. 1423-1447, 1999.

[7] Bäck, T., Fogel, D. & Michalewicz, Z., *Handbook of Evolutionary Computation*. Institute of Physics Publishing and Oxford University, 1997.

[8] Ormerud, P., Extracting information from noisy time series data. *Technical report*, Volterra Consulting Ltd, 2004.

This page intentionally left blank

Power Coefficient – a non-parametric indicator for measuring the time series dynamics

B. Pecar
University of Gloucestershire, UK

Abstract

Time series resume various shapes and exhibit varying dynamics when presented as graphs. Partially, this dynamic is measured via the Hurst exponent and a related metric called the fractal dimension. Both metrics are not sufficiently sensitive to discriminate variations in time series dynamics. To address this need, a non parametric indicator, called the Power Coefficient (the PC coefficient) is suggested. The coefficient was calculated for a number of artificially generated and the real life time series to identify how selective it is. The results indicate that, although it is sensitive to the length of the time series, it is far superior to the existing metrics.

Keywords: time series analysis, dynamics metrics, the Hurst exponent, the fractal dimension, the Power Coefficient, the PC coefficient.

1   Introduction

A time series is a snapshot, contained in a time window, of the historical behavioural pattern of the observed variable. Various methods are used as descriptive and predictive tools related to time series, and we do not intend to focus on them (for further discussion, see Mills [1], Brockwell and Davis [2], Dunis [3], Enders [4], Kantz and Schreiber [5]). In order to apply a particular method, certain assumptions need to be made. These assumptions are explicitly, or implicitly, linked with the properties of the time series. The usual tests, or assumptions made, are mainly related to the linearity issues, stationarity, serial correlation, process memory and other similar assumptions. Nevertheless, it remains the fact that the general appearance of the time series will influence the choice of the method. The question associated with this is: how do we measure the appearance of the time series? This paper will focus on visual properties of

time series and will attempt to classify them in accordance with an indicator that was specifically designed for such a purpose.

2   Background

Any conventional textbook in quantitative methods provides a variety of descriptive statistics that characterise a data set and which can, by default, be applied to time series, which are just one specific form of data sets. However, most of these statistics (the mean, variance, skewness, kurtosis, etc.) are fixed and applicable to stationary time series. The majority of the real-life time series fall into a non-stationary category, rendering effectively these descriptive statistics of little value. However, several other attempts were made to encapsulate complex time series behaviour into a single coefficient. One of the most successful attempts can be linked with the work H. E. Hurst conducted towards the middle of the last century in Egypt.

In 1951 Hurst [6] published a seminal paper in which he tried to establish the correct storage required in the Great Lakes of the Nile River Basin. His objective was to determine the capacity of a reservoir guaranteeing a minimum discharge when the water intake is low. To determine the optimum capacity, he examined the maxima and minima of a number of reservoirs over a number of periods and concluded that there is an optimum range required to maintain the average discharge. When this range was measured against different time horizons, he discovered strong scaling properties. This scaling property is estimated as the slope of the line on a log plot showing the changes in the rescaled range versus the number of cases used for the range (Range is a function of deviations from the mean. Rescaling the ranges means that they are divided by their appropriate standard deviation, i.e. R/σ.). Contrary to his expectation that this slope should be equivalent to 0.5 (which would be normal for random events), he discovered that the number is around 0.7 (to be precise it is 0.73). In other words, the scaling of certain natural phenomena (reservoirs included) does not follow chance, but shows an increasing amount of persistence and long process memory.

Benoit Mandelbrot [7] took Hurst's ideas as a foundation for making a further breakthrough in 1965. He renamed the rescale analysis exponent to the Hurst exponent and changed its original notation from K, as used by Hurst, to H. More importantly, he linked it with the fractal dimension and established that they are related as $d_F = 2 - H$, where d_f is the fractal dimension.

Mandelbrot established that (in particular for self-afine data sets) the smaller the value of the Hurst exponent, the rougher the surface of the curve describing the data set. As the Hurst exponent takes values $0<H<1$, this implies that for H close to zero, the series will closely resemble a white noise process, whilst for H close to unity, the series will be still jumping up and down, but with the element of persistence built into it. The Hurst exponent, therefore, indicates both how long the series memory is and how rough the series is. As Mandelbrot [8] put it "… fractional noises with a high value of H are the most violently fluctuating among fractional noises.". Until today the Hurst exponent and the fractal

dimension remain the only two complementary methods of measuring the roughness of the data set and its capacity to fill in the plane.

Although Mandelbrot provided the relationship to estimate the Hurst exponent from the fractal dimension (usually it is other way round), it is also possible to infer the Hurst exponent from the spectral density exponent β. Spectral densities S_x have the approximate value of:

$$S_x \propto \frac{1}{f^{\beta}} \qquad (1)$$

where, f are the frequencies and the exponent β = 2H + 1, provides a link with the Hurst exponent. However, there are several problems with calculating the Hurst exponent. First of all, the calculation algorithms are quite time-consuming and not so easy to implement without introducing some complex looping in the code. A very good freeware, called Fractan, for calculating several invariant measures, including the Hurst exponent, was created by Vyacheslav Sychyov and it is available from: http://impb.psn.ru/~sychyov/html/cv_e.shtml. The second point is that the Hurst exponent is not calculated explicitly, it is estimated. This fundamentally can cause some difficulties and misinterpretations.

As the Hurst exponent is one of the ways of estimating the fractal dimension d_F, we can say that essentially neither the Hurst exponent nor the fractal dimension discriminate enough the appearances of the time series. Essentially, as stated by Mandelbrot [7], for time series it is expected that the fractal dimension d_F will take values between one and two. The closer d_F is to one (i.e. the closer H to one is), the smoother the line. Conversely, the closer d_F is to two (i.e. the closer H to zero is), the more dynamic the line is. If the value of $d_F = 2$, we say that the time series is equivalent to the white noise process. For the value of $d_F = 1.5$ we have a special case called ordinary Brownian motion. Any other value of d_F indicates fractional Brownian motion, i.e. for the values of $1.5 < d_F < 2$, the process is anti-persistent and for the values of $1 < d_F < 1.5$, the process is persistent, exhibiting long memory (Persistent process implies autocorrelation function following exponential decay, whilst anti-persistent process implies negative correlation between observations, i.e. every subsequent observation is more likely to go in the opposite direction from the previous one rather than follow the trend).

These concepts can also be translated into a visual appearance of a time series. The closer the value of H to 1 and the closer the value of d_F to 1, the less dynamic (erratic) the time series. The closer the value of H to 0 and the closer the value of d_F to 2, the more dynamic the time series and the more 'jumpy' it looks.

However, even some elementary experimentation can quickly demonstrate that the Hurst exponent and the fractal dimension are not sensitive enough and discriminating enough, unless we handle random walks. In addition to this, the

value of the Hurst exponent has meaningful interpretation only for the random processes. For any other type of series it is inconsistent and ambiguous.

Although both coefficients (the Hurst exponent and the fractal dimension) can be used to describe some elementary properties of time series, this would constitute the most elementary feature descriptor if a time series was used as an element in some sort of repository, such as a Case Based Reasoning (CBR) system. Unfortunately, the emphasis should be on the word "elementary", as it is impossible to classify time series precisely in accordance with different values of the Hurst exponent. The case features have to be of higher resolution in order to secure better sensitivity and diagnostic/predictive properties. This paper will explore an alternative, a simple and non-parametric method for measuring the roughness of a time series.

3   The Power Coefficient (PC)

To suggest a measure of a time series appearance, we start with intuitive reasoning. In a time series the first and the last observation in a data set constitute the boundaries of the time window, which characterises certain behaviour of the data set. The shortest distance between these two points is a smooth, straight line. However, in reality a time series will exhibit some dynamics in this observed time window and will seldom follow the straight line. Very often, the first and the last observation have no effect on the overall behavioural pattern of the variable in this time window. It is more likely that, moving from period to period, the actual observations will deviate from the shortest path and will, in a way, wander or rumble through the space, creating a wiggly curve that characterises the behaviour of this variable in a given time window. The amount of curve wiggling will, in a way, show the 'energy' of the time series.

As the curve wanders, the implication is that the observations will travel certain distance which will inevitably be greater that the shortest path that connects the end points of the time series. The total distance that the variable will travel in a time window is the cumulative value of all the observation differences. Dividing it by the number of observations in a data set will give us the average travel distance between any two periods for this particular time window.

We can also observe that in a time series, at least one observation will inevitably be the maximum value for the given time window. Equally, at least one observation will be the minimum value, unless we deal with a horizontal line, in which case both the max and the min are the same. If a data set displays maximum level of dynamics for the given time window, then it would expect it to 'jump' from the minimum to maximum and back between any two periods in the series. In other words, by jumping up and down from minimum to maximum, the series would display maximum possible dynamics, or, maximum amount of energy embedded in its character. This means that the difference between the maximum and minimum value of the time series is a hypothetical average distance that this time series would travel if it had the most dynamic history.

These two values, the actual average distance and the hypothetical average distance, can be used to calculate a measure of energy the time series exhibits, and we will, therefore, name it tentatively as the Power Coefficient (PC). This coefficient is simply calculated as:

$$PC = \frac{\left| \dfrac{\sum_{i=1}^{k} m_i}{n} \right|}{\left| x_{max} - x_{min} \right| + \varepsilon} \times 100 \qquad (2)$$

where,
x_i = actual observation
m_i = distance between two observations i.e. observation difference $m_i = x_i - x_{i-1}$
x_{min} = minimum value in the time series
x_{max} = maximum value in the time series
ε = arbitrarily small constant
k = number of observations in an interval, $2 < k < n$, (some reasonable number)
n = total number of observations

The Power Coefficient (PC) will take various values for different patterns and theoretically if a variable oscillated from minimum to maximum between any two periods, its maximum value is 100. On the other hand, zero is the minimum value that the PC will show, which applies to horizontal straight line only. Strictly speaking this is not true because in this case both max and min value are the same, yielding ∞. To avoid this inconvenience, present only in exceptionally small number of cases, an arbitrarily and infinitesimally small constant ε was introduced in the denominator.

In order to demonstrate how the value of the PC changes as the dynamics of the time series changes, we conducted experiments with 41 different time series. The list of all 41 time series and their PC values are given in Table 1. However, before we draw any conclusions, we need explain how some of the series were generated.

4   Data sets and results

The PC coefficient was experimentally applied to a number of time series. The experiments involved 41 different time series, such as: smooth curves, chaotic attractors, artificially generated random walks, periodic curves (sinusoids, see-saw curves and Wierestrasse curve), periodic curves with added random elements, stationary random processes (the white noise and ARCH) and the actual time series from the New York Stock Exchange (NYSE).

Four companies were arbitrarily picked from the NYSE, and they are BP, IBM, Coca Cola (KO) and Pfizer (PFE). Their daily and minute closing values

were captured, varying in appearance from the white noise look to a random walk type of behaviour. Their returns for both daily and minute stock values were also calculated, generating time series that resemble either a white noise, or an ARCH process. The returns on closing values of stocks were calculated as:

$$R_t = \log X_{t+\Delta t} - \log X_t \qquad (3)$$

where, X_t are stock values (either daily or minute values) and Δt is the time difference (a lag). This implies that for $\Delta t = 1$, i.e. for lag 1, we get short term returns (i.e. returns after one minute or one day, depending on the resolution of the data set) and for larger number of lags, some longer return value is obtained.

As a reference point, the Hurst exponents were also estimated for all 41 data sets and their fractal dimensions inferred. Essentially, the fractal dimensions d_F fell into three characteristic regimes:

1) d_F around 1 – All artificially generated random walks, all stock closing values (both daily and minute) and most of the smooth curves

2) d_F around 1.5 – All stock returns (both daily and minute), Lorenz attractor, one smooth curve, see-saw curve with added random component, sunspot numbers, the Wierestrass curve and the white noise process

3) d_F around 2 – All the sinusoids, MinMax series, ARCH process, Rossler and Henon attractor

Visual inspection of the graphs depicting the data sets that fall in the same d_F category indicates that the fractal dimension (or the Hurst exponent) does not discriminate these sufficiently and that it does not provide meaningful clues about the fundamental features of every time series, to say the least. The PC values, on the other hand, provide much better resolution for this problem.

The results in Table 1 indicate certain tendencies, and we have sorted the values in the ascending order of the PC coefficient. For all the smooth curves, the value of the PC is always 0.1 (for the horizontal line the value is zero and for all other smooth curves the part behind the decimal point depends on the value of the constant ε), and for the series number 41, which oscillates between its maximum and its minimum value every two periods, the PC is 100, as expected.

In general, the PC coefficients show seven characteristic regimes:

A) PC close to zero – All smooth curves

B) PC between 0 and 1 – All *closing* stock values and all artificially generated random walks

C) PC between 1 and 3 – The Weirestrasse curve, nonstationary see-saw curve and two stock minute return series (BP and Coca Cola)

D) PC between 3 and 5 – Majority of the *stock return* series (Coca Cola and IBM daily returns, and IBM and Pfeizer minute returns)

E) PC between 5 and 10 – The Lorenz and Rossler attractor, ARCH process, low frequency sinusoids and two daily return stock series (BP and Pfeizer)

F) PC between 10 to 20 – Majority of sinusoids and various see-saw curves
G) PC above 30 – The white noise process and the Henon attractor. MinMax series, as an exception, has the maximum value of 100.

Table 1:   Calculation of the PC values for a number of time series.

No	Series	Observations	PC	No	Series	Observations	PC
1	Horizontal Line	1000	0.000	21	Weierestrass curve	6001	2.038
2	Linear Nonstationary Trend	1000	0.100	22	See-Saw nonstationary	114	2.564
3	Power curve	1000	0.100	23	IBM minute returns	5861	3.630
4	Expontial curve	1000	0.100	24	Coca Cola daily returns	8422	3.634
5	Logarithmic curve	1000	0.100	25	IBM daily returns	10404	4.279
6	Coca Cola daily stock value	8423	0.257	26	Pfeizer minute returns	5472	4.568
7	IBM daily stock values	10405	0.306	27	Sunspots avg monthly num	3042	4.840
8	BP daily stock values	6653	0.391	28	Low frequency (LF) sinusо	873	5.561
9	Coca Cola minute stock va	5473	0.398	29	ARCH	5200	6.854
10	Pfeizer daily stock values	5391	0.429	30	Rossler attractor	1000	6.896
11	Random walk 3	8000	0.474	31	Pfeizer daily returns	5390	6.954
12	Random walk 5	8000	0.512	32	Lorenz attractor	1000	7.821
13	Random walk 1	8000	0.589	33	BP daily returns	6652	8.030
14	Random walk 2	8000	0.593	34	See-Saw stationary	1009	11.111
15	BP minute stock values	5458	0.615	35	LF sinusoid with white noi	873	11.520
16	Pfeizer minute stock values	5473	0.687	36	HF sinusoid with white noi	873	15.358
17	US Discount rates (52.8 ye	634	0.740	37	High frequency (HF) sinus	963	16.677
18	Random walk 4	8000	0.795	38	See-Saw stationary + rand	1009	18.507
19	BP minute returns	5457	1.509	39	White noise	1500	33.531
20	Coca Cola minute returns	5472	1.745	40	Henon attractor	1000	38.229
				41	MinMax curve	1000	100

Table 1 shows that, for practical reasons, we were not consistent with the sample sizes of the analysed data sets and that the lengths vary dramatically. In addition to this, some inconsistencies are visible, for example, the nonstationary see-saw series seems to be in a completely different category than the other see-saw series. It is very likely that the number of observations taken into account play a major role. To address this question we briefly explored how sensitive this coefficient is to the sample size, i.e. the length of the data set.

5   Sensitivity results

To explore this issue experimentally, we used only several data sets (see Table 2) and calculated their PC values for varying sample sizes. The sample size 1 in Table 2 implies that the PC value was calculated for the full length of the given time series. The sample size 0.5 means that the series was split into two half and the average PC value was calculated on the basis of these two half's PC values. The sample size 0.2 means that the series was split into five sub-samples and the average PC value was calculated on the basis of their PC values. And finally, the sample size 0.1 means that the series was split into ten sub-samples and the average PC value was calculated on the basis of their PC values.

The factors in the table imply that as the sample size goes down, the value of the PC goes up. To use an example of the Weirestrass curve, we can see that the value of the PC for the whole data set was 2.038. The samples that are ten times smaller (column under 0.1), show the value of PC 2.133 larger than the large data set, i.e. their estimated PC is 4.349, as opposed to 2.038.

As expected, for stationary random processes (ARCH and white noise) the sample size will not affect the value of PC. On the other end of the spectrum, for processes with long memory (Wiener process, daily and minute stock values), if the sample size is one tenth of what it actually should be, then the PC value could be overestimated between 4 to 5 times. For any other process in between these two extremes (in this experiment the daily and minute returns and the Weirestrasse curve), the overestimation of the PC value in case of a sample that is 10% of the actual series, is approximately 2 to 3 times.

We recall from the d_F regimes that Rossler and Henon attractors, as well as the ARCH process, fell into the same category. By observing the visual appearance of these time series we instinctively see that this does not make sense. On the other hand according to the PC values, Henon attractor is by far the most dynamic, even more dynamic than the white noise, whilst the Lorenz and Rossler attractors are clearly in a different league with much lower level of dynamics exhibited. This is much more in line with the visual appearance of these time series.

Table 2:   Factor by which the PC value changes as the sample size reduces.

Series	No of observations	PC value for all observat.	Sample	size		
			1	0.5	0.2	0.1
White noise	1500	33.531	1	1.000	0.985	1.001
ARCH	5200	6.854	1	1.000	1.001	1.006
Weierestrass curve	6001	2.038	1	1.000	1.995	2.133
IBM min returns	5861	3.63	1	1.242	1.525	2.635
KO daily returns	8422	3.634	1	1.535	2.354	2.965
PFE min values	5473	0.687	1	1.246	2.253	3.428
Random walk	8000	0.474	1	1.747	2.792	4.084
BP daily values	6653	0.393	1	1.934	3.365	4.842

It is interesting to see how the actual stock values series vs. stock returns series were classified. Fractal dimension d_F clearly differentiates between the return values as opposed to actual stock values. The PC does the same. In fact, the PC discriminates between three regimes of returns. Some minute returns fall into the same category as the Weirestrasse curve and nonstationary see-saw type of curves. Some daily returns are close to chaotic attractors (the PC values close to the Lorenz and Rossler attractor). A majority of both minute and daily returns fall in the category between these two regimes.

6   Conclusions

The coefficient we suggested in this paper, the PC coefficient, is a nonparametric measure of the dynamics of the time series. It discriminates different time series in accordance with the amount of distance the time series travels from first to the last observation. For nonstationary time series resembling random walks, the coefficient is very sensitive to the length of the series, i.e. number of

observations. Nevertheless, the coefficient provides a much better classification properties than the Hurst exponent and/or the fractal dimension.

References

[1] Mills, T.C., *The Econometric Modelling of Financial Time Series*. 1994, Cambridge: Cambridge University Press. 247

[2] Brockwell, P.J. and R.A. Davis, *Introduction to Time Series and Forecasting*. 2002, New York: Springer-Verlag

[3] Dunis, C., *Forecasting Financial Markets*. 1996, Chichester: John Wiley. 292

[4] Enders, W., *Applied Econometric Time Series*. 2004, Honoken: John Wiley.

[5] Kantz, H. and T. Schreiber, *Nonlinear Time Series Analysis*. 2000, Cambridge: Cambridge University Press. 304

[6] Hurst, H.E., *Long-term storage capacity of reservoirs.* Transactions of the American Society of Civil Engineers, 1951. **116**: p. 770-808

[7] Mandelbrot, B., *The Fractal Geometry of Nature*. 1983, New York: W H Freeman & Co. 468

[8] Mandelbrot, B.B., *Gaussian Self-Affinity and Fractals*. 2002, New York: Springer Verlag. 654

This page intentionally left blank

Author Index

WITPRESS

Risk Analysis V

Computer Simulation in Risk Analysis and Hazard Mitigation V

Edited by: V. POPOV, Wessex Institute of Technology, UK, C. A. BREBBIA, Wessex Institute of Technology, UK

Covering a series of important topics, which are of current research interest and have practical applications, this book examines all aspects of risk analysis and hazard mitigation, ranging from specific assessment of risk to mitigation associated with both natural and anthropogenic hazards. Originally presented at the Fifth International Conference on Computer Simulation in Risk Analysis and Hazard Mitigation, the papers cover topics such as: Risk Mitigation; Estimation of Risk; Hazard Prevention; Management and Control; Data Collection and Analysis; Information Society Technologies in Risk; Man-made Risk; Seismic Hazard; Marine and Maritime Risk; Landslides and Slope Movements; Floods and Droughts; Soil, Water and Air Contamination; Health Issues; Policy and Decision Making; Risk and Sustainability and Operational Issues such as Energy Response; Risk Communication; Risk Perception.

WIT Transactions on Ecology and the Environment, Volume 91
**ISBN: 1-84564-172-8   2006   apx 400pp
apx £145.00/US$265.00/€217.50**

Data, Text and Web Mining and their Business Applications and Management Information Engineering

Edited by: A. ZANASI, TEMIS Text Mining Solutions SA, Italy,
C. A. BREBBIA, Wessex Institute of Technology, UK, N. F. F. EBECKEN, COPPE/UFRJ, Brazil

This book publishes papers from the Seventh International Conference on Data Mining and Information Systems. The book brings together state-of-the-art research results and practical development experiences from researchers and application developers from many different areas. The book covers topics as diverse as: Data Mining Themes such as Text Mining; Web Content, Structure and Usage Mining; Clustering Technologies; Categorisation Methods; Link Analysis; Data Preparation; Applications in Business, Industry and Government; Customer Relationship Management; Competitive Intelligence; Applications in Science and Engineering; Mining Geospatial Data; Business Process Management (BPM); Data Mining Inspired by Nature; National Security, and also Management Information Engineering such as Enterprise Information Systems; Applications of GIS and GPS; Applications of MIS; Remote Sensing; Information Systems Strategies and Methodologies; Hydro and Geo Informatics, Transportation; Bio Informatics; Biodiversity Information Systems.

WIT Transactions on Information and Communication Technologies, Volume 37
**ISBN: 1-84564-178-7   2006   apx 500pp
apx £180.00/US$325.00/€270.00**

Data Mining in E-Learning

Edited by: C. ROMERO, Universidad de Cordoba, Spain, S. VENTURA, Universidad de Cordoba, Spain

The development of e-learning systems, particularly, web-based education systems, has increased exponentially in recent years. Following this line, one of the most promising areas is the application of knowledge extraction. As one of the first of its kind, this book presents an introduction to e-learning systems, data mining concepts and the interaction between both areas.
Series: Advances in Management Information Vol 4
ISBN: 1-84564-152-3 2006 apx 315pp apx £110.00/US$197.00/€165.00

The Internet Society II

Advances in Education, Commerce and Governance

Edited by: K. MORGAN, University of Bergen, Norway, C. A. BREBBIA, Wessex Institute of Technology, UK, J. M. SPECTOR, Florida State University, USA

This book explores the technical, social and cultural implications of the emerging Information and communication technologies, addressing the technological and scientific development within education, commerce, governance, and security with a special emphasis on the impact on individuals, culture and society. Bringing together papers from the Second International Conference on Advances in Education, Commerce & Governance: Technology's Impact on Individuals, Culture and Society, the text will be of interest to researchers and academics working in areas related to the social, psychological and cultural impact of information communications technology (ICT). Specifically the book addresses a wide range of topics as diverse as: E-Commerce and E-Governance; Data and Information Privacy; Psychology; Gender; Culture; New Learning.
WIT Transactions on Information and Communication Technologies, Volume 36
ISBN: 1-84564-170-1 2006 apx 450pp apx £165.00/US$295.00/€247.50

Implementing Data Mining Algorithms in Microsoft SQL Server

C. L. CUROTTO, CESEC/UFPR Civil Engineering Graduate Program, , N. F. F. EBECKEN, COPPE/UFRJ, Brazil

Data mining technology is growing rapidly and, in many industries, has become commonplace. The main reasons for this are advances in computer technology that acquire, store and retrieve enormous amounts of data about everything from everywhere.
In 2000, Microsoft introduced a data mining feature in their Microsoft® SQL Server™ 2000 Analysis Services. This book covers all the practicalities required to integrate a third-party data mining algorithm into SQL Server 2000.
This book is designed for use by data mining researchers and information technology workers. It will also be an ideal text for IT master and doctorate courses.
Series: Advances in Management Information Vol 3
ISBN: 1-84564-037-3 2005 160pp+CD-ROM £89.00/US$142.00/€133.50

WIT*PRESS*

Handling Missing Data

Applications to Environmental Analysis

Editors: G. LATINI and G. PASSERINI, Universita Politecnica delle Marche, Italy

Featuring a wide range of techniques for analysing and filling gaps in time series data, this book contains recent research by the Air Pollution and Environmental Fluid Dynamics Group at Marche University of Technology, Ancona. The book is accompanied by a CD-ROM which includes two programs, a Toolbox for Air Quality Data Filling and a MATLAB© Neural Network Tool Collection.
Series: Advances in Management Information, Vol 1
**ISBN: 1-85312-992-5   2004
200pp+CD-ROM
£91.00/US$145.00/€136.50**

Data Mining VI

Data Mining, Text Mining and their Business Applications

*Edited by: A. ZANASI, TEMIS Text Mining Solutions SA, Italy,
C. A. BREBBIA, Wessex Institute of Technology, UK, N. F. F. EBECKEN, COPPE/UFRJ, Brazil*

This book contains most of the papers presented at the Sixth International Conference on Data Mining held in Skiathos, Greece. Twenty-five countries from all the continents are represented in the papers published in the book, offering a real multinational and multicultural range of experiences and ideas.
There has been an explosion of interest in data mining applications to unstructured data

and this is reflected in a large increase in text mining. Consequently it is not difficult to forecast that in the next months several applications will appear dedicated to analysis of content coming from the billions of available web pages, newsgroups, emails, chat lines and message boards, and that the interest in unstructured data mining and text mining will grow amongst researchers, OEM and system integrators working in sectors as information retrieval, semantic web, linguistics, and knowledge management. In business applications, the most promising areas are those regarding National Security, Competitive Intelligence and Customer Relationship Management. They can all be incorporated into the area of Intelligence Analysis, of especial interest for textual data mining and text mining applications.
Contents include: Data mining; Text mining; Neural networks and decision trees; Link analysis; Clustering and categorization; Consumer and strategic intelligence; Applications in science, engineering and life sciences, and Applications in business, industry and government.
WIT Transactions on Information and Communication Technologies, Volume 35
**ISBN: 1-84564-017-9   2005   568pp
£199.00/US$318.00/€298.50**

WIT eLibrary

Home of the Transactions of the Wessex Institute, the WIT electronic-library provides the international scientific community with immediate and permanent access to individual papers presented at WIT conferences. Visitors to the WIT eLibrary can freely browse and search abstracts of all papers in the collection before progressing to download their full text.

Visit the WIT eLibrary at
http://library.witpress.com