New Economic Windows

Frédéric Abergel
Bikas K. Chakrabarti
Anirban Chakraborti
Nivedita Deo
Kiran Sharma *Editors*

New Perspectives and Challenges in Econophysics and Sociophysics

Springer

New Perspectives and Challenges in Econophysics and Sociophysics

New Economic Windows

Frédéric Abergel · Bikas K. Chakrabarti ·
Anirban Chakraborti · Nivedita Deo ·
Kiran Sharma
Editors

New Perspectives and Challenges in Econophysics and Sociophysics

Springer

Editors
Frédéric Abergel
Centrale Supélec
Gif-sur-Yvette, France

Bikas K. Chakrabarti
Saha Institute of Nuclear Physics
Kolkata, India

Anirban Chakraborti
Jawaharlal Nehru University
New Delhi, India

Nivedita Deo
University of Delhi
New Delhi, India

Kiran Sharma
Jawaharlal Nehru University
New Delhi, India

ISSN 2039-411X ISSN 2039-4128 (electronic)
New Economic Windows
ISBN 978-3-030-11363-6 ISBN 978-3-030-11364-3 (eBook)
https://doi.org/10.1007/978-3-030-11364-3

Library of Congress Control Number: 2018967429

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

The search for truth should be the goal of our activities; it is the sole end worthy of them.

—Henri Poincaré

Preface

This volume contains essays that were mostly presented in the joint international workshop entitled "Econophys-2017" and "APEC-2017," held at the Jawaharlal Nehru University and University of Delhi, New Delhi, during November 15–18, 2017. For the first time, the Econophys series and the Asia Pacific Econophysics Conference (APEC) series merged together to have a great workshop, which was organized jointly by the Jawaharlal Nehru University, University of Delhi, Saha Institute of Nuclear Physics, and CentraleSupélec. We received great support and encouragement from the steering committee of the APEC.

Economic and financial markets appear to be in a permanent state of flux. Billions of agents interact with each other, giving rise to complex dynamics of economic quantities at the micro- and macro-levels. With the availability of huge data sets, researchers are able to address questions at a more granular level than were possible earlier. Fundamental questions of aggregation of action and information, coordination, complexity, and evolution of economic and financial networks have received significant importance in the current research agenda of the Econophysics literature. In parallel, the Sociophysics literature has focused on large-scale social data and their inter-relations. Empirical approach has become a front-runner in finding short-lived patterns within the data. The essays appearing in this volume include the contributions of distinguished experts and researchers and their co-authors from varied communities—economists, sociologists, financial analysts, mathematicians, physicists, statisticians, and others. A positive trend for this interdisciplinary track is that more and more sociologists, economists, and statisticians have started interacting with the physicists, mathematicians, and computer scientists! Evidently, most have reported their recent works and reviews on the analyses of economic and social behaviors. A few papers have been included that were accepted for presentation but were not presented at the meeting since the contributors could not attend due to unavoidable reasons. The contributions are organized into three parts. The first part comprises papers on "Econophysics". The papers appearing in the second part include studies in "Sociophysics". Finally, the third part is Miscellaneous, containing a proposal for an Interdisciplinary research center, and an "Epilogue", which discusses the advent of "Big data" research.

We are grateful to all the local organizers and volunteers for their invaluable roles in organizing the meeting, and all the participants for making the workshop a success. We acknowledge all the experts for their contributions to this volume. We also thank Vishwas Kukreti, Arun Singh Patel, Hirdesh Kumar Pharasi, and Amrita Singh for their help in the LaTeX compilation of the articles. The editors are also grateful to Mauro Gallegati and the Editorial Board of the New Economic Windows series of the Springer-Verlag (Italy) for their continuing support in publishing the Proceedings in their esteemed series.[1] The conveners (editors) also acknowledge the financial support from Jawaharlal Nehru University, University of Delhi, and CentraleSupélec.

Gif-Sur-Yvette, France — Frédéric Abergel
Kolkata, India — Bikas K. Chakrabarti
New Delhi, India — Anirban Chakraborti
New Delhi, India — Nivedita Deo
New Delhi, India — Kiran Sharma
September 2018

[1] Past volumes:

1. Econophysics and Sociophysics: Recent Progress and Future Directions, Eds. F. Abergel, H. Aoyama, B. K. Chakrabarti, A. Chakraborti, N. Deo, D. Raina and I. Vodenska, New Economic Windows, Springer-Verlag, Milan, 2017.
2. Econophysics and Data Driven Modelling of Market Dynamics, Eds. F. Abergel, H. Aoyama, B. K. Chakrabarti, A. Chakraborti, A. Ghosh, New Economic Windows, Springer-Verlag, Milan, 2015.
3. Econophysics of Agent-based models, Eds. F. Abergel, H. Aoyama, B. K. Chakrabarti, A. Chakraborti, A. Ghosh, New Economic Windows, Springer-Verlag, Milan, 2014.
4. Econophysics of Systemic Risk and Network Dynamics, Eds. F. Abergel, B. K. Chakrabarti, A. Chakraborti and A. Ghosh, New Economic Windows, Springer-Verlag, Milan, 2013.
5. Econophysics of Order-driven Markets, Eds. F. Abergel, B. K. Chakrabarti, A. Chakraborti, M. Mitra, New Economic Windows, Springer-Verlag, Milan, 2011.
6. Econophysics & Economics of Games, Social Choices and Quantitative Techniques, Eds. B. Basu, B. K. Chakrabarti, S. R. Chakravarty, K. Gangopadhyay, New Economic Windows, Springer-Verlag, Milan, 2010.
7. Econophysics of Markets and Business Networks, Eds. A. Chatterjee, B. K. Chakrabarti, New Economic Windows, Springer-Verlag, Milan 2007.
8. Econophysics of Stock and other Markets, Eds. A. Chatterjee, B. K. Chakrabarti, New Economic Windows, Springer-Verlag, Milan 2006.
9. Econophysics of Wealth Distributions, Eds. A. Chatterjee, S. Yarlagadda, B. K. Chakrabarti, New Economic Windows, Springer-Verlag, Milan, 2005.

Contents

Part I
Econophysics

Chapter 1
Strategic Behaviour and Indicative Price Diffusion in Paris Stock Exchange Auctions

Damien Challet

Abstract We report statistical regularities of the opening and closing auctions of French equities, focusing on the diffusive properties of the indicative auction price. Two mechanisms are at play as the auction end time nears: the typical price change magnitude decreases, favoring underdiffusion, while the rate of these events increases, potentially leading to overdiffusion. A third mechanism, caused by the strategic behavior of traders, is needed to produce nearly diffusive prices: waiting to submit buy orders until sell orders have decreased the indicative price and vice-versa.

Introduction

Research in market micro-structure has focused on the dynamical properties of open markets [5, 9]. However, main stock exchanges have been using auction phases when they open and close for a long time.[1] Auctions are known to have many advantages, provided that there are enough participants: for example, auction prices are well-defined, correspond to larger liquidity, and decrease price volatility (and bid-ask spreads) shortly after the opening time and before closing time (see e.g. [8, 10, 11]).

Only a handful of papers are devoted to the dynamics of auction phases, i.e., periods during which market participants may send limit or market orders specifically for the auction. Reference [6] investigates when fast and slow traders send their orders during the opening auction phase of the Paris Stock Exchange and find markedly different behaviors: the slow brokers are active first, while high-frequency traders are mostly active near the end of auctions. In the same vein, [3] shows how and when

[1] London Stock Exchange and XETRA (Germany) recently added a mid-day short auction phase.

D. Challet (✉)
Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes, Centralesupélec, Université Paris Saclay, Saint-Aubin, France
e-mail: damien.challet@centralesupelec.fr

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_1

low-latency traders (identified as high-frequency traders) add or remove liquidity in the pre-opening auction of the Tokyo Stock Exchange. Accordingly, [12] finds typical patterns of high-frequency algorithmic trading in the auctions of XTRA. Finally, [7] analyzes anonymous data from US equities and compute the response functions of the final auction price to the addition or cancellation of auction orders as a function of the time remaining until the auction, which have strikingly different behaviors in the opening and closing auction phases.

Auctions, Data and Notations

The opening auction phase of Paris Stock Exchange starts at 7:15 and ends at 9:00 while the closing auction phase is limited to the period 17:30 to 17:35. The auction price maximises the matched volume.

From the Thomson Reuters Tick History, we extract auction phase data for the 2013-04-16 components of the CAC40 index. This database contains all the updates to either the indicative match price or the indicative matched volume in the 2010-08-02 to 2017-04-12 period, which amounts to 8,095,524 data points for the opening auctions and 15,007,048 for the closing auctions. Note that the closing auction phase has about twice as many updates despite being considerably shorter.

For each asset α, we denote the indicative price of auction $x \in \{\text{open, close}\}$ of day d at time t by $\pi^x_{\alpha,d}(t)$, the time of auction x by t^x and the auction price by $p^x_{\alpha,d}$. Dropping the index α since this paper focuses on a single asset at a time, the i-th indicative price change occurs at physical time $t^x_{i,d}$ and its log-return equals $\delta p^x_{i,d} = \log \pi^x_d(t^x_{i,d}) - \log \pi^x_d(t^x_{i-1,d})$. It is useful to work in the time-to-auction (TTA henceforth) time arrow: setting $\tau = t^x - t$, the log-return between the final auction price and the current indicative is then $\Delta p^x_d(\tau) = \log p^x_d - \log \pi^x_d(t)$.

Similarly, the indicative matched volume is written as $W^x_d(t)$, while the final volume is V^x_d. Finally, when computing averages over days, since updates occur at random times, we will use time coarsening by $\delta\tau$ seconds, i.e. compute quantity averages over days within time slices of $\delta\tau$ seconds.

Figure 1.1 illustrates why auctions deserve attention: the relative importance of the closing auction volume has more than doubled in the last 10 years. Note that the relative opening auction volume of French equities is quite small (typically around 1%) and has stayed remarkably constant.

From Collisions in Event Time to Diffusion in Physical Time

It is useful to consider the price as the position of a uni-dimensional random walker and assume that each price change is caused by a collision: if collision i shifts the price p by δp_i, after N collisions the mean square displacement equals

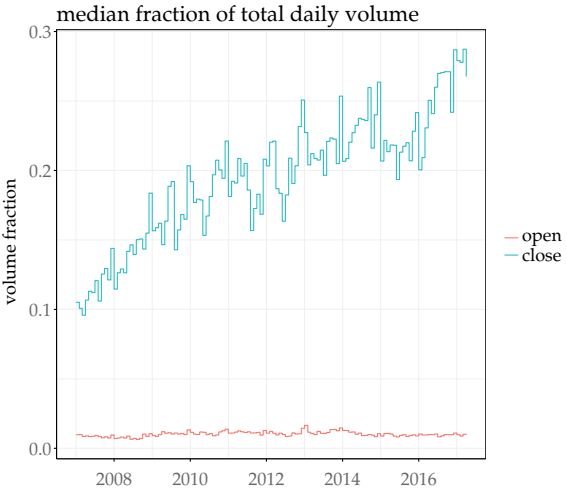


Fig. 1.1 Opening and closing fraction of the total daily volume (median computed over all the tickers) since 2007, showing the global increase of the relative importance of the closing auction, but not of the opening auction. Medians over assets of monthly medians for single assets

$$E\left(\left[\sum_{i=1}^{N}\delta p_i\right]^2\right) \propto N \tag{1.1}$$

if the increments δp_i are i.i.d, a straightforward consequence of the central limit theorem. This corresponds to standard diffusion. In addition, if the collisions occur at a constant rate ρ, then time is homogeneous and $E(t_i) = i\rho$. As we shall see, none of these assumptions is true during auctions, which makes them quite interesting dynamical systems.

Event Rates

In the case of indicative auction prices, the event rate is not constant: the activity usually increases just before the auction time. This finding is a generic feature of auctions with fixed end time [4], and more generally of human procrastinating nature when faced with a deadline, be it conference registration [1] or paying its fee [2].

Let us denote by $N_d^x(t) = \sum_{i,\ 0<t_{i,d}^x\le t} 1$ the number of price events (changes) having occurred up to time t on day d for auction x. The activity pattern of day d can be measured by the ratio between the number of events up to time t on day d and the total number of events which happened that day, defined as $\nu_d(t) = N_d^x(t)/N_d^x(t^x)$. The average and median $\nu(t) = M(\nu_d)(t)$, where M stands for either average or median over days, can be seen in Fig. 1.2. One similarly defines the fraction between the indicative matched volume at time t and the auction volume $\omega_{d(t)} = M(W_d^x(t)/V_d^x)$, reported in the same figure.

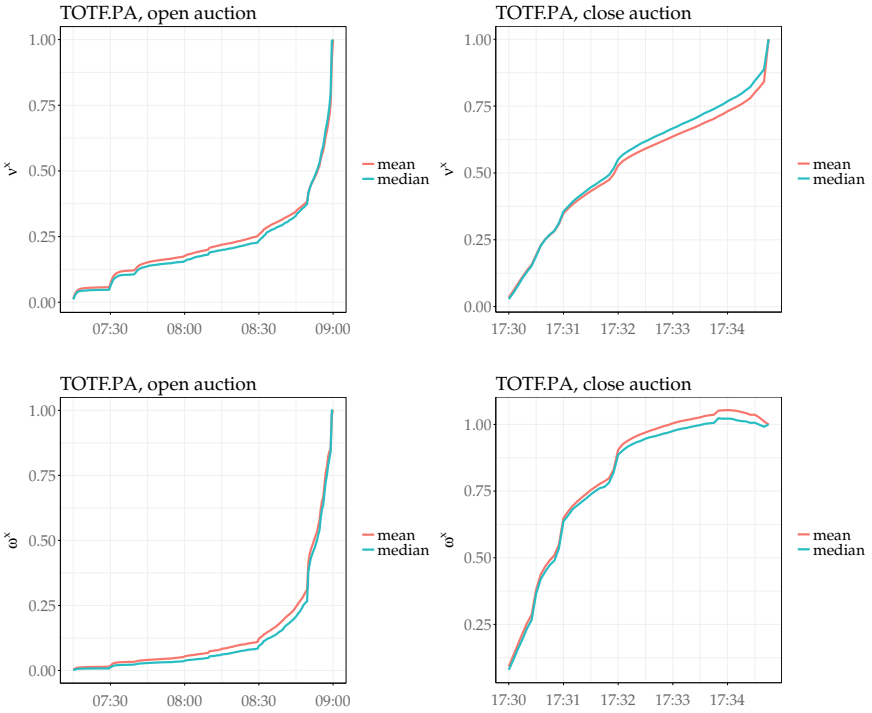


Fig. 1.2 Average activity patterns for opening and closing auctions (left and right plots, respectively) for the most active asset (Total). Upper plots: scaled price change events ν at a function of physical time t. Lower plots: scaled indicative volume fraction ω as a function of t. $\delta\tau = 30\,\mathrm{s}$ for opening auctions and 5 s for closing auctions

There are clear peaks of changes for both ν and ω at unimaginative physical times such as 7:30, 8:30, etc., and at round minutes and multiples of 30 s during the closing auction. This, of course, denotes a regular behavior of some investors. If each peak is systematically caused by a single trader, there are reasons to think that this regularity does inject information and that it will be exploited by more flexible traders. However, sending orders at the same time as other traders is a rational behavior as it allows one to hide in the crowd, unless one's orders are systematically of the same imbalance sign as the aggregate volume at that time. Thus, from a game theoretical point of view, the emergence of activity peaks is self-organized and stable. Nothing constrains the number of peaks and their locations, which are hence instances of emerging, self-organized conventions. The closing auction being much shorter than the opening one, it is natural that the peaks should appear at round minutes, as this somehow provides more obvious peak locations than the opening auction. When the closing auction lasts for a much longer time, e.g. for US equities, there are much fewer price activity peaks [7].

The global pattern of price changes and total volume matched clearly differs between both types of auctions. During opening auctions, the price change rate increases much, starting from a low baseline. During closing auctions, the opposite happens: price change activity is first large, slows down during the first 2–3 minutes and then picks up again just before the cut-off time (17:34:45). The average relative matched volume $\omega(t)$ behaves similarly as $\nu(t)$ during the opening auctions, probably because prices changes are mostly caused by the arrival of new matchable volume, not cancellations. Indeed, half of the open auction events typically happen in the last 10 minutes for most assets, and half of the volume is matched in the last minutes. Closing auctions display a different behavior: more than half of the volume is matched during the first minute, and 80% during the first two minutes. For a few assets (TOTF.PA, UNBP.PA, for ones), there is a peak of indicative matched volume up to 10% larger than the auction volume about one minute before the end of the auction; the same behavior is found in US equity markets [7].

Activity Acceleration

The acceleration pattern of price change rate follows some regularity. To characterize it in a simpler way, it is useful to work in Time-To-Auction τ frame. Since the latter reverts the time arrow, the activity decelerates as a function of τ. Let us denote the average event rate $\rho^x(\tau)$ so that the expected number of event in the period τ to $\tau + \delta\tau$ is $n^x(\tau) = E[N_d^x(\tau + \delta\tau) - N_d^x(\tau)] = \rho^x(\tau)\delta\tau$. Figure 1.3 shows $n^{\text{open}}(\tau)$ of several assets, together with the smoothed version of n^x, denoted by $n^x_{\text{smoothed}} = N^x(\tau)/\tau$: if $n^x \propto \tau^{-\beta}$, so does n^x_{smoothed} but with much less noise, which helps assessing the presence of a power-law visually. We shall drop the x superscript when no confusion is possible.

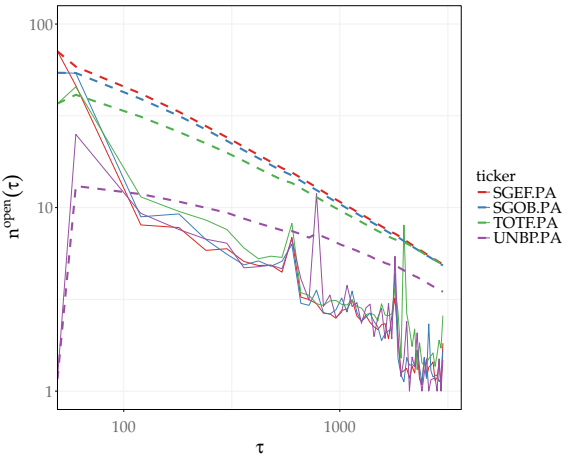


Fig. 1.3 Average number of price changes as a function of the time to auction τ, in seconds, for the opening auction. Dashed lines refer to n_{smoothed}. Time coarsing factor $\delta\tau = 60$s

Assuming that $n(\tau) \propto \tau^{-\beta}$, we perform a robust linear fit of $\log n(\tau) = cst - \beta\tau$ for $\tau \in [100, 300]$ seconds and only keep the fits whose t-statistics associated with β is larger than 5. This particular choice of interval for τ corresponds to a typical period during which the autocorrelation of δp_i at one lag is roughly constant (see section "Diffusive Properties of Indicative Prices"). In addition, for each asset, we only keep days during which there were at least 50 price changes.

If the typical absolute value of price change σ does not depend on τ and is still i.i.d., Eq. (1.1) becomes

$$E\left(\left[\sum_{i=1}^{N} \delta p_i\right]^2\right) = \sum_{i:\ t_i <= \tau} E(\delta p_i)^2 \propto \sigma \tau^{1-\beta} \tag{1.2}$$

hence the Hurst exponent in τ time, denoted by h, equals $(1 - \beta)/2$: the price change rate influences the diffusive pattern in a simple way, given the above approximations. It is worth noting at this juncture that in the normal time frame the price is overdiffusive if σ does not depend on τ and if $\beta > 0$, i.e., if the rate of price changes increases near the auction end time and the Hurst exponent in the normal time arrow is $H = (1 + \beta)/2$.

Typical Price Change

When the indicative price changes, it jumps to the next non-empty tick of the auction order book. Thus, the typical indicative price change reflects the density of the latter, which increases as the auction time nears. As a consequence, the typical price change magnitude σ is not constant but decreases near the auction end time, or equivalently increases as a function of τ. Once again, for opening auctions, we find an approximate power-law relationship $\sigma(\tau) \propto \tau^{\alpha}$ (see Fig. 1.4). We apply the same method as for $n(\tau)$ to estimate α: we only keep days during which there were at least 50 price changes for a given asset; robust fits of $\log \delta p(\tau) = cst + \alpha\tau$ for $\tau \in [100, 300]$ are carried out. Only fits whose t-statistics associated with α are larger than 5 are kept.

Diffusive Properties of Indicative Prices

It is easy to see why the increase of activity and decrease of the typical magnitude of price changes have antagonistic and purely mechanistic effects on the diffusive properties of the indicative auction price in the simplest case: neglecting the autocorrelations and cross-correlations of both $n(\tau)$ and δp_i, Eq. (1.2) becomes indeed

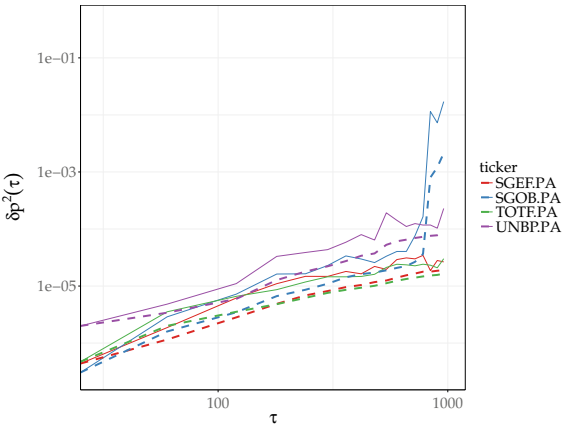


Fig. 1.4 Average scale of the log price increment as a function of the time to auction τ, in seconds, for the opening auction. Dashed lines refer to the smoothed quantity

$$E\left(\Delta p^2\right)(\tau) \simeq \sum_{\tau'<=\tau} E(n(\tau')\delta p^2(\tau')) \propto \tau^{h_0} \tag{1.3}$$

$$\simeq \sum_{\tau'<=\tau} E(n(\tau'))E(\delta p^2(\tau')) \propto \tau^{1+\alpha-\beta} = \tau^{h_0^{(\alpha\beta)}}, \tag{1.4}$$

The first approximation assumes that all δp_i within a time slice are i.i.d, while the second one assumes no correlation between n and δp^2. The relative merits of both approximations can be assessed in Fig. 1.5. The first approximation corresponds to the continuous black line and the second one to the black dots. Both curves are close together; however neglecting the dependence between n and $|\delta p|$ underestimates the typical magnitude of Δp. The same figure makes it clear that something is wrong even in the first approximation, as $\sum_{\tau'<=\tau} E(n(\tau')\delta p^2(\tau'))$ is about 10 times larger than $E(\Delta p^2)$. This discrepancy is mainly due to the bouncing behavior of π for large τ: a large δp_i is typically followed by large δp_{i+1} of opposite sign, which inflates $E(\delta p^2)$ and does not correspond to significant price change as the latter reverts immediately to a value close to that before event i. This is why trimmed means, which removes a given fraction of the largest δp_i for each time slice and each day, decrease much this discrepancy. The latter is also due in part to a simple strategic behavior: during the auction phase, negative indicative price change triggers the sending of buy orders and vice-versa, causing an intrinsically smaller than expected $\Delta p(\tau)^2$ (see below for a more detailed discussion) (Fig. 1.6).

Let us now compare the TTA Hurst exponents of the above quantities, plotted in Fig. 1.7 for the 6 stocks whose fits of both α and β are deemed significant. Two features stand out. First, h_0 overestimates h, even when accounting for the fairly large error bars. This implies that the dynamics caused by the interplay between typical price change shrinking and the acceleration of the activity is more subtle than the

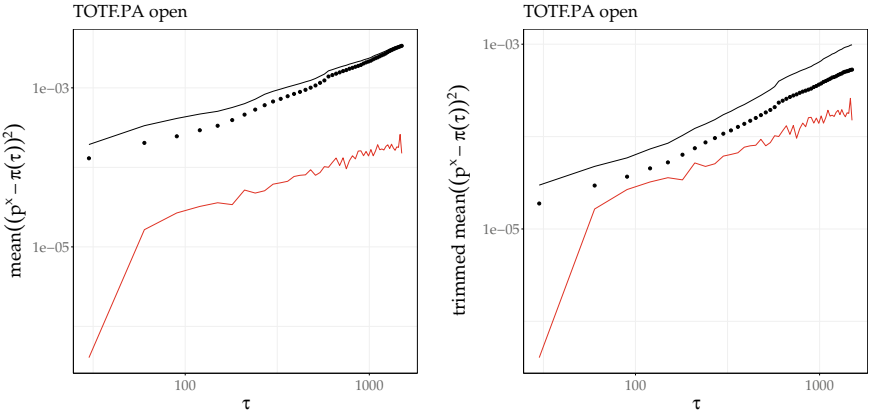


Fig. 1.5 Average square difference between the auction price and the indicative price τ seconds before the opening auction. Continuous red lines (bottom of the figure) refer to $E(\Delta p^2)(\tau)$, The upper black continuous line is $\sum_{\tau'<=\tau}[n(\tau')\delta p^2(\tau')]$, and the black dots are $\sum_{\tau'<=\tau} E[n(\tau')]E[\delta p^2(\tau')]$. Left plot: plain averages over all values of δp_i; right plot: trimmed means where the 20% largest (in absolute value) δp_i for each day and each time slice of $\delta\tau = 30\,\mathrm{s}$ have been removed in the computation of the averages of quantities based on δp_i

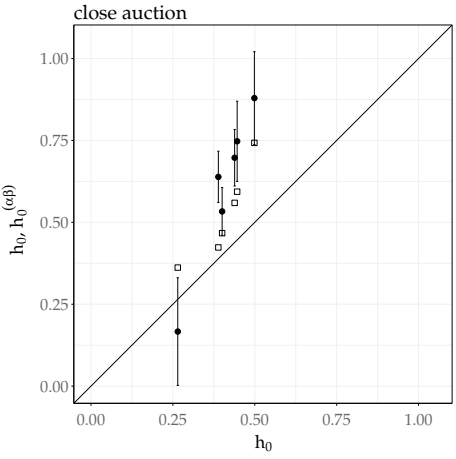


Fig. 1.6 Autocorrelation between two consecutive price changes within time slices of $\delta\tau = 60\,\mathrm{s}$, averaged over all days for TOTF.PA

simple approximation above. In fact, interestingly, $h_0^{\alpha\beta}$ also overestimates the Hurst h_0 exponent: this emphasizes the fact that the δp_i are not i.i.d.

Indeed, in practice, even linear autocorrelation of both $\rho(\tau)$ and δp_i and the cross-correlation between them are not negligible. Let us focus on the autocorrelation of δp_i, denoted by $C_{\delta p}(\delta i)$. For each time slice $[\tau, \tau + \delta\tau[$, we average $C_{\delta p,d}(1)$ over all the days for a given asset. Figure 1.6 plots this quantity versus τ for TOTF.PA, the most active asset in our dataset. Generally, $C_{\delta p}(1) < 0$; even more, it becomes more and more negative near the auction time, i.e., for small τ. Since the price changes become

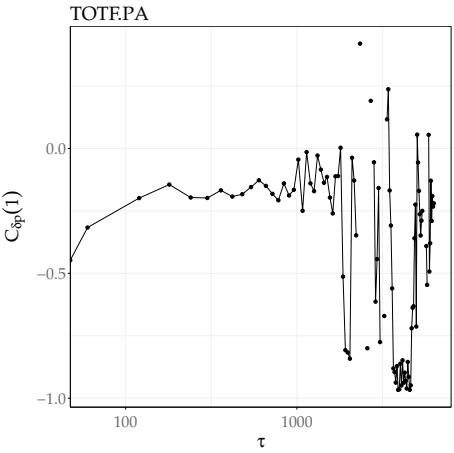


Fig. 1.7 Hurst exponents h_0 and $h_0^{(\alpha\beta)}$ versus the actual Hurst exponent for the 6 assets of the CAC40 that yield good power-law fits of both $\sigma^2 \propto \tau^{\alpha}$ and $n \propto \tau^{-\beta}$; open auction, $\delta\tau = 60[s]$; Time-To-Auction arrow. Error bars correspond to one standard deviation. When no error bar is visible, the error is at most as large as the symbol

relatively smaller in that limit, this reflects a purposeful bounce of the indicative auction price between two close price ticks; the large negative autocorrelation points to strategic behavior, by which traders try to decrease the immediate impact of their auction orders by submitting their orders after other orders of the opposite sign (hence to hide their actions); in fact, the autocorrelation of the sign of δp, $C_{\text{sign}\,\delta p}(1)$ is even smaller than $C_{\delta p}(1)$ for small τ. For large τ, this auto-correlation also tends to have very small values, which is reinforced by the fact that an outstandingly large δp_i is often followed by a similarly large δp_{i+1} of opposite sign. Thus strategic behavior is more common for small τ.

When $C_{\delta p}(1)$ does not depend on τ, it only modifies the prefactor of τ in Eq. (1.3) by a factor of the order $\frac{1+C_{\delta p}(1)}{1-C_{\delta p}(1)}$, not the Hurst exponent, and thus explains in part the discrepancy between $E(\Delta p^2)(\tau)$ and $\sum_{\tau'=1}^{N} E[n(\tau')\delta p(\tau')^2]$. The dependence of $C_{\delta p}(1) < 0$ on τ modifies the apparent Hurst exponent in a nontrivial way. This is why we measured h for $\tau \in [100, 300]$, i.e., in a region where $C_{\delta p}(1) < 0$ is the most constant.

Discussion

Indicative auction prices display non-trivial properties due in part to the antagonistic effects of both the acceleration of activity and the reduction of the typical price change magnitude. However, the indicative price is much less over-diffusive than what these two effects alone imply. In other words, the deviation from purely mechanistic effect points to a more subtle dynamics. This makes sense, as the traders have a clear incentive to minimize their easily detectable impact. Their strategic behavior results in often alternatively positive and negative indicative price changes, i.e., in a clearly

anti-correlated price changes. Quite tellingly, this negative auto-correlation is more and more pronounced as the auction end nears.

So far, we have used a basic data type, which nevertheless has a rich behavior. More detailed data, such as data from the auction book, will allow us to characterize order strategic placement, the evolution of the average auction book density and the price impact of new orders and order cancellations much before the auction time, in the spirit of the response function of [7], but accounting for both the volume of new auction orders and their immediate impact on the auction order book.

References

1. Alfi, V., Parisi, G., Pietronero, L.: Conference registration: how people react to a deadline. Nat. Phys. **3**(11), 746 (2007)
2. Alfi, V., Gabrielli, A., Pietronero, L.: How people react to a deadline: time distribution of conference registrations and fee payments. Open Phys. **7**(3), 483–489 (2009)
3. Bellia, M., Pelizzon, L., Subrahmanyam, M.G., Uno, J., Yuferova, D.: Low-latency trading and price discovery: Evidence from the Tokyo stock exchange in the pre-opening and opening periods (2016)
4. Borle, S., Boatwright, P., Kadane, J.B.: The timing of bid placement and extent of multiple bidding: an empirical investigation using eBay online auctions. Stat. Sci. 194–205 (2006)
5. Bouchaud, J.-P., Farmer, J.D., Lillo, F.: How markets slowly digest changes in supply and demand. In: Hens, T., Schenk-Hoppè, K.R. (eds.) *Handbook of Financial Markets: Dynamics and Evolution*, pp. 57–160. Elsevier (2009)
6. Boussetta, S., Lescourret, L., Moinas, S.: The role of pre-opening mechanisms in fragmented markets (2016)
7. Challet, D., Gourianov, N.: Dynamical regularities of US equities opening and closing auctions. Submitted to Market Microstructure and Liquidity (2018). arXiv:1802.01921
8. Chelley-Steeley, P.L.: Market quality changes in the London stock market. J. Bank. Financ. **32**(10), 2248–2253 (2008)
9. O'Hara, M.: Market Microstructure Theory. Wiley, New York (1997)
10. Pagano, M.S., Peng, L., Schwartz, R.A.: A call auction's impact on price formation and erder routing: Evidence from the NASDAQ stock market. J. Financ. Mark. **16**(2), 331–361 (2013)
11. Pagano, M.S., Schwartz, R.A.: A closing call's impact on market quality at Euronext Paris. J. Financ. Econ. **68**(3), 439–484 (2003)
12. Yergeau, G.: Machine learning and high-frequency algorithms during batch auctions (2018)

Chapter 2
Complex Market Dynamics in the Light of Random Matrix Theory

Hirdesh K. Pharasi, Kiran Sharma, Anirban Chakraborti and Thomas H. Seligman

Abstract We present a brief overview of random matrix theory (RMT) with the objectives of highlighting the computational results and applications in financial markets as complex systems. An oft-encountered problem in computational finance is the choice of an appropriate epoch over which the empirical cross-correlation return matrix is computed. A long epoch would smoothen the fluctuations in the return time series and suffers from non-stationarity, whereas a short epoch results in noisy fluctuations in the return time series and the correlation matrices turn out to be highly singular. An effective method to tackle this issue is the use of the power mapping, where a non-linear distortion is applied to a short epoch correlation matrix. The value of distortion parameter controls the noise-suppression. The distortion also removes the degeneracy of zero eigenvalues. Depending on the correlation structures, interesting properties of the eigenvalue spectra are found. We simulate different correlated Wishart matrices to compare the results with empirical return matrices computed using the S&P 500 (USA) market data for the period 1985–2016. We also briefly review two recent applications of RMT in financial stock markets: (i) Identification of "market states" and long-term precursor to a critical state; (ii) Characterization of catastrophic instabilities (market crashes).

H. K. Pharasi
Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México,
62210 Cuernavaca, México
e-mail: hirdeshpharasi@gmail.com

K. Sharma · A. Chakraborti (✉)
School of Computational and Integrative Sciences, Jawaharlal Nehru University,
New Delhi 110067, India
e-mail: anirban@jnu.ac.in

K. Sharma
e-mail: kiransharma1187@gmail.com

T. H. Seligman
Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México,
62210 Cuernavaca, México
e-mail: seligman@icf.unam.mx

Centro Internacional de Ciencias, 62210 Cuernavaca, México

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_2

Introduction

With the advent of the "Big Data" era [9, 13], large data sets have become ubiquitous in numerous fields—image analysis, genomics, epidemiology, engineering, social media, finance, etc., for which we need new statistical and analytical methods [3, 5, 6, 15, 29]. Empirical correlation matrices are of primal importance in big data analyses, since various statistical methods strongly rely on the validity of such matrices in order to isolate meaningful information contained in the "observational" signals or time series [2]. Often the time series are of finite lengths, which can lead to spurious correlations and make it difficult to extract the signal from noise [11, 26]. Hence, it is very important to understand quantitative effects of finite-size time series in determination of empirical correlations [8, 11, 26, 33].

Random matrix theory (RMT) tries to describe statistics of eigenvalues of random matrices, often in the limit of large dimensions. The subject came up first in a celebrated paper of Wishart [39] in 1929 where he proposed that the correlation matrix of white noise time series was an adequate prior for correlation matrices. E. Cartan proposed the classical random matrix ensembles in an important but little known paper [4]. After that there was increasing interest in the subject among which it is important to mention work by L.G. Hua, who published the first monographs on the subject in 1952; an English translation is available [12].

Wigner introduced RMT to physics, based on the assumption that the interactions between the nuclear constituents were so complex that they could be modeled as random fluctuations in the framework of his R-matrix scattering theory [36]. This culminated in the presentation of the Hamiltonian $\hat{H}$ as a large random matrix, such that the energy levels of the nuclear system could be approximated by the eigenvalues of this matrix, and indeed the spacings between the energy levels of nuclei could be modeled by the spacing of eigenvalues of the matrix [37, 38]. The use of RMT has spread over many fields from molecular physics [14] to quantum chromodynamics [28]. Lately, RMT has become a popular tool for investigating the dynamics of financial markets using cross-correlations of empirical return time series [25, 30].

In this chapter, we present recent techniques of random matrix theory (RMT) mainly focused on computational results and applications of correlations in financial markets viewed as complex systems [1, 10, 30, 31]. A central problem that often arises in computational finance is the choice of the epoch size over which the empirical cross-correlation return matrix needs to be computed. A very long epoch would smoothen the fluctuations in return time series and also the time series suffers from the problem of non-stationarity [19], whereas a short-time epoch would result in noisy fluctuations in return time series and the correlation matrix turns out to be highly singular (with many zero eigenvalues) [8]. Among others, an effective method to tackle this issue has been the use of the power mapping [8, 11, 26, 33], where a non-linear distortion is applied to a short epoch correlation matrix. Here, we demonstrate how the value of distortion parameter controls the noise-suppression. It also removes the degeneracy of the zero eigenvalues (which for very small values of the distortion

parameter leads to a well separated "emerging spectum" near zero). Depending on the correlation structures, interesting properties of the eigenvalue spectra are found. Correlation matrices constructed from white noise were introduced by Wishart and their eigenvalue spectrum gets a shape of Marčenko-Pastur distribution [16]; there are significant deviations when a correlation structure is introduced [7]. We simulate different correlated Wishart matrices [18, 39] to compare the results with empirical return matrices computed using S&P 500 (USA) market data for the period 1985–2016 [8]. We also briefly review two recent applications of RMT in financial stock markets: (i) Identification of "market states" and long-term precursor to a critical state [23]; (ii) Characterization of catastrophic instabilities (market crashes) [8].

This chapter is described as follows. Section "Data Description, Methodology and Results" discusses the data description, methodology and results in details. Section "Recent Applications of RMT in Financial Markets" contains applications of RMT in financial markets. Finally, section "Concluding Remarks" contains concluding remarks.

Data Description, Methodology and Results

Data Description

We have used the database of Yahoo finance [40], for the time series of adjusted closure prices for S&P 500 (USA) market, for the period 02/01/1985–30/12/2016 ($T = 8068$ days); number of stocks $N = 194$, where we have included the stocks that are present in the index for the entire duration. The sectoral abbreviations are given in Table 2.1.

Methodology and Results

Correlations between different financial assets play fundamental roles in the analyses of portfolio management, risk management, investment strategies, etc. However, one only has finite time series of the assets prices; hence, one cannot calculate the exact

Table 2.1 Abbreviations of ten different sectors for S&P 500 index

Labels	Sectors	Labels	Sectors
CD	Consumer discretionary	ID	Industrials
CS	Consumer staples	IT	Information technology
HC	Health care	MT	Materials
EG	Energy	TC	Technology
FN	Financials	UT	Utilities

correlation among assets, but only an approximation. The quality of the estimation of the true cross-correlation matrix strongly depends on the ratio between the length of the financial price time series T and the number of assets N. The larger the ratio $Q = T/N$, the better the estimation is; though for practical limitations, the ratio can be even smaller than unity. However, such correlation matrices are often too noisy, and thus need to be filtered from noise. To build the correlation matrices, we first calculate the return r_i from the daily price P_i of stocks $i = 1, \ldots, N$, at time t (trading day):

$$r_i(t) = \ln P_i(t) - \ln P_i(t-1), \tag{2.1}$$

where $P_i(t)$ denotes the price of stock i at time t. Since different stocks have varying levels of volatility, we define the equal-time Pearson cross-correlation coefficient as

$$C_{ij}(\tau) = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sigma_i \sigma_j}, \tag{2.2}$$

where $\langle \ldots \rangle$ denotes the time average and σ_k denotes the standard deviation of the return time series $r_k, k = 1, \ldots, N$, computed over an epoch of M trading days ending on day τ. The elements C_{ij} are restricted to the domain $-1 \leq C_{ij} \leq 1$, where $C_{ij} = 1$ corresponds to perfect correlations, $C_{ij} = -1$ to perfect anti-correlations, and $C_{ij} = 0$ to uncorrelated pairs of stocks. The difficulties in analyzing the significance and meaning of the empirical cross-correlation coefficients C_{ij} are due to several reasons, which include the following:

1. Market conditions change with time and the cross-correlations that exist between any pair of stocks may not be stationary if an epoch chosen is too long.
2. Too short epoch, for estimation of cross-correlations, introduces "noise", i.e., fluctuations.

For these reasons, the empirical cross-correlation matrix $\boldsymbol{C}(\tau)$ often contains "random" contributions plus a part that is not a result of randomness [22, 24]. Hence, the eigenvalue statistics of $\boldsymbol{C}(\tau)$ are often compared against those of a large random correlation matrix—a correlation matrix constructed from mutually uncorrelated time series (white noise) known as Wishart matrix.

We first reproduce the basic results of RMT, e.g., the Marčenko-Pastur distribution, or Marčenko-Pastur law, which describes the asymptotic behavior of eigenvalues of square random matrices [16]. Then, we present a study of time evolution of the empirical cross-correlation structures of return matrices for N stocks and the eigenvalues spectra over different time epochs, and try to extract some new properties or information about the financial market [8, 23].

Wishart and Correlated Wishart Ensembles

Let us construct a large random matrix $\boldsymbol{B}$ arising from N random time series each of length T, where the entries of a time series are real independent random variables

drawn from a standard Gaussian distribution with zero mean and variance σ^2, such that the resulting matrix $\boldsymbol{B}$ is $N \times T$. Then the Wishart matrix can be constructed as

$$\boldsymbol{W} = \frac{1}{T}\boldsymbol{B}\boldsymbol{B}'. \tag{2.3}$$

In RMT, the ensemble of Wishart matrices is known as the *Wishart orthogonal ensemble*. In the context of a time series, $\boldsymbol{W}$ may be interpreted as the *covariance* matrix, calculated over N stochastic time series, each with T statistically independent variables. This implies that on average, $\boldsymbol{W}$ does not have cross-correlations.

A correlated Wishart matrix can be constructed as

$$\boldsymbol{W} = \frac{1}{T}\boldsymbol{G}\boldsymbol{G}', \tag{2.4}$$

where $\boldsymbol{G} = \boldsymbol{\zeta}^{1/2}\boldsymbol{B}$, is a $N \times T$ matrix; $\boldsymbol{G}'$ is the $T \times N$ transpose matrix of $\boldsymbol{G}$, and the $N \times N$ positive definite symmetric matrix ζ controls the actual correlations. If $\boldsymbol{\zeta}$ is a diagonal matrix with the diagonal entries as unity and off-diagonal entries as zero (i.e., $\zeta = \mathbb{1}$, the identity matrix), then the resulting matrix $\boldsymbol{W}$ reduces to one of the former *Wishart orthogonal ensemble*. If the diagonal entries of $\boldsymbol{\zeta}$ are unity and off-diagonal elements are non-zero and real, then the resulting matrices form the *correlated Wishart orthogonal ensemble*. For simplicity, in this chapter, we have generated and used $\boldsymbol{\zeta}$ for which all the off-diagonal elements are same (equal to a constant U, which lies between zero and unity).

The spectrum of eigenvalues for the Wishart orthogonal ensemble can be calculated analytically. For the limit $N \to \infty$ and $T \to \infty$, with $Q = T/N$ fixed (and bigger than unity), the probability density function of the eigenvalues is given by the Marčenko-Pastur distribution:

$$\bar{\rho}(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda}, \tag{2.5}$$

where σ^2 is the variance of the elements of $\boldsymbol{G}$, while λ_{min} and λ_{max} satisfy the relation:

$$\lambda_{min}^{max} = \sigma^2\left(1 \pm \frac{1}{\sqrt{Q}}\right)^2. \tag{2.6}$$

For $Q \leq 1$, positive semi-definite matrices $\boldsymbol{W}$, the density $\bar{\rho}(\lambda)$ in the above Eq. 2.5 is normalized to Q and not to unity. Therefore, taking into account the $(N - T)$ zeros, we have

$$\bar{\rho}(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda} + (1 - Q)\delta(\lambda). \tag{2.7}$$

First, we have generated a Wishart matrix $\boldsymbol{W}$ (with $\zeta = \mathbb{1}$) of size $N \times N$ constructed from N time series of real independent Gaussian variables, each of finite

length T, zero mean and unit variance ($\sigma^2 = 1$). Figure 2.1 shows the effect of finite sizes of the sets of parameters N and T on the probability distributions of the elements W_{ij} of the Wishart ensemble and the corresponding eigenvalue spectra. Figure 2.1a shows the probability distribution of the elements of the Wishart matrix of dimensions, where $N = 1024$ and $T = 10240$. Figure 2.1d shows the corresponding density of eigenvalues $\bar{\rho}(\lambda)$, which takes the shape of the theoretical Marčenko-Pastur distribution (red dashed line) [16]. Similarly, Fig. 2.1b, c show the respective probability distributions of the elements of Wishart matrices generated using the sets of parameters $N = 10240$ and $T = 102400$, and $N = 30720$ and $T = 307200$. We can see that with increase in system size (both N and T) the shape of the distribution becomes narrower, implying that the amount of spurious cross-correlations decreases. Ideally, the distribution should be a Dirac-delta at zero, since true cross-correlations do not exist. The eigenvalue spectra are less sensitive to the parameters N and T, as can be seen in Fig. 2.1e, f, which show the corresponding eigenvalue spectra. For all of the above simulations, we find the simulated data agree closely with the theoretical Marčenko-Pastur distributions (red dashed lines) with $\lambda_{max} = 1.732$ and $\lambda_{min} = 0.468$ (theoretically calculated using Eq. 2.6, and $Q = 10$).

As we have mentioned earlier, the assumption of stationarity fails for a very long return time series, so it is often useful to break one long time series of length T into n shorter epochs, each of size M (such that $T/M = n$). The assumption of stationarity then improves for each of the shorter epochs. However, if there are N return time series, such that $N >> M$, then the corresponding cross-correlation matrices are highly singular with $N - M + 1$ zero eigenvalues, which lead to poor eigenvalue statistics. We use the power map technique [11, 34] to break the degeneracy of eigenvalues at zero. In this method, a non-linear distortion is given to each element (W_{ij}) of the Wishart matrix $\boldsymbol{W}$ (or later in each correlation coefficient C_{ij} of the empirical cross-correlation matrix $\boldsymbol{C}$) of short epoch by:

$$W_{ij} \rightarrow (\text{sign } W_{ij})|W_{ij}|^{1+\varepsilon}, \tag{2.8}$$

where ε is a noise-suppression parameter. For very small distortions, e.g., $\varepsilon = 0.001$ (as used here), we get an "emerging spectrum" of eigenvalues, arising from the degenerated eigenvalues at zero which is well separated from the original spectrum. The power mapping method suppresses noise present in the correlation structure of short-time series (see e.g., Refs. [8, 17, 21, 23, 32] for recent studies and applications). Later in this chapter, we study different aspects of the power mapping method by varying the value of distortion ε from 0 to 0.8.

In Fig. 2.2, we have studied the effect of non-linear distortion on the behavior of Wishart ensemble ($U = 0$), where $N >> M$. The top row of Fig. 2.2 shows semi-log plots of the ensembles with parameters: (a) $N = 1024$ and $M = 512$, and (b) $N = 1024$ and $M = 64$. Then small non-linear distortions with $\varepsilon = 0.001$ are given to the ensembles to display the emerging spectra, shown in Fig. 2.2c, d. Interestingly, the shape of the emerging spectrum changes from a semi-circle to a strongly distorted one, as M becomes shorter. Also, note that emerging spectrum shifts towards the left

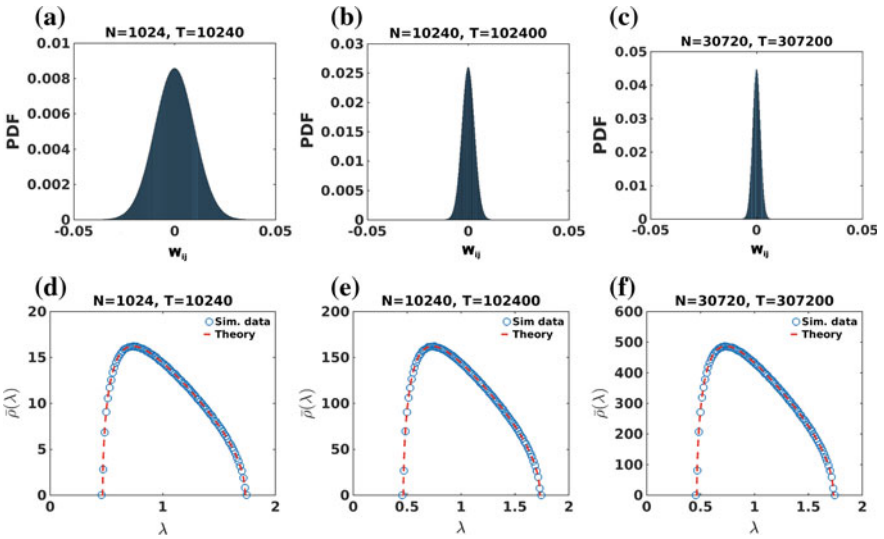


Fig. 2.1 **a–c** show the effect of finite size on true correlations with the dimensions of $\boldsymbol{B}$ (N = finite, T = finite and $Q(= T/N) = 10$). The probability distribution of elements (W_{ij}) of the Wishart ensemble of size, constructed from N time series, each with real independent Gaussian random variables of length T with zero mean and variance σ^2. The variance of the distribution of W_{ij} decreases with the increase of N and T and reduces to zero for $N \to \infty$ and $T \to \infty$ with $\frac{T}{N} =$ finite. **d–f** show the density of eigenvalues $\bar{\rho}(\lambda)$ of Wishart ensemble, which are numerically fitted with the Marčenko-Pastur distributions [16] (red dash lines) for all N and T. The numerical values of $\lambda_{max} = 1.732$ and $\lambda_{min} = 0.468$ of the spectra also match exactly with the results theoretically calculated from Eq. 2.6. Numerical results for the probability distributions of the elements (W_{ij}) and densities of the eigenvalues ($\bar{\rho}(\lambda)$) have been generated using averages up to 200 ensembles

side as M becomes shorter. For smaller values of M, some of the eigenvalues of emerging spectrum become negative. The number of negative eigenvalues depend on the size of the epoch M, the distortion parameter ε and the mean correlation in the case of a correlated Wishart ensemble [21].

Figure 2.3 shows the effect of a constant correlation with strength U on the eigenvalue spectra and the emerging spectra of correlated Wishart ensembles with parameters $N = 1024$ and $M = 64$. Figure 2.3a–c show the eigenvalue distributions, on the semi-log scales, for the correlated Wishart ensembles with correlations $U = 0.1$, $U = 0.3$, and $U = 0.8$, respectively. Insets show the densities of non-zero eigenvalues, which are closely described by the Marčenko-Pastur distributions in all cases. In the bottom row, Fig. 2.3d–f show the densities of the corresponding emerging spectra arising from non-linear distortion of the degenerate eigenvalues at zero. The shapes of the emerging spectra change from distorted semi-circle to Lorentzian-like, as the constant correlation values increase for the correlated Wishart ensembles.

Next, we present the effect of the distortion (or noise-suppression) parameter ε on the eigenvalue spectra in Fig. 2.4. Figure 2.4a–f show the distributions of eigenvalues for the correlated Wishart ensembles with parameters $N = 1024$ and $M = 64$, and

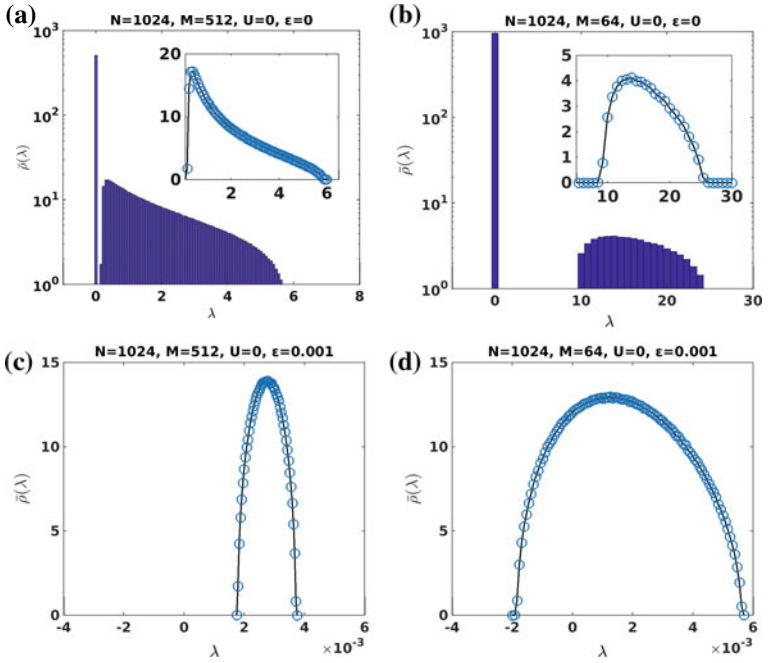


Fig. 2.2 Semi-log plot of the eigenvalue distribution of Wishart matrix $\boldsymbol{W}$, using the set of parameters **a** $N = 1024$ and $M = 512$; **b** $N = 1024$ and $M = 64$. For short epochs ($N > M$), the eigenvalue spectra have $N - M + 1$ zero eigenvalues and the remaining eigenvalues of the spectra show a distributions similar to the Marčenko-Pastur distribution. Insets show the zoomed in views of remaining $M - 1$ eigenvalues. **c** and **d** show the emerging spectra, generated using the power map technique with $\varepsilon = 0.001$, which are deformed semi-circular. Numerical results for densities of eigenvalues have been generated using the averages over 1000 ensemble members. Note that the emerging spectrum shifts towards left for smaller values of M, and also some of its eigenvalues become negative at smaller values of M

varying distortion parameter values: $\varepsilon = 0.0, 0.1, 0.2, 0.4, 0.6$ and 0.8, keeping a constant correlation ($U = 0.1$) among all off-diagonal elements in $\boldsymbol{\zeta}$. The densities of non-zero eigenvalues are closely described by the Marčenko-Pastur distributions, but the emerging spectra move towards the main spectra as the value of ε increases. The emerging spectra is absent at $\varepsilon = 0$, while it merges with the main spectrum at high values of distortion parameter, e.g., $\varepsilon = 0.8$.

Eigenvalue Decomposition of the Empirical Cross-Correlation Matrix

We also analyze $N = 194$ adjusted daily closure price time series of the stocks of S&P 500 (USA) index from the Yahoo finance database [40]. As discussed in the methodology subsection, we construct the empirical cross-correlation matrix $\boldsymbol{C}(\tau)$ for an epoch of $M = 200$ trading days, ending on trading day τ. In Fig. 2.5a, e, we choose two correlation matrices for the time series from 07/03/2011 to 16/12/2011 (high

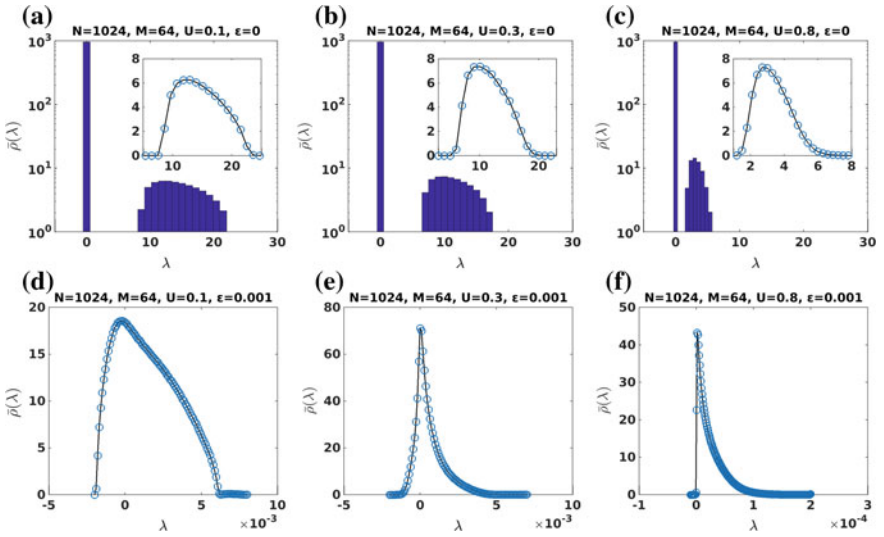


Fig. 2.3 Eigenvalue spectra of correlated Wishart ensembles with parameters, $N = 1024$ and $M = 64$, shown on semi-log scales with constant correlations: **a** $U = 0.1$, **b** $U = 0.3$, and **c** $U = 0.8$. Insets show the corresponding densities of non-zero eigenvalues, which are closely described by the Marčenko-Pastur distributions. **d–f** show the densities of the emerging spectra, when non-linear distortions (with $\varepsilon = 0.001$) are applied to the same matrices. Note that the shape of the emerging spectrum changes from distorted semi-circle to a Lorentzian-like with the increase of constant correlation strength U

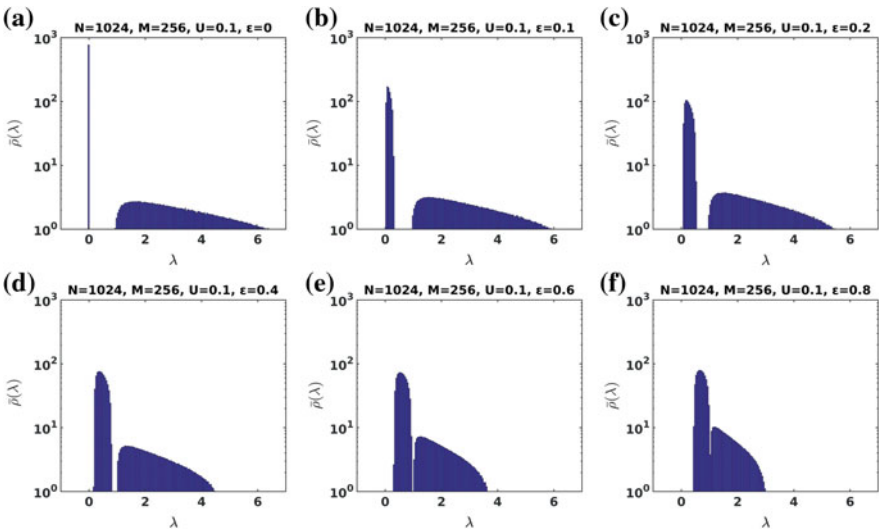


Fig. 2.4 Semi-log plots of the eigenvalue spectra for the correlated Wishart ensemble $\boldsymbol{W}$ with parameters $N = 1024$ and $M = 256$ at a constant correlation with $U = 0.1$, and distortion parameters of: **a** $\varepsilon = 0$, **b** $\varepsilon = 0.1$, **c** $\varepsilon = 0.2$, **d** $\varepsilon = 0.4$, **e** $\varepsilon = 0.6$, and **f** $\varepsilon = 0.8$. For $\varepsilon = 0.1$, the emerging spectrum is well separated from non-zero eigenvalues but with the increase of the distortion parameter ε the emerging spectrum starts moving towards the remaining non-zero eigenvalues spectra, and eventually merges with it at higher values, e.g., $\varepsilon = 0.8$

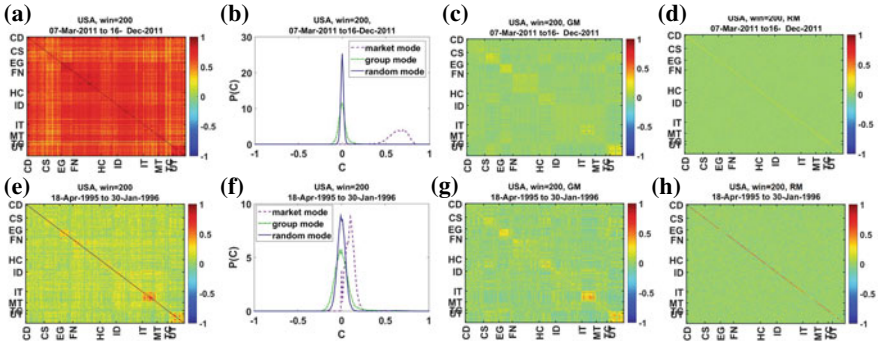


Fig. 2.5 **a** and **e** show the cross-correlation matrices of 194 stocks of S&P 500 for $M = 200$ days during: **a** 07/03/2011 to 16/12/2011; **b** 18/04/1995 to 30/01/1996. The stocks are arranged according to their industrial groups (abbreviations are given in Table 2.1). The blocks along the diagonal show the correlations within the same industrial groups; the color-bar shows the amount of correlation among the stocks. **a** shows the correlation matrix with high mean correlation and (**e**) with low mean correlation. **b** and **f** show the eigenvalue decomposition of the correlation matrix into the market mode, group modes and random modes. The market mode captures the mean market correlation, which corresponds to the dominant eigenvalue of the matrix. The group modes give the sectoral behavior of the market characterized by the subsequent 15 eigenvalues for a correlation matrix (**a**) and the next 62 eigenvalues for a correlation matrix (**e**) of the market. The rest of the eigenvalues show random behavior. **c** and **g** are the correlation matrix after removing the market mode and random modes from the correlation matrix; thus the matrix is composed of group modes only. We can visualize the block structure which shows the correlation among sectors. **d** and **h** show the correlation matrix after removing the market mode and group modes from the correlation matrix; so the matrix is composed of random modes only

mean correlation) and 18/04/1995 to 30/01/1996 (low mean correlation), respectively. The color-bar shows the amount of correlation among the stocks. The stocks are arranged according to their industrial groups (abbreviations are given in Table 2.1). The blocks along the diagonal show the correlations within the same industrial group. Figure 2.5b, f show the eigenvalue decomposition of the correlation matrices into the respective market mode, the group modes and the random modes. From such a segregation/decomposition, it is also possible to reconstruct the contributions of different modes to the aggregate correlation matrix as we show below.

The largest eigenvalue of the correlation matrix, corresponds to a market mode reflects the aggregate dynamics of the market common across all stocks, and strongly correlated to the mean market correlation. The group modes capture the sectoral behavior of the market, which are 15 eigenvalues subsequent to the largest eigenvalue of the correlation matrix of Fig. 2.5c, and the 62 subsequent eigenvalues for correlation matrix of Fig. 2.5g. Remaining eigenvalues capture the random modes behavior of the market (see Fig. 2.5d, h). By using the eigenvalue decomposition, we can thus filter the true correlations (coming from the signal) and the spurious correlations (coming from the random noise). For this, we first decompose the aggregate correlation matrix as

$$\boldsymbol{C} = \sum_{i=1}^{N} \lambda_i a_i a_i', \tag{2.9}$$

where λ_i and a_i are the eigenvalues and eigenvectors, respectively, of the correlation matrix $\boldsymbol{C}$. An easy way of handling the reconstruction of the correlation matrix is to sort the eigenvalues in descending order, and then rearranging the eigenvectors in corresponding ranks. This allows one to decompose the matrix into three separate components, viz., market, group and random

$$C = C^M + C^G + C^R, \tag{2.10}$$

$$= \lambda_1 a_1 a_1' + \sum_{i=2}^{N_G} \lambda_i a_i a_i' + \sum_{i=N_G+1}^{N} \lambda_i a_i a_i', \tag{2.11}$$

where N_G is taken to be 15 for the high mean correlated matrix (Fig. 2.5a) and 62 for the low mean correlation (Fig. 2.5e), i.e., corresponding to the 15 (or 62) eigenvalues after the largest one, for two chosen correlation matrices. It is worth noting that the result is not extremely sensitive to the exact value of N_G. As mentioned above, the eigenvectors from 2 to N_G describe the sectoral dynamics.

Figure 2.5c, g show the correlation matrices after removing the market mode and random modes from the respective correlation matrices; so the matrices show group modes only. We can see the block structures, which exhibit the correlations among the sectors. Figure 2.5d, h show the correlation matrices after removing the market mode and group modes; so the matrices display the random modes only.

An important observation is that the market mode shifts towards the right with the increment of the mean correlation. The group modes almost coincide with the random modes but with higher variance. Thus, the sectoral dynamics are almost absent whereas the market mode is very strong (similar to what was observed in Ref. [27]).

Figure 2.6a shows the average cross-correlation matrix of $N = 194$ stocks of S&P 500 for the entire duration 1985–2016 ($T = 8068$ trading days). We decomposed the average cross-correlation matrix into the market mode, group modes and random modes. As usual, the market mode captures the mean market correlation corresponding to the maximum eigenvalue, which is separate from rest of the eigenvalues (see Ref. [35] for the comparison of the behavior of maximum eigenvalues in correlated Wishart ensembles). The group modes, which tell about the sectoral behavior of the market, largely coincide with the random modes and correspond to the random behavior of the stocks. The resulting eigenvalue distribution (shown in Fig. 2.6c) thus has part that is a Marčenko-Pastur distribution [16] (see Fig. 2.6c and its inset) and some deviations. As $N << T$ so we do not get any zero eigenvalues. The maximum eigenvalue ($\lambda_{max} = 55.72$) of the spectra dominates the whole market. The next 19 eigenvalues correspond to the group modes, and the rest behave as random modes. The smallest eigenvalue of the spectrum $\lambda_{min} = 0.22$.

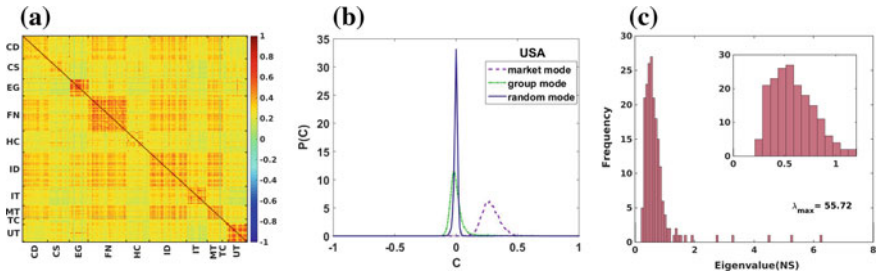


Fig. 2.6 **a** Average cross-correlation matrix of 194 stocks of S&P 500 in 32-years period from 1985 to 2016. The stocks are arranged according to their industrial groups (abbreviations are given in Table 2.1). The diagonal blocks show the correlations within the same industrial groups and off diagonal elements show correlations with other industrial groups. **b** Eigenvalue decomposition of the average correlation matrix into market mode, group modes and random modes. The market mode captures the mean market correlation. The group modes give the sectoral behavior of the market. The random modes of the correlation matrix yield the Marčenko-Pastur distribution. **c** Eigenvalue spectrum of the correlation matrix, evaluated for the *long* return time series for the entire period of 32-years, with the maximum eigenvalue of the normal spectrum $\lambda_{max} = 55.72$. The largest eigenvalue is well separated from the 'bulk'. Inset shows the random part of the spectrum, with the smallest eigenvalue of the normal spectrum $\lambda_{min} = 0.22$

Figure 2.7a shows the cross-correlation matrices constructed from *surrogate* data ($N = 194$ correlated Gaussian noises, each of length $T = 10000$) such that the matrix has 10 diagonal blocks of different correlations (equal to the average correlations of different sectors of the S&P 500 market). Figure 2.7d shows the *surrogate* cross-correlation matrix ($N = 194$; $T = 10000$) but now with one big block and 6 smaller blocks. The mean correlation of the big block is equal to the mean correlation of four sectors (**CD**, **FN**, **ID** and **MT** of Fig. 2.6a) and they show high inter-sectorial correlation in S&P 500 market in 32 years. Eigenvalue spectra of the correlation matrices are shown in Fig. 2.7b, e, each of which consists of the Marčenko-Pastur distributions (see insets), followed by 10 (and 7) eigenvalues corresponding to 10 (and 7) blocks (similar to sectors), respectively. Figure 2.7c, f show the $3D$ MDS plots, where the points (representing stocks) are scattered based on the correlations among the 10 and 7 blocks, respectively. In the MDS maps, more correlated stocks are placed nearby and anti-correlated are placed far apart (see also Ref. [23]). The k-means clustering performed on the surrogate data matrices, with $k = 10$ and $k = 7$, yield 10 and 7 different clusters (represented in different colors), respectively.

Dynamics of the Correlation Structure of US Market

Next, we study the time evolution of the market correlations computed with the daily returns of $N = 194$ stocks of S&P 500 over the period of 32-year (1985–2016, with $T = 8068$ trading days).

Figure 2.8a, b show plots of mean of correlation coefficients ($< C_{ij} >$), mean of absolute values of correlation coefficients ($< |C_{ij}| >$) and the difference of the

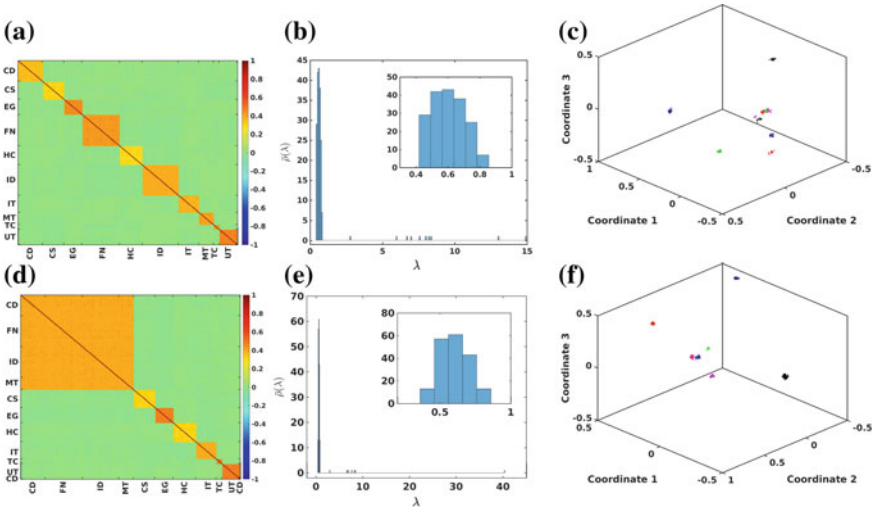


Fig. 2.7 **a** Cross-correlation matrices constructed from the correlated Gaussian time series with 10 diagonal blocks of different correlations (equal to the average correlation of each sector in Fig. 2.6a). **d** shows the same cross-correlation matrix but with one big block and 6 smaller blocks. The mean correlation of the big block is equal to the mean correlation of four sectors (**CD**, **FN**, **ID** and **MT** of Fig. 2.6a). They have high inter-sectorial correlation over the last 32 years in S&P 500 market. **b** and **e** show the eigenvalue spectra of the correlation matrices, which consist of the Marčenko-Pastur distributions followed by 10 group modes corresponding to 10 sectors and 7 group modes corresponding to 7 sectors, respectively. Insets show the enlarged pictures of the random part of the spectrum. **c** and **f** show plots of 10 and 7 different clusters, respectively, drawn in different colors using 3-dimensional k-means clustering technique. The clustering was performed on 3-D multidimensional scaling (MDS) map of 194 stocks. Each point on the MDS map represents a stock of the market. The points are scattered in the map, based on the cross-correlations among the stocks—more correlated stocks are placed nearby and less correlated are placed far apart (see also Ref. [23])

absolute mean and the mean of correlation coefficients $df = < |C_{ij}| > - < C_{ij} >$ for short epochs of $M = 20$ days, with shifts of: $\Delta\tau = 1$ day (95% overlap) and $\Delta\tau = 10$ days (50% overlap), respectively. Shifts toward the positive side of correlations are pointing toward periods of market crashes (with very high mean correlation values). The values of df are anti-correlated with the values of the mean of correlation coefficients. During a market crash when mean of correlation coefficient is high, there are very little anti-correlations among the stocks, then the value of df is extremely small, indeed near to zero (see Ref. [17]). It may act as an indicator of a market crash, as we observe that there is a high anti-correlation between the values of df and $< C_{ij} >$, with leads of one or two days (ahead of the market crashes). Similarly, Fig. 2.8c, d show the plots of variance, skewness, and kurtosis of the correlation coefficients C_{ij} as functions of time with shifts of $\Delta\tau = 1$ day and $\Delta\tau = 10$ days, respectively. The mean correlation is anti-correlated to variance and skewness of $\boldsymbol{C}$, i.e., when the mean correlation is high then both variance and skewness are low. Kurtosis is highly correlated to the mean correlation. These observations are seen in

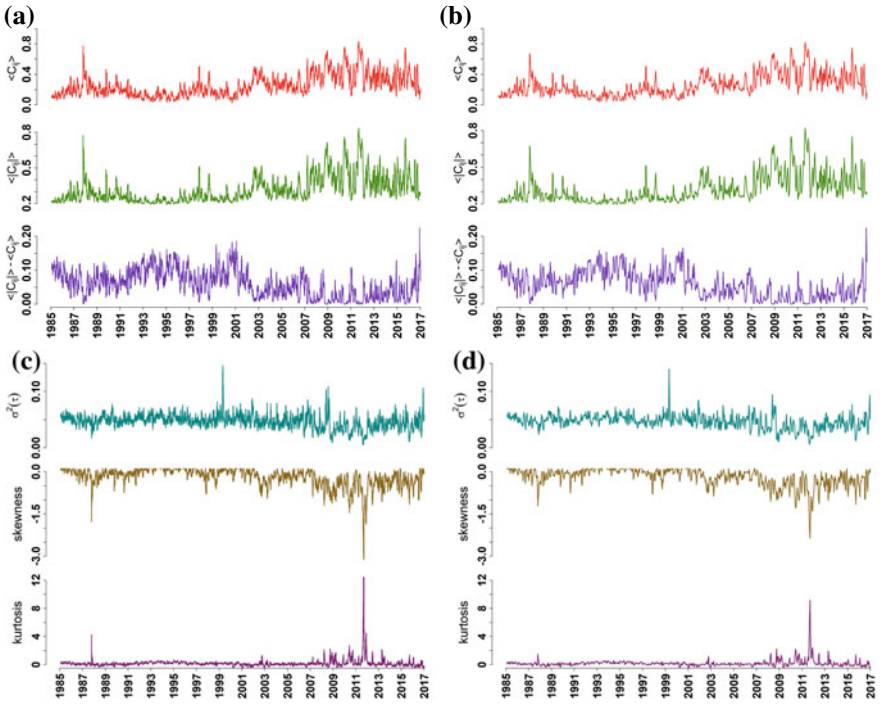


Fig. 2.8 Plots of mean of correlation coefficients ($< C_{ij} >$), mean of absolute values of correlation coefficients ($< |C_{ij}| >$) and the difference ($df =< |C_{ij}| > - < C_{ij} >$) as functions of time, for short epochs of $M = 20$ days, and shifts of: **a** $\Delta\tau = 1$ day and **b** $\Delta\tau = 10$ days. We find that during crashes (when mean correlation is very high), the difference $df =< |C_{ij}| > - < C_{ij} >$ show minima (close to zero) (see Ref. [17]). Plots of variance (σ^2), skewness, and kurtosis of the correlation coefficients as functions of time, for short epochs of $M = 20$ days, and shifts of: **c** $\Delta\tau = 1$ day and **d** $\Delta\tau = 10$ days

the dynamical evolution of the market with epochs of $M = 20$ days, and shifts of $\Delta\tau = 1, 10$ day(s).

The scatter plots between $< C_{ij} >$ and $< |C_{ij}| >$, and $< C_{ij} >$ and $df (=< |C_{ij}| > - < C_{ij} >)$ for different time lags (no-lag, lag-1, lag-2, and lag-3) of empirical correlation matrices $\boldsymbol{C}(\tau)$, with 194 stocks of S&P 500 and epochs of $M = 20$ days, and shift of $\Delta\tau = 1$ day, are shown in Fig. 2.9a, b, respectively. Here lag-1, lag-2, and lag-3 represent time lags of 1 day, 2 days, and 3 days, respectively. The color-bar shows the time period from 1985 to 2016 in years. The scatter plots show the correlations among $< C_{ij} >$ versus $< |C_{ij}| >$ and $< C_{ij} >$ versus df, at different time lags. The variances of the scatter plots increase with the increase of time lag, keeping the value of linear correlation coefficient almost similar. The strong linear correlation between $< C_{ij} >$ and $< |C_{ij}| >$ may give us an early information about a crash up to 3 days ahead (from the result of lag-3). Similar linear correlations are also visible in Fig. 2.9c, d, between $< C_{ij} >$ and $< |C_{ij}| >$, and $< C_{ij} >$ and df, at different time lags (no-lag, lag-1, lag-2, and lag-3) for a shift of $\Delta\tau = 10$ days. Here,

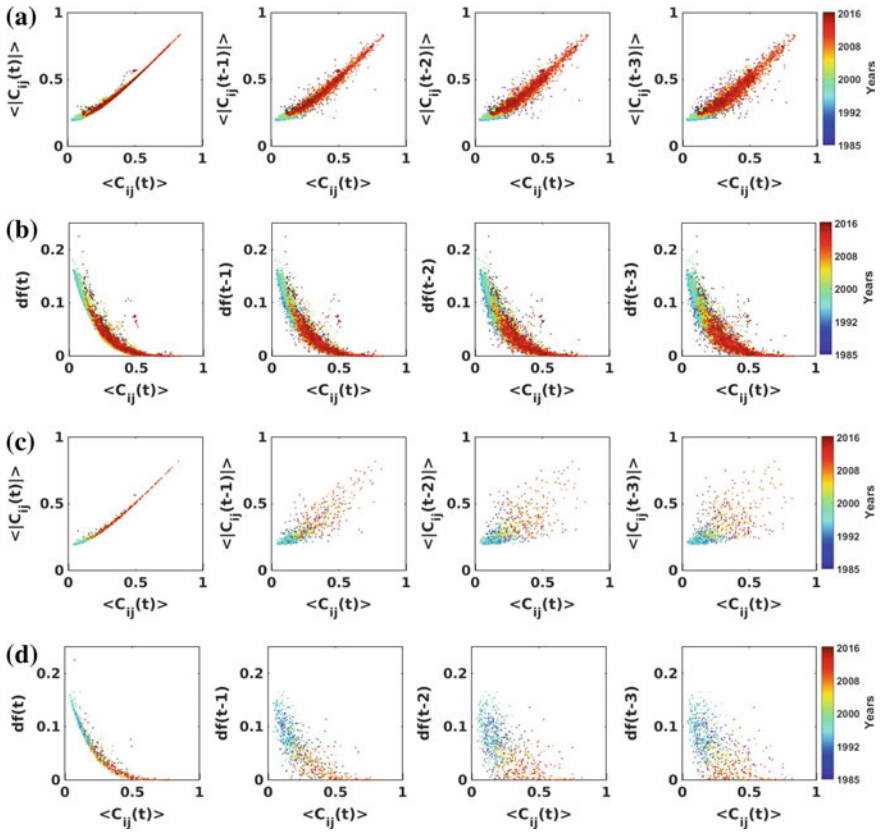


Fig. 2.9 Scatter plots of $< C_{ij} >$ versus $< |C_{ij}| >$ and $< C_{ij} >$ versus $df =< |C_{ij}| > - < C_{ij} >$, for different time lags (No lag, 1-day, 2-days and 3-days) for the correlation matrix of epoch 20 days, with shifts of: **a**–**b** $\Delta\tau = 1$ day; **c**–**d** $\Delta\tau = 10$ days. The color-bar shows the time period in years

obviously lag-1, lag-2, and lag-3 represent time lags of 10 days, 20 days, and 30 days, respectively. The large variances in scatter plots indicate that it is hard to detect and extract information about a crash, e.g., 30 days in advance.

Figure 2.10a shows the temporal variation of mean correlation ($< C_{ij} >$), maximum eigenvalue (λ_{max}), number of negative eigenvalues ($\#-ve\ EV$) and smallest eigenvalue (λ_{min}) of the emerging spectra with a shift of $\Delta\tau = 1$ day. Using a small distortion ($\varepsilon = 0.01$), we break the degeneracy of eigenvalues at zero and get the "emerging spectra" of eigenvalues which contain some interesting infromation about the market. The effect of the small distortion parameter $\varepsilon = 0.01$ is negligible on non-zero eigenvalues of the spectrum including λ_{max}. We observed high correlation between $< C_{ij} >$ and λ_{max}. But the other properties of emerging spectrum ($\#-ve\ EV$ and λ_{min}) are less correlated with mean correlation$< C_{ij} >$ [21]. Figure 2.10b shows the same for the shift of $\Delta\tau = 10$ days.

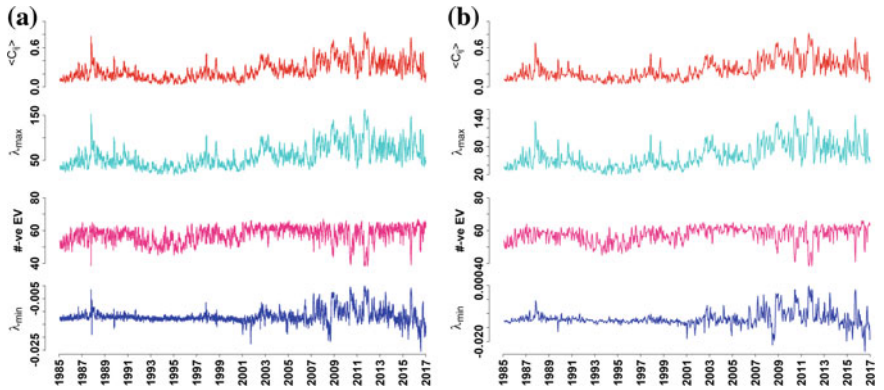


Fig. 2.10 Plots for mean of correlation coefficients ($< C_{ij} >$), maximum eigenvalue (λ_{max}), number of negative eigenvalues ($\# - ve\ EV$) and smallest eigenvalue (λ_{min}) of the spectrum as a function of time for an epoch of 20 days at $\varepsilon = 0.01$ with shifts of: **a** $\Delta\tau = 1$ day and **b** $\Delta\tau = 10$ days. The correlation between $< C_{ij} >$ and λ_{max} is high, but two other properties of the "emerging spectrum" ($\# - ve\ EV$ and λ_{min}) are less correlated to mean correlation $< C_{ij} >$

Recent Applications of RMT in Financial Markets

Identification of Market States and Long-Term Precursors to a Crash State

The study of the critical dynamics in any complex system is interesting, yet it can be very challenging. Recently, Pharasi et al. [23] presented an analysis based on the correlation structure patterns of S&P 500 (USA) data and Nikkei 225 (JPN) data, with short time epochs during the 32-year period of 1985–2016. They identified "market states" as clusters of similar correlation structures which occurred more frequently than by pure chance (randomness).

They first used the power mapping to reduce noise of the singular correlation matrices and obtained distinct and denser clusters in three dimensional MDS map (as shown in Fig. 2.11a). The effects of noise-suppression were found to be prominent not only on a single correlation matrix at one epoch, but also on the similarity matrices computed for different correlation matrices at different short-time epochs, and their corresponding MDS maps. Using 3D-multidimensional scaling maps, they applied k-means clustering to divide the clusters of similar correlation patterns into k groups or market states. One major difficulty of this clustering method is that one has to pass the value of k as an input to the algorithm. Normally, there are several proposed methods of determining the value of k (often arbitrary). Pharasi et al. [23] showed that using a new prescription based on the cluster radii and an optional choice of the noise suppression parameter, one could have a fairly robust determination of the "optimal" number of clusters.

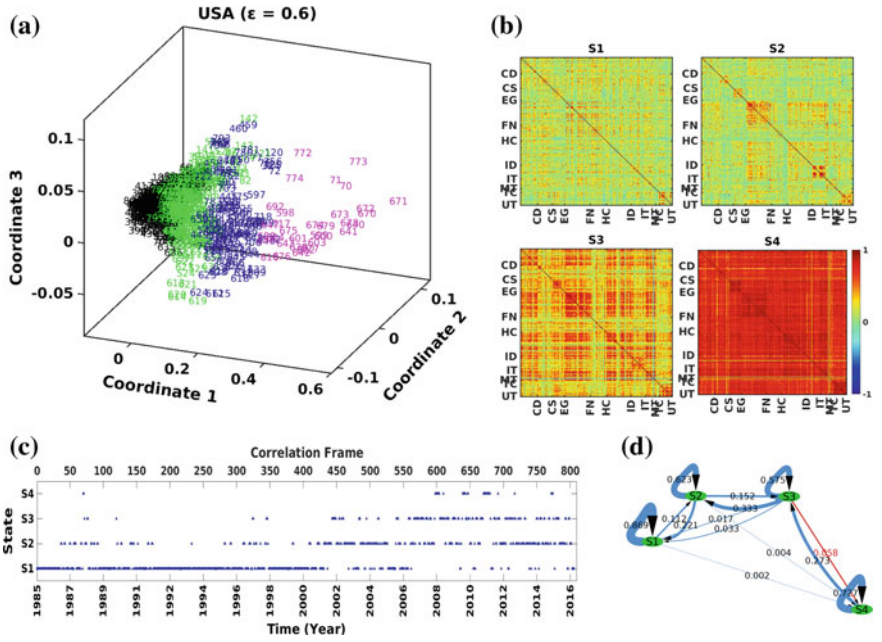


Fig. 2.11 **a** Classification of the US market into four typical market states. k-means clustering is performed on a MDS map constructed from noise suppressed ($\varepsilon = 0.6$) similarity matrix [20]. The coordinates assigned in the MDS map are the corresponding correlation matrices constructed from short-time series of $M = 20$ days and shifted by $\Delta\tau = 10$ days. **b** shows the four different states of the US market as S1, S2, S3 and S4, where S1 corresponds to a calm state with low mean correlation, and S4 corresponds to a critical state (crash) with high mean correlation. **c** Temporal dynamics of the US market in four different states ($S1$, $S2$, $S3$ and $S4$) for the period of $1985 - 2016$. **d** Network plot for transition probabilities of paired market states (MS). The transition probability of paired market states going from $S1$ and $S2$ to $S4$ is much lesser than 1% but from $S3$ to $S4$ is 6%. Figure adapted from Ref. [23]

In the new prescription, they measured the mean and the standard deviation of the intra-cluster distances using an ensemble of fairly large number (about 500) of different initial conditions (choices of random coordinates for the k-centroids or equivalently random initial clustering of n objects); each set of initial conditions usually results in slightly different clustering of the n objects representing different correlation matrices. If the clusters of points are very distinct in the coordinate space, then even for different initial conditions, the k-means clustering method yields same results, producing a small variance of the intra-cluster distance. However, the problem of allocating the matrices into the different clusters becomes problematic, when the clusters are very close or overlapping, as the initial conditions can then influence the final clustering of the different points; so there is a larger variance of the intra-cluster distance for the ensemble of initial conditions. Therefore, a minimum variance or standard deviation for a particular number of clusters implies the robustness of the clustering. For optimizing the number of clusters, Pharasi et al. proposed that one should look for *maximum* k, which has the *minimum variance* or standard

deviation in the intra-cluster distances with different initial conditions. Thus, based on the modified prescription of finding similar clusters of correlation patterns, they characterized the market states for USA and JPN.

Here, in Fig. 2.11b, we reproduce the results for the US market, showing four typical market states. The evolution of the market can be then viewed as the dynamical transitions between market states, as shown in Fig. 2.11c. Importantly, this method yields the correlation matrices that correspond to the critical states (or crashes). They correspond to the well-known financial market crashes and clustered in market state $S4$. They also analyzed the transition probabilities of the paired market states, and found that (i) the probability of remaining in the same state is much higher than the transition to a different states, and (ii) most probable transitions are the nearest neighbor transitions, and the transitions to other remote states are rare (see Fig. 2.11d). Most significantly, the state adjacent to a critical state (crash) behaved like a long-term "precursor" for a critical state, serving an early warning for a financial market crash.

Characterization of Catastrophic Instabilities

Market crashes, floods, earthquakes, and other catastrophic events, though rarely occurring, can have devastating effects with long term repurcussions. Therefore, it is of primal importance to study the complexity of the underlying dynamics and signatures of catastrophic events. Recently, Sharma et al. [8] studied the evolution of cross-correlation structures of stock return matrices and their eigenspectra over different short-time epochs for the US market and Japanese market. By using the power mapping method, they applied the non-linear distortion with a small value of distortion parameter $\varepsilon = 0.01$ to correlation matrices computed for any epoch, leading to the *emerging spectrum* of eigenvalues.

Here, we reproduce some of the significant findings of the paper [8]. Interestingly, it is found that the statistical properties of the emerging spectrum display the following features: (i) the shape of the emerging spectrum reflects the market instability (see Fig. 2.12a, b), (ii) the smallest eigenvalue (in a similar way as the maximum eigenvalue, which captured the mean correlation of the market) indicated that the financial market had become more turbulent, especially from 2001 onward (see Fig. 2.12c), and (iii) the smallest eigenvalue is able to statistically distinguish the nature of a market turbulence or crisis—internal instability or external shock (see Fig. 2.12c). In certain instabilities the smallest eigenvalue of the emerging spectrum was positively correlated with the largest eigenvalue (and thus with the mean market correlation) while in other cases there were trivial anti-correlations. They proposed that this behavioral change could be associated to the question whether a crash is associated to intrinsic market conditions (e.g., a bubble) or to external events (e.g., the Fukushima meltdown). A lead-lag effect of the crashes was also observed through the behavior of λ_{min} and mean correlation, which could be examined further.

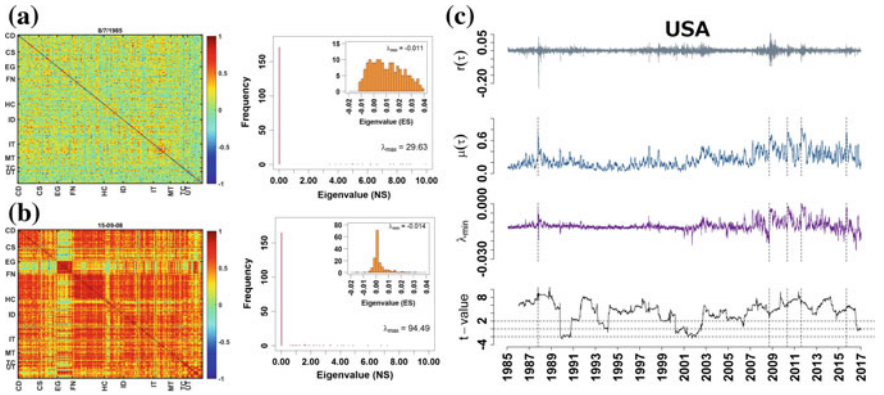


Fig. 2.12 **a** Non-critical (normal) period of the correlation matrix and its eigenvalue spectrum, evaluated for the *short* return time series for an epoch of $M = 20$ days ending on 08-07-1985, with the maximum eigenvalue of the normal spectrum $\lambda_{max} = 29.63$. Inset: Emerging spectrum using power map technique ($\varepsilon = 0.01$) is a deformed semi-circle, with the smallest eigenvalue of the emerging spectrum $\lambda_{min} = -0.011$. **b** Critical (crash) period of the correlation matrix and its eigenspectrum, evaluated for an epoch of $M = 20$ days ending on 15-09-2008, with the maximum eigenvalue of the normal spectrum $\lambda_{max} = 94.49$. Inset: Emerging spectrum using power map technique ($\varepsilon = 0.01$) is Lorentzian, with the smallest eigenvalue of the emerging spectrum $\lambda_{min} = -0.014$. **c** USA (i) market return $r(t)$, (ii) mean market correlation $\mu(t)$, (iii) smallest eigenvalue of the emerging spectrum (λ_{min}), and (iv) t-value of the t-test, which tests the statistical effect over the lag-1 smallest eigenvalue $\lambda_{min}(t-1)$ on the mean market correlation $\mu(t)$. The mean of the correlation coefficients and the smallest eigenvalue in the emerging spectra are correlated to a large extent. Notably, the smallest eigenvalue behaves differently (sharply rising *or* falling) at the same time when the mean market correlation is very high (crash). The vertical dashed lines correspond to the major crashes, which brewed due to internal market reactions. Note that, the smallest eigenvalue of the US market indicates that the financial market has become more turbulent from 2001 onward. Figure adapted from Ref. [8]

Concluding Remarks

We have presented a brief overview of the Wishart and correlated Wishart ensembles in the context of financial time series analysis. We displayed the dependence of the length of the time series on the eigenspectra of the Wishart ensemble. The eigenspectra of large random matrices are not very sensitive to $Q = T/N$; however, the amount of spurious correlations is dependent on it. To avoid the problem of non-stationarity and suppress the noise in the correlation matrices, computed over short epochs, we applied the power mapping method on the correlation matrices. We showed that the shape of the emerging spectrum depends on the amount of the correlation U of the correlated Wishart ensemble. We also studied the effect of the non-linear distortion parameter ε on the emerging spectrum.

Then we demonstrated the eigenvalue decomposition of the empirical cross-correlation matrix into market mode, group modes and random modes, using the

return time series of 194 stocks of S&P 500 index during the period of 1985-2016. The bulk of the eigenvalues behave as random modes and give rise to the Marčenko-Pastur. We also created surrogate correlation matrices to understand the effect of the sectoral correlations. Then we studied the eigenvalue distribution of those matrices as well as k-means clustering on the MDS maps generated from the correlation matrices. Evidently, if we have 10 diagonal blocks (representing sectors) then we get 10 clusters on a MDS map. Similarly, when we merged the four blocks to one and had 7 diagonal blocks then again we got 7 clusters on the MDS map.

Further, we studied the dynamical evolution of the statistical properties of the correlation coefficients using the returns of the S&P 500 stock market. We computed the mean, the absolute mean, the difference between absolute mean and mean, variance, skewness and kurtosis of the correlation coefficients C_{ij}, for short epochs of $M = 20$ days and shifts of $\Delta\tau = 1$ day and $\Delta\tau = 10$ days. We also showed the evolution of the mean of correlation coefficients, maximum eigenvalue of the correlation matrix, as well as the number of negative eigenvalues and smallest eigenvalue of the emerging spectrum, for the same epoch and shift.

Finally, we discussed the applications of RMT in financial markets. In an application, we demonstrated the use of RMT and correlation patterns in identifying possible "market states" and long-term precursors to the market crashes. In the second application, we presented the characterization of catastrophic instabilities, i.e., the market crashes, using the smallest eigenvalue of the emerging spectra arising from correlation matrices computed over short epochs.

Acknowledgements The authors thank R. Chatterjee, S. Das and F. Leyvraz for various fruitful discussions. A.C. and K.S. acknowledge the support by grant number BT/BI/03/004/2003(C) of Govt. of India, Ministry of Science and Technology, Department of Biotechnology, Bioinformatics division, University of Potential Excellence-II grant (Project ID-47) of JNU, New Delhi, and the DST-PURSE grant given to JNU by the Department of Science and Technology, Government of India. K.S. acknowledges the University Grants Commission (Ministry of Human Resource Development, Govt. of India) for her senior research fellowship. H.K.P. is grateful for postdoctoral fellowship provided by UNAM-DGAPA. A.C., H.K.P., K.S. and T.H.S. acknowledge support by Project CONACyT Fronteras 201, and also support from the project UNAM-DGAPA-PAPIIT IG 100616.

References

1. Bar-Yam, Y.: General Features of Complex Systems. Encyclopedia of Life Support Systems (EOLSS). UNESCO, EOLSS Publishers, UK (2002)
2. Bendat, J.S., Piersol, A.G.: Engineering Applications of Correlation and Spectral Analysis, p. 315. Wiley-Interscience, New York (1980)
3. Bouchaud, J.P., Potters, M.: Theory of Financial Risk and Derivative Pricing: from Statistical Physics to Risk Management. Cambridge University Press, Cambridge (2003)
4. Cartan, É.: Sur les domaines bornés homogènes de lespace den variables complexes. In: Abhandlungen aus dem mathematischen Seminar der Universität Hamburg, vol. 11, pp. 116–162. Springer, Berlin (1935)

5. Chakraborti, A., Muni Toke, I., Patriarca, M., Abergel, F.: Econophysics review: I. empirical facts. Quant. Financ. **11**(7), 991–1012 (2011)
6. Chakraborti, A., Muni Toke, I., Patriarca, M., Abergel, F.: Econophysics review: II. agent-based models. Quant. Financ. **11**(7), 1013–1041 (2011)
7. Chakraborti, A., Patriarca, M., Santhanam, M.: Financial time-series analysis: a brief overview. In: Econophysics of Markets and Business Networks, pp. 51–67. Springer, Berlin (2007)
8. Chakraborti, A., Sharma, K., Pharasi, H.K., Das, S., Chatterjee, R., Seligman, T.H.: Characterization of catastrophic instabilities: market crashes as paradigm (2018). arXiv:1801.07213
9. Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014)
10. Gell-Mann, M.: What is complexity? Complexity **1**, 16–19 (1995)
11. Guhr, T., Kälber, B.: A new method to estimate the noise in financial correlation matrices. J. Phys. A: Math. Gen. **36**(12), 3009 (2003)
12. Hua, L.: Harmonic Analysis of Functions of Several Complex Variables in the Classical Domains, vol. 6. American Mathematical Society (1963)
13. Jin, X., Wah, B.W., Cheng, X., Wang, Y.: Significance and challenges of big data research. Big Data Res. **2**(2), 59–64 (2015)
14. Leviandier, L., Lombardi, M., Jost, R., Pique, J.P.: Fourier transform: a tool to measure statistical level properties in very complex spectra. Phys. Rev. Lett. **56**(23), 2449 (1986)
15. Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge (2007)
16. Marčenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Math. USSR-Sbornik **1**(4), 457 (1967)
17. Martinez, M.M.R.: Caracterización estadistica de mercados europeos. Master's thesis, UNAM (2018)
18. Mehta, M.L.: Random Matrices. Academic (2004)
19. Mikosch, T., Stărică, C.: Nonstationarities in financial time series, the long-range dependence, and the igarch effects. Rev. Econ. Stat. **86**(1), 378–390 (2004)
20. Münnix, M.C., Shimada, T., Schäfer, R., Leyvraz, F., Seligman, T.H., Guhr, T., Stanley, H.E.: Identifying states of a financial market. Sci. Rep. **2**, 644 (2012)
21. Ochoa, S.: Mapeo de Guhr-Kaelber aplicado a matrices de correlación singulares de dos mercados financieros. Master's thesis, UNAM (2018)
22. Pandey, A., et al.: Correlated Wishart ensembles and chaotic time series. Phys. Rev. E **81**(3), 036202 (2010)
23. Pharasi, H.K., Sharma, K., Chatterjee, R., Chakraborti, A., Leyvraz, F., Seligman, T.H.: Identifying long-term precursors of financial market crashes using correlation patterns. New J. Phys. **20**, 103041 (2018). arXiv:1809.00885
24. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T., Stanley, H.E.: Random matrix approach to cross correlations in financial data. Phys. Rev. E **65**(6), 066126 (2002)
25. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Stanley, H.E.: Universal and nonuniversal properties of cross correlations in financial time series. Phys. Rev. Lett. **83**(7), 1471 (1999)
26. Schäfer, R., Nilsson, N.F., Guhr, T.: Power mapping with dynamical adjustment for improved portfolio optimization. Quant. Financ. **10**(1), 107–119 (2010)
27. Sharma, K., Shah, S., Chakrabarti, A.S., Chakraborti, A.: Sectoral co-movements in the Indian stock market: a mesoscopic network analysis, pp. 211–238 (2017)
28. Shuryak, E.V., Verbaarschot, J.: Random matrix theory and spectral sum rules for the Dirac operator in QCD. Nuclear Phys. A **560**(1), 306–320 (1993)
29. Sinha, S., Chatterjee, A., Chakraborti, A., Chakrabarti, B.K.: Econophysics: an Introduction. Wiley, New York (2010)
30. Utsugi, A., Ino, K., Oshikawa, M.: Random matrix theory analysis of cross correlations in financial markets. Phys. Rev. E **70**(2), 026110 (2004)
31. Vemuri, V.: Modeling of Complex Systems: An Introduction. Academic, New York (1978)

32. Vinayak, Prosen, T., Buča, B., Seligman, T.H.: Spectral analysis of finite-time correlation matrices near equilibrium phase transitions. Europhys. Lett. **108**(2), 20006 (2014)
33. Vinayak, Schäfer, R., Seligman, T.H.: Emerging spectra of singular correlation matrices under small power-map deformations. Phys. Rev. E **88**(3), 032115 (2013)
34. Vinayak, Seligman, T.H.: Time series, correlation matrices and random matrix models. In: AIP Conference Proceedings, vol. 1575, pp. 196–217. AIP (2014)
35. Vyas, M., Guhr, T., Seligman, T.H.: Multivariate analysis of short time series in terms of ensembles of correlation matrices (2018). arXiv:1801.07790
36. Wigner, E.: Ep wigner. Ann. Math. **53**, 36 (1951)
37. Wigner, E.P.: On the distribution of the roots of certain symmetric matrices. Ann. Math. 325–327 (1958)
38. Wigner, E.P.: Random matrices in physics. SIAM Rev. **9**(1), 1–23 (1967)
39. Wishart, J.: The generalised product moment distribution in samples from a normal multivariate population. Biometrika 32–52 (1928)
40. Yahoo finance database. https://finance.yahoo.co.jp/ (2017). Accessed 7 July 2017, using the R open source programming language and software environment for statistical computing and graphics

Chapter 3
A Few Simulation Results of Basic Models of Limit Order Books

Ioane Muni Toke

Abstract We use a simplified framework for the modeling of limit order books, in which only the best quotes (prices and volumes) are monitored. Within this framework we test models in which the flows of limit and market orders are modeled by Poisson processes, Hawkes processes, or processes with state-dependent intensities. We provide simulation results to compare some distributions of interest, such as volumes, price, spread, autocorrelation of orders signs, etc.

General Framework

Modeling of limit order books has been of interest in the financial microstructure community for some time. References [1, 3, 4] provide some review elements.

Dynamics of the Limit Order Book

We consider a simplified framework modeling the limit order book, in which we focus on the best quotes. The dynamics of X is defined by three main types of events, namely limit orders, market orders and cancellations. All types of orders can be submitted either on the ask side or on the bid side. We now specify the characteristics of these orders.

A limit order can either be submitted at the best quote or inside the spread (*aggressive* limit order) if the spread is large enough at the time of submission ($s(t-) > 1$). This is modeled by drawing a random variable δ with conditional distribution $\pi^{S-}_{s(t-)=s}$ with integer support $\{0, \ldots, s-1\}$, i.e. depending on the spread at the time of submission. If $\delta = 0$, then the limit order is submitted at the best quote: the quantity

I. Muni Toke (✉)
Mathématiques et Informatique pour la Complexité et les Systèmes, CentraleSupélec, Université Paris-Saclay, 91190 Gif-Sur-Yvette, France
e-mail: ioane.muni-toke@centralesupelec.fr

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_3

Table 3.1 Summary of the possible effects of a limit order with size σ^L and placement δ on the limit order book X

Side	Placement	Dynamics			
		$b(t)$	$a(t)$	$p^B(t)$	$s(t)$
Bid	$\delta = 0$	$b(t-) + \sigma^L$	$a(t-)$	$p^B(t-)$	$s(t-)$
	$\delta > 0$	σ^L	$a(t-)$	$p^B(t-) + \delta$	$s(t-) - \delta$
Ask	$\delta = 0$	$b(t-)$	$a(t-) + \sigma^L$	$p^B(t-)$	$s(t-)$
	$\delta > 0$	$b(t-)$	σ^L	$p^B(t-)$	$s(t-) - \delta$

Table 3.2 Summary of the possible effects of a market order with size σ^M on the limit order book X. δ, q_B and q_A are random variables with distribution π^{S+}, $\pi^{B,2}$ and $\pi^{A,2}$

Side	Size	Dynamics			
		$b(t)$	$a(t)$	$p^B(t)$	$s(t)$
Bid	$\sigma^M < b(t-)$	$b(t-) - \sigma^M$	$a(t-)$	$p^B(t-)$	$s(t-)$
	$\sigma^M \geq b(t-)$	q_B	$a(t-)$	$p^B(t-) - \delta$	$s(t-) + \delta$
Ask	$\sigma^M < a(t-)$	$b(t-)$	$a(t-) - \sigma^M$	$p^B(t-)$	$s(t-)$
	$\sigma^M \geq a(t-)$	$b(t-)$	q_A	$p^B(t-)$	$s(t-) + \delta$

available at the best quote is increased by the size of the limit order σ^L. If $\delta > 0$, then the limit order is submitted inside the spread: the spread is reduced by δ (hence the notation π^{S-}) and the volume of the best quote is reset to the size of the limit order denoted by σ^L. Table 3.1 summarizes the effect of a limit order on X.

If an ask market order with size σ^M is submitted, two situations are possible. If $\sigma^M < a(t-)$, then the best quote is reduced by σ^M: $a(t) = a(t-) - \sigma^M$, and the rest of the order book is unchanged. But if $\sigma^M \geq a(t-)$ (i.e. in the case of an aggressive market order that moves the price), then the size of the best quote is reset to a random quantity q_A with some distribution $\pi^{A,2}$ and the spread is increased by some random quantity δ with distribution π^{S+} with values in $\mathbb{N}^*$. Symmetric effects occur on the bid side. Table 3.2 summarizes the dynamics of X in case of a market order.

Finally, cancellations have the same effects as market orders, since they also consist in decreasing the quantities available. When a cancellation occurs, one standing order is selected. If its size σ^C is strictly less than the current best quote, then prices do not change and the volume decreases by σ^C. If not, then a reset occurs as in the case of an aggressive market order. Table 3.3 summarizes the dynamics of X in case of a cancellation.

Common Notations

All models described below focus on bid market orders (MB), ask market orders (MA), bid limit orders (LB) and ask limit orders (LA). Let $\mathcal{T} \triangleq \{MB, MA, LB, LA\}$ be the set of types of orders under investigation. For $T \in \mathcal{T}$, $(N^T(t))_{t\geq 0}$ is the

Table 3.3 Summary of the possible effects of a cancellation of an order of size σ^C on the limit order book X. δ, q_B and q_A are random variables with distribution π^{S+}, $\pi^{B,2}$ and $\pi^{A,2}$

Side	Best quote size	Dynamics			
		$b(t)$	$a(t)$	$p^B(t)$	$s(t)$
Bid	$b(t-) > \sigma^C$	$b(t-) - \sigma^C$	$a(t-)$	$p^B(t-)$	$s(t-)$
	$b(t-) = \sigma^C$	q_B	$a(t-)$	$p^B(t-) - \delta$	$s(t-) + \delta$
Ask	$a(t-) > \sigma^C$	$b(t-)$	$a(t-) - \sigma^C$	$p^B(t-)$	$s(t-)$
	$a(t-) = \sigma^C$	$b(t-)$	q_A	$p^B(t-)$	$s(t-) + \delta$

counting process of all orders of type T, and $\lambda^T(t)$ denotes its intensity process. All subsequent models specify the four-dimensional intensity vector $\boldsymbol{\lambda} = (\lambda^{MB}, \lambda^{MA}, \lambda^{LB}, \lambda^{LA})$ of the vector process $\boldsymbol{N} = (N^{MB}, N^{MA}, N^{LB}, N^{LA})$.

Besides limit and market orders, bid and ask cancellations are counted by the process $(N^{CB}(t), N^{CA}(t))_{t\geq 0})$. All models investigated in this study assume that cancellations are proportional to the volume available in the book : $\lambda^{CB}(t) \triangleq \theta^B b(t)$ and $\lambda^{CA}(t) \triangleq \theta^A a(t)$.

Distributions of the sizes of orders are the same in all models. Distribution π^{S+} and family of conditional distributions $\left(\pi^{S-}_{s(t-)=n}\right)_{n\in\mathbb{N}}$ of spread movements are the same in all models. Reset distributions $\pi^{B,2}$ and $\pi^{A,2}$ are also the same in all models.

The behavior of the order book is assumed to be symmetric with respect to the bid and ask sides.

Models

Model 1—Poisson Processes

In this model, counting processes of bid/ask limit market orders are assumed to be homogeneous Poisson processes with constant intensity. Given the assumed bid/ask symmetry, the intensity vector is written

$$\forall t \geq 0, \boldsymbol{\lambda}(t) \triangleq (\mu^M, \mu^M, \mu^L, \mu^L) \in \left(\mathbb{R}^*_+\right)^4. \tag{3.1}$$

This Poisson model specification has thus 2 parameters to be estimated.

Model 2—Hawkes Processes

In this model, counting processes of bid/ask limit and market orders are assumed to form a four-dimensional Hawkes process with exponential kernels:

$$\forall t \geq 0, \boldsymbol{\lambda}(t) \triangleq \boldsymbol{\lambda}_0 + \int_0^t (\boldsymbol{\Phi} * d\boldsymbol{N})(s), \tag{3.2}$$

where $\boldsymbol{\lambda}_0 \in \left(\mathbb{R}_+^*\right)^4$, and $\boldsymbol{\Phi}$ is the 4×4-matrix with elements $\Phi_{T,U} : [0, \infty) \to \mathbb{R}_+, t \mapsto \alpha_{TU} e^{-\beta_{TU}(t)}$ for $T, U \in \mathscr{T}$. Again we assume a bid/ask symmetry, so that $\boldsymbol{\lambda}_0 \triangleq (\mu_0^M, \mu_0^M, \mu_0^L, \mu_0^L)$, and the matrices $\boldsymbol{\alpha} = (\alpha_{TU})_{T,U \in \mathscr{T}}$ and $\boldsymbol{\beta} = (\beta_{TU})_{T,U \in \mathscr{T}}$ have a specific cross structure per block:

$$\begin{pmatrix} \alpha_{MM}^s & \alpha_{MM}^o & \alpha_{ML}^s & \alpha_{ML}^o \\ \alpha_{MM}^o & \alpha_{MM}^s & \alpha_{ML}^o & \alpha_{ML}^s \\ \alpha_{LM}^s & \alpha_{LM}^o & \alpha_{LL}^s & \alpha_{LL}^o \\ \alpha_{LM}^o & \alpha_{LM}^s & \alpha_{LL}^o & \alpha_{LL}^s \end{pmatrix} \text{ and } \begin{pmatrix} \beta_{MM}^s & \beta_{MM}^o & \beta_{ML}^s & \beta_{ML}^o \\ \beta_{MM}^o & \beta_{MM}^s & \beta_{ML}^o & \beta_{ML}^s \\ \beta_{LM}^s & \beta_{LM}^o & \beta_{LL}^s & \beta_{LL}^o \\ \beta_{LM}^o & \beta_{LM}^s & \beta_{LL}^o & \beta_{LL}^s \end{pmatrix}, \tag{3.3}$$

where the exponents s and o stand for "same side" and "opposite side". This Hawkes model specification has thus 18 parameters to be estimated.

Model 3—Point Processes with Cox-Type State-Dependent Intensities

In this model, counting processes of bid/ask limit and market orders are assumed to be point processes with Cox-type state dependent intensities:

$$\forall t \geq 0, \lambda^T(t) \triangleq \exp\left(\sum_{k=0}^{K_T} \theta_k^T X_k^T(t)\right), \tag{3.4}$$

where $X_k^T, k = 1, \ldots, K_T$ are covariates describing the state of the order book, $T \in \mathscr{T}$. Given the assumed bid/ask symmetry, we have $K_{MB} = K_{MA} = K_M$, $K_{LB} = K_{LA} = K_L$ and for any i:

$$\theta_i^{MB} = \theta_i^{MA} = \theta_i^M, \text{ and } \theta_i^{LB} = \theta_i^{LA} = \theta_i^L. \tag{3.5}$$

More specifically, we follow [9] and assume that the state of the order book is described by the spread $s(t)$ and the volume of the best quote $q_1^T(t)$ (equal to $b(t)$ for bid orders, $a(t)$ for ask orders), so that intensities may be written in the following way:

$$\begin{aligned} \lambda^T(t) = \exp\big(&\theta_0^T + \theta_1^T \log s(t) + \theta_2^T \log q_1^T(t) + \theta_{11}^T (\log s(t))^2 \\ &+ \theta_{22}^T \left(\log q_1^T(t)\right)^2 + \theta_{21}^T \log s(t) \log q_1^T(t)\big), \end{aligned} \tag{3.6}$$

where $T \in \mathscr{T}$. This Cox model specification has thus 12 parameters to be estimated.

Models Estimation

Likelihood Analysis of the Intensity Process

Partial log-likelihood of the point process $\{N^T(t), t \in [0, \tau]\}$, with $T \in \mathscr{T}$, is written:

$$\mathscr{L}_\tau^T(\boldsymbol{\vartheta}) = \int_0^\tau \log \lambda^T(t)\, dN_t^M - \int_0^\tau \lambda^T(t)\, dt, \tag{3.7}$$

where $\boldsymbol{\vartheta}$ is a generic notation for the parameter vector of the model. For numerical purposes, let $\{t_i^T\}_{i=1,\ldots,n_\tau^T}$ be the set of all events of type $T \in \mathscr{T}$ in the sample $[0, \tau]$. In the case of Model 1, maximization of the partial log-likelihood straightforwardly leads to the estimators

$$\hat{\lambda}_0^T = \frac{n_\tau^T}{\tau}. \tag{3.8}$$

In the case of Model 2, the partial log-likelihood is written as a function of the parameter vector $(\boldsymbol{\lambda}_0, \boldsymbol{\alpha}, \boldsymbol{\beta})$:

$$\begin{aligned}\mathscr{L}_\tau^T(\boldsymbol{\lambda}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}) = &- \lambda_0^T \tau - \sum_{U \in \mathscr{T}} \sum_{i=1}^{n_\tau^U} \frac{\alpha_{TU}}{\beta_{TU}} \left(1 - e^{-\beta_{TU}(\tau - t_i^U)}\right) \\ &+ \sum_{i=1}^{n_\tau^T} \log \left[\lambda_0^m(t_i^T) + \sum_{U \in \mathscr{T}} \sum_{t_k^U < t_i^T} \alpha_{TU} e^{-\beta_{TU}(t_i^T - t_k^U)} \right].\end{aligned} \tag{3.9}$$

Recursive formulas are available for efficient computation of this quantity, and subsequent maximization. Details can be found in [8] and references therein. Finally, in the case of Model 3, if $\{t_i^s\}$ is the set of times of jumps of the spread process s, and $\{t_i^{q_1}\}$ is the set of times of jumps of the first limit process q_1 (equal to b for bid orders, a for ask orders), then the log-likelihood on the sample is numerically computed as follows:

$$\begin{aligned}\mathscr{L}_\tau^T(\boldsymbol{\theta}^T) = &\ \theta_0^T n_\tau^T + \theta_1^T \sum_{t_i^T} \log s(t_i^T -) + \theta_{11}^T \sum_{t_i^T} [\log s(t_i^T -)]^2 \\ &+ \theta_2^T \sum_{t_i^T} \log q_1(t_i^T -) + \theta_{22}^T \sum_{t_i^T} [\log q_1(t_i^T -)]^2 \\ &+ \theta_{12}^T \sum_{t_i^T} \log s(t_i^T -) \log q_1(t_i^T -) \\ &- \sum_{t_i \in \{t_i^S\} \cup \{t_i^{q_1}\}} (t_i - t_{i-1}) \exp \Big[\theta_0^T + \theta_1^T \log s(t_i -) + \theta_{11}^T [\log s(t_i -)]^2 \\ &+ \theta_2^T \log q_1(t_i -) + \theta_{22}^T [\log q_1(t_i -)]^2 + \theta_{12}^T \log s(t_i -) \log q_1(t_i -) \Big].\end{aligned} \tag{3.10}$$

Details about the estimation of Model 3 can be found in [9].

Estimation of Other Distributions

Sizes of Limit and Market Orders

Empirical sizes of orders are normalized by a common factor $\bar{\sigma}$ and rounded to the nearest integer. If the nearest integer is 0, then the volume is rounded above to 1. We model the distributions $(\pi^L(n))_{n\in\mathbb{N}^*}$ and $(\pi^M(n))_{n\in\mathbb{N}^*}$ of the normalized sizes of respectively limit and market orders with a one-parameter distribution:

$$\pi^T(1) = 1 - e^{-2\mu^T} \tag{3.11}$$

$$\pi^T(n) = e^{-(n-1)\mu^T} - e^{-n\mu^T}, \quad n \geq 2, \tag{3.12}$$

where $T \in \{M, L\}$ and μ^T is the parameter. This corresponds to a discretized exponential distribution, and μ^T can be estimated by the inverse of the mean size of orders of type T. Figure 3.1 shows an example of the calibration of $(\pi^L(n))_{n\in\mathbb{N}^*}$ and $(\pi^M(n))_{n\in\mathbb{N}^*}$ to empirical data.

Reset Distributions Upon Spread Increase

When the first limit drops to zero because of a trade or a cancellation, the volume of the best quote is reset to $\pi^{B,2}$ or $\pi^{A,2}$, which are equal by the bid-ask symmetry assumption. We model the normalized reset volumes by a basic geometric distribution with support on $\mathbb{N}^*$:

$$\forall n \in \mathbb{N}^*, \pi_n^{B,2} = \pi_n^{A,2} = q(1-q)^{n-1}, \tag{3.13}$$

where the parameter q is straightforwardly estimated. Figure 3.2 shows an example of the calibration of $(\pi_n^{A,2})_{n\in\mathbb{N}^*}$ and $(\pi_n^{B,2})_{n\in\mathbb{N}^*}$ to empirical data.

The distribution π^{S+} of the spread increase is also modeled by a basic geometric distribution with support on $\mathbb{N}^*$. Figure 3.3 shows an example of the calibration of $(\pi^{S+}(n))_{n\in\mathbb{N}^*}$ to empirical data.

Spread Decrease

When a limit order is submitted, spread is decreased by a quantity δ drawn from a distribution with support $\{0, \ldots, n-1\}$ given that $s(t-) = n$, denoted by $\pi^{S-}_{s(t-)=n}$. δ (resp. $-\delta$) is thus in ticks the placement of a bid (resp. ask) limit order relatively to the best quote. We model each conditional distribution with a two-parameter (a_n, p_n) modification of a geometric distribution that accounts for a greater activity at the best quote:

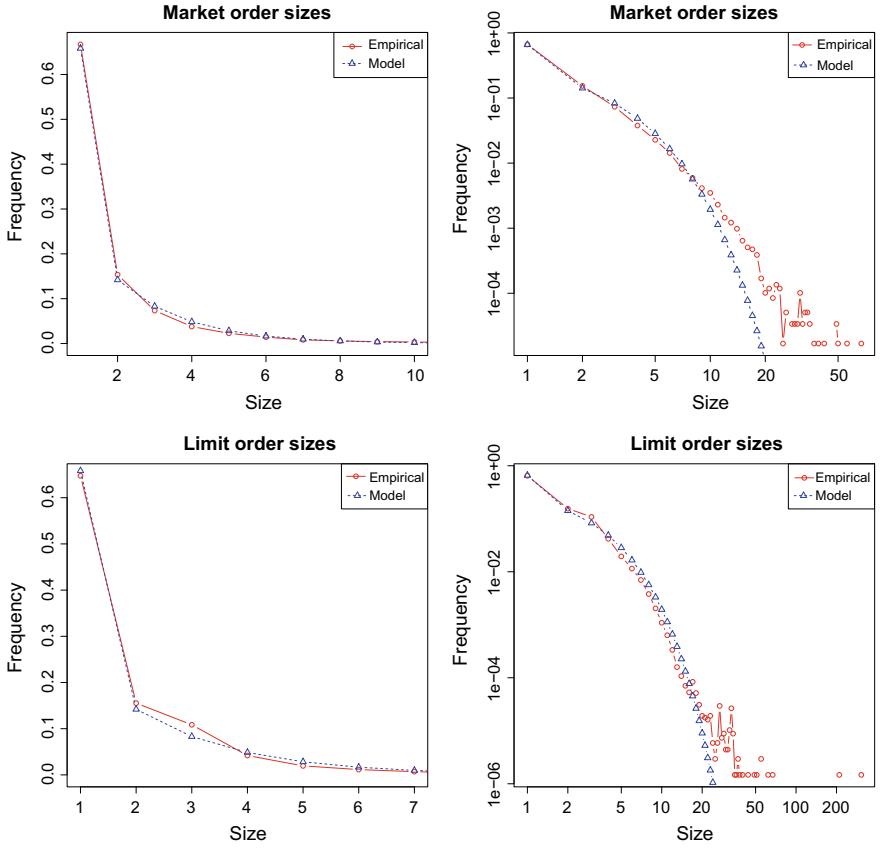


Fig. 3.1 Empirical and fitted distribution $(\pi^L(n))_{n\in\mathbb{N}}$ and $(\pi^M(n))_{n\in\mathbb{N}}$ of the normalized sizes of respectively limit orders (bottom) and market orders (top), in natural (left) and semi-log (right) scales. Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm. Normalizing size $\bar{\sigma} = 150$

$$\forall k \in \{0, \ldots, n-1\}, \pi^{S-}_{s(t-)=n}(k) = \frac{a_n \mathbf{1}_{k==0} + p_n(1-p_n)^k}{a_n + 1 - (1-p_n)^n}, \tag{3.14}$$

for $n \geq 2$. When $n = 1$ all limit orders are obviously submitted at the best quote. Each conditional distribution is separately fitted by likelihood maximization. Figure 3.4 shows an example of the calibration of the distributions $\pi^{S-}_{s(t-)=n}$ to empirical data.

Cancellation

The expected lifetime θ^{-1} of a pending order is the same in all models. Its estimator is computed such that, in the estimated Poisson model, the empirical rate of share arrival in the limit order book is equal to the empirical rate of share removal

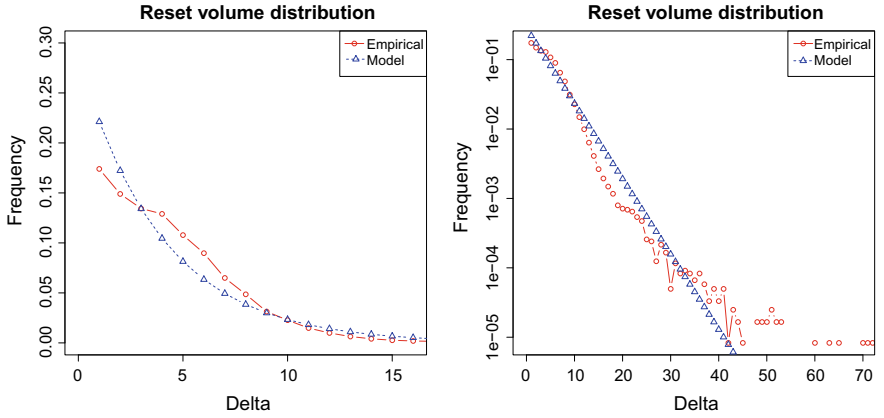


Fig. 3.2 Empirical and fitted distributions $\pi^{B,2} = \pi^{A,2}$ of the normalized sizes of the best quote upon spread increase, in natural (left) and semi-log (right) scales. Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm. Normalizing size $\bar{\sigma} = 150$

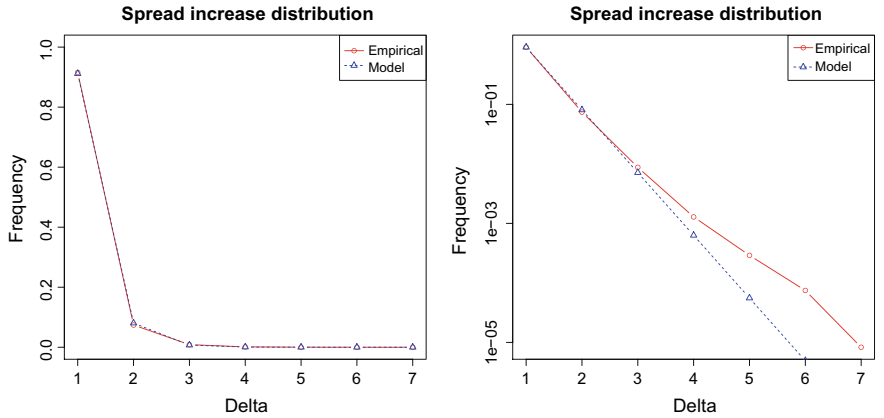


Fig. 3.3 Empirical and fitted distribution π^{S+} of the size of the spread increase when the best quote drops to zero, in natural (left) and semi-log (right) scales. Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm

$$\hat{\theta} = \frac{\hat{\lambda}_0^L - \hat{\lambda}_0^M}{\tau \bar{X}}, \tag{3.15}$$

where $\hat{\lambda}_0^L$ and $\hat{\lambda}_0^M$ are the estimated Poisson rates of arrival of respectively limit and market orders, τ is the sample length (horizon) and $\bar{X}$ is the empirical average normalized order book size.

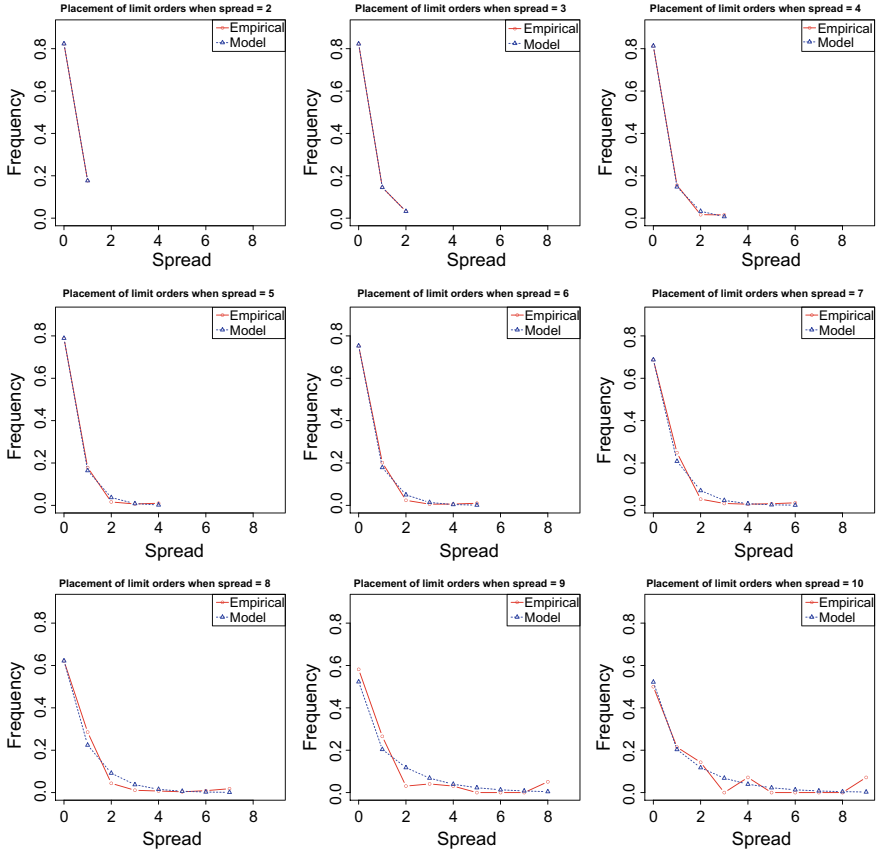


Fig. 3.4 Empirical and fitted distributions $\pi^{S-}_{s(t-)=n}$ of the placement of limit orders, for a spread size n ranging from 2 to 10 (top left to bottom right, horizontally). Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm

Simulation Results

Using TRTH data we reconstruct the order flows for the stock BNPP.PA on April 2016 using the method described in [7]. For each of the 21 trading days of this month, we extract a two-hour sample ranging from 10 am to 12 pm. All models are fitted on each daily sample, and each sample is treated as an independent realization. We then run simulations to generate a two-hour sample for each model. We compare empirical distributions of interest and the ones generated by each model.

As expected, Models 2 and 3 are able to reproduce the properties that were specifically targeted by their design. On the one hand, Model 2, based on Hawkes processes, handles the distributions of orders durations much better than the Poisson (Model 1) or state-dependent models (Model 3), as it is illustrated on Fig. 3.5. The clustering

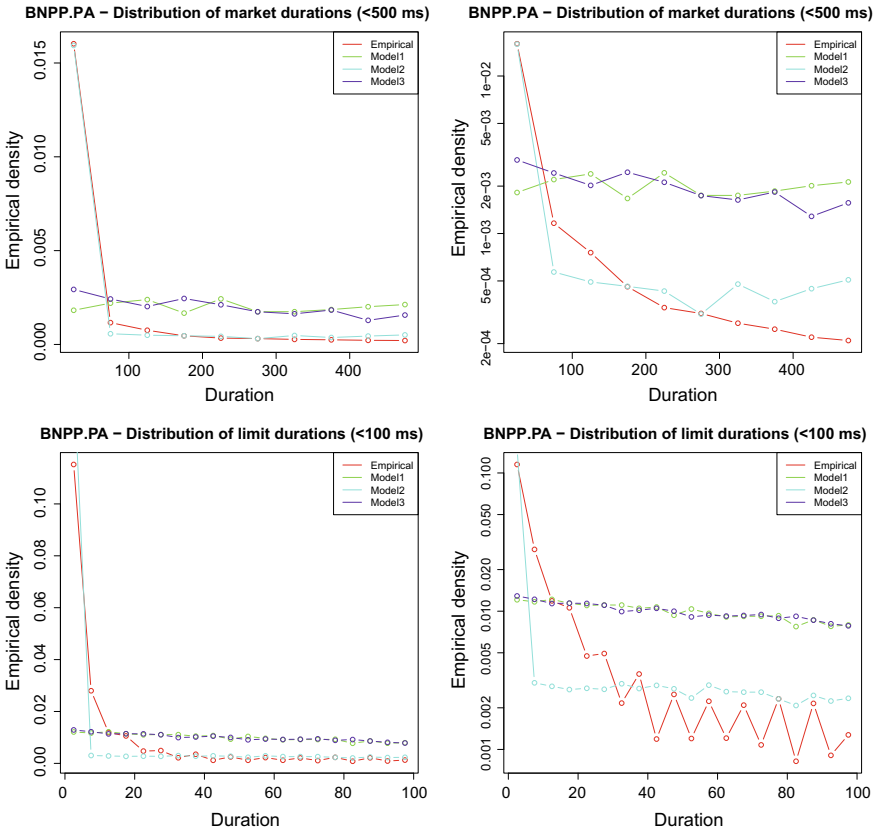


Fig. 3.5 Empirical and model distribution of the durations of market orders (top) and limit orders (bottom), in natural (left) and semi-log (right) scales. Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm

property of orders (here a probability peak for small durations) is partially captured by the Hawkes setting, but not at all by the other models, which basically simulate exponentially distributed durations. On the other hand, Model 3 provides realistic state-dependent intensities of submissions of orders, while Poisson (Model 1) and Hawkes (Model 2) intensities are insensitive to the order book state, as it is illustrated on Fig. 3.6 (On these plots, the size of the markers is proportional the frequency of observation of the value of the abscissa, either the spread s or the size of the best quote q_1).

Performances of the models can be assessed by comparing their ability to simulate realistic order books, and realistic prices. Figure 3.7 shows that all models perform similarly with respect to the reproduction of the distribution of the size of the simulated order book (all books are a bit too furnished). However, when focusing on the spread distribution, illustrated on Fig. 3.8, it appears that the Poisson model

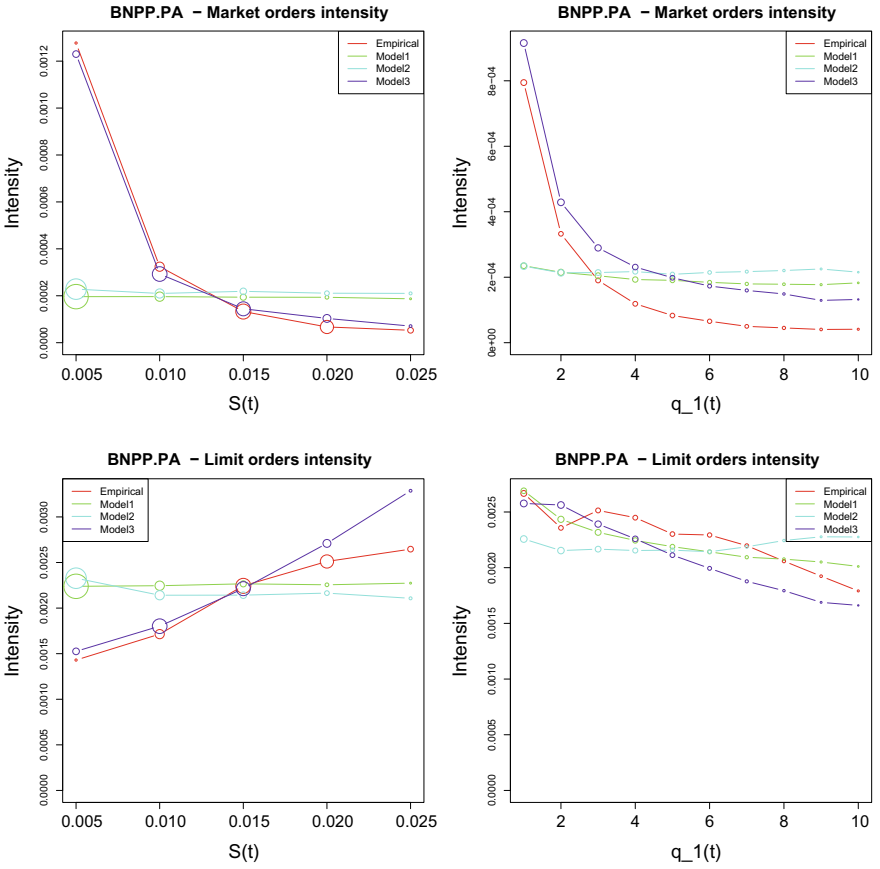


Fig. 3.6 Empirical and model conditional intensities of market orders (top) and limit orders (bottom), as functions of the spread (left) and order book size (right). Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm

(Model 1) generates an order book "stuck" with a low spread while the Hawkes model (Model 2) provides a spread distribution only marginally better. The state-dependent model is the only one producing a spread distribution with its maximum close to three ticks, as in the empirical data. Similarly, the state-dependent model (Model 3) provides slightly better looking price trajectories. As an illustration, Fig. 3.9 plots the increments distributions of the price trajectories sampled at 5 and 20 s. It appears that all models fail to accurately simulate the variance of the price increments, but that the state-dependent model (Model 3) seems to perform slightly better than the Hawkes one (Model 2) and much better than the Poisson one (Model 1) with respect to this property.

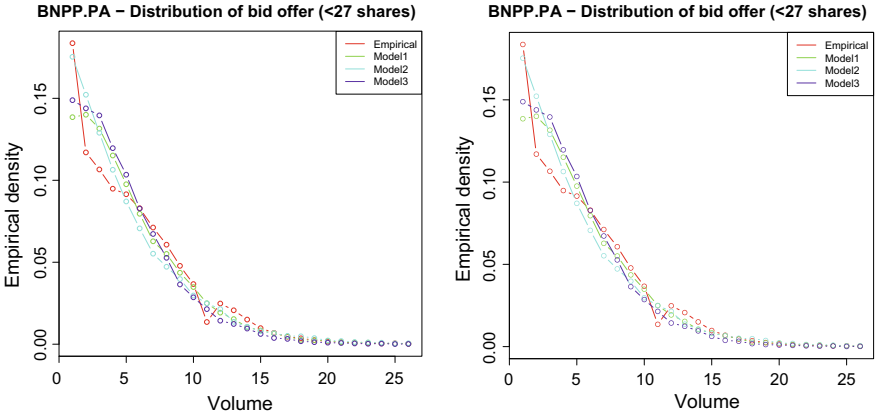


Fig. 3.7 Empirical and model distribution of the order book size, in natural (left) and semi-log (right) scales. Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm

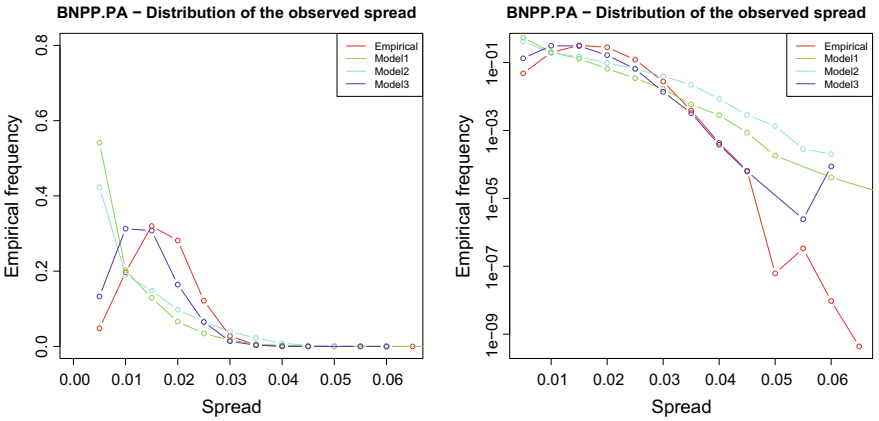


Fig. 3.8 Empirical and model distribution of the spread, in natural (left) and semi-log (right) scales. Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm

Further Improvements

This basic investigation of the simulated outputs of toy models of limit order books has obviously many limitations. It uses basic, non-tuned, versions of models, with schematic input distributions. It nonetheless shows that state-dependent modeling of limit order book is needed for realistic simulations purposes. In particular, this type of model is very flexible, and many covariates may be investigated to improve the results. As an illustration of potential improvements, let us mention the use

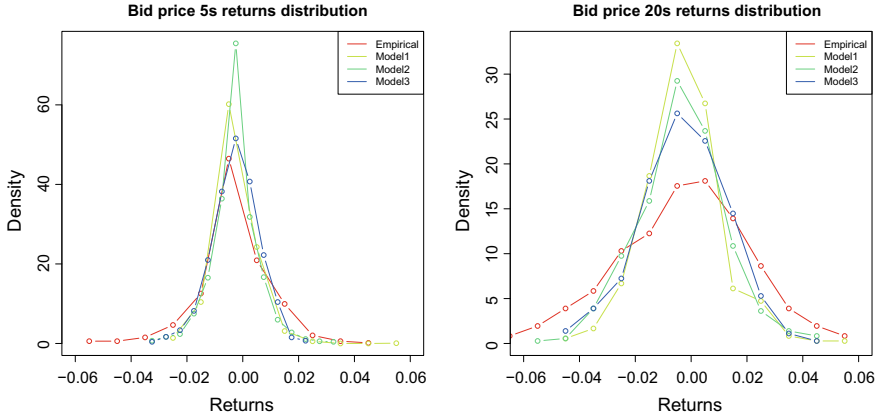


Fig. 3.9 Distributions of the increments of the bid price process $b(t)$ sampled at 5 s (left) and 20 s (right). Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm

of the imbalance and the last traded sign ($+1$ for ask market orders, -1 for bid market orders) as crucial elements influencing the submission of bid/ask market orders (see e.g. [2, 5] for the last trade sign, and [6] for the imbalance). Let us define a state-dependent model (hereafter Model 3b) in which market orders intensities depend on the spread $s(t)$, the imbalance $i(t)$, and the last traded sign $\varepsilon(t)$ as follows:

$$\lambda^T(t) = \exp\left(\theta_0^T + \theta_1^T s(t) + \theta_2^T i(t) + \theta_3^T \varepsilon(t) + \theta_4^T s(t)\varepsilon(t)\right), \tag{3.16}$$

where $T \in \{MA, MB\}$. For the sake of bid and ask symmetry, $\theta_k^{MA} = \theta_k^{MB}$ for $k = 0, 1$ and $\theta_k^{MA} = -\theta_k^{MB}$ for $k = 2, 3, 4$. Limit orders are submitted as in Model 3. Likelihood estimation is obviously similar to the one of Model 3.

Model 3b performs very similarly to Model 3 as for the previous properties investigated (for the sake of brevity and readability we do not reproduce updated figures). Moreover, it turns out that the inclusion of ε as a covariate in Model 3b is sufficient to provide a more realistic simulation of the autocorrelation of trade signs, without having to include an exogenous mechanism (e.g. an autoregressive process to model ε). Figure 3.10 plots the autocorrelation of trades up to lag 5, and shows that some short term memory is simulated, which is not the case for other models. This does not seem sufficient to simulate a long memory mechanism, but it nonetheless indicate that investigations on significant covariates are potentially fruitful for realistic limit order book parametric simulations.

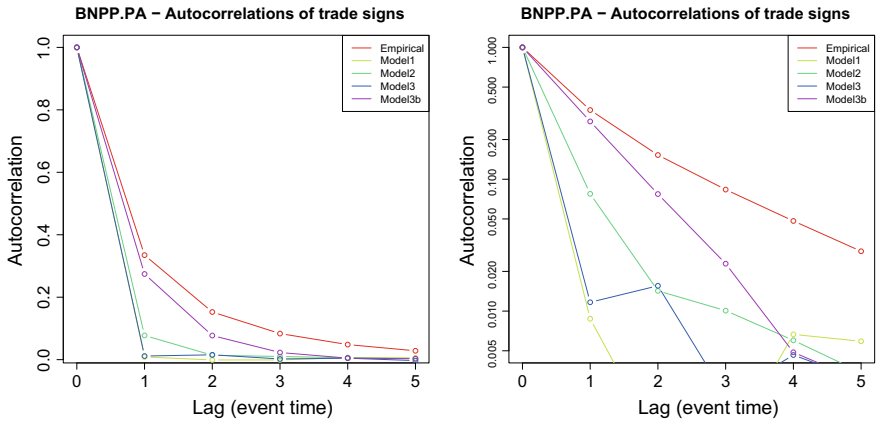


Fig. 3.10 Empirical and model autocorrelation of trades. Data : BNPP.PA, April 2016, all trading days from 10 am to 12 pm

Acknowledgements The author thanks Nakahiro Yoshida for useful discussions.

References

1. Abergel, F., Anane, M., Chakraborti, A., Jedidi, A., Muni Toke, I.: Limit Order Books. Cambridge University Press, Cambridge (2016)
2. Bouchaud, J.-P., Gefen, Y., Potters, M., Wyart, M.: Fluctuations and response in financial markets: the subtle nature of 'random' price changes. Quant. Financ. **4**(2), 176–190 (2004)
3. Chakraborti, A., Muni Toke, I., Patriarca, M., Abergel, F.: Econophysics review: I empirical facts. Quant. Financ. **11**(7), 991–1012 (2011)
4. Gould, M.D., Porter, M.A., Williams, S., McDonald, M., Fenn, D.J., Howison, S.D.: Limit order books. Quant. Financ. **13**(11), 1709–1742 (2013)
5. Lillo, F., Farmer, J.D.: The long memory of the efficient market. Stud. Nonlinear Dyn. Econ. **8**(3) (2004)
6. Lipton, A., Pesavento, U., Sotiropoulos, M.G.: Trade arrival dynamics and quote imbalance in a limit order book (2013), arXiv:1312.0514
7. Muni Toke, I.: Reconstruction of order flows using aggregated data. Mark. Microstruct. Liq. **02**(02), 1650007 (2016)
8. Muni Toke, I., Pomponio, F.: Modelling Trades-Through in a Limit Order Book Using Hawkes Processes. Econ.: Open-Access, Open-Assess. E-J. **6**, 1–23 (2012)
9. Muni Toke, I., Yoshida, N.: Modelling intensities of order flows in a limit order book. Quant. Financ. **17**(5), 683–701 (2017)

Chapter 4
Optimizing Execution Cost Using Stochastic Control

Akshay Bansal and Diganta Mukherjee

Abstract We devise an optimal allocation strategy for the execution of a predefined number of stocks in a given time frame using the technique of discrete-time Stochastic Control Theory for a defined market model. This market structure allows an instant execution of the *market* orders and has been analyzed based on the assumption of discretized geometric movement of the stock prices. We consider two different cost functions where the first function involves just the fiscal cost while the cost function of the second kind incorporates the risks of non-strategic constrained investments along with fiscal costs. Precisely, the strategic development of constrained execution of K stocks within a stipulated time frame of T units is established mathematically using a well-defined stochastic behaviour of stock prices and the same is compared with some of the commonly-used execution strategies using the historical stock price data.

Introduction

The problem of cost-efficient execution of a given stock with a lesser known distribution of its price is highly correlated with the fundamental difficulty of forecasting stock prices. A practical solution to any one of these two would bring some insight to solve the other. Investors and professional analysts frequently try to model stock prices with the help of the available information and certain noise factors whose distribution depends on various market aspects such as inflationary rates, financial status of the company and its competitive workforce. We attempt this exercise below. The rest of this section presents a formulation of the problem at hand. Section

A. Bansal · D. Mukherjee (✉)
Indian Statistical Institute, Kolkata, India
e-mail: diganta@isical.ac.in

A. Bansal
e-mail: akshaybansal14@gmail.com

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_4

"Defining Cost-Efficient Execution Strategy" discusses the execution strategy. The optimal strategy derivations are detailed in section "Optimal Investment Strategy for Instantaneous Stock Execution". This section also discusses the algorithm for the methodology and numerical results. Finally section "Conclusion" concludes.

Problem Formulation

Mathematically, the execution problem can be reformulated as follows:

Determine cost-efficient policy

$$\boldsymbol{\pi}^* = \{\mu_0^*(x_0, R_0), \mu_1^*(x_1, R_1) \ldots \mu_N^*(x_N, R_N)\}$$

such that

$$\begin{aligned} x_{k+1} &= g(x_k, u_k, \varepsilon_k)\ \forall k \in \mathbb{Z} \\ u_k &= \mu_k^*(x_k, R_k)\ \forall k \in \{0, 1, \ldots, N\} \\ \sum_{r=0}^{N} u_r &= K \end{aligned}$$

where $g(x, u, \varepsilon)$ is a known function which updates itself at each of the N equispaced time points in the time duration T, R_k is the stock position held at time point t_k and x_k is the stock price at time t_k.

Bertsimas and Lo [2] devised one such policy by partitioning the entire time frame into N intervals of equal length and performing the transaction of buying K/N shares at the start of each interval. In order to analyze the expected investment cost of such policy, Bertsimas utilized the discrete form of Arithmetic Brownian Motion (ABM) ($x_t = x_{t-1} + h(u_t) + \eta\varepsilon_t$) to periodically update the stock price. The major drawback with ABM model for stock price updation is that the non-negative behavior of stock price prevails only for shorter time frames T and the resultant optimal action (no. of shares bought out of the remaining stock pool) at each transaction point remained independent of any current/previous state information. Almgren and Chriss [1] extended the Bertismas model for limit order markets by incorporating the variance associated with the execution shortfall in the objective function. More recently, application of some of data-driven statistical techniques based on Reinforcement Learning [8] by Kakade et al. [5] and Nevmyvaka et al. [7] have resulted in significant improvement over simpler execution strategies such as submit and leave. In 2014, Cont and Kukanov [3] developed a more generalized mathematical framework for optimal order execution in limit order markets by incorporating targeted execution size due to bounded execution capacity of limit orders.

Defining Cost-Efficient Execution Strategy

The uncertainty factor (ε) involved in the state-updation function of stock price leads us to one such pathway of determining a cost-efficient policy (satisfying the conditions of (4.1)) by minimizing the expected future cost leading to the application of well-established theory of Stochastic Control. Mathematically, the exact optimization problem reduces to determining optimal policy $\pi^* = \{\mu_0^*(x_0, R_0), \mu_1^*(x_1, R_1) \ldots \mu_N^*(x_N, R_N)\}$ for the objective

$$\min_{\{\pi\}} \mathbb{E}_0\left[\sum_{r=0}^{N} u_r x_r\right] \tag{4.1}$$

Subject to the conditions:

$$\begin{aligned} u_k &= \mu_k(x_k, R_k)\, \forall k \in \{0, 1, \ldots, N\} \\ R_{k+1} &= R_k - u_k \\ x_{k+1} &= g(x_k, u_k, \varepsilon_k) \forall k \\ \sum_{r=0}^{N} u_r &= K \\ u_k &\geq 0\ \forall k \end{aligned} \tag{4.2}$$

where x_k is the stock price at time point t_k, R_k is the stock position held at time t_k and u_k is the appropriate action (investment strategy).

Reduction to Finite Horizon Problem for Integral States

The optimal policy for the problem formulated in (4.2) is devised by reducing it to a finite horizon problem where the discrete time-investment is a many-to-one mapping from the tuple of stock price and remaining stocks to the countable and finite set of non-negative integers. The proposed solution to (4.2) is described as following:

Given a uniform partition $\Pi(T) = \{t_0, t_1, \ldots, t_N\}$ with X being the finite set of all possible stock prices and $P = \{r \in \mathbb{Z}^+ \mid r \leq K\}$ the set of all possible stock positions. Then at any given time point $t \in \Pi(T)$, the state vector $(x, R) \in X \times P$. If the function $f(x, u, R)$ computes the *instantaneous* cost for the current state (x, R) and action u, the optimal policy $(\pi^* = \{\mu_0^*(x_0, R_0), \mu_1^*(x_1, R_1) \ldots \mu_N^*(x_N, R_N)\})$ for the objective function (4.1) can be computed dynamically for each discrete time point using Bellman's principle of optimality [4]. Precisely, to determine the time t_k policy function $\mu_k^*(x_k, R_k)$, optimal action u_k^{opt} is tabulated as a function of all $(x_k, R_k) \in X \times P$ using the *adaptive* cost objective

$$J_k(x_k, R_k) = \min_{\{u_k\}} \sum_{i}^{\infty} \Pr(\varepsilon_k^i | \mathscr{F}_k)[f(x_k, u_k, R_k) + J_{k+1}(g(x_k, u_k, \varepsilon_k^i), R_k - u_k)] \tag{4.3}$$

where $\mathscr{F}_k$ is the t_k-filtration (information contained till time t_k).

At the final time point t_N, the optimal action would be to buy all the remaining R_N. Thus $J_N(x_N, R_N)$ simply reduces to $f(x_N, R_N, R_N)$.

If the uncertainty parameter (ε_k) is independent of information $\mathscr{F}_k$, then (4.3) further simplifies to

$$J_k(x_k, R_k) = \min_{\{u_k\}} \sum_{i}^{\infty} \Pr(\varepsilon_k^i)[f(x_k, u_k, R_k) + J_{k+1}(g(x_k, u_k, \varepsilon_k^i), R_k - u_k)] \tag{4.4}$$

At any time point $t_k \in \Pi(T)$, the optimal action u_k^{opt} and $J_k(x_k, R_k)$ can be dynamically computed using (4.4) for each of the state element $(x_k, R_k) \in X \times P$.

Pitfalls of Reduction to Integral Finite Horizon Case

1. The numerical algorithm for its implementation mandates the construction of a three-dimensional matrix where each two-dimensional sub-matrix corresponds to a unique time point. Therefore its space complexity is of the order $\Omega(x_{max} K N)$.
2. The method imposes an additional restraint of the finiteness and countability of the set of all possible states $(X \times P)$.
3. The numerical search for the optimal integral solution can at best be accomplished using branch and bound algorithm [6] whose worst case complexity is still K (initial stock position). Thus the eventual time complexity for this algorithm is $\Omega(x_{max} K^2 N)$.

Optimal Investment Strategy for Instantaneous Stock Execution

In this section we will develop an investment strategy based on the idea of Stochastic Control Theory discussed in section "Reduction to Finite Horizon Problem for Integral States" for the market structure which sanctions the investor to buy any number of stocks on an instant basis at the current *market price* (mid-point of bid-ask spread). Unlike Almgren and Chriss [1], we have modelled stock prices using discretized Geometric Motion as the Bachelier's model ($x_t = x_{t-1} + h(u_t) + \eta\varepsilon_t$) would eventually return negative stock prices with non-zero probability in the limit of longer time duration. The discrete time stock price model we intend to use in our analysis is given by:

$$x_{k+1} = x_k(1 + \beta u_k + \varepsilon_k) \tag{4.5}$$

here x_{t+1} is the stock price at time t_{k+1}, ε_k is a random noise with $\mathbb{E}[\varepsilon_t] = 0$ and βu_k is the drift in stock price due to the buying action of u_k no. of stocks with β being some kind of *prominence factor* which varies according to one's influence in the stock market. For our case, we'll assume β belonging to the range $[10^{-5}, 10^{-4}]$. In the following subsections, we'll establish the general nature of some of the investment strategies for different kinds of *instantaneous* cost functions and compare their performance with some well-established policies.

Allocation Policy for Fiscal Cost Function

For this particular case, the *instantaneous* cost function is exclusively monetary i.e $f(x_k, u_k, R_k)$ (as in section "Reduction to Finite Horizon Problem for Integral States") is simply given by

$$f(x_k, u_k, R_k) = x_k u_k \tag{4.6}$$

where $u_k \leq R_k$.

Accordingly, the expression for optimal expected cost (4.3) modifies to

$$J_k(x_k, R_k) = \min_{\{u_k\}} \left[x_k u_k + \sum_i^{\infty} Pr(\varepsilon_k^i) J_{k+1}(g(x_k, u_k, \varepsilon_k^i), R_k - u_k) \right] \tag{4.7}$$

On rewriting the above expression for penultimate time point ($t = t_{N-1}$) by modelling the stock price using 4.5, the objective simplifies to

$$\begin{aligned} J_{N-1}(x_{N-1}, R_{N-1}) = & \min_{\{u_{N-1}\}} [x_{N-1}u_{N-1} + x_{N-1}(R_{N-1} - u_{N-1})(1 + \beta u_{N-1})] \\ & (\because \mathbb{E}[\varepsilon_{N-1}] = 0, J_N(x_N, R_N) = x_N R_N) \end{aligned} \tag{4.8}$$

leading to the following deduction.

Deduction 1 *When the nature of the instantaneous cost function is completely fiscal i.e. $f(x, u, R) = xu$ and the stock price is modeled using (4.5), the optimal investment policy due to stochastic control (Problem 4.1) simply converges to the purchase of the entire stock block of size K at time $t = t_N$. In general, the result holds for any stock price updation function of the form $x_{t+1} = x_t(1 + h(u_t) + \varepsilon_t)$ where $h(u_t)$ is a non-decreasing drift with $h(0) = 0$.*

Proof On rearranging the terms of penultimate time objective for the drift $h(u)$, (4.8) modifies to

$$J_{N-1}(x_{N-1}, R_{N-1}) = \min_{\{u_{N-1}\}} [x_{N-1}R_{N-1} + x_{N-1}(R_{N-1} - u_{N-1})h(u_{N-1})] \tag{4.9}$$

Table 4.1 Comparison of total expenditure between Bertsimas' (B) and One-Time(OT) policy based on their daily opening price spanning a total of 100 working days (Feb'16–Jun'16)

Stock	Investment cost (B)	Investment cost (OT)	Ratio (OT:B)
GOOG	$719770.69	$738000	1.02532
AAPL	$97670.42	$106636.21	1.09179
QCOM	$48983.12	$48808.40	0.99643
NVDA	$36247.86	$35704.41	0.98500
LXS.DE	€39972.39	€40986.76	1.02537

As $(R_{N-1} - u_{N-1})h(u_{N-1}) \geq 0$, the optimal action (u_{N-1}^{opt}) results in zero with $J_{N-1}(x_{N-1}, R_{N-1}) = x_{N-1}R_{N-1}$. By recursively calculating u_k^{opt} and $J_k(x_k, R_k)$ using the functional form of $J_{k+1}(x_{k+1}, R_{k+1})$ (4.7), it's trivial to observe the identical nature of the objective function for all $0 \leq k \leq N-1$. Hence the above deduction follows.

Resultant Policy and Comparison with Bertsimas' Model

Deduction 1 can be further generalized by observing the degenerate nature of the objective function at the penultimate time point i.e. both 0 and R_{N-1} are the optimal solutions to the objective (4.9). Henceforth, the optimal allocation policy modifies to the total investment for the entire stock block (K) at any one of the time point $t \in \{t_0, t_1, \ldots, t_N\}$.

Tabulated below is the total expenditure resulting from Bertsimas' policy and one-time investment at the midpoint $T/2$ (Table 4.1).[1]

As evident from the data above, the one-time investment policy may frequently fail to perform better than the distributed investment policy (due to Bertsimas).

Allocation Policy for Constrained Cost Function

Due to the possibility of positive accumulation of random noise (ε_t) over large no. of discrete time steps, the allocation policy devised in the last section has a tendency of resulting in a greater investment cost compared to the policy of distributed trading over the same no. of time steps. Thus, we've made an attempt to modify the *instantaneous* cost $f(x, u, R)$ by incorporating non-negative penalty in addition to the fiscal cost if the current action (u_k) violates certain market specific bounds. Specifically, a pre-determined set of bounds—an upper bound (UB) and a lower bound (LB) restricts the fractional consumption (u_k/R_k) at every time point t_k. The effect of penalty imposed for the case when the fractional consumption goes below

[1] As per the stock data obtained from Yahoo Finance.

the lower bound (LB) is less pronounced at initial time points compared to the later ones as the opportunistic time window to minimize the total expenditure decreases gradually with the passage of another transaction opportunity. The non-existence of such a restriction would eventually result in the investor holding a large fraction of his initial stock position at later time points with fewer opportunities to improve his total investment cost. Similarly, by restricting the investor to buy a large fraction of his current stock position (exceeding the upper bound (UB)) at the earlier time points of the transaction window, one instructs the investor to employ a distributed investment strategy till the near end of the transaction window where this constraint is liberalized. Mathematically, these two kind of restrictions can be summarized by modifying *instantaneous* cost ($f(x, u, R)$) using the logarithmic barrier resulting in the functional form:

$$\begin{aligned} f(x_k, u_k, R_k) = x_k u_k &- x_k C_l \left(\frac{t_k}{t_N}\right)^{\gamma} \log\left(1 - \max\left(0, LB - \frac{u_k}{R_k}\right)\right) \\ &- x_k C_u \left(\frac{t_N}{t_k}\right)^{\gamma} \log\left(1 - \max\left(0, \frac{u_k}{R_k} - UB\right)\right) \\ &\forall k \in \{0, 1, 2, \ldots, N-1\} \end{aligned} \tag{4.10}$$

Here C_l, C_u and γ are positive market specific constants with $C_l \gg C_u$.

The Bellman's criteria for optimality (4.4) can now be applied for the *instantaneous* cost $f(x, u, R)$ given by (4.10) resulting in another useful deduction.

Deduction 2 *Let X be the set of all possible stock prices and the instantaneous cost $f(x_k, u_k, R_k)$ be taken of the form given by (4.10). Then the adaptive cost objective ($J_k(x_k, R_k)$) given by (4.4) is linearly dependent on x_k ($\forall x_k \in X$).*

Proof Let $P(n)$ be the proposition that the cost $J_k(x_k, R_k)$ is linearly dependent on x_k $\forall k \geq n$.

Base Case: The objective function at penultimate time point (t_{N-1}) is given by

$$\begin{aligned} J_{N-1}(x_{N-1}, R_{N-1}) = x_{N-1}u_{N-1} &- x_{N-1}C_l \left(\frac{t_{N-1}}{t_N}\right)^{\gamma} \log\left(1 - \max\left(0, LB - \frac{u_{N-1}}{R_{N-1}}\right)\right) \\ &- x_{N-1}C_u \left(\frac{t_N}{t_{N-1}}\right)^{\gamma} \log\left(1 - \max\left(0, \frac{u_{N-1}}{R_{N-1}} - UB\right)\right) \\ &+ x_{N-1}(1 + \beta u_{N-1})(R_{N-1} - u_{N-1}) \end{aligned}$$

which is evidently linearly dependent on x_{N-1}. Thus $P(N-1)$ holds true.

Inductive Step: Let $P(k+1)$ holds true for some $k \leq N-1$. Then

$$\begin{aligned} J_k(x_k, R_k) = x_k u_k &- x_k C_l \left(\frac{t_k}{t_N}\right)^{\gamma} \log\left(1 - \max\left(0, LB - \frac{u_k}{R_k}\right)\right) \\ &- x_k C_u \left(\frac{t_N}{t_k}\right)^{\gamma} \log\left(1 - \max\left(0, \frac{u_k}{R_k} - UB\right)\right) \\ &+ \mathbb{E}[J_{k+1}(x_k(1 + u_k + \varepsilon_k), R_k - u_k)] \end{aligned}$$

From the induction hypothesis, $J_{k+1}(x_k(1+u_k+\varepsilon_k), R_k - u_k)$ is linearly dependent on $x_k(1+u_k+\varepsilon_k)$ thus $J_k(x_k, R_k)$ is linearly dependent on x_k. Hence $P(n)$ holds $\forall 0 \leq n \leq N-1$.

This computationally useful corollary follows trivially from the previous deduction.

Corollary 1 *Let X be the set of all possible stock prices and the instantaneous cost be taken of the form given by (4.10). Then the optimal action u_k for the objective (4.4) is independent of $x_k \, \forall \, k \in \{0, 1, \ldots, N-1\}$.*

Numerical Algorithm for Policy Evaluation

The Deduction 2 (and thus Corollary 1) is extremely advantageous to develop an efficient algorithm for determining the policy as the optimal action resulting from the theory of stochastic control is independent of stock price x. Hence all future computations can be performed by assuming stock price to be *unity*.

Algorithm 1 An efficient algorithm to compute optimal policy for constrained cost

while $r \in \{0, 1, \ldots, InitialSize\}$ **do**
  $J[N][r] \leftarrow r$ ▷ Optimal Cost at time t_N for $x = 1$
  $U[N][r] \leftarrow r$ ▷ Optimal action at t_N (ind. of x from last corollary)
while $i \in \{N-1, N-2, \ldots, 0\}$ **do** ▷ To evaluate u_{opt} and $J_i(1, r)$ at each time point t_i
  $J[i][0] \leftarrow 0$ ▷ Optimal Cost when stock position is null
  $U[i][0] \leftarrow 0$ ▷ Optimal Action when stock position is null
  while $r \in \{1, 2, \ldots, InitialSize\}$ **do** ▷ To determine the optimal action for each possible stock position
    $u_{opt} \leftarrow 0$
    $val_{opt} \leftarrow f(1, u_{opt}, r) + (1 + \beta u_{opt}) J[i+1][R - u_{opt}]$
    while $u \in \{1, 2, \ldots, r\}$ **do** ▷ Brute force search to determine optimal action dynamically
      $val_u \leftarrow f(1, u, r) + (1 + \beta u) J[i+1][R - u]$
      if $val_u \leq val_{opt}$ **then**
        $u_{opt} \leftarrow u$
        $val_{opt} \leftarrow val_u$
    $J[i][r] \leftarrow val_{opt}$
    $U[i][r] \leftarrow u_{opt}$

Resultant Policy for Constrained Objective

The optimal allocation vector (in row-major form) with initial stock position of 1000 shares for $\beta = 5 \times 10^{-5}$, $C_l = 1000$, $C_u = 10$, $\gamma = 2$, $LB = 0.2$, $UB = 0.6$ and different number of time points is depicted as under

- $N = 10$

$$\mathbf{u}^{opt} = \begin{bmatrix} 600 & 240 & 96 & 38 & 15 & 6 & 1 & 2 & 1 & 1 \end{bmatrix}$$

- $N = 30$

$$\mathbf{u}^{opt} = \begin{bmatrix} 0 & 600 & 110 & 58 & 47 & 37 & 30 & 24 & 19 & 15 \\ 12 & 10 & 8 & 6 & 5 & 4 & 3 & 3 & 2 & 2 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- $N = 50$

$$\mathbf{u}^{opt} = \begin{bmatrix} 0 & 0 & 0 & 600 & 39 & 72 & 58 & 46 & 37 & 30 \\ 24 & 19 & 15 & 12 & 10 & 8 & 6 & 5 & 4 & 3 \\ 3 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- $N = 100$

$$\mathbf{u}^{opt} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 119 \\ 176 & 141 & 113 & 90 & 72 & 58 & 46 & 37 & 30 & 24 \\ 19 & 15 & 12 & 10 & 8 & 6 & 5 & 4 & 3 & 3 \\ 2 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

With a steady increment in the number of available time points N for transaction in the fixed interval T, the resultant allocation policy follows a strategy of smaller stock acquisition towards the beginning and end of the interval T whereas bigger transactions are made towards the middle. Intuitively, this kind of allocation behaviour can be explained the observing the effect of the drift βu which has a tendency to increase the stock price resulting in a larger investment cost. Therefore, it is advantageous to make small transactions towards the beginning in such a way that the stock prices have a little tendency to drift upwards and at the same time a noticeable fraction of the initial stock position is also fulfilled followed by a major acquisition towards the middle. The resultant hefty drift would eventually have a little effect on the total investment cost as the remaining stocks constitute a small fraction of the initial stock position K.

Table 4.2 Comparison of total expenditure between Bertsimas' (B) and Cost with Risks (WR) based on their daily opening price spanning a total of 100 working days (Feb'16–Jun'16)

Stock	Investment cost (B)	Investment cost (WR)	Ratio (WR:B)
GOOG	$719770.69	$699576.13	0.97194
AAPL	$97670.42	$94117.02	0.96361
QCOM	$48983.12	$45666.70	0.93229
NVDA	$36247.86	$28670.91	0.79096
LXS.DE	€39972.39	€35319.80	0.88360

Conclusion

The policy resulting from the analysis performed in section "Allocation Policy for Constrained Cost Function" by incorporating several risk-factors has shown considerable improvement over the Bertsimas' policy with its total expenditure tabulated as under[2]:

In summary, the non-performance of one-time investment policy (Table 4.1) and significant improvement of the policy resulting from the modified cost function (Table 4.2) by incorporating market risks can be safely established for the average case analysis of market model-I keeping in mind the existence of a non-zero probability of the occurrence of a case scenario where the above deduction fails to hold.

The *instantaneous* cost objective (4.10) could be improved further by factoring constraints in a rational manner such that the penalty levied upon their violation does not undermine or overestimate the effective fiscal cost. Another way to improve the cost objective is by estimating the effect of current stock price before converging to any possible action. For instance, if the bounds on the possible stock prices and its probability distribution throughout the entire time duration T is already known, then one can possibly make use of this information by tuning the penalty functions appropriately as a significantly lower stock price and higher probability density would result in a net reduced risk for the case when one intends to invest in a large fraction even at the earlier time points. Similarly, a higher price (close to upper bound) would levy a high penalty even when one is within the bounds of the imposed constraints. These kind of formulations would bring in the dependence of the stock price resulting in improved policies but with a slight trade-off of an increased time and space complexity.

Another possible way to improve the performance of the resulting control action is by utilizing a more general form of the stock price updation function based on the theory of Linear Price Impact with Information as suggested in i.e. the stock price at each successive time point can now be modeled as:

[2] As per the stock data obtained from Yahoo Finance.

$$\begin{aligned} x_{t+1} &= f(x_t, u_t, Z_t, \varepsilon_t) \\ Z_t &= g(Z_{t-1}, \eta_t) \end{aligned} \tag{4.11}$$

Acknowledgements We thank the conference participants at Statfin2017 for their insightful comments and suggestions which has helped us to improve our work.

References

1. Almgren, R., Chriss, N.: Value under liquidation. Risk **12**(12), 61–63 (1999)
2. Bertsimas, D., Andrew, W.Lo.: Optimal control of execution costs. J. Financ. Mark. **1**(1), 1–50 (1998)
3. Cont, R., Kukanov, A.: Optimal order placement in limit order markets (2013)
4. Dimitri, P.: Bertsekas. Dynamic Programming and Stochastic Control. Academic Press, New York (1976)
5. Even-Dar, E., Kakade, S., Mansour, Y.: Reinforcement learning in POMDPs without resets (2005)
6. Land, A.H., Doig, A.G.: An automatic method of solving discrete programming problems. Econ.: J. Econ. Soc. 497–520 (1960)
7. Nevmyvaka, Y., Feng, Y., Kearns, M.: Reinforcement learning for optimized trade execution. In: Proceedings of the 23rd International Conference On Machine Learning, pp. 673–680. ACM (2006)
8. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, vol. 1. MIT press, Cambridge (1998)

Chapter 5
Hierarchical Financial Structures with Money Cascade

Mahendra K. Verma

Abstract In this paper we show similarities between turbulence and financial systems. Motivated by similarities between the two systems, we construct a multiscale model for hierarchical financial structures that exhibits a constant cascade of wealth from large financial entities to small financial entities. According to our model, large and intermediate scale financial institutions have a power law distribution. However, it exhibits Maxwellian distribution at individual scales.

Introduction

A financial system is quite complex due to its multiscale and time-dependent nature. Its complexity is accentuated by features such as saving, banking, corruption, stock-market, natural calamities, etc. Despite such complicated structures, scientists have attempted to model a financial systems using simple ideas. One of the leading questions in this field is how to model the wealth and income distributions of individuals and companies [1]. In this paper we will address this question.

Earlier models of income distribution of individuals are motivated by equilibrium statistical mechanics. In such models, individuals are mapped to particles in a thermodynamic system, and economic activities to scattering among particles. Following this analogy, it is expected that the income distribution follows Maxwellian or Gibbs distribution, similar to the distribution of kinetic energy in a gas container.

The above distribution however holds only for low income groups. Pareto [2] and others showed that the individual income in a large income group exhibits power law distribution. There have been many attempts to model this power law distribution using nonequilibrium nature of the system. See Chakrabarti et al. [1] and references there in.

In a financial system, wealth cascades from large financial entities to smaller ones. This cascade is somewhat similar to the cascade of kinetic energy in a turbu-

M. K. Verma (✉)
Indian Institute of Technology Kanpur, Kanpur, India
e-mail: mkv@iitk.ac.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_5

lent system. In addition, a well-developed financial system contains income groups with a wide range of distribution. Also, note that these income groups interact with each other. Motivated by the above similarities between turbulence and finance, we construct a model for a hierarchical financial system which is quite similar to Kolmogorov's model for α turbulent flow.

The structure of the paper is as follows: In section "Equilibrium Model", we describe a generic equilibrium model of a financial system. Section "Multiscale Model of Turbulence" contains a brief description of Kolmogorov's model for turbulence. In section "A Model of Hierarchical Financial Entities" we construct a model for hierarchical financial system; this model is analogous to the Kolmogorov's model of turbulence. We conclude in section "Discussions and Conclusion".

Equilibrium Model

In this section, we describe an equilibrium model of wealth distribution [3]. Before that we discuss thermodynamics of an isolated gas reservoir in which gas molecules interact with each other via collisions. Under thermodynamic approximation, all the molecules in the gas have approximate equal energy. The variation in the energy of the molecules is given by Maxwell or Gibbs distribution [4]:

$$P(E) = \exp(-E/k_B T) \tag{5.1}$$

where $E = mv^2/2$ is the kinetic energy of a molecule of mass m, and $k_B T = \langle mv^2 \rangle /2$ is the average kinetic energy of all the molecules. Note that this system has a single energy scale $k_B T$. Also, the system is in equilibrium, and it obeys principle of detailed energy balance. As a result, there is no energy transfer from one region to another, both in real and Fourier space.

Now we are ready to describe an equilibrium model of wealth distribution [1, 3]. In the past, several researchers have shown connections between economic systems and equilibrium thermodynamics (e.g., a gas reservoir described above) [3]. The individuals or economic entities are analogous to the gas molecules, and wealth to the kinetic energy of the molecules. Refer to Table 5.1 for a detailed comparison. Using this analogy, researchers deduced that the wealth distribution $P(W)$ in an economy follows Maxwell or Gibbs distribution:

$$P(W) = \exp(-W/\langle W \rangle), \tag{5.2}$$

where $\langle W \rangle$ is the average individual wealth in the economic system. More refined models yield log-normal distribution [1].

Table 5.1 Analogies between an equilibrium economic model and a thermodynamic system

Thermodynamics	Economics
Thermodynamic system	Economy
Gas molecules	Economic entities (individuals)
Individual kinetic energy	Individual wealth
Collisions	Economic interactions
Average kinetic energy	Average wealth

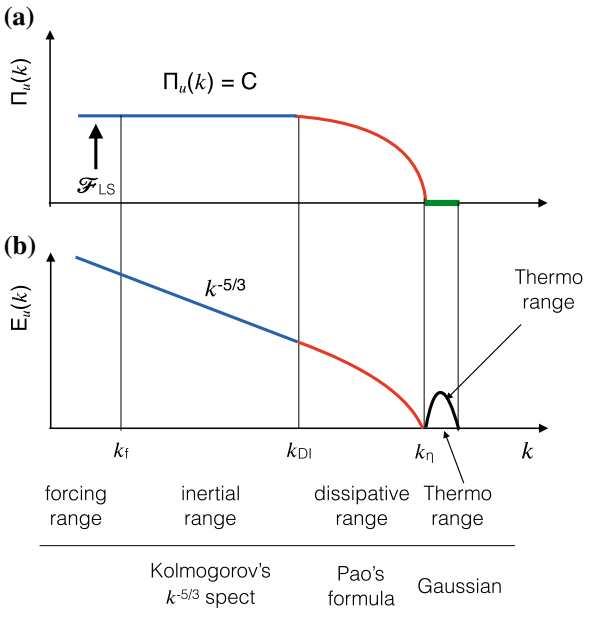


Fig. 5.1 Kolmogorov's picture of hydrodynamic turbulence. The flow is forced at large scale with an energy injection rate of $\mathscr{F}_{LS}$. **a** The energy flux is constant in the inertial range, and it decays in the dissipative range. Flux is zero in the thermodynamic range. **b** The energy spectrum exhibits $k^{-5/3}$ spectrum in the inertial range. In the thermodynamic range, the molecules of the fluid exhibit Maxwellian distribution (black curve)

Multiscale Model of Turbulence

Many nonequilibrium systems have properties very different from that of the gas reservoir described above [5]. For example, consider a turbulent fluid stirred at large length scaless. The kinetic energy at large scales cascades to intermediate scale, and then to small scales. The kinetic energy flux Π_u is constant in the inertial range, and then it decreases in the dissipation range. Due to the energy cascade, principle of detailed balance is broken in such a system. The energy flux is zero at microscopic scale where we expect thermodynamic principles to hold. This is Kolmogorov's picture of hydrodynamic turbulence [6–9]. See Fig. 5.1 for an illustration of the energy flux and energy spectrum.

The energy spectrum $E_u(k)$ of a turbulent flow has been derived using dimension analysis. Using

$$[E_u(k)] = [E_u/k] = [L^3/T^2]; \quad [\Pi_u] = [E_u/T] = [L^2/T^3]; \quad [k] = [L]^{-1}, \tag{5.3}$$

we derive the following formula for the kinetic energy spectrum:

$$E_u(k) = K_{\rm Ko}\epsilon_u^{2/3}k^{-5/3}, \tag{5.4}$$

where $K_{\rm Ko}$ is Kolmogorov's constant, and ϵ_u is the kinetic energy dissipation rate. Pao [10] extended the above formula to the dissipation range, and obtained

$$\Pi_u(k) = \epsilon_u \exp\left(-\frac{3}{2}K_{\rm Ko}(k/k_d)^{4/3}\right), \tag{5.5}$$

$$E_u(k) = K_{\rm Ko}\epsilon_u^{2/3}k^{-5/3}\exp\left(-\frac{3}{2}K_{\rm Ko}(k/k_d)^{4/3}\right), \tag{5.6}$$

where k_d is Kolmogorov's wavenumber [11]. The fluid kinetic energy vanishes beyond the dissipation range, i.e., for $k > k_d$. The above function describes the inertial and dissipative ranges (blue and red curves of Fig. 5.1) quite well. We expect thermodynamic ideas to work beyond this scale. The energy of the molecules would follow Maxwell's or Gibbs' distribution, as shown by the black curve of Fig. 5.1b.

Shell model is a popular model of hydrodynamic turbulence. In one version of the shell model, called GOY shell model of turbulence,

$$\frac{d}{dt}u_n + \nu k_n^2 u_n = -i(a_1 k_n u_{n+1}^* u_{n+2}^* + a_2 k_{n-1} u_{n+1}^* u_{n-1}^* + a_3 k_{n-2} u_{n-1}^* u_{n-2}^*), \tag{5.7}$$

where u_n is a complex number representing the velocity field at length scale $k_n = 2^n$, a_1, a_2, a_3 are constants, and ν is the kinematic viscosity [12]. In the shell model, the kinetic energy follows

$$E(k_n) = \frac{|u_n|^2}{2k_n} \sim k_n^{-5/3}, \tag{5.8}$$

in accordance with Kolmogorov's theory of turbulence. Our finance model has a similar form as the above shell model, as we will describe in the next section.

A Model of Hierarchical Financial Entities

We construct a model for a hierarchical finance system in a similar lines as Kolmogorov's picture for hydrodynamic turbulence. In this model, we assume that the wealth is generated at the largest scale, and then it flows from larger financial structures to smaller structures in a steady manner. We also assume that the financial entities of similar sizes interact with each other. This is similar to the local interactions in turbulence. In addition, in the absence of financial pilferage, we expect the cascade of money from large structures to smaller structures to be a constant. This

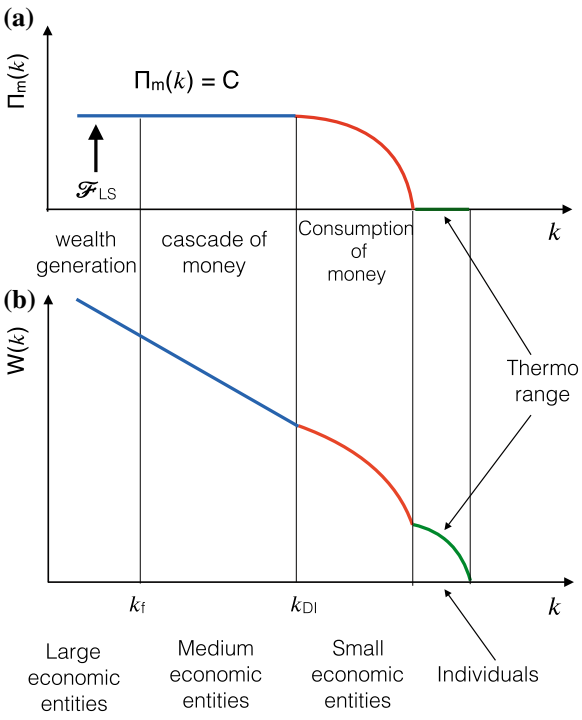


Fig. 5.2 In a hierarchical finance model, **a** the flux of money $\Pi_m(k)$, and **b** the wealth distribution of financial entities. The large and medium economic entities have constant money flux and power law distribution of wealth. Small economic entities exhibit exponential distribution, while the thermodynamic range exhibits Maxwellian wealth distribution and zero money flux

Table 5.2 Analogies between turbulence and hierarchical financial system

Turbulence	Financial system
Fluid structures	Financial entities
Multiscale	Multiscale
kinetic energy of a structure	Wealth of a financial entity
Constant energy flux	Constant money supply
Power law $E_u(k)$ at intermediate scales	Power law for large income entities
Exponential $E_u(k)$ at small scales	Expect similar scaling
Random motion beyond k_d	Gibbs distribution at individual scale

is same as the assumption of constant energy cascade in hydrodynamic turbulence. See Fig. 5.2 for an illustration, and Table 5.2 for a listing of similarities between a turbulent system and a hierarchical financial system.

We place these financial entities in a two-dimensional wavenumber grid.[1] Let us denote the financial asset of a financial entity at the wavenumber **k** as $W(\mathbf{k})$. The number of mesh points on a 2D disc of radius k is

[1] Dimensionality of a hierarchical financial system is an undetermined parameter. Here we choose $d = 2$ using an observation that these structures reside on the surface of the Earth.

$$n(k) = 2\pi k. \tag{5.9}$$

We solve for the wealth distribution as a function of n. To illustrate, there are fewer financial entities at small k, corresponding to financial giants (e.g. Google and Apple of today). Large number of modes at large k correspond to small units like small companies or individuals.

Motivated by the shell model of turbulence, we construct the following model for the hierarchical financial entities:

$$\frac{dW_k}{dt} = ak^{\alpha} W_{k-1} W_{k+1} - bk^{\beta} W_k + Q_{k,1}, \tag{5.10}$$

where a, b, α and β are constants, and W_k is analogous to the shell spectrum in turbulence. Hence,

$$W_k = 2\pi k W(\mathbf{k}). \tag{5.11}$$

In Eq. (5.10), the first term in the RHS represents the interactions among financial entities at scales k, $k-1$ and $k+1$, while the second term represents financial losses at scale k (e.g., recurring expenses, electricity bills). The third term $Q_{k,1}$ represents the wealth generation at the largest scale, $k = 1$.

This is a very simple model because it ignores nonlocal interactions, as well as other complex things like loans, savings, banks, generation of wealth at the intermediate and small scales, etc. Further, we assume a steady state in which money flows from larger structures to smaller structures. The wealth is finally consumed at the smallest structures of the system.

First, we focus on the large and intermediate scale where we expect a power law scaling. We also assume that the financial losses at these scales are negligible. Under a steady state,

$$\Pi = \frac{dW_k}{dt} \sim k^{\alpha} W_k^2, \tag{5.12}$$

where Π is the cascade of money. We invert Eq. (5.12) that yields

$$W_k \sim \Pi^{1/2} k^{-\alpha/2}. \tag{5.13}$$

Now using Eqs. (5.9, 5.11), we obtain

$$n(W) \sim \Pi^{\frac{1}{\alpha+2}} W^{-\frac{2}{\alpha+2}}, \tag{5.14}$$

where we write $W(\mathbf{k})$ as W. The above formula yields the number of financial entities $n(W)$ with wealth W. Clearly, $\alpha = -1$ gives

$$n(W) \sim \Pi W^{-2}, \tag{5.15}$$

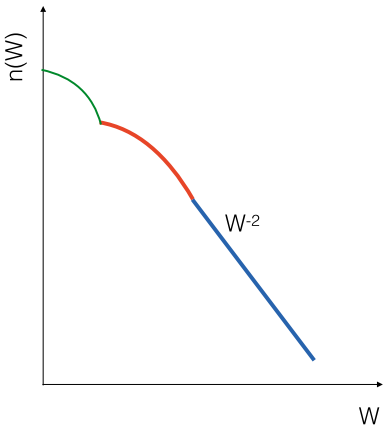


Fig. 5.3 Plot of wealth distribution $n(W)$ versus W that indicates number of financial entities with wealth W

which is similar to the Pareto's law for the wealth distribution [1, 2] of the large financial entities. Note however that the exponent depends quite crucially on the choice of α. In Fig. 5.3, we exhibit the inverted form of Fig. 5.9b, or the plot of wealth distribution $n(W)$ versus W.

The wealth cascades down to smaller scales, and it finally gets consumed at the dissipation scales (individual level). It could be in the form of consumption of food and basic needs. Following the popular equilibrium model [3], the wealth distribution at this scale follows Maxwellian or Gibbs distribution. We also expect an income group between the power law regime and the Gibbs distribution. This regime may follow a law similar to that Pao's model for turbulence, which was discussed in section "Multiscale Model of Turbulence".

Several cautionary remarks are in order. Our model describes the wealth of financial entities. Pareto's law however is stated for individual incomes. In free market, a financial entity is essentially owned or controlled by several individuals or a group of individuals. Therefore, it is reasonable to assume that the wealth distribution of financial entities also reflects the wealth or income distribution of individuals. Also, a large financial entity contains smaller entities, thus forming hierarchical structures.

A corollary to the above model is as follows. Let us consider finance distribution in a country. The central government transfers resources to various states who distributes it to lower levels in a hierarchical manner, e.g.,

$$\text{states} \rightarrow \text{districts} \rightarrow \text{villages}. \tag{5.16}$$

Following the same line of arguments as before, we deduce that the financial resources at hierarchical level must be a power law. If there is no corruption, then the money supply at different levels is constant.

The above model is very simple. It ignores many important ingredients such as savings, stocks, banking, pilferage of wealth, nonlocal interactions, etc. This model however has certain novelty. It emphasises on multiscale nature of financial systems,

cascade of money at different scales, and nonequilibrium nature of the financial system.

Discussions and Conclusion

Our finance model, though simple, captures multiscale economic transaction among financial entities and explains coexistence of power law and Maxwellian distribution for the wealth [1]. The model has other predictions as well. Note that the model has a free parameter α. The present multiscale model has many assumptions that need to be studied in detail for applicability in real financial system. For example, we need to include savings, banking, variable energy flux, etc. in a more refined model. In addition, we need to contrast the present model with the existing financial models, some of which are described in [1, 13–15].

A small financial system without hierarchy may exhibit detailed balance and Maxwellian distribution for the wealth distribution. As soon as a financial system becomes sufficiently large and it follows a free-market economy, we expect wealth inequalities to develop based on individual abilities and ambitions. Such a system will exhibit power law distribution. Strong economic regulations may suppress the inequality and make the power law shallower.

We believe that the finance model presented here shed important insights into financial systems. Yet, it is a preliminary model and it needs more work and refinements.

Acknowledgements I am thankful to Anirban Chakraborti, Supratik Banerjee, Andre Sukhanovskii, Rodion Stepanov, Franck Plunian, Abhishek Kumar, and Kiran Sharma for very useful discussions and ideas.

References

1. Chakrabarti, B.K., Chakraborti, A., Chakravarty, S.R., Chatterjee, A.: Econophysics of Income and Wealth Distributions. Cambridge University Press, Cambridge (2013)
2. Pareto, V.: Cours d'économie politique. Librairie Droz, Paris (1964)
3. Saha, M., Srivastava, B.N.: A Treatise on Heat, 5th edn. Indian Press, Kolkata (1950)
4. Landau, L.D., Lifshitz, E.M.: Statistical Physics. Course of Theoretical Physics, 3rd edn. Elsevier, Oxford (1980)
5. Ma, S.-K.: Statistical Mechanics. World Scientific, Singapore (1985)
6. Kolmogorov, A.N.: Dissipation of energy in locally isotropic turbulence. Dokl. Acad. Nauk SSSR **32**(1), 16–18 (1941)
7. Kolmogorov, A.N.: The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. Dokl. Acad. Nauk SSSR **30**(4), 301–305 (1941)
8. Lesieur, M.: Turbulence in Fluids. Springer, Dordrecht (2008)
9. Verma, M.K.: Physics of Buoyant Flows: From Instabilities to Turbulence. World Scientific, Singapore (2018)

10. Pao, Y.-H.: Transfer of turbulent energy and scalar quantities at large wavenumbers. Phys. Fluids **11**(6), 1371–1372 (1968)
11. Leslie, D.C.: Developments in the Theory of Turbulence. Clarendon Press, Oxford (1973)
12. Ditlevsen, P.D.: Turbulence and Shell Models. Cambridge University Press, Cambridge (2010)
13. Chakrabortiand, A., Chakrabarti, B.K.: Statistical mechanics of money: how saving propensity affects its distribution. Eur. Phys. J. B **17**, 167–170 (2000)
14. Patriarca, M., Chakraborti, A., Kaski, K.: Statistical model with a standard Γ distribution. Phys. Rev. E **70**, 016104 (2004)
15. Chakraborti, A., Patriarca, M.: Variational Principle for the Pareto Power Law. Phys. Rev. Lett. **103**, 228701 (2009)

Chapter 6
Effect of Tobin Tax on Trading Decisions in an Experimental Minority Game

Dipyaman Sanyal

Abstract James Tobin (The new economics one decade older: the Elliot Janeway lectures in honor of Joseph Schumpeter. Princeton University, Princeton, 1974, [1]) proposed a transactions tax for currency trading to reduce volatility in this highly speculative market. We conduct a 40-period experiment using a financial market mechanism and introduce a tax after 20 periods. While earlier experimental studies have used double auctions to study the Tobin Tax, our experiment uses a completely speculative market design (the minority game) which better emulates global currency markets. We find that trading volumes fall after the tax is imposed supporting results from existing studies, but in contrast to these studies, we observe a significant decrease in volatility (without any effect on market size). Our experimental findings are largely in line with a simulation model using the minority game developed by Bianconi et al. (J Econ Behav Organ 70(1–2):231–240, 2009, [2]) where the authors find that "the introduction of Tobin taxes in agent-based models of currency markets can lead to a reduction of both speculative trading and the magnitude of exchange rate fluctuations."

Introduction

The simple Minority Game (MG) model, which is similar to market entry or congestion games in economics with multiple potential equilibriums generated a lot of interest in the physics community with hundreds of papers and more than a dozen books published that analyse how equilibrium is achieved, oscillations around the equilibrium, the learning process of agents and various extensions to the basic game (see Kets, for a survey of the literature [3]). A significant portion of this research stems from modeling possibilities that utilize statistical mechanics, especially focusing on its potential application to financial markets. In a financial market, a participant buy-

D. Sanyal (✉)
School of Social Sciences, Centre for Economic Studies and Planning, Jawaharlal Nehru University, New Delhi 110067, India
e-mail: deep@dono.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_6

ing or selling an instrument (i.e., going long or short) is betting against the market price and hopes to be in the minority that has accurately judged the value of the instrument. He/she is essentially questioning the collective wisdom of the market.

The Minority Game was designed by Challet and Zhang [4], inspired by Brian Arthur's El Farol Bar Problem [5]. To summarize Arthur's problem: El Farol, a Santa Fe bar, has only 60 seats for its live music shows on Thursdays, and yet, 100 people are keen to be there. This, of course, leads to a mixed-strategy Nash equilibrium; however, it seems unlikely that each of the 100 players will model out every possible behavior of the other 99—as they should in a game-theoretic problem. Additionally, there is no a-priori best strategy, because if there was, all agents would have used it and the suboptimal outcome of 'everyone at home' or 'everyone at the bar' would have occurred. To overcome these issues, Arthur posits that agents faced with similar problems which entail full information of a large number of players' preferences choose inductive over deductive thinking. This might be a suboptimal decision-making tool, but it is certainly more probable that agents take decisions based on some rules and update their beliefs after every experience.

However, even in the inductive framework, if agents base their reasoning on M past observations and there are N agents, the agent would need to analyze $(N+1)^M$ possible combinations, which is increasing in N. To simplify matters, Challet and Zhang devised the Minority Game (MG), which simplified the analysis to only a binary option: go to the bar or not go to the bar based on one simple predictor—whether there were more than or less than the optimal number in the past M periods, since the agents do not care about the actual number of people who show up at the bar but only if it is above or below the optimal. This reduces the choice set to 2^M, which is increasing in M (number of periods of history) but not in N. This makes the inductive reasoning framework an even more plausible representation of human behavior in these circumstances due to the significant lower need for mental modeling in decision making (Challet and Zhang [6]; Challet [7]; Challet and Marsili [8], Challet et al. [9]).

The MG thus provides a simple model of financial markets with N players taking trading decisions based only on past outcomes of success or failure (Gou) [10–12]. This design of the market mechanism is a good approximation of the currency markets since, unlike a double auction market, the underlying asset in this market is primarily speculative and does not trade around fundamental values. The currency market is considered to be an extremely speculative market with significant disconnects from underlying fundamentals (Frankel and Rose [13]).

In this paper, we have designed and conducted a lab experiment to test for the outcomes of a particular inference from the MG literature, that is the introduction of a Tobin tax in a currency market and its impact on trading behavior and volatility.

The Economics and Econophysics of Tobin Tax

James Tobin proposed to introduce a tax on every international currency transaction to lower speculation in these markets. Proponents of deregulation have, of course, argued that the tax may be prohibitively expensive to impose and may, inadvertently, lead to higher volatility.

Among developed economies, the empirical effect of the Tobin tax was studied extensively between 1984 and 1986 when Sweden imposed the tax on equities transactions and then on bonds and derivatives, and finally, increased the tax rate. The Swedish experiment was overall disastrous and a large number of transactions and equity offering moved from the Swedish markets to other financial centers like London (Umlauf [14]). It should be noted here that the Tobin tax was strictly planned for a highly speculative currency market and not for other markets like equities or bonds. Unfortunately, there are no examples of bilateral Tobin taxes to understand the effects of these taxes on currency pairs, which are traded against each other.

To study the potential effects of these taxes on a speculative market, Bianconi et al. [2] introduced a Tobin tax in the canonical minority game and used it to study a simulated currency market. Their model concludes that the Tobin tax may be useful in "a reduction of speculative trading and reduce the magnitude of exchange rate fluctuations at intermediate tax rates." This outcome has serious policy implications, especially after the financial meltdown of 2008. The imposition of a financial transactions tax is under consideration and 11 European countries are seeking to potentially introduce the proposal of a Tobin tax. Although a number of countries impose a Securities Transactions Tax (STT) that is effectively similar to a Tobin tax, forex markets are amongst the most speculative markets of scale (see survey by Frankel and Rose, on 'disconnect' between economic fundamentals and currency fluctuations [13]). Thus, effects of an STT may not be similar to the effects of such a tax on speculative transactions like those in the currency markets. To explore the relationship between the tax and the markets further, we replicate a similar, and yet, simpler market with human agents in the lab to study whether the theoretical conclusion of Tobin [1] and the conclusions of the agent-based models of Bianconi et al. [2] hold true.

Experimental Minority Game

The minority game provides a large body of theoretical and simulated models, which are ripe for empirical testing. However, there is a significant difference between empirical analysis of minority games from financial markets data and from laboratory experiments data. Empirical econophysics addresses the challenge of testing these models by analyzing naturally-occurring data from complex systems like financial markets and there is a large and growing literature on empirical econophysics and minority games, where the patterns of the data-generating process in complex systems (especially financial markets) have been extensively studied (Chalet et al.

[7]; Chakraborti et al. [15]). But the problem faced in such analyses is the variety of effects that are simultaneously at play in financial markets which makes it extremely difficult to conclusively comment on the effect of individual factors. To control for the effects of potential confounding of results (due to multiple factors changing at the same time) has been a major reason for the growth of experimental methods in social sciences and financial economics, despite the existence of large scale data from the field. Experimental methods allow for a controlled environment, which allows the scientist to make changes in single variables that help in understanding the effect of that particular variable in a system without the confounding effects of myriad other factors that act upon a 'real-world' market. Although a significant amount of theoretical and simulation-based research has been done on MG, there are very few papers (Bottazzi et al. [16, 17]; Patkowski and Ramsza [18]; Laureti et al. [19], Berg et al. [20]) which use the methods of experimental economics to provide support (or not) to the results generated from theory and there is only one text book on experimental econophysics (Huang [21]).

The Tobin Tax Minority Game Experiment

The Minority Game provides an ideal setting for testing the original concepts of the Tobin tax. Since there is no underlying fundamental value of the 'asset' in a Minority Game it is a completely speculative market where market participants gain only when they are in the minority and there is no other mechanism to affect the demand (and thus, price) of the asset. Additionally, due to a completely neutral nature of the asset definition itself, there is potentially limited impact of any biases that we might observe (anchoring to names of stocks, bias towards value stocks due to name recognition etc.). Past experiments on the Tobin tax (see Hanke et al. for an example and a survey of the literature [22]) use the double auction mechanism to study the effects of the tax on asset prices, trading behaviour and volatility. While double auctions might be closer to actual securities market trading mechanisms, the currency markets are arguably better emulated by a minority game due to its inherent speculative nature. In the Minority Game the only process of 'winning' is by being in the minority and it has no relationship to better estimating the fundamental value of the asset (which is true for double auctions and securities markets).

We conducted a 40-period experiment of the Minority Game (as developed by Challet and Zhang [4] with 143 subjects and introduced a Tobin tax after 20 periods of the game. Three sessions were conducted with subjects who were second-year MBA students at the Institute of Management Technology, Ghaziabad, India, enrolled in Business Analytics and Risk Analytics courses. Table 6.1 shows the pay-off matrix as shown to the participants. Every 100 tokens were converted to INR 10 and an additional extra credit (of 8 grade points) was offered—with a total maximum potential payout of INR 400 and 8% extra credit points (in classes where average grades were approximately 70%). The experiments were designed and performed using an online survey service and pen and paper.

Table 6.1 Payout matrix for the experimental minority game

Periods 1–20	Minority	Not in minority
Trade	100	0
Do not trade	100	0
Periods 21–40	Minority	Not in minority
Trade	95	0
Do not trade	100	0

The subjects were offered a binary decision choice: trade or not trade. They could win 100 tokens if they were in the minority in any period in Rounds 1–20. A Tobin tax of 5% was imposed from Round 21. Thus, from Rounds 21–40, if subjects did not trade and were in the minority, they would still make 100 tokens. However, if they did trade and were in the minority, they would make (100–5%) 95 tokens. Additionally, if the subjects were not in the minority, they would make 0 in any period in either phase of the experiment.

The set-up of the payoff structure in the game is similar to a transactions tax in completely speculative markets, and thus, akin to what a Tobin tax would be like on currency markets. The sole difference in this pay off structure is that the tax is implemented only in the gain frame and not in the loss frame by using a relative (versus an absolute) tax rate, to avoid any effects of biases relating to loss aversion. If the tax had been implemented on the loss frame, Trade-Not-in-Minority would have yielded a payoff of negative 5 tokens in periods 21–40 since the tax is on the transaction and not only on gains (similar to an income tax on realized capital gains). Due to a negative number in the Trade-Not-in-Minority state, subjects would have a negative payoff, which may lead a strong aversion towards trading and this might confound the effects that we would see in running the experiment with a tax only on the positive frame. Additionally, in most jurisdictions, loss making trades allow for tax offsets which would counterbalance any impact of a small absolute tax.

Results

We utilize three metrics from the financial markets that can be considered in the Minority Game structure to check if the taxes have the directional effect that is expected from it from a policy perspective. First, we try to find the effect of the Tobin tax on aggregate trading behavior—that is, do subjects tend to trade less due to the imposition of a tax. Secondly, we try to quantify the effect of the tax on excess demand in the market. Excess demand measures the behavior of the individual traders (and not just the aggregate trading choices) according to their decision to trade or not. And finally, we try to measure the effect of the tax on the volatility and compare it across the trading and non-trading periods.

Table 6.2 Number of periods in which 'Trade' was Minority

Periods 1–20	Periods 21–40
48%	63%

Trading Choice

We find a statistically significant effect on the imposition of the Tobin tax on the trading behavior of the experimental subjects. In Table 6.2, the numbers represent the number of periods in which Trade was the minority choice. If Trade is in the minority in equal number of instances in Periods 1–20 and 21–40, it would imply that the Tobin tax has had no behavioral effect on subjects and choices have remained the same. After the tax is imposed, Trade is a minority in significantly more periods; thus, showcasing a fall in the number of subjects choosing to trade as an effect of the Tobin tax. Between periods 1–20, in 29 out of the 60 non-tax periods (as mentioned above, three sessions were conducted with a total of 60 trading periods without taxes and 60 with taxes), that is 48% of the periods, Trade was the minority choice. Whereas in the tax frame (periods 21–40), Trade was in minority in 63% of the periods. This reduction is statistically significant and demonstrates a sharp fall in the propensity to trade due to tax imposition.

Excess Demand

In each period of the game, a player provides one signal: trade or not. Thus, for every period, Bid (b_i) is defined as:

$$b_i(t) \in \{-1, 1\}$$

and the resultant excess demand in the market is given by:

$$\sum_{i=1}^{N} A(t) = b_i(t).$$

If the taxes had no effect on excess demand, the sum value of the bids would remain similar in the trading and non-trading periods. In the experimental data, the average (per period) excess demand decreases from 0.13 in the periods with tax to -1.2 in periods without tax, implying a significant fall in excess demand in the market.

Volatility

Policy makers have primarily considered implementing the Tobin tax due to its potential effects on volatility. Popular macroeconomic models of currency exchange rates are based on concepts of purchasing power parity (Cassel, [23]) or relative interest rates differentials (Frankel, [24]) in case of floating exchange rates. However, actual currency exchange rates are far more volatile than either of these theories would support. Inflation data is released no more than once a month (or, in most countries, once a quarter) and interest rates are modified by central banks infrequently; however, speculative demand and supply dynamics generate fairly volatile currency exchange rates.

To understand the effect of the tax on volatility, we use the measurement process used for the MG that is calculated from the bids, if we do not assume any price setting mechanism and is defined as:

$$\sigma^2 = \frac{< A(t)^2 >}{N}$$

where $\langle . \rangle$ is a time average in the stationary state of the model. The normalization to the number of agents (N) is added to guarantee a finite value of σ^2 in the infinite system limit. In our experimental setting, the introduction of the Tobin tax leads to a 9.35% reduction in variance.

This result contrasts with existing empirical results using the securities transactions taxes analyzed by Capelle-Blancard and Havrylchyck [25] or earlier papers like Lanne and Vesala [26]. We believe that this difference arises from two sources: firstly, the studies often analyze broader securities transactions taxes which differ from the Tobin tax which was meant to be introduced in a speculative market like currency markets; and secondly, the tax is treated as an increase in transactions cost. Huber et al. mention that most papers in the existing literature "...consider the Tobin tax as a particular type of transactions costs on currency markets. Therefore, they investigate the impact of the size of transactions costs on trading volume and volatility. Using an innovative approach to derive transactions costs from futures prices, they show that higher transactions costs are associated with higher volatility and lower trading volume on foreign exchange markets. Similar results are presented in Hau (2006). Hence, there is no general agreement on the consequences of a Tobin tax on price volatility." We agree with that statement and propose that there is a larger behavioral impact of a tax than a simple increase in transactions cost.

The primary shortcoming of the current mechanism is the lack of an explicit price setting mechanism in the Minority Game. This allows us to study volume volatility of transactions volume in the market but not explicitly the volatility of prices. To generate a price setting mechanism, Bianconi et al. [2] use the Grand Canonical Minority Game (GCMG) in their paper. However, in the experimental study we study the original form of the game. The primary difference between the two is the option in the GCMG to sell or short sell the 'asset', thus allowing for three options to the participants: buy, sell or hold. We do not use that mechanism, since the primary

focus of this experiment is to study the effects of the taxes on trading volume and volatility and not prices. Additionally, the binary option form of the game does not present any theoretical problems in emulating a financial market in its simplest form, but aids in observing the direct effects of taxation on the volatility of the asset.

Conclusion and Future Work

The minority game provides an ideal framework to study the currency markets due its completely speculative design. This is the first experimental study which uses the minority game to study the effects of a Tobin tax on a currency market. Using commonly calculated measures for the minority game as a model of financial markets, we observe a significant reduction in excess demand, volatility and the number of trades due to the imposition of a Tobin tax.

In ongoing research, we are attempting to understand the behavioral aspects that lead to the outsized effect on volatility from the 5% tax rate that was imposed in this experiment. That is, we are attempting to understand why do markets tend to overreact to an imposition of a Tobin tax? And in conjunction, we are analyzing individual heterogeneity in trading behavior due to the imposition of the tax that might allow us to understand differential trader types in the market, their learning process and their strategies under a tax regime.

References

1. Tobin, J.: The New Economics One Decade Older: The Elliot Janeway Lectures in Honor of Joseph Schumpeter, p. 88. Princeton University, Princeton (1974)
2. Bianconi, G., Galla, T., Marsili, M., Pin, P.: Effects of tobin taxes in minority game markets. J. Econ. Behav. Organ. **70**(1–2), 231–240 (2009)
3. Kets, W.: Learning with fixed rules: the minority game. J. Econ. Surv. **26**(5), 865–878 (2012)
4. Challet, D., Zhang, Y.C.: Emergence of cooperation and organization in an evolutionary game. Phys. A Stat. Mech. Appl. **246**(3–4), 407–418 (1997)
5. Arthur, W.B.: Inductive reasoning and bounded rationality. Am. Econ. Rev. **84**(2), 406–411 (1994)
6. Challet, D., Zhang, Y.C.: On the minority game: analytical and numerical studies. Phys. A Stat. Mech. Appl. **256**(3–4), 514–532 (1998)
7. Challet, D.: Modelling markets dynamics: minority games and beyond. Ph.D. thesis, Verlag Nicht Ermittelbar (2000)
8. Challet, D., Marsili, M.: Relevance of memory in minority games. Phys. Rev. E **62**(2), 1862 (2000)
9. Challet, D., Marsili, M., Ottino, G.: Shedding light on EL Farol. Phys. A Stat. Mech. Appl. **332**, 469–482 (2004)
10. Gou, C.: Dynamic behaviors of mix-game model and its applications (2005). arXiv:physics/0504001
11. Gou, C.: Simulation of financial markets by agent-based mix-game model (2005)
12. Gou, C.: Agents play mix-game. Econophysics of Stock and Other Markets, pp. 123–132. Springer, Berlin (2006)

13. Frankel, J.A., Rose, A.K.: Empirical research on nominal exchange rates. Handbook of International Economics, vol. 3, pp. 1689–1729. Elsevier Science, Amsterdam (1995)
14. Umlauf, S.R.: Transaction taxes and the behavior of the swedish stock market. J. Financ. Econ. **33**(2), 227–240 (1993)
15. Chakraborti, A., Toke, I.M., Patriarca, M., Abergel, F.: Econophysics review: I. Empirical facts. Quant. Financ. **11**(7), 991–1012 (2011)
16. Bottazzi, G., Devetag, G., Dosi, G.: Adaptive learning and emergent coordination in minority games. Simul. Model. Pract. Theory **10**(5–7), 321–347 (2002)
17. Bottazzi, G., Devetag, G.: A laboratory experiment on the minority game. Phys. A Stat. Mech. Appl. **324**(1–2), 124–132 (2003)
18. Płatkowski, T., Ramsza, M.: Playing minority game. Phys. A Stat. Mech. Appl. **323**, 726–734 (2003)
19. Laureti, P., Ruch, P., Wakeling, J., Zhang, Y.C.: The interactive minority game: a web-based investigation of human market interactions. Phys. A Stat. Mech. Appl. **331**(3–4), 651–659 (2004)
20. Berg, J., Marsili, M., Rustichini, A., Zecchina, R.: Statistical mechanics of asset markets with private information (2001)
21. Huang, J.P.: Experimental Econophysics. Springer, Berlin (2015)
22. Hanke, M., Huber, J., Kirchler, M., Sutter, M.: The economic consequences of a Tobin tax an experimental analysis. J. Econ. Behav. Organ. **74**(1–2), 58–71 (2010)
23. Cassel, G.: Abnormal deviations in international exchanges. Econ. J. **28**(112), 413–415 (1918)
24. Frankel, J.A.: On the mark: a theory of floating exchange rates based on real interest differentials. Am. Econ. Rev. **69**(4), 610–622 (1979)
25. Capelle-Blancard, G., Havrylchyk, O.: The impact of the French securities transaction tax on market liquidity and volatility. Int. Rev. Financ. Anal. **47**, 166–178 (2016)
26. Lanne, M., Vesala, T.: The effect of a transaction tax on exchange rate volatility. Int. J. Financ. Econ. **15**(2), 123–133 (2010)

Chapter 7
Migration Network of the European Union: Quantifying the Effects of Linguistic Frictions

Aparna Sengupta and Anindya S. Chakrabarti

Abstract Immobility puzzle in the European Union (EU) takes the form of observed level of migration within the EU being substantially less than what is expected in a union allowing free labor mobility. We use a dynamic general equilibrium model of migration in a multi-region setting with heterogeneity in sectoral compositions, productivity and endowments of productive inputs, to construct a flow network of migrants (Chakrabarti and Sengupta, Econ. Model. 61:156–168 (2017)) ([7], Economic Modelling). When tested on the US data which we consider to be a benchmark for institutional homogeneity compared to Europe, this model explains substantial part of the variation in both the nominal and relative flows of state-to-state migration under suitable calibration. On the other hand, this model explains the relative flow network of the EU well but predicts a higher nominal flow than is seen in the data, thus illustrating and quantifying the puzzle. Following the hypothesis that institutional heterogeneity across the EU countries induces frictions on such labor reallocation process, we use dyadic regression to analyze the effects of pair-wise institutional distances which capture a broad spectrum of socio-cultural and political differences between countries, on the estimated missing mass of migrants. Linguistic differences appear to be the key factor explaining the missing mass of migrants.

Introduction

Migration is an outcome of a large number of people making choices over locations, based on consideration of multiple pecuniary and non-pecuniary factors. Over and above the pecuniary motives, institutional heterogeneity between the source and the destination might play an important role in the outcome of the decision-making

A. Sengupta (✉)
Bates White Economic Consulting, Washington DC 20005, USA
e-mail: aparna.sengupta@bateswhite.com

A. S. Chakrabarti
Economics Area, Indian Institute of Management, Ahmedabad, India
e-mail: anindyac@iima.ac.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_7

process. Within the US, the average flow of migrants across all states was about 2% in the last 20 years whereas in the European Union (EU) this rate is far less even after allowing free labor mobility across member states with the formation of the union. Reference [4] described this low level of migration as the 'European immobility puzzle'. This has often been attributed to cross-region variations in institutions which are much larger in the EU than in the US, indicating that institutional heterogeneity induces substantial friction on the labor relocation process (see e.g. [2, 12]). In this paper, we use a dynamic general equilibrium framework proposed earlier by [7] to map the whole migration network across all constituent states of an union on its macroeconomic fundamentals. This allows us to study the deviations of actual migration network from the equilibrium network, to pin down the effects of institutional heterogeneity between the source and the destination on the flow of migrants. As we will show, linguistic differences appear to be more important in explaining the migration gap in the EU than a multitude of other factors including heterogeneity in politics, customs or social attitudes. Thus our work provides a structural explanation of the missing mass of migrants.

Purely economic incentives e.g. higher wage or productivity in one country vis-a-vis another [3, 13], creates economic motivation for migration. Institutional factors also impact this decision-making process. For example, [2] argued that differences in cultures or customs present an impediment in the process. We conjecture that from a purely economic point of view, the phenomena of migration between countries with similar economic characteristics can be thought of as an adjustment process or reallocation of labor resulting from uneven productivity shocks. However, various socio-cultural and political factors can induce frictions on that mechanism reducing the extent of reallocation. Thus observed migration contains effects of both types of factors, economic as well as non-economic, which may work in opposite directions.

In the following, we first present a dynamic model with N-regions, two-sectors augmented with sector and region-specific idiosyncratic productivity shocks. This model had previously been developed by [7] to understand migration in presence of multiple destinations in a frictionless world. The basic objective is to build a migration network across countries, which can be compared with yearly data. We consider a model with T periods and N regions (N countries belonging to the EU or N states of the US). Within a year, all regions receive idiosyncratic productivity shock for T times. Each region is populated by a continuum of workers who consume and can move to different regions depending on the relative productivity across regions. Capital stock is fixed across regions. Workers consume tradables and non-tradables, both of which are produced via a two-stage production process. Following [5], we assume that trade occurs only at the intermediate stage. Due to productivity shocks, there will be productivity differences across states. Simultaneously, productivity shocks will have spill-over effects on other regions through the trade channel. Thus there would be net productivity differences across regions creating opportunities for workers to move across regions. This flow of workers from one region to another is interpreted as bilateral migration. When all regions experience such productivity shocks in one-period $t \in T$, some regions will be net donors and some regions will be

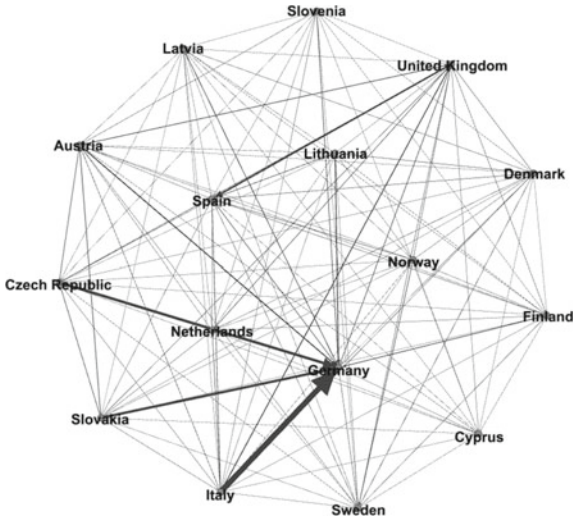


Fig. 7.1 Migration network (directed and weighted) of a subset of European union (year 2000)

net receivers. Figure 7.1 for example, shows the net migration network and Fig. 7.2 shows the spatial distribution of migration from actual data. By aggregating the flow network over T periods, we can construct a migration network across countries. By calibrating the productivity shocks to the yearly tfp shocks across countries, we interpret the migration network to be representative of the yearly migration network.

Under suitable calibration, the model is consistent with the US data in terms of the labor network generated as well as the total mass of migrants [7]. It is important to stress here that the model generates the full network of labor flow across all

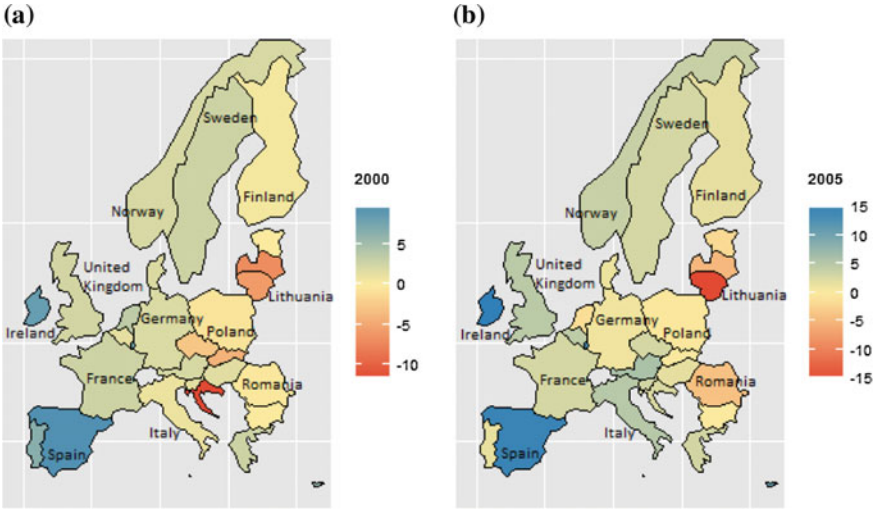


Fig. 7.2 Net migration in (a subset of) European Union. Panel **a**: year 2000. Panel **b**: year 2005

constituent regions, i.e. if there are N regions, it generates $N \times N$ flow matrix. Hence it provides a systemic view to study the aggregate flow across all pairs of regions as well as relative flows across specific pairs of regions. Since the US is taken to a frictionless case with sufficient institutional homogeneity, a good fit of the model to the US data indicates that the model captures economic incentives arising due to productivity, to a reasonable degree. When calibrated to the EU, the same model captures the relative flow of labor across the EU well but predicts higher aggregate flow than is seen in data. This gap can be thought of representing missing migrants in the European immobility puzzle. Such quantification of the missing mass of migrants is the main objective of using the model in the present context.

After quantifying the missing mass of migrants across every pair of countries, we explain the gap by using an array of various institutional factors (more than 50 variables, full list provided in Table 7.9). We see that linguistic differences is the most persistent factor across years explaining the gap. Entry into Euro area increases bilateral migration across countries. Presence of informal labor markets and Hofstede indices (socio-cultural distance) can explain the gap partially.

The rest of the paper is structured as follows. In section "A Structural Model of Migration Network (Chakrabarti and Sengupta, 2017)", we describe the basic model. In section "The Migration Network of Europe", we construct the simulated migration network in the EU and contrast it with the actual network. Finally, in section "European Immobility Puzzle", we carry out regression analysis and show that linguistic differences are important factors for explaining the missing mass of migrants. Then we conclude.

A Structural Model of Migration Network (Chakrabarti and Sengupta, 2017)

In this section, we briefly describe a model to generate a migration network across regions in a frictionless economy. This model was proposed in [7] with an application to the US interstate migration data. The mathematical structure is a simplified version of the model proposed by [5], who used it for modeling propagation of productivity shocks across regions. In the present context, we have used the model to capture annual bilateral migration between different pairs of countries within EU upon suitable calibration. We do not give the full derivation of the model here to avoid repetition. For brevity of exposition, we broadly define the structure of the model below and move on to the estimation in next section.

Households' Problem

In each region a continuum of households who supply labor for the local production process. There are two final goods, tradables (M) and non-tradables (S). For the sake of exposition, we assume manufacturing industries to constitutes the tradables

sector and the service producing industries to constitute the non-tradables sector. The households' optimization problem in the nth region at a generic time-point t is to choose consumption of manufactured (C_{nt}^M) and service goods (C_{nt}^S) so as to

$$\max_{\{C_{nt}^M,\ C_{nt}^S\}} U_{nt} = (C_{nt}^M)^{\alpha}(C_{nt}^S)^{(1-\alpha)} \quad \text{subject to} \quad P_{nt}^M C_{nt}^M + P_{nt}^S C_{nt}^S \leq r_{nt}\frac{K_{nt}}{L_{nt}} + w_{nt}, \tag{7.1}$$

where the term on the right hand side denotes per-capita income which is the sum of rental income earned from fixed capital stock (K) and wage (w). Interest rate is denoted by r and labor by L and wage rate by w. If the migrants travel through a sequence of regions $\{n\}_{1,\ldots,T}$, her T periods' utility is given by

$$U^T = \mathbb{E}\left(\sum_{\tau=1}^{T} U_{n\tau}\right). \tag{7.2}$$

Production Process

Following [5], we impose a two-tier production structure with trade in the intermediate goods. Manufactured goods and the service products are final goods which are produced by bundling a continuum of intermediate goods. These intermediates are produced by combining local labor and capital stock.

Intermediates' Production

There are two sectors $j \in \{M, S\}$ in each region n, each producing a continuum of varieties of intermediate goods. Each sector has an i.i.d. productivity shock process, ξ_{nt}^j and also receives shocks at the productivity level Z_{nt}^j. The shock process ξ_{nt}^j follows a Frechet distribution characterized by parameter θ^j. The production functions are symmetric across sectors ($j \in \{M, S\}$) and defined as

$$q_{nt}^j = \xi_{nt}^j Z_{nt}^j (k_{nt}^j)^{\beta}(l_{nt}^j)^{1-\beta}, \tag{7.3}$$

where lowercase letters l and k denote the demand for labor and capital respectively by a representative firm. The shock process Z_{nt}^j follows a multiplicative random walk,

$$Z_{nt}^j(t+1) = \psi_{it} Z_{nt}^j(t) \quad \text{where } \psi_i \sim N(1, \sigma_i) \text{ and } i \in \{M, S\}. \tag{7.4}$$

Final Goods' Production

Final goods are produced by bundling continuum of intermediates $\tilde{q}$ in both sectors ($j \in \{M, S\}$),

$$Q_{nt}^{j} = \left[\int (\tilde{q}_{nt}^{j}(\xi^{j}))^{\gamma_{nt}^{j}} \phi^{j}(\xi^{j}) d\xi^{j} \right]^{1/\gamma_{nt}^{j}}, \tag{7.5}$$

where $\phi(.)$ denotes the distribution of shocks,

$$\phi^{M}(\xi^{M}) = exp\left(- \sum_{n=1}^{N} (\xi_{nt}^{M})^{-\theta^{M}} \right), \tag{7.6}$$

$$\phi^{S}(\xi^{S}) = exp\big(- (\xi_{nt}^{S})^{-\theta^{S}} \big). \tag{7.7}$$

The aggregate price level for the final goods can be written as

$$P_{nt}^{j} = \left[\int \left(p_{nt}^{j}(\xi^{j}) \right)^{\frac{\gamma_{nt}^{j}}{\gamma_{nt}^{j}-1}} \phi^{j}(\xi^{j}) d\xi^{j} \right]^{\frac{\gamma_{nt}^{j}-1}{\gamma_{nt}^{j}}}. \tag{7.8}$$

Market Clearing Conditions for Factors of Production

Since final goods are non-tradable in all sectors, the corresponding market clearing condition would just state that domestic demand equals domestic supply. Moreover, even for non-tradables, the same condition would hold. Following [5], the trade cost between regions n and m (in units of good produced in location n) is given as

$$\tau_{nm}^{M} \geq 1 \ \text{ and } \ \tau_{nm}^{S} = \infty, \tag{7.9}$$

implying that service goods are non-tradable where as manufactured goods are tradables.

For productive inputs, total inputs i.e. labor or capital, must be equal to the sum of the sectoral allocation,

$$L_{nt}^{M} + L_{nt}^{S} = L_{nt} \ \text{ and } \ K_{nt}^{M} + K_{nt}^{S} = K_{nt} \ \ \forall n \leq N. \tag{7.10}$$

Note that since capital is immobile, we do not have market clearing condition for capital at the aggregate level. For labor, the corresponding condition would be $\sum_{n} L_{nt} = \bar{L}$, where $\bar{L}$ is the total labor endowment in the whole economy.

Equilibrium

Given labor endowments $\{L_{nt}\}$ and the capital endowment $\{K_{nt}\}_{nt}$, a competitive equilibrium is an utility level $\bar{U}$, input prices $\{r_{nt}, w_{nt}\}_{nt}$, labor allocation $\{L_{nt}\}_{nt}$, expenditure on manufactured goods and services $\{X^M_{nt}, X^S_{nt}\}_{nt}$, consumption $\{C^M_{nt}, C^S_{nt}\}_{nt}$, prices of manufactured goods and services $\{P^M_{nt}, P^S_{nt}\}_{nt}$ and pairwise regional intermediate expenditure share in manufactured goods and services $\{\pi^M_{nit}, \pi^S_{nit}\}_{nit}$ such that all markets clear in all regions $n \in N$.

At every time point t, all regions receive shocks $\hat{Z}^j_{nt}$. Given the values for $\{\theta^j, \alpha, \beta\}^N_{n,j=\{S,M\}}$ and data for the initial allocation $\{I_{nt}, L_{nt}, \pi^j_{ni}, \hat{Z}^j_{nt}\}^{N,N}_{n,i,j=\{S,M\}}$, we can find solution for changes in all real quantities $\{\hat{w}_{nt}, \hat{L}_{nt}, \hat{X}^j_{nt}, \hat{P}^j_{nt}, X'^j_{nt}, \pi'^j_{ni}\}^{N,N}_{n,i,j=\{M,S\}}$ where $\hat{x} = x_{new}/x_{old}$.

Simulating the Migration Network

After the realization of productivity shocks across regions at a time point $t \in T$, workers will move across regions giving rise to changes in labor allocation as $\{\hat{L}_{nt}\}_{n\in N}$. Thus, net change in labor allocation can be found as $\{(\hat{L}_{nt} - 1)L_{nt}\}_{n\in N}$. Reference [7] defined net flow of workers across the $\{i, j\}$th pair at time t as

$$F^t_{ji} = \left(\frac{(\hat{L}_{jt} - 1)L_{jt}}{\sum_{n\in\mathcal{N}^{out}}(\hat{L}_{nt} - 1)L_{nt}}\right)(\hat{L}_{it} - 1)L_{it}, \tag{7.11}$$

where $\mathcal{N}^{out}$ is the set of countries from which labor migrates to other countries and $j \in \mathcal{N}^{out}$. Thus at each period $t \in T$, we can construct the labor flow network by considering pairwise inflows and outflows for all pairs of regions $\{i, j\} \in N^2$. By summing up labor flow networks across T periods, we can construct the aggregate yearly migration network. See [7] for a complete derivation and discussion.

Calibration Exercise

We calibrate the parameters (see Table 7.1 for the numerical values) of the theoretical model for 16 of the countries in Europe. We also present the analysis for the US to make a comparison. The regions in both the cases have inherent homogeneity in terms of economic factors. However, institutional frictions should be much clearer in the EU countries. In the following, we discuss the calibration exercise and then compare the results obtained from the theoretical model with the real data.

The shocks to the productivity Z Eq. (7.4) is normally distributed and scaled by the time horizon T,

Table 7.1 Calibrated parameter values

Description	parameter	value
Service goods' share in cost	1-α	0.6
Capital's share in cost	β	0.3
Dispersion parameter (intermediates): Manf.	θ_m	8
Dispersion parameter (intermediates): Serv.	θ_s	2
Std. dev. of aggregate shocks (US, 2007): Manf.	σ_M	0.038
Std. dev. of aggregate shocks (US, 2007): Serv.	σ_S	0.005
Std. dev. of aggregate shocks (EU, 2000): Manf.	σ_M	0.027
Std. dev. of aggregate shocks (EU, 2001): Manf.	σ_M	0.023
Std. dev. of aggregate shocks (EU, 2002): Manf.	σ_M	0.028
Std. dev. of aggregate shocks (EU, 2003): Manf.	σ_M	0.034
Std. dev. of aggregate shocks (EU, 2004): Manf.	σ_M	0.028
Std. dev. of aggregate shocks (EU, 2005): Manf.	σ_M	0.067
Std. dev. of aggregate shocks (EU, 2006): Manf.	σ_M	0.042
Std. dev. of aggregate shocks (EU, 2007): Manf.	σ_M	0.105
Std. dev. of aggregate shocks (EU, 2000): Serv.	σ_S	0.014
Std. dev. of aggregate shocks (EU, 2001): Serv.	σ_S	0.019
Std. dev. of aggregate shocks (EU, 2002): Serv.	σ_S	0.010
Std. dev. of aggregate shocks (EU, 2003): Serv.	σ_S	0.011
Std. dev. of aggregate shocks (EU, 2004): Serv.	σ_S	0.009
Std. dev. of aggregate shocks (EU, 2005): Serv.	σ_S	0.020
Std. dev. of aggregate shocks (EU, 2006): Serv.	σ_S	0.014
Std. dev. of aggregate shocks (EU, 2007): Serv.	σ_S	0.024
Length of simulation	T	200
# simulations averaged	–	O(10)

$$\psi_{it} \sim N\left(\frac{1}{T}, \frac{\sigma_i}{T}\right). \tag{7.12}$$

The standard deviation of TFP (σ) of sectors in the EU in one particular year is matched with the cross-sectional standard deviation of the same sector across the countries in the EU for that year. Table 7.1 presents the calibrated values of the parameters. The values which we take to be common across regions and time, are given at the beginning. For others, we mention the relevant unions as well as the sectors and years. For each year, the standard deviations of the productivity shocks are averaged over all constituents regions of the union for the purpose of simulation over T periods.

From data, we calculate the pairwise migration as

$$y_k = \frac{m_{ij}^{data} + m_{ji}^{data}}{\sum_{nt} L_{nt}^{data}}. \tag{7.13}$$

whereas the model predicted pairwise migration is calculated as

$$x_k = \frac{m_{ij}^{model} + m_{ji}^{model}}{\sum_{nt} L_{nt}^{model}}. \tag{7.14}$$

We also control for contiguity through a dummy variable. The basic specification for nominal flow network is

$$y_k = \alpha_0 + \alpha_1 x_k + \alpha_2 D_{cont.} + \varepsilon_k \tag{7.15}$$

where $D_{cont.}$ is a dummy for contiguity and ε_k is an $i.i.d.$ error term. If $\hat{\alpha_0} = 0$ and $\alpha_1 = 1$, we can say that the model captures the nominal flow network well. As [7] had shown, the predicted total mass of migrants for US match pretty well with the data. Calibrating the model we see that the total flow should be around 2%. From ACS data (Table 7.8) we do get the overall migration to be around 2%. Thus the orders of the nominal flows as is seen in the data and derived from the model, are arguably comparable. We also consider the relative flows of migrants across regions by constructing a new dependent variable $\tilde{y}_k = y_k / \sum_k y_k$ and a new explanatory variable is $\tilde{x}_k = x_k / \sum_k x_k$. The relative network of migration in the US, which was taken as the closest approximation to a frictionless place in terms of both economic and non-economic factors, is also described well by a model emphasizing only economic incentives behind migration.

The Migration Network of Europe

We look into migration data from 2000 to 2007 for 16 countries (See Appendix "Sources of Data" for details on sources of data) which gives us the full 16×16 migration matrix depicting the bilateral flow. Our objective is to build the complete matrix from the theoretical model and compare each element with the data. However, there is incompleteness in the available data showing the bilateral flow of labor as a few countries do not report the migration statistics at all, some countries stop reporting after a period of time and some start only after a time point. So we extract the maximum amount of data available and compare it with the results that the theoretical model provides. Table 7.2 provides a summary of the data available.

From the model we get that due to TFP differences net migration in the 16 countries should be around 2%. Next, we regress the dyad specific bilateral migrations from actual data on the TFP driven migration results (from the theoretical model). Table 7.3 contains results of regressing data on model-predicted migration. In Tables 7.4 and 7.6 we present the panel results for the 16 countries over 2000–2007.

We find from Table 7.3 that though the coefficient of TFP driven migration is much lower than 1 which should have been the case if the model matches data perfectly, but it is significant and in each year the model has sufficiently high R^2. This is an

Table 7.2 Descriptive summary for bilateral migration within Europe (16 countries)

Year	Obeservations	Mean	Std. Dev	Min	Max
2000	66	5056.924	8813.484	0	45439
2001	66	5231.076	9290.369	0	43375.5
2002	66	5377.379	9570.473	2	41312
2003	66	5203.114	9455.308	6	49670
2004	66	5608.924	10292.27	3	59337
2005	66	5830.758	10729.29	7	57652
2006	69	5239.217	10345.33	8	56612
2007	66	4307.53	7815.329	16	34417

Table 7.3 Regression results with robust errors for the EU—Nominal

	TFP driven migration (Rob Std Err)	Contiguity (Rob Std Err)	Intercept (Rob Std Err)	Adj. R^2
2000	0.05836^c	0.00001^c	0.00000	0.7808
	(0.00781)	(0.00001)	(0.00000)	
2001	0.05870^c	0.00002^b	0.00000	0.7700
	(0.00683)	(0.00001)	(0.00000)	
2002	0.06118^c	0.00001^c	0.00000	0.7794
	(0.00613)	(0.00000)	(0.00000)	
2003	0.05456^c	0.00001^b	0.00000	0.6468
	(0.00980)	(0.00001)	(0.00000)	
2004	0.05709^c	0.00001^b	0.00000	0.5938
	(0.01216)	(0.00001)	(0.00000)	
2005	0.06132^c	0.00001^b	0.00000	0.6788
	(0.00921)	(0.00001)	(0.00000)	
2006	0.06376^c	0.00001^b	0.00000	0.6927
	(0.01324)	(0.00001)	(0.00000)	
2007	0.06030^c	0.00001^b	0.00000	0.7450
	(0.00512)	(0.00001)	(0.00000)	

Note: ^{a}p<0.1, ^{b}p<0.05, ^{c}p<0.01, $N = 68$

interesting finding as it basically suggests that the total mass of migrants predicted by the model is much higher than what is seen in the data. Table 7.4 presents a panel estimate of the same. The estimated coefficients indicate a similar conclusion.

Next, we regress the relative weights of edges of the data on model. The results are presented in Table 7.5. Clearly, after normalization the estimated coefficient increases to about 0.8 which is much closer to 1. Note that $\tilde{y}, \tilde{x} \in [0, 1]$ making them comparable in order. So in relative sense the theoretical model does quite well in explaining the migration in Europe. However, the it does not match the total migration; in fact predicts a much higher value. Table 7.6 presents a panel estimate on the relative flows.

Table 7.4 Panel regression result for the EU—Nominal

Variable	Coefficient (Rob Std Err)
TFP driven migration	0.061 ***
	(0.010)
Intercept	0.000
	(0.000)
N	528
R^2 overall	0.6629
$\chi^2_{(2)}$	34.82

Table 7.5 Regression results with robust errors for the EU—Relative

	TFP driven migration (Rob Std Err)	Contiguity (Rob Std Err)	Intercept (Rob Std Err)	Adj. R^2
2000	0.84328[c]	0.01700[c]	–0.00046	0.7808
	(0.11289)	(0.00612)	(0.00116)	
2001	0.83180[c]	0.01779[b]	–0.00042	0.7700
	(0.09676)	(0.00675)	(0.00105)	
2002	0.85237[c]	0.01418[c]	–0.00013	0.7794
	(0.08539)	(0.00522)	(0.00100)	
2003	0.79403[c]	0.01494[b]	0.00063	0.6468
	(0.14261)	(0.00591)	(0.00113)	
2004	0.77543[c]	0.01344[b]	0.00116	0.5938
	(0.16522)	(0.00600)	(0.00120)	
2005	0.81675[c]	0.01351[b]	0.00052	0.6788
	(0.12271)	(0.00606)	(0.00101)	
2006	0.75724[c]	0.01441[b]	0.00122	0.6927
	(0.15722)	(0.00702)	(0.00127)	
2007	0.71356[c]	0.01980[b]	0.00104	0.7450
	(0.06064)	(0.00894)	(0.00122)	

Note: [a]p<0.1, [b]p<0.05, [c]p<0.01, $N = 68$

Table 7.6 Panel regression result for the EU—Relative

Variable	Coefficient (Rob Std Err)
TFP driven migration	0.650 ***
	(0.097)
Intercept	0.005 ***
	(0.002)
N	528
R^2 overall	0.6611
$\chi^2_{(2)}$	44.746

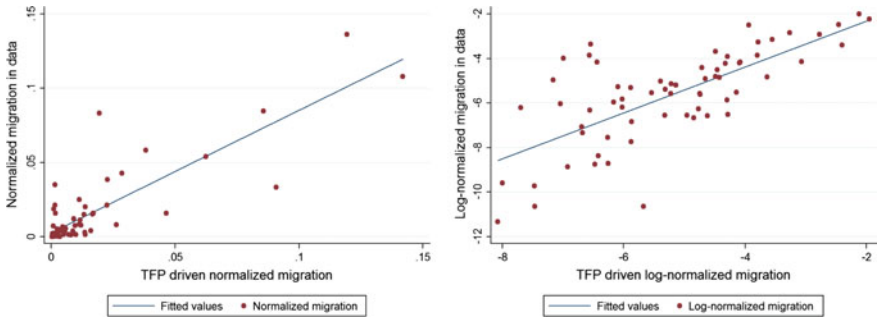


Fig. 7.3 Scatter plots showing the normalized actual dyad migration data on TFP simulated results for the European countries for year 2000

In the left panel of Fig. 7.3, we plot the normalized bilateral migration data on the y-axis and the predicted values of the same on the x-axis. In the right panel we take the natural log of both variables to reduce the effects of the outliers. Each point on the scatter plot denotes the real data and the prediction for a dyad.

European Immobility Puzzle

One of the basic principles behind the formation of the European Union was to ensure freedom of movement of productive inputs. In particular, it was supposed to reduce the barriers to labor flow making the market more flexible. Multilateral gravity equation helps us to pin down the relative strengths of the edges of the migration network. However, as is clear from the above results, the model shows that under reasonable parameterization the predicted mass of migrants are in the order of 100 times more than what is seen in Europe for the period we considered (2000–07). This refers to the puzzle that even after the legal and political barriers have been systematically removed thus potentially reducing economic frictions on the labor allocation process, people did not respond immediately to the existing incentives. This problem has attracted attentions both from a theoretical and policy-making point of view. In particular, [2] ascribe this role to the negative effects of cultural differences indicating that such distances can induce an extremely low migratory response if properly addressed. In this paper, we complement this analysis using many other types of frictions ranging from social to political along with the obvious factor, linguistic differences. In this section, we look into a list of fine-grained measures of institutional differences between the 16 European countries and argue that these substitute some of the TFP driven migration instead of complementing and thus, provide "frictions" opposing the incentives.

Distances in Institution and Culture

We look at a broad list of variables which could ideally be considered as frictions. We start with historical links between countries. We used the [6] data to determine colonial links between countries or whether the two countries in the dyad were the same country historically.

One of the hypothesis could be that language barrier is one of the reasons which stops people from migrating easily. To control for this we looked into several language indices. From the CEPII, bilateral data on whether two countries speak the same official language, native language, language proximity index and common language index was obtained. In Table 7.7, LangIndex is the common language index. This index gives an approximate distance between two countries due to language. If the index is higher that means the two countries have fewer language barriers. We also looked into ethnologue language statistics [8]—country-specific data on total number of languages used as the first language, immigrant languages in the country and probability that two people selected at random will have different mother tongues (Greenberg's diversity index).

Differences in culture could be another barrier to migration—to study this effect, we use the Hofstede's cultural indices [11]. This is a rich set of index encompassing cultural aspects such as individualism versus collectivism in the economy, uncertainty avoidance, power distance (strength of social hierarchy), masculinity-femininity (task orientation versus person-orientation), long-term orientation and indulgence

Table 7.7 Regression results with robust errors for the EU—frictions

	Contiguity	LangIndex	Indivi	Pragm	Euro	ShadowEco	Intercept	Adj. R^2
2000	0.23^a	0.43^b	-0.12^c	-0.11^b	–0.06	-0.01^b	0.72^c	0.4901
	(0.13)	(0.19)	(0.04)	(0.05)	(0.05)	(0.01)	(0.25)	
2001	0.26^a	0.44^b	-0.13^c	-0.12^b	–0.06	-0.01^b	0.77^c	0.4923
	(0.14)	(0.19)	(0.05)	(0.06)	(0.05)	(0.01)	(0.26)	
2002	0.21	0.45^b	-0.13^c	-0.12^b	–0.07	-0.01^b	0.78^c	0.4798
	(0.13)	(0.20)	(0.05)	(0.06)	(0.05)	(0.01)	(0.27)	
2003	0.21	0.46^b	-0.13^c	-0.12^b	–0.07	-0.02^b	0.78^c	0.4765
	(0.13)	(0.20)	(0.05)	(0.05)	(0.06)	(0.01)	(0.26)	
2004	0.18	0.49^b	-0.13^b	-0.13^b	–0.07	-0.01^a	0.78^c	0.4599
	(0.13)	(0.20)	(0.05)	(0.06)	(0.05)	(0.01)	(0.29)	
2005	0.18	0.44^b	-0.11^b	-0.10^b	–0.06	-0.01^a	0.66^c	0.4522
	(0.12)	(0.18)	(0.04)	(0.05)	(0.05)	(0.01)	(0.23)	
2006	0.18	0.53^b	-0.08^b	-0.09^a	–0.09	-0.01^a	0.57^b	0.3977
	(0.13)	(0.23)	(0.04)	(0.05)	(0.06)	(0.01)	(0.22)	
2007	0.15	0.59^b	-0.09^b	-0.10^a	-0.15^b	-0.02^b	0.67^c	0.4431
	(0.13)	(0.23)	(0.04)	(0.05)	(0.07)	(0.01)	(0.23)	

Note: ^{a}p<0.1, ^{b}p<0.05, ^{c}p<0.01, $N = 68$

versus self-restraint. In Table 7.7 'Indiv' refers to Individualism and 'Pragm' refers to Pragmatism and they are two of the Hofstede cultural index. These indices are country specific. For dyad level regression we considered the numerical differences between these indices for the two countries as a proxy of their 'distance' in the corresponding category. So a higher value in distance for 'individualism' between a pair of countries would mean that one country in the pair believes in individualistic society as a way of life and the other country believes in a relatively less individualistic society which is another way of saying that the country believes in a more collective/family-oriented way of life.

Next, we considered several stability indices broadly related to the polity. All data were collected from various reports compiled and made publicly available by World Bank. We looked into government stability, democracy index, ethnic tensions, religious conflicts, military in politics and external conflicts to understand the political stability in the economy. For each of these risk rating available on country level we considered the 'distance' between the ratings between two countries for dyad regressions. For socio-economic stability, we looked into corruption index, freedom of press, socio-economic conditions and voice and accountability. Distance between financial stability indices like financial risk, investment profile and existence of shadow economy are also included as controls. Distance in shadow economy index would mean in the dyad one of the countries has a huge underground economy and the other one does not. We also looked into some of the Europe specific dummies—such as using euro or not and entry into European Union. In the next section we look into the regression results on all the mentioned distance variables.

A general rule we followed is that since many of these frictional variables are extremely correlated especially so when they belong to the same family. We use stepwise regression methodology to pin down the predictors. Most of the considered 'friction' variables under an umbrella term broadly defining similar characteristics are correlated. Given the high level of correlation in the data, we do not consider all variables simultaneously as that will not increase the predictive power. The point is that many of the frictional variables that turn out to be important in explaining the puzzle, are not unique. They often have some other measures, almost similarly defined and hence very correlated, that can be almost equally effective in explaining the same phenomena.

Explaining the Missing Flow: Effects of Institutional Factors

For the 16 countries in Europe, we computed all the institutional distance measures. As a response variable we consider the ratio of actual bilateral migration data to TFP driven migration. We regress this variable on various institutional measures. The results are tabulated in Table 7.7. The reason we took the ratio of the data to the model prediction (y_k/x_k as defined in Eqs. (7.13) and (7.14) resp.) as the variable to be explained is that this way we get rid of the gravity terms which are driven solely by economic causes. Thus the left over variations would be driven by other non-

economic factors. Two methodological points are to be noted. One, some variation in the data could be due to misreporting which we cannot rectify and two, we are considering the model to capture the economic incentives completely and in the gravity equation set up, the proportionality term captures all institutional effects, magnifying or lessening the flow. Consider any pair of regions $\{i, j\}$ and call it dyad k. Given this notation, we see that $y_k = C_{ij}^{data}.L_i.L_j/d_{ij}^{\eta}$ and $x_k = C_{ij}^{model}.L_i.L_j/d_{ij}^{\eta}$ and numerical solutions indicate $C_{i,j}^{model}$ is roughly a constant, independent of the specific dyad considered (i.e. $C_{i,j}^{model} = C$). Hence, we have

$$\frac{y_k}{x_k} = \left(\frac{C_{ij}^{data}}{C}\right) d_{ij}. \tag{7.16}$$

Thus after taking ratios, the gravity terms wash out and we get the pair-specific constants capturing the socio-economic and political distances. The idea is that a low value of the variable (y_k/x_k) indicates that less migration occurred between a pair of countries consisting the dyad k in reality than in the model. Therefore, a negative value of the coefficient of a suitably defined distance metric would indicate presence of a friction. Alternatively, in presence of similarities in any dimension e.g. linguistic, we would expect a higher flow.

Therefore, following the notation in Eqs. (7.13) and (7.14), the regression specification is

$$\frac{y_k}{x_k} = \delta_0 + \delta V_k + \delta_1 D_{cont} + \varepsilon_k' \tag{7.17}$$

where V_k is a vector of distances measured for multiple socio-political and economic attributes, D_{cont} is a dummy for contiguity and ε' is an error term. Table 7.7 shows the regression results for the European country dyads. For each year, from 2000 to 2007, we regress ratio of actual bilateral migration data to TFP driven migration on Euro currency dummy and distance between-language index, Hofstede index of individualism (vs. collectivism) and pragmatism, financial risk index and shadow economy, controlling for contiguity. We also tested for a bunch of other variables including various social and political factors which did not turn out to be significant. Types and sources of all variables considered are presented in Tables 7.8 and 7.9.

Table 7.8 Data sources

Data	Source
Migration—EU	[10]
Trade—EU	[10] with [14] and [15]
Contiguity—EU	[6]
TFP—EU	[16]
Cultural indices	[11]
Economic indicators	[17]

Table 7.9 Friction variables with source

List of variables	Type	Source
Power distance	Cultural	[11]
Individualism	Cultural	[11]
Masculinity	Cultural	[11]
Uncertainty avoidance	Cultural	[11]
Pragmatism	Cultural	[11]
Indulgence	Cultural	[11]
Diversity index	Cultural	[8]
Total number of living languages	Cultural	[8]
% of all living languages	Cultural	[8]
Indigenous languages	Cultural	[8]
Immigrant languages	Cultural	[8]
Common official language	Cultural	[6]
Common spoken language	Cultural	[6]
Common native language	Cultural	[6]
Unadjusted value of linguistic proximity (Tree)	Cultural	[6]
Adjusted value of linguistic proximity (Tree)	Cultural	[6]
Unadjusted value of linguistic proximity (ASJP)	Cultural	[6]
Adjusted value of linguistic proximity (ASJP)	Cultural	[6]
Common language Index based on our log specification	Cultural	[6]
Common language Index based on a level specification	Cultural	[6]
Entry in EU	Institutional	[9]
Euro currency use	Institutional	[9]
Two countries are contiguous or not	Institutional	[6]
Common language by 9% of the population in both countries	Cultural	[6]
Colonial link	Cultural	[6]
Common colonizer after 1945	Cultural	[6]
Currently in a colonial relationship	Cultural	[6]
Colonial relationship after 1945	Cultural	[6]
Were or are the same country	Institutional	[6]
Geodesic distances	Institutional	[6]
Geographic coordinates of the capital cities	Institutional	[6]
Civil liberties score and the political rights score	Political	[17]
Freedom of the press	Political	[17]
Institutionalized democracy	Political	[17]
Polity score	Political	[17]
Voice and accountability	Political	[17]
Political Stability	Political	[17]
Government effectiveness	Political	[17]
Regulatory quality	Political	[17]

(continued)

Table 7.9 (continued)

List of variables	Type	Source
Rule of law	Political	[17]
Corruption control	Political	[17]
Transparency in public sector	Political	[17]
Age of leadership	Political	[17]
Excluded population in total politically relevant population	Political	[17]
Power sharing groups	Political	[17]
Ethnic groups	Political	[17]
Discriminated population	Political	[17]
Powerless population	Political	[17]
Regional power	Political	[17]
Junior partner in power sharing arrangement	Political	[17]
Senior parter in power sharing arrangement	Political	[17]
Monopoly power	Political	[17]
Shadow economy	Political	[17]

The signs of the coefficients have meaningful interpretation—for example having similar language helps in migration (positive signs of the LangIndex) and different cultures act as an impediment to migration (negative signs for distance between cultural index). This exercise shows that there are factors which encourage or discourage migration, over and above mere economic incentives. We have done robustness checks in Appendix "Frictional Variables and Additional Plots" in terms of partial regressions. The partial residual plot for language is also shown. All results agree with the prior interpretation. Th finding that language has an important role to play determining the level of international migration is also corroborated by the empirical exercise of [1]. They focus mostly on the skill and ability of people in learning languages whereas we complement our findings by incorporating related variables on other distance measures of languages (see Table 7.9). Qualitatively similar findings prevail.

Summary and Conclusion

We have presented a structural interpretation of the European immobility puzzle by quantifying the mass of missing migrants in the EU in presence of relative heterogeneity of productivity across countries. We show that linguistic differences are important factors for explaining the missing mass of migrants. Finally, we can ask a seemingly obvious question: why did we take social distances as a friction? Would it be possible to imagine a scenario where a higher social distance actually complements migratory responses rather than substituting it. The answer is, it is possible. In south-to-north migration this may in fact provide an incentive to migrate.

For example, people would migrate from low income countries to comparatively prosperous ones but only selectively. Along with economic incentives, migrants also weigh their chances on the socio-political conditions of the receiving countries.[1] Thus a higher distance between a donor country and a receiver country may compel individuals to migrate. However, when the countries are more-or-less similar in these respects, this might hinder the labor reallocation process as is found in case of the European countries. Although we have not established causality in the present exercise, we may state that the findings here are indicative of the idea that higher linguistic homogeneity contributes to more migration, which is in sync with the findings of [1].

Appendix

Sources of Data

For the European Union we looked into bilateral migration from 2000 to 2007 within Austria, Belgium, Denmark, Spain, Germany, Czech Republic, Finland, France, Italy, Hungary, Ireland, Netherlands, Sweden, Slovenia, United Kingdom and Norway (see Table 7.8). Migration is defined as movement across different countries of residence in one year. More specifically, if a person was in a different country of residence in the previous year than this year, then we count that person as a migrant.

Frictional Variables and Additional Plots

Below we present Table 7.10 showing correlation among a few institutional variables. High correlation is apparent indicating possible multicollinearity problems in the OLS estimation if we use a large number of variables simultaneously. We have used stepwise regression to find the important variables from the large set of variables considered in Table 7.9.

We would like to understand the influence of each variable which is used as friction (see Table 7.9). We use the post-estimation tool partial regression plot for this. In the dyadic regression the dependent variable is the ratio of bilateral migration as seen in data to bilateral migration which is TFP driven (simulated). We try to understand the importance of each variable, for example language index—for this we first regress the dependent variable on the remaining regressors (not including language index) and plot the residuals on the Y-axis. Next we regress language index on the remaining regressors and plot the residuals on the X-axis. These plots show relation between the dependent variable and each friction variable (Fig. 7.4).

[1] We are, of course, not considering forced migration due to political and social instability like the Syrian crisis.

Table 7.10 Correlation matrix for political stability indices

	Voiceacc	Polstab	Govteffec	Reg quality	Ruleoflaw	Corrupt	Transparency
Voiceacc	1.00						
Polstab	0.83	1.00					
Govteffec	0.83	0.72	1.00				
Regulation quality	0.56	0.53	0.72	1.00			
Ruleoflaw	0.91	0.84	0.94	0.63	1.00		
Corruptcont	0.93	0.84	0.91	0.66	0.96	1.00	
TransparencyCPI	0.91	0.76	0.85	0.56	0.91	0.91	1.00

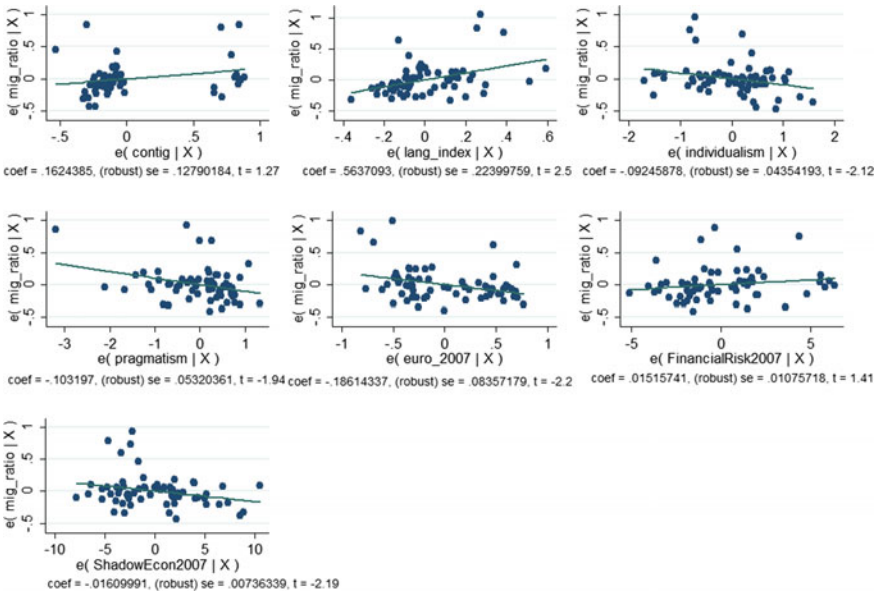


Fig. 7.4 The partial regression plot for all the variables in 2007

We use the component plus residual plot (partial residual plot) to get more clarity on the functional form of the relation between the dependent variable and friction variables 1-by-1. For example to understand the relation between the ratio of bilateral migration in data to TFP driven bilateral migration (y-variable) to language index we first regress y on all the x variables. Then we subtract the effect of all the other regressors (not language index) from the y-variable and plot that on the Y-axis. We call this component plus residual. We compare it with the language index which is plotted on the X-axis in Fig. 7.5.

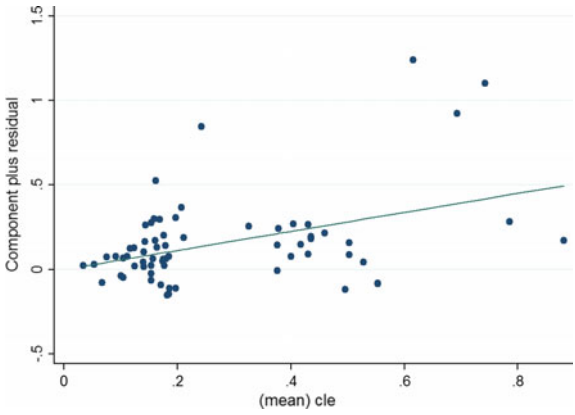


Fig. 7.5 The partial residual plot for language index in 2007

References

1. Adsera, A., Pytlikova, M.: The role of language in shaping international migration. Econ. J. **125**, 49–81 (2015)
2. Belot, M., Ederveen, S.: Cultural barriers in migration between OECD countries. J. Popul. Econ. **25**, 1077–1105 (2012)
3. Bertoli, S., Moraga, J.F., Ortega, F.: Crossing the border: self-selection, earnings and individual migration decisions. J. Dev. Econ. **101**, 75–91 (2013)
4. Braunerhjelm, P., Faini, R., Norman, V., Ruane, F., Seabright, P.: Monitoring European Integration, vol. 10. CEPR, London (2000)
5. Caliendo, L., Parro, F., Rossi-Hansberg, E., Sarte, P.G.: The impact of regional and sectoral productivity changes on the U.S. economy. In: Federal Reserve Bank of Richmond, Working paper, pp. 13–14 (2014)
6. CEPII (2006). http://www.cepii.fr/cepii/en/bdd_modele/bdd.asp
7. Chakrabarti, A.S., Sengupta, A.: Productivity differences and inter-state migration in the U.S.A multilateral gravity approach. Econ. Model. **61**, 156–168 (2017)
8. Ethnologue (2014). https://www.ethnologue.com/
9. Europa, E.U (2014). http://europa.eu/index_en.htm
10. Eurostat (2014). http://ec.europa.eu/eurostat/data/database
11. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations: Software of the Mind. McGraw-Hill, USA, (2010)
12. Kaplan, Z.: The eu's internal market and the free movement of labor: Economic effects and challenges. In: Handbook of Research on Unemployment and Labor Market Sustainability in the Era of Globalization. IGI Global (2017)
13. Kennan, J., Walker, J.R.: The effect of expected income on individual migration decisions. Econometrica **79**(1), 211–251 (2011)
14. OECD (2014). https://data.oecd.org/
15. Penn World Table (2014). http://cid.econ.ucdavis.edu/pwt.html
16. The Maddison-Project (2013). http://www.ggdc.net/maddison/maddison-project/home.htm
17. World Bank Reports (2011). http://data.worldbank.org/data-catalog/wdr2011

Chapter 8
Interdependence, Vulnerability and Contagion in Financial and Economic Networks

Irena Vodenska and Alexander P. Becker

Abstract Financial and economic networks are neither static nor independent of one another, but are rather quite interconnected with a high level of dependence and mutual influence. In light of global economic convergence, countries depend on one another through trade relations, foreign direct investments, flow of funds in international capital markets, bank borrowing and lending operations, or foreign exchange trading. As economic entities and financial markets become increasingly intertwined, a shock in a financial network can provoke significant cascading failures throughout the global economic system. Here we attempt to understand potential sources of future shocks and whether bubbles and systemic risk build-up in financial networks can be anticipated. We review approaches to study global financial markets and bank networks to uncover system characteristics and relationships that might increase the vulnerability of economic networks. The efficiency of regulatory responses motivated by crisis and the proper level of regulation are extremely important for a sound global economic system.

Introduction

The Global Financial and Economic System

We live in an increasingly complex and interconnected world. A major challenge regarding the state of the financial and economic system is its vulnerability. The stability of production, global trade, capital transfer and financial markets depend on proper functioning and reliability of the global economic system. Another aspect

I. Vodenska
Administrative Sciences Department, Boston University Metropolitan College, 1010 Commonwealth Avenue, Boston, MA, USA
e-mail: vodenska@bu.edu

A. P. Becker (✉)
Department of Physics, Boston University, Boston, MA, USA
e-mail: apbecker@bu.edu

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_8

of stability is the role of regulation and policy in light of evolving complexity of financial products such as credit default swaps (CDSs) and mortgage-backed securities (MBOs) or collateralized debt obligations (CDOs) that contributed to the most recent global financial crisis of 2008. Regulations usually react to crises, which often happen after a bubble formation followed by a burst. The future of economic and financial systems depends greatly on the actions taken by the regulators. Why is regulation important? Why can't we just embrace innovation and "enjoy" the ride? Why can't the banks behave as they have been and worry about problems when they arise? There are more questions than answers, and here we attempt to shed light on some underlying conditions that could potentially create instabilities in financial systems. The financial crisis of 2008 was a period when we had more questions than answers. What happened? Why were some banks like Lehman Brothers allowed to fail and others were saved with injection of $700 billion of Troubled Asset Relief Program (TARP) funds to bail out financial institutions and billions of dollars in aid for non-financial institutions in form of low interest loans offered by the U.S. Treasury and the Federal Reserve [1]? Maybe the banks were saved in 2008 to avoid the scenario following the Great Depression of 1929 when the banks were allowed to fail and the crisis deepened during the decade after, with unemployment in the U.S. reaching unprecedented levels of 20%, in a crumbling economy. While President Hoover believed that government should not intervene in the economy and should not be concerned with creating jobs, after President Roosevelt was elected in 1932, he was dedicated to economic reforms during his presidency. Under his watch, the Federal Deposit Insurance Corporation (FDIC) was created to regain the trust of people in the banking system. Another notable institution created by the Securities Exchange Act of 1934 was the Securities and Exchange Commission (SEC) put in charge of regulating financial markets. An important research question to address though is not whether the government should be saving the banks or letting them fail, but rather improving macro-prudential regulation to create a stable financial system that can absorb shocks and survive without much outside intervention. Inspired by financial crises, we study the underlying characteristics of economic and financial systems and the impact of risk propagation throughout the system. We analyze the interdependencies of this complex network of networks including global stock markets, foreign exchange markets, bank networks, real estate, specific bank loans and their characteristics as well as sovereign debt networks. Interdependent networks are shown to be more vulnerable to shocks with rapid damage propagation throughout the networks, compared to single or isolated networks [2, 3]. While financial systems may function seamlessly in calm economic periods, at an onset of crisis, the network dynamics changes drastically, often resulting in severe crippling of the entire system.

Financial crises can cause harm not only locally, but can also spread to other countries in a region like in the case of the 1997 Asian financial crisis or more globally like the 2008 subprime mortgage crisis. Similarly, crises that start in one sector of the economy can spread to other sectors, i.e., a banking crisis can significantly affect the main economy, stalling production, which in turn could interrupt trade and affect negatively economic outputs and national incomes. The economic system is composed of various institutions, individuals, corporations, and markets that act as

agents with different roles and objectives. Stock markets are commonplace where investors obtain and divest specific investments, reflecting the market values of such investments. The stock market is an important leading economic indicator that can serve as an early warning signal to a possible economic downturn and it is widely followed by market participants. The foreign exchange market is the largest financial market with a daily turnover of over $5 trillion, and it is active 24 h a day except for weekends [4]. Needless to say that currency trading and exchange rates, depreciations and appreciations among world currencies are important indicators of countries' stabilities, economic changes, living standards, purchasing power, and global trade. Their importance in financing production and growth in the global economy puts banks naturally in the center of the economic network of networks. Here we survey literature that analyzes risk propagation in the banking system based on shared asset portfolios and studies the effect of loss propagation throughout the network.

The rest of this paper is organized as follows: in section "Interconnectedness Across Scales" we review literature that studies the interconnectedness of financial markets, focusing on community structure and lead-lag relationships in stock markets and the foreign exchange market. Section "Systemic Risk and Shock Propagation" summarizes literature regarding systemic risk and shock propagation in financial systems. In section "Conclusion" we offer our conclusions and an outlook towards open problems in network approaches to financial stability.

Interconnectedness Across Scales

An increase in global trade and technological advances have brought the world economies closer together, and in the last decades they have become more integrated than ever before. Many manufacturers rely on global supply chains, corporations and financial institutions act across national borders, and global trade accounted for almost 30 percent of the world GDP, as of 2016 [5]. All this has lead to a greater interconnectedness on many scales. Advances in technology have not only allowed for greater specialization in production, they have also facilitated trading in overseas financial markets in real time.

Today trade and business relations link and intertwine national economies and financial markets. Developments in one part of the world have the potential to influence other parts through a variety of channels. Since commerce and trade are facilitated through the exchange of currency, understanding the dynamics of the foreign exchange market is essential to monitoring developments in the global economy. In combination with national equity markets, the foreign exchange market forms a complex economic network which is subject to a multitude of outside influences. They include but are not limited to economic and political shocks as well as shocks arising from long-term shifts in macroeconomic trends (e.g. inflation or unemployment).

With the increase in global trade, the financial sector has grown significantly in the last decades as well. In the United States, for example, it makes up about 20 percent of GDP [6]. The interdependence of banks and other financial institutions has also

become larger and comes in many forms. Banks may directly interact with each other through unsecured interbank lending, their portfolios may show significant overlap due to similar strategies, securitization, or regulation. Similarly, they may be exposed to the same business partners and clients whose performance influences their ability to repay loans.

Correlation and Communities of Global Financial Markets

The increase of global trade and business as well as technological advances have lead to stronger interaction between financial markets, such as stock markets and foreign exchange markets. Through trade relations, currency policies, financial contracts, and cross-country investments, a financial crisis in one place has the potential to spill over to other countries. Such spillover is not a recent phenomenon. The Russian bond crisis affected investors worldwide and significantly affected American markets, leading to the collapse of Long-Term Capital Management in 1998 [7].

Financial contagion is reminiscent of disease transmission. Just like individuals can be considered to be either in a state of illness or health, financial markets can generally be classified to be in a crisis state or a non-crisis state. While links between individuals in disease transmission usually describe some form of contact or interaction that would allow for the infection to spread, links in a financial network are typically approximated through the similarity of characteristics among the financial entities. Stock markets serve as proxies for the health and robustness of the underlying economies, and foreign exchange markets encode macroeconomic fundamentals like GDP growth, inflation rate and unemployment. Therefore, studying the network structure of global financial markets becomes critical to anticipate how a shock may propagate through the system. It is generally assumed that a high degree of correlation suggests a larger likelihood of crisis spreading from one node to another. In particular, tightly knit communities are more likely to react to a given shock in a similar way, and therefore identifying these communities may help to anticipate the dynamics of the shock throughout the network.

Community Analysis of Financial Markets

The relationships of financial markets is typically quantified using Pearson's correlation coefficient ρ. Correlations change over time, and financial markets tend to cluster into distinguishable groups which can be attributed to a variety of reasons. For example, while the stock markets in Asia trade simultaneously, there is no overlap in their trading hours with American markets. Some economies may be closely connected with each other because of geographic proximity which invites strong trade and business relationships. In financial networks based on correlation measures, a community structure emerges. Changes in correlations change over time translate to changes in the community structure of financial networks. It has been well stud-

ied that correlations among individual stocks as well as equity markets in different countries generally increase in times of crisis [8, 9].

These findings are confirmed in [10] for equity markets where the non-crisis period 2002–2006 is compared with the crisis period 2007–2012. However, the foreign exchange market exhibits the opposite behavior. In the calm period the correlation among currencies is larger than during financial turmoil. Likewise, the correlation between equity and foreign exchange decreases with the onset of the global financial crisis of 2008.

Authors in [10] study the community structure differences between the crisis and the non-crisis period for a multi-layer network comprised of equity markets and the foreign exchange markets. The change in clustering of equity markets can be seen Fig. 8.1. Most strikingly, the troubled eurozone countries Italy, Spain, Greece and Portugal form their own cluster during the crisis period.

When using correlation to infer links in financial networks, a typical challenge is filtering noise. Due to spurious relations and the finiteness of the data, even non-related markets may exhibit a correlation that is different from 0. In [10], this problem is addressed by constructing a planar maximally filtered graph (PMFG), introduced in [11]. Whereas the minimum spanning tree connects all nodes with the smallest number of links possible, the PMFG preserves more relational information while maintaining the planarity of the graph. The algorithm starts by sorting the empirical correlations from largest to smallest. Then, one by one, all nodes are connected until adding another link would make the graph non-planar.

Alternatively extracting the essential information can be achieved through a combination of principal component analysis (PCA) and random matrix theory (RMT). Originally introduced to physics to study the spectra of nuclei, physicists have successfully used random matrix theory to analyze correlation matrices in financial markets [12, 13]. A correlation matrix of size $N \times N$ derived from uncorrelated time series of length L has non-zero entries even if $N \to \infty$ and $L \to \infty$ (as long as $N/L > 1$ and constant). The eigenvalue distribution of such a random matrix is analytically tractable. This approach can be extended to finite time series and allows for the comparison between the eigenvalue spectrum of empirical data and of a comparable random matrix. In particular, the analytical solution defines a cutoff above which eigenvalues and their corresponding corresponding eigenvalues cannot be randomness alone. For larger data sets, this RMT procedure can be computationally expensive. Instead of classical PCA and RMT, in [14] the authors use the complex Hilbert principal component analysis (CHPCA), which is able to better explain the variance in the data as well as identify lead-lag relationships. Additionally CHPCA is less computationally costly and therefore better suited for large data sets.

Using data from equity markets as well as foreign exchange markets from 1999 to 2012, six significant eigenvectors emerge. The method is, for example, able to single out the Icelandic banking crisis and its effects on both the krona and the Icelandic stock market, as the corresponding eigenvalue peaks during this time, as shown Fig. 8.2.

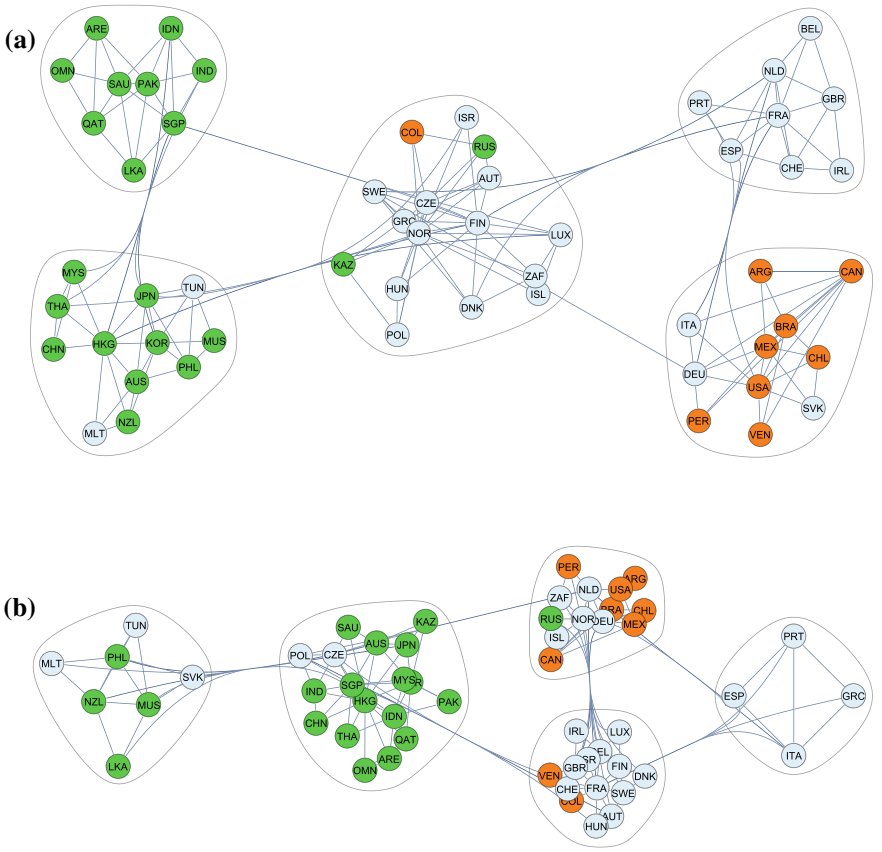


Fig. 8.1 Planar maximally filtered graph (PMFG) for the stock markets during **a** the economically calm period from 2002 to 2006 and **b** crisis period from 2007 to 2012. The countries are denoted by their three-letter symbols and are color-coded according to their geographical locations: green for Asia, light blue for Europe and orange for the Americas. During the calm period, five large clusters appear which are mostly geographically divided. The clusters significantly change during the crisis period. Most notably the troubled eurozone countries, Italy, Spain, Greece and Portugal, form their own cluster. The Asian countries form one larger cluster centered around Hong Kong and Singapore, the major financial centers of Southeast Asia

Lead-Lag Relationships in Financial Markets

Daily returns are calculated by comparing the closing prices of two consecutive trading days. Since foreign exchange markets trade around the clock, daily returns for any currency pair relate to the same instance of time. Opening and closing times of equity markets, however, are different due to various timezones. Although Asian markets are the first to open on any given date, this does not make them leading markets. Instead US markets set the pace for other financial markets, driven by the reach and influence of American corporations, the importance of the US dollar for

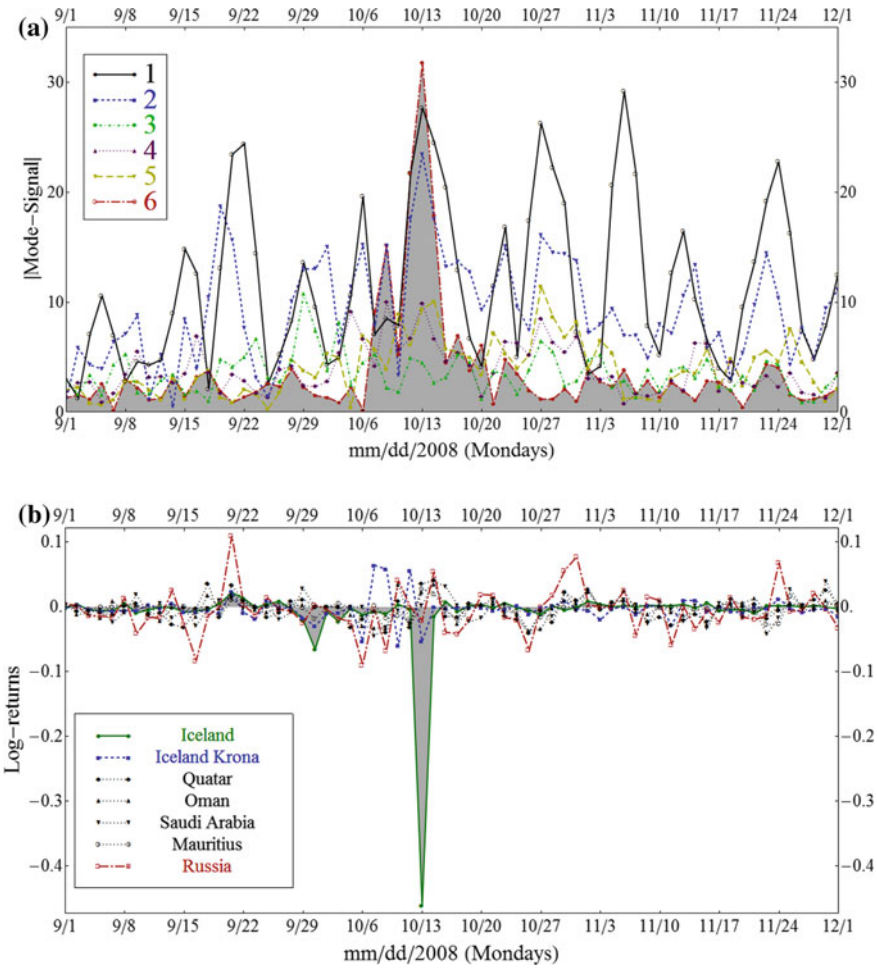


Fig. 8.2 The correlation matrix of international financial markets yields six significant eigenvectors, corresponding to different market modes. The sixth eigenvector is driven by Iceland as well as the Russian and Polish stock market and negatively correlated to stock markets of some major oil exporters from the Middle East. Part (a) shows a big spike in the signal of this eigenvector in October 2013 which stems from the Icelandic financial collapse, illustrated in part (b)

global trade and the size of the market. Measured by total market capitalization of stocks traded, the New York Stock Exchange (NYSE) and the NASDAQ are the two largest exchanges in the world. Both [10, 14] study significant lead-lag relationships among global financial markets, highlighting, for example, that the US and the German markets are predictive of the performance of other markets. In [14], the authors further conclude that "general currency appreciations lead or contribute to positive equity market returns". Similarly strong depreciations correspond to depressed stock markets in times of severe crisis, particularly in emerging economies. Such a down-

turn often follows a rapid appreciation of the currencies of emerging economies as a result of large capital inflow and investments. Examples of this are the Argentinian crisis in the early 1990s or, a few years later, Mexico [15]. In the aftermath of the European sovereign debt crisis, however, the correlation between the euro and European stock markets has been observed to be strongly negative. These various types of relationship signify that there are more underlying causes for certain financial market and currency dynamics. This calls for identifying a method to study the position of a currency within the foreign exchange network.

Monetary Policy and the Foreign Exchange Market

If the underlying currencies for two stock markets are different, this needs to be taken into consideration when computing the correlation between them. Similarly, to find the correlation between two currencies, both have to be expressed in the same base currency. However, the base currency introduces a bias in the correlation since many currency pairs exhibit idiosyncratic behavior. For example, the Swiss franc is generally more correlated to the euro than to the US dollar. Choosing the euro as base currency therefore yields different results for the correlation of the Swiss franc with other currencies compared to choosing the US dollar as base currency. Additionally, the currency which is chosen as the base currency has to be omitted from the correlation analysis. This limits the ability to interpret and study the foreign exchange market as a network.

One approach to mitigate this issue, as employed in [10, 14], is to use the International Monetary Fund's Special Drawing Rights (SDR). The SDR is a basket of currencies, weighted according to international trade share and foreign exchange reserves. Currently five currencies are included in this basket, the US dollar, the euro, the British pound and the Japanese yen as well as the Chinese yuan as the most recent addition. Quoting in SDR allows for inclusion of all currencies of interest as opposed to selecting one currency as the base currency, hence excluding it from the analysis. However, since the SDR is a linear combination of a subset of the pool of analyzed currencies, it still introduces some bias.

Foreign exchange markets are strongly influenced by monetary policy. The foreign exchange market is a zero-sum game where the appreciation of one currency corresponds to the depreciation of another. One example is the international Fisher effect, which links the interest rate differentials of two countries to the expected exchange rates of their respective currencies. Currency correlations can also change significantly because of central bank interventions, and the action of one central bank may propagate through the network of currencies. Using this perspective that an impetus to one currency will reverberate through the system by appreciations and depreciations of other currencies, [16] develops a new methodology to address this challenge. The foreign exchange market consists of triads: if currency A appreciates, that is, rises, against currency B, and currency B appreciates against currency C, it has to follow that currency A appreciates against C and at a larger rate than B. This allows a ranking of the currencies: first A, then B, and then C. In [16] the ranking is

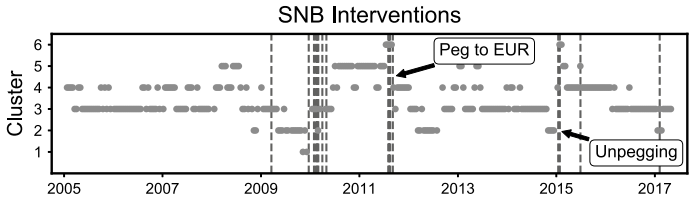


Fig. 8.3 In blue the different roles which the Swiss franc has played within the foreign exchange market are shown. Cluster 1 corresponds to a currency with very stable behavior like a reference currency. Cluster 6 corresponds to a currency with extremely wild swings and volatile behavior. Following the financial crisis, the Swiss franc became very volatile due to a large amount of appreciation until the Swiss National Bank intervened (indicated by red lines) and started enforcing a cap with respect to the euro. Once the cap was lifted in a surprise move, the wild swings resumed for a while until the Swiss franc returned to its pre-crisis behavior

extended to the full market as the "symbolic performance". The highest ranked currency at any given time is the currency that appreciates against all other currencies, while lowest lowest ranked currency is the currency that depreciates in relation to all other currencies. Clustering algorithms allow to classify currencies according to their ranking distributions, giving insight into the roles they play within the market. The euro and the US dollar are reference currencies exhibiting a low degree of variance in their swings, while currencies of countries which rely heavily on commodity exports prove to be more volatile. The paper shows that central bank interventions and sudden changes in monetary policy may disrupt these roles. Case in point: the Swiss franc and its pegging to the euro between 2011 and 2015, as Fig. 8.3 illustrates.

Shared Portfolios and Lending Relationships

Banks are public institutions, traded on financial markets, and their trading prices reflect the health of their financial statements. In a sense the stock price is therefore a bird's eye view of the balance sheet consisting of investments, loans, real assets, liabilities and equity. If financial institutions fail, this is likely due to the deterioration of value of specific balance sheet items. This is why it is important to study in detail specific assets and liabilities to understand systemic risk. In addition to being able to quantify specific balance sheet items, proper modeling is required to assess the vulnerabilities of networks of financial institutions. Financial markets show high degrees of correlation during a crisis, and while many mechanisms contribute to this interconnectedness, in the following we explore the impact of portfolio overlap between financial institutions as well as their lending relationships.

As [17] points out, counter-party risk, roll-over risk, and common asset holdings are the three major contagion channels of financial stress. Exposure to the same asset classes or derivatives written on these asset classes could contribute to a widespread effect of a financial crisis, such as the subprime mortgage crisis of 2008 starting in the

US and spreading around the globe. A large overlap in asset portfolios of financial institutions poses two major risks. One, such an overlap induces a correlation in the performance of the banks' portfolios, and two, financial institutions may have to react similarly to market developments, potentially putting more pressure on a distressed asset. Consider the subprime mortgage crisis when many financial institutions were exposed to mortgage backed securities (MBSs) and collateralized debt obligations (CDOs). MBSs are derivate instruments deriving their value from the underlying pool of mortgages. If homeowners are unable to make mortgage payments, the corresponding MBS is adversely effected. Collateralized debt obligations are derivate instruments with underlying fixed income securities whose value directly affects the CDO. Due to inaccurate credit ratings the likelihood of default of these instruments was underestimated by investors. As market conditions deteriorated, many institutions found themselves over-leveraged and over-exposed to toxic assets, that is, assets whose risk was significantly misrepresented. This resulted in fire sales and further downward pressure on the value of these troubled assets.

To understand the overlapping portfolio of the banks, [17] represents the holdings of financial institutions by a bipartite network of banks in one layer and assets in the other as a stylized model of a financial system. The authors show that diversification within a finite set of assets may lead to instability and global cascades. We discuss such models of shock propagation and their implications for systemic risk in the following section.

Using a bipartite network comprised of banks and assets as well, [18] empirically studies the structural changes in a financial network, namely the network of credit relationships in Japan surrounding the banking crisis following a collapse in asset prices in the early 1990s. Analyzing the distribution of the weights of the links in the network, differences in bank portfolios emerge which could be early warning signals to the failure of specific banks in the late 1990s.

Systemic Risk and Shock Propagation

Research of financial networks following the global financial crisis of 2008 and the European sovereign debt crisis of 2011 has focused on systemic risk, that is, the risk of a system to collapse in its entirety provoked by a shock to a part of the system, possibly as small as just one entity or institution. Approaches to the regulation of financial networks can be categorized as micro-prudential and macro-prudential. While micro-prudential regulation focuses of the health of individual banks and their balance sheets, macro-prudential regulation aims to provide stability for the banking system as a whole. A major distinction between the two approaches is the emphasis on the importance of correlations and connectedness within the financial system. While both micro-prudential and macro-prudential regulation consider the health of banks as an important factor for financial stability, micro-prudential regulation looks at the banks as isolated entities, while macro-prudential regulation incorporates the connectivity among the banks as a source of systemic risk.

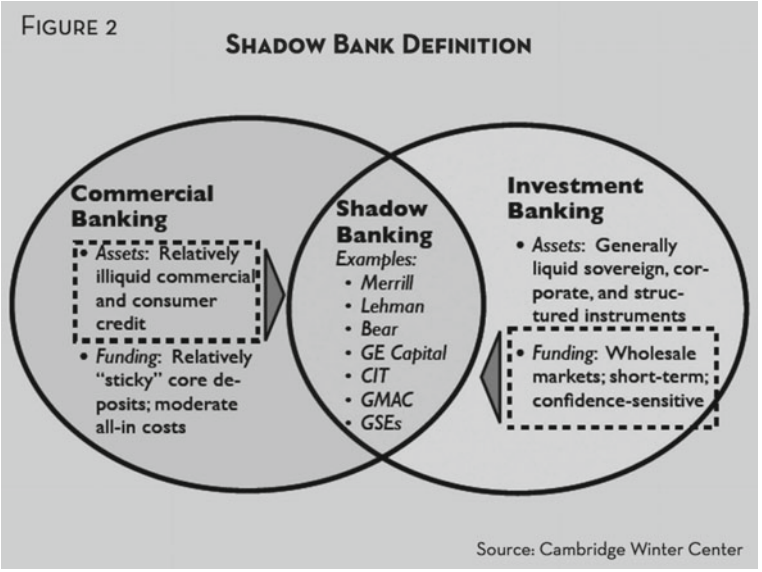


Fig. 8.4 The banking system is comprised of commercial banks whose main business is consumer-oriented and investment banks whose main business is investing and supporting its client with corporate actions. Shadow banking describes the overlap where one institutions is engaged in both, potentially with conflicting interests for their customers

The need for a macro-prudential approach became apparent following the global financial crisis of 2008 which highlighted a high degree of interdependence among commercial banks, investment banks and shadow banking institutions (see Fig. 8.4).

Inspired by [19], researchers and policy makers have debated the network topology and conduits of shock propagation and their consequences for systemic risk. As [20] summarizes,

> Two polar views on this relationship between the structure of the financial network and systemic risk] have been suggested in the academic literature and the policy world. The first maintains that the "incompleteness" of the financial network can be a source of instability, as individual banks are overly exposed to the liabilities of a handful of financial institutions [...] The second view, in stark contrast, hypothesizes that it is the highly inter- connected nature of the financial system that contributes to its fragility, as it facilitates the spread of financial distress and solvency problems from one bank to the rest in an epidemic-like fashion.

In [21] the authors argue that it is "difficult to generate contagion solely through spillover losses". They further find that there exist bounds to contagion and loss amplification which are independent of the network structure. In other words, [21] shows that knowing the size of the initial shock is enough to estimate the worst outcome for the financial system, regardless of degree distribution and the location of the initial shock. It is important to point out that the network structure does matter when including other effects like bankruptcy cost or loss of trust. While [21] derives the analytical results within the framework of payment clearing vectors by [19], more

realistic models have to take into account mark-to-market losses. In the following section we discuss cascading failure models and fire sale mechanics which include mark-to-market losses and the loss of trust.

Cascading Failure Models

How do losses spread across a financial system? Let's consider a bipartite network in which we have banks in one layer and assets in the other layer. A connection between a bank and an asset exists if the bank has that asset on its books, that is, the bank is exposed to the asset. Using balance sheets of US commercial banks in 2007, [22] proposes a cascading failure model describing how losses spread through the financial system in case of a crisis. The balance sheets of US commercial banks, among others, includes exposures from residential and commercial real estate loans as well as loans for land development and agriculture. If debtors fail to deliver the agreed upon payments, they are in default, and the bank takes a loss on this loan. From a network perspective, a node in the asset layer is distressed, and it transmits this shock via its links to banks. If banks are highly leveraged or overexposed to a given asset or group of assets, they are at risk of failing. This occurs when their overall assets, after taking the initial loss, fall below a threshold, their liabilities. The model then marks down all assets that the failed bank owned, as their value deteriorates. This leads to another round of asset depreciation and potential bank failures, as Fig. 8.5 shows. Empirically testing this model and comparing it to actual defaults in the global financial crisis of 2008, [22] concludes that this model is suitable for systemic (macro-prudential) stress-tests of the banking system. Similarly [18] aims to devise an early warning system for financial institutions, showing the structural differences of bank holdings in Japan as the asset bubble burst in the early 1990s.

Fire Sale Mechanics

Many cascading failure models (e.g. [17, 18, 22, 23]) include a liquidity parameter or a market impact function. This specification allows to account for market conditions when a bank has to re-balance their portfolio, modeling the mark-to-market losses it incurs. This does not model, however, the behavior of banks in response to a shock. Since banks are subject to constraints on their leverage or their capital to risk-weighted assets ratio, the so-called tier 1 capital ratio, they have to act when a significant part of their portfolio is in distress. One way for banks to de-leverage or to reduce their tier 1 capital ratio is to sell loans or bonds which they have on their books. This negatively effects the price and thus the value of these loans or bonds. In [24] the liquidity parameter on the asset side is complemented by a bank response function on the bank side, which models the selling behavior. Any disposal of assets by the bank leads to a value depreciation in this simple model; however, the

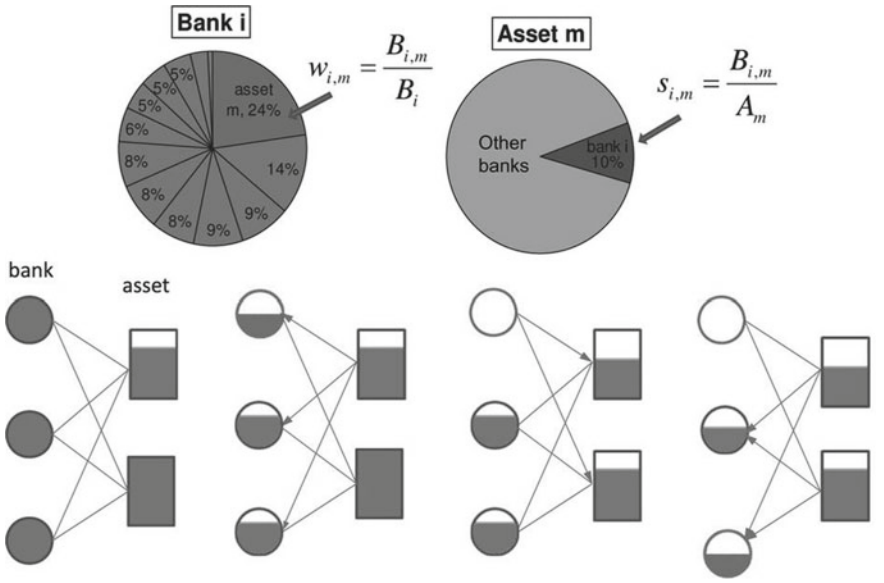


Fig. 8.5 Cascading failure model as proposed by [22]. A shock to an asset may lead to defaults on the bank level which in puts further stress on the market, propagating the crisis

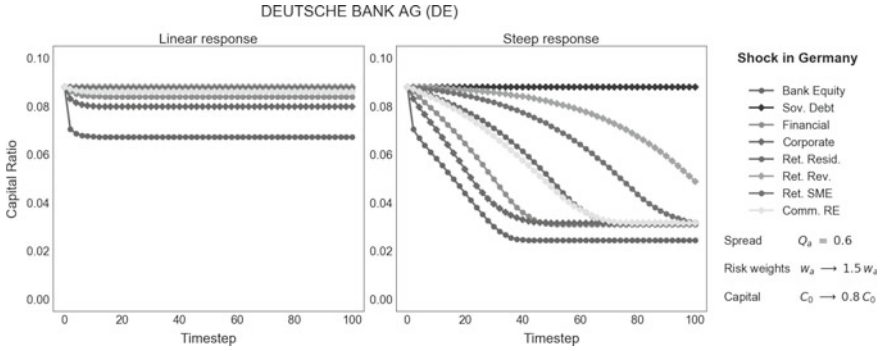


Fig. 8.6 The impact of fire sale dynamics on one German bank after a shock to different asset classes in its home country, such as sovereign debt, corporate loans or residential real estate. The shock corresponds to an increase of 50 percent in risk weights which banks use to determine their capital ratios and moderate liquidity for the assets. The left panel shows the scenario in which banks are risk neutral (linear response), and the right panel the scenario in which they are risk-averse (steep response). In the risk-neutral case the initial shock quickly dissipates and does not spread. In the risk-averse case, however, a cascade of fire sales severely impacts the financial system

combination of liquidity parameter and bank response function determines whether an initial shock can affect the entire system.

Figure 8.6 shows that even if the initial shock is large and the liquidity of the affected assets is limited, a spillover does not occur when the banks react in a risk-

neutral fashion. However, when banks are averse to losses, as is arguably more realistic, initial shocks are easily enhanced through the interconnectedness of the network. Therefore [24] suggests that rational behavior and trust within the banking system are necessary to maintain stability in times of crisis, making a point for regulatory or government intervention when liquidity and/or trust erode.

Conclusion

Severe financial crises usually inflict a high cost on global societies. The development of the European sovereign debt crisis, for example, is linked to the global financial crisis of 2008 which started as a subprime mortgage crisis in the United States. Wrong estimates of credit risk, excessive leverage of financial institutions and inadequate regulation allowed the US housing bubble to grow to such extent that it threatened the future of the global financial system when it burst. "Too big too fail" quickly became "too interconnected to fail", and in order to avoid a reprise of the Great Depression of 1929, governments stepped in to save the banks. Some European governments, being too indebted themselves, required bailout assistance and had to implement strong austerity measures. Therefore, either through bail-out efforts for banks or due to excessive sovereign debt, tax payers had to, and still have to foot the bill for saving the banks and supporting governments.

The consequences of these financial crises have inspired a great research effort to understand the roots of systemic risk and to build more robust economic networks.

The economy is a network of networks, and as such it is much more fragile than any single network. Through community analysis or the study of lead-lag relationships of financial markets, researchers are able to extract network characteristics and infer indicators of potential contagion channels of distress. The dynamics of currency relations, for instance, are affected by many factors such as macroeconomic indicators, market speculation, government intervention and political shocks like the Brexit vote. On the other hand, bipartite network of banks and assets can offer insights into fire sale dynamics, where such fire sales depend on market condition (e.g. asset liquidity), the behavior of banks (e.g. their regulatory constraints and risk-aversion), or the reaction of regulators (e.g. providing liquidity in a credit crunch). The literature shows the large risks that financial institutions may face through contagion and spillover effects due to their investment and loan portfolios. This, however, forms only one layer of a multiplex network of banks, in which other layers contain bank exposures to foreign exchange, stock markets, or derivative securities.

In summary, three lessons emerge: one, the global economy is strongly interdependent and can be appropriately represented as multiplex network; two, tools from network science enable researchers to uncover the multiplex network structure; and three, network-based spreading and propagation models can be used to study distress dynamics in financial and economic networks. Understanding such multiplex networks will be crucial to improving regulation. While considerable progress has

been made to investigate the sources and spreading of systemic risk, building a better regulatory framework for financial networks is far from being done.

References

1. Piketty, T.: Why Save the Bankers?: And Other Essays on Our Economic and Political Crisis. Houghton Mifflin Harcourt, Boston (2016)
2. Buldyrev, S.V., Shere, N.W., Cwilich, G.A.: Interdependent networks with identical degrees of mutually dependent nodes. Phys. Rev. E **83**(1), 016112 (2011). https://doi.org/10.1103/PhysRevE.83.016112
3. Gao, J., Buldyrev, S.V., Stanley, H.E., Havlin, S.: Networks formed from interdependent networks. Nat. Phys. **8**(1), 40–48 (2011). https://doi.org/10.1038/nphys2180
4. Bank for international settlements: triennial central bank survey of foreign exchange turnover in 2013 - preliminary results released by the BIS (2013). http://www.bis.org/press/p130905.htm(accessedon10April2016)
5. WTO: World Trade Statistical Review (Bernan Distribution / World Trade Organization, 2017), p. 109 (2017)
6. Federal reserve bank of St. Louis: FRED, gross domestic product by industry. https://fred.stlouisfed.org/release/tables?rid=331&eid=211.Cited29April2018
7. Lowenstein, R.: When Genius Failed: The Rise and Fall of Long-Term Capital Management. Random House, New York (2000)
8. Sandoval Jr., L., Franca, I.D.: Correlation of financial markets in times of crisis. Physica A **391**(1–2), 187–208 (2012). https://doi.org/10.1016/j.physa.2011.07.023
9. Silvennoinen, A., Thorp, S.: Financialization, crisis and commodity correlation dynamics. J. Int. Financ. Mark. Inst. Money **24**, 45–65 (2013). https://doi.org/10.1016/j.intfin.2012.11.007
10. Vodenska, I., Becker, A.P., Zhou, D., Kenett, D.Y., Stanley, H.E., Havlin, S.: Community analysis of global financial markets. Risks **4**(2), 13 (2016). https://doi.org/10.3390/risks4020013
11. Tumminello, M., Aste, T., Di Matteo, T., Mantegna, R.N.: A tool for filtering information in complex systems. Proc. Natl. Acad. Sci. **102**(30), 10421–10426 (2005). https://doi.org/10.1073/pnas.0500298102
12. Laloux, L., Cizeau, P., Bouchaud, J.P., Potters, M.: Noise dressing of financial correlation matrices. Phys. Rev. Lett. **83**(7), 1467 (1999)
13. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T., Stanley, H.E.: Random matrix approach to cross correlations in financial data. Phys. Rev. E **65**(6), 066126 (2002). https://doi.org/10.1103/PhysRevE.65.066126
14. Vodenska, I., Aoyama, H., Fujiwara, Y., Iyetomi, H., Arai, Y.: Interdependencies and causalities in coupled financial networks. PLOS ONE (2016). https://doi.org/10.1371/journal.pone.0150994
15. Sachs, J., Tornell, A., Velasco, A.: Financial crises in emerging markets: the lessons from 1995. NBER working Paper No. 5576 (1996). https://doi.org/10.3386/w5576
16. Wollschläger, M., Becker, A.P., Vodenska, I., Stanley, H.E., Schäfer, R.: Economic and political effects on currency clustering dynamics. Quantitative Finance (under review) (2018)
17. Caccioli, F., Shrestha, M., Moore, C., Farmer, J.D.: Stability analysis of financial contagion due to overlapping portfolios. J. Bank. Financ. **46**, 233–245 (2014). https://doi.org/10.1016/j.jbankfin.2014.05.021
18. Sakamoto, Y., Vodenska, I.: Systemic risk and structural changes in a bipartite bank network: a new perspective on the Japanese banking crisis of the 1990s. J. Complex Netw. **5**(2), 315–333 (2017)
19. Eisenberg, L., Noe, T.H.: Systemic risk in financial systems. Manag. Sci. **47**(2), 236–249 (2001)
20. Acemoglu, D., Ozdaglar, A., Tahbaz-Salehi, A.: Systemic risk and stability in financial networks. Am. Econ. Rev. **105**(2), 564–608 (2015)

21. Glasserman, P., Young, H.P.: How likely is contagion in financial networks? J. Bank. Financ. **50**, 383–399 (2015)
22. Huang, Z., Vodenska, I., Havlin, S., Stanley, H.E.: Cascading failures in Bi-partite graphs: model for systemic risk propagation. Sci. Rep. **3**, 1219 (2013). https://doi.org/10.1038/srep01219
23. Roukny, T., Bersini, H., Pirotte, H., Caldarelli, G., Battiston, S.: Default cascades in complex networks: topology and systemic risk. Sci. Rep. **3**, 2759 (2013). https://doi.org/10.1038/srep02759
24. Vodenska, I., Aoyama, H., Becker, A.P., Fujiwara, Y., Iyetomi, H., Lungu, E.: Network approach to understanding the fragility of financial systems. Journal of Financial Stability (under review) (2018)

Chapter 9
Multi-layered Network Structure: Relationship Between Financial and Macroeconomic Dynamics

Kiran Sharma, Anindya S. Chakrabarti and Anirban Chakraborti

Abstract We demonstrate using multi-layered networks, the existence of an empirical linkage between the dynamics of the financial network constructed from the market indices and the macroeconomic networks constructed from macroeconomic variables such as trade, foreign direct investments, etc. for several countries across the globe. The temporal scales of the dynamics of the financial variables and the macroeconomic fundamentals are very different, which make the empirical linkage even more interesting and significant. Also, we find that there exist in the respective networks, core-periphery structures (determined through centrality measures) that are composed of similar set of countries—a result that may be related through the 'gravity model' of the country-level macroeconomic networks. Thus, from a multi-lateral openness perspective, we elucidate that for individual countries, larger trade connectivity is positively associated with higher financial return correlations. Furthermore, we show that the Economic Complexity Index and the equity markets have a positive relationship among themselves, as is the case for Gross Domestic Product. The data science methodology using network theory, coupled with standard econometric techniques constitute a new approach to studying multi-level economic phenomena in a comprehensive manner.

K. Sharma · A. Chakraborti (✉)
School of Computational and Integrative Sciences,
Jawaharlal Nehru University, New Delhi 110067, India
e-mail: anirban@jnu.ac.in

K. Sharma
e-mail: kiransharma1187@gmail.com

A. S. Chakrabarti
Economics Area, Indian Institute of Management, Ahmedabad, India
e-mail: anindyac@iima.ac.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_9

Introduction

Financial networks are major vehicles for transmitting shocks across different economic entities, which lead to complex dynamics. A well known phenomenon is that financial variables are considerably more volatile than macroeconomic variables with a much higher frequency, whereas, macroeconomic variables tend to show a much slower dynamics. A simple inspection of data suggests, there is wide variation even in intra-day stock returns, whereas macroeconomic variables move by a perceptible magnitude only over quarters or years if not longer time horizon. Thus these two types of variables differ both in frequency as well as the magnitude of oscillation. A directly related observation is that the magnitude of fluctuations of the financial variables often seems decoupled from the fluctuations in the underlying macroeconomic variables. This is formally known as the excess volatility puzzle. In an aggregate sense, growth rates of macroeconomic entities like firm-size variables shows bi-exponential distributions [1]. But the corresponding financial indices typically have a power law structure which indicates much wider dispersion than exponential distributions. Thus, although the financial indices should reflect movements in underlying macroeconomic factors, it seems unlikely that the dynamics of individual financial time series can be readily explained by the dynamics of the corresponding economic variable.

In this chapter, we follow a complementary approach. In the finance literature, researchers have focused on factor models to relate economic variables to financial ones. We propose in the following that rather than looking at the time-series properties, a more useful approach could be to analyze the cross-sectional variation in the return structure and to find if there is any macro variable that explains the variation. In particular, we posit that the aggregate financial network across countries are in sync with the dynamics of underlying macroeconomic fundamentals. The main idea stems from the work of Sharma et al. [2], which showed that at the sectoral level, there is a one-to-one mapping between the economic size of the sectors and centrality in the corresponding financial network. We extrapolate that idea to the country level. The novelty of the present approach lies in two factors. First, the earlier paper considered economic size of sectors measured by three indices (total market capitalization, revenue and employment) to be the underlying factors. Here, at the country level we extend the analysis by constructing country-to-country macroeconomic networks which underlie the financial network. In particular, we analyze the foreign direct investment network and trade network. Thus, it allows us to actually create a multi-layered network [3, 4] rather than just focusing on the size effect. Second, given that the Gravity equations are good models to understand country-level macroeconomic networks (see e.g. the works [5, 6] among others), we have an explanatory model of the relationship of this multi-layered network through the gravity model. Wang et al. [7] constructed and analyzed a cross-country financial network. They analyzed the topological properties of the network with different clustering algorithms. We differ significantly from their approach with our emphasis on country-level fundamentals and their connections with the financial network. Main problem-wise the

closest work to ours is of Qadan and Yagil [8], who analyzed a very similar problem with econometric techniques. But they did not explicitly consider network topology. Hence, our results complement their findings. Finally, Bookstaber and Kenett [9] constructed a multi-layered map of the financial system and analyzed its topology. Our usage of multi-layered network was motivated by that paper, but our emphasis on the macroeconomic variables provide new and different features of the data.

The main points of this chapter are as follows: First, there is a relationship between centrality measures of financial return correlation network across countries and the same for trade and FDI networks. Second, from a multi-lateral openness perspective, we show that even for individual countries, larger trade connectivity is positively associated with higher financial return correlations. Third, we analyze the network architecture by using different clustering algorithms, which in turn allows us to identify the countries that are at the core or at the periphery.

Macroscopic View

In this section, we study the relationship between financial indices return network, trade and foreign direct investment (FDI) networks as a multiplex network for 18 European countries. Next, we study the world stock market and relationship of macro variables and indicators like economy size, Economic Complexity Index, etc.

Data Description

For the macro-level analysis, we have used the data of the adjusted closing price for 18 European countries downloaded from Thomson Reuters Eikon database [10], within the time period of 2001–2009. The countries are: (1) **AUT**—Austria (2) **BEL**—Belgium (3) **CZE**—Czech Republic (4) **DEU**—Germany (5) **DNK**—Denmark (6) **ESP**—Spain (7) **FRA**—France (8) **GBR**—United Kingdom (9) **HUN**—Hungary (10) **IRL**—Ireland (11) **ITA**—Italy (12) **LVA**—Latvia (13) **NLD**—The Netherlands (14) **POL**—Poland (15) **PRT**—Portugal (16) **ROU**—Romania (17) **SVK**—Slovak Republic and (18) **SWE**—Sweden. Data for Foreign Direct Investment (FDI) and international trade for same 18 European countries is downloaded from *External and intra-EU trade, A statistical yearbook*, 2011 edition published by *eurostat*. To study the evolution of world stock markets, we have used the adjusted closing price of 51 market indices across the globe downloaded from the Thomson Reuters Eikon database, within the time period of 2001–2015. The countries: (1) **USA**—The United States of America (2) **CAN**—Canada (3) **BRA**—Brazil (4) **ARG**—Argentina (5) **MEX**—Mexico (6) **CHL**—Chile (7) **VEN**—Venezuela (8) **PER**—Peru (9) **JPN**—Japan (10) **SGP**—Singapore (11) **CHN**—China (12) **AUS**—Australia (13) **HKG**—Hong Kong (14) **KOR**—Korea (15) **IND**—India (16) **IDN**—Indonesia (17) **MYS**—Malaysia (18) **THA**—Thailand (19) **PHL**—Philippines (20)

PAK—Pakistan (21) **LKA**—Sri Lanka (22) **GBR**—United Kingdom (23) **FRA**—France (24) **ITA**—Italy (25) **ESP**—Spain (26) **RUS**—Russia (27) **NLD**—The Netherlands (28) **CHE**—Switzerland (29) **SWE**—Sweden (30) **POL**—Poland (31) **BEL**—Belgium (32) **NOR**—Norway (33) **AUT**—Austria (34) **DNK**—Denmark (35) **GRC**—Greece (36) **PRT**—Portugal (37) **HUN**—Hungary (38) **IRL**—Ireland (39) **TUR**—Turkey (40) **ROU**—Romania (41) **SVK**—Slovak Republic (42) **HRV**—Croatia (43) **CZE**—Czech Republic (44) **LVA**—Latvia (45) **DEU**—Germany (46) **QAT**—Qatar (47) **SAU**—Saudi Arabia (48) **OMN**—Oman (49) **KWT**—Kuwait (50) **TUN**—Tunisia and (51) **ZAF**—South Africa, spread across the continent of Latin America, Asia, Europe and Africa. Economic complexity data is downloaded from Atlas of Economic Complexity [11] for the period 2001–2015. GDP (per capita by country) data is downloaded from Knoema world data Atlas [12].

Some Basic Measures

We consider aggregate stock market indices $\{P_{it}\}_{i\in N,t\in T}$ for N countries and T periods. We construct the return series by taking simple log differences of the prices levels

$$\{r_{it}\}_{i\in N,t\in T-1} = \log\left(\{P_{it}\}_{i\in N,t\in T}/\{P_{i(t-1)}\}_{i\in N,t\in T}\right). \tag{9.1}$$

Next, we construct the correlation matrix $\rho_{N\times N}$ from the N time-series. We have used eigenvector centrality (we refer to the measure as EVC) to measure centrality of different nodes in a given network. EVC is defined by a vector $e_{N\times 1}$ which solves

$$\lambda e = \rho e, \tag{9.2}$$

where λ is an eigenvalue of the matrix ρ. EVC is defined as the eigenvector e corresponding to

$$\lambda = \max_{\{i\in N\}}\{\lambda_i\}. \tag{9.3}$$

For all variables x, we construct the z-score of the same as

$$x_z = (x - E(x))/\sigma_x. \tag{9.4}$$

The Relationship Between Financial Indices, International Trade, and Foreign Direct Investment

In this section, we try to find the FDI-trade linkage between host and home countries [13, 14], and their effect on financial indices in the form of a multiplex network. Here, we show empirical evidence for 18 European countries (see data description in

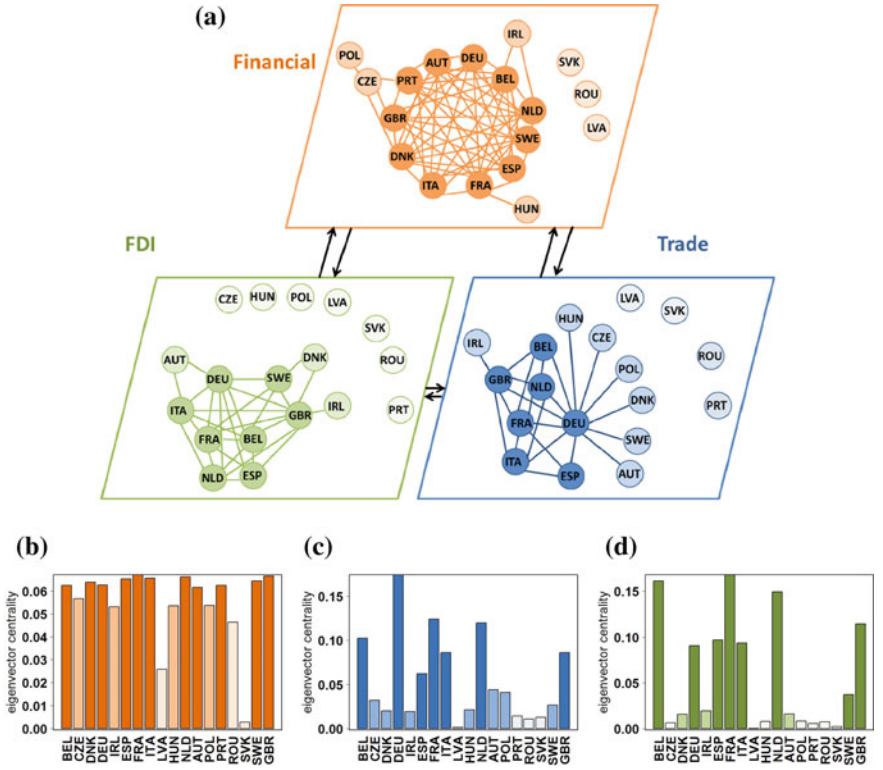


Fig. 9.1 **a** Multiplex network for 18 European countries for year 2008. Financial indices are the top-level network and macroeconomic entities (Trade and FDI) are base-level networks. Eigenvector centrality for **b** Financial indices, **c** Trade and **d** Foreign direct investment (FDI). We divided the EVC in all the networks with three different shades (light to dark)

section "Data Description") whether financial indices and international trade of these nations are substitute or compliments, i.e., whether a great market index held by a nation is associated with decreases or increases of its export and imports. The effect of FDI on trade is always a concern for the policymakers. So we studied the effect of FDI on international trade and financial indices. For this analysis, we have chosen both the developed and developing countries of European continent. The literature on FDI and trade generally points to a positive growth relationship.

In Fig. 9.1, we present a multi-layered network view of the 18 European countries. In panel (a), we construct that the base-level networks formed across countries in terms of trade flow and FDI flow. Both of these two networks capture the connections through economic variables. The top-layer, on the other hand, has been constructed from the financial indices. Here, we examine the relationship between the upper layer of financial network and lower levels of FDI and trade networks. The countries occupying central positions in the correlation network are also central in the corresponding

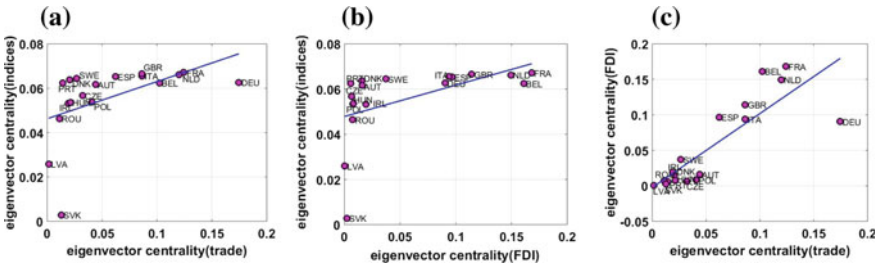


Fig. 9.2 Linear regression of normalized eigenvector centrality between **a** trade and financial indices having $\beta = 0.17 \pm 0.07$ with $p - value = 0.03$, **b** FDI and financial indices having $\beta = 0.15 \pm 0.06$ with $p - value = 0.03$, **c** Trade and FDI having $\beta = 1.04 \pm 0.17$ with $p - value = 0.00001$, of 18 European countries for year 2008

trade and FDI network. In panels (b), (c) and (d), we show the eigenvector centralities of the corresponding countries for these three variables. We cut down the EVC at three levels and that is reflected in the network of financial indices, trade and FDI. In all the networks three countries: SVK, ROU and LVA are forming no link with other countries. PRT is not forming any link in trade and FDI network. CZE, HUN, POL are not forming links in trade network with rest of the countries. Germany is Europe's one of the developed country and strongest economy due to its highly skilled labor force, high quality of life for its resident, etc. as visible in Fig. 9.1a trade network. We computed the eigenvector centrality (normalized) of financial indices, international trade and FDI. Then, we regress these three variables as shown in Fig. 9.2. EVC's of trade and FDI points to a positive growth relationship having $\beta = 1.04 \pm 0.17$ with $p - value = 0.00001$. Germany is an outlier. EVC's of trade and financial indices are also showing positive slope having $\beta = 0.17 \pm 0.07$ with $p - value = 0.03$. Latvia and Slovakia are outliers. EVC's of FDI and financial indices are showing a mildly positive slope having $\beta = 0.15 \pm 0.06$ with $p - value = 0.03$.

To see the co-evolution of trade and financial indices, we regress the EVC's of indices and trade for the period 2003, 2005, 2007 and 2009 as shown in Fig. 9.3. The positive slopes of the best fit line indicates that higher centrality in the financial network is occupying more central positions in the trade network. This pattern holds true for all four time periods, both before and immediately after the financial crisis. Thus, we show that there exists a mapping between the financial network and the trade network. The co-movement of three countries: SVK, LVA and ROU is traced. In the year 2009, ROU came closer to the rest of the countries (as seen in Fig. 9.3h). Germany is always an outlier.

We also conduct a microscopic study of the relation between trade flow and co-evolution of financial indices at the country level. For illustrative purpose, we have chosen two reasonably large European economies viz. Germany (DEU) and France (FRA). In Fig. 9.4, we plot the nominal trade flow as a function of index correlations for pairs of countries, where we fix the origin country. In Fig. 9.4, the left column shows the analysis for Germany (DEU), whereas the right column shows

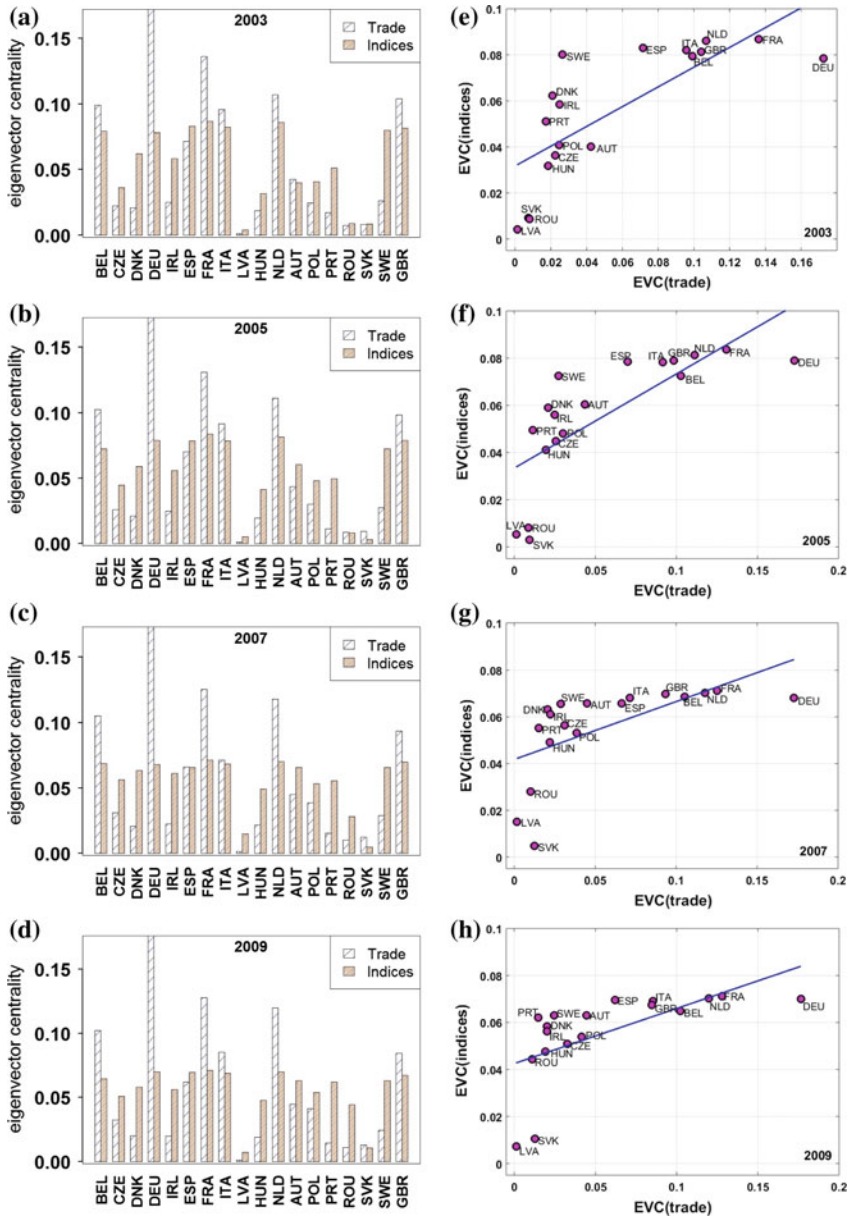


Fig. 9.3 Mapping between the EVC's of the financial network and the trade network for the years: 2003, 2005, 2007, 2009 across 18 European countries. The left panel (**a**, **b**, **c** and **d**) shows bar charts of the normalized EVC's of financial indices and trade. The right panel (**e**, **f**, **g** and **h**) shows the scattered plots of the normalized EVC's of financial indices and trade along with the best fit line having slope, **e** 0.43 ± 0.09 for 2003, **f** 0.40 ± 0.09 for 2005, **g** 0.25 ± 0.08 for 2007 and **h** 0.23 ± 0.07 for 2009. The positive slopes of the best fit line indicate that higher centrality in the financial network is correlated with occupying more central positions in the trade network. SVK, LVA and ROU always evolving together. DEU is an outlier

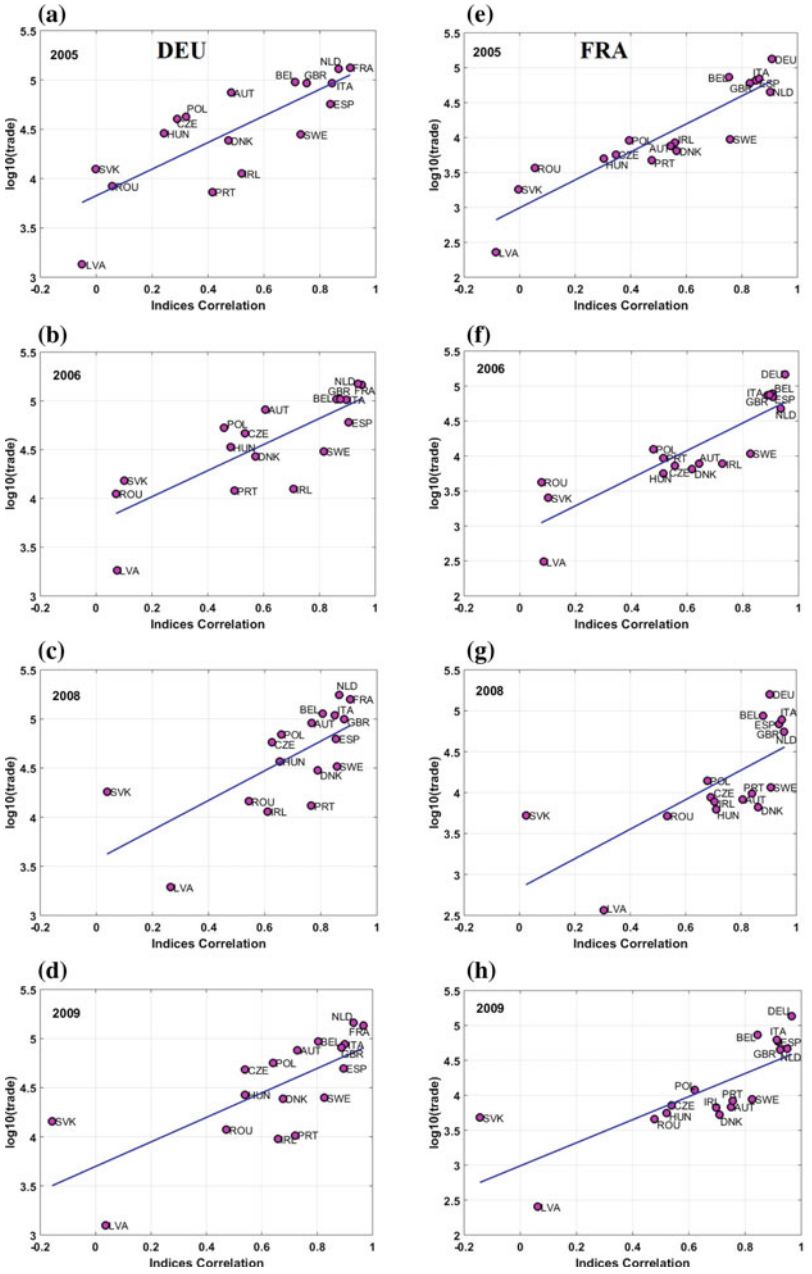


Fig. 9.4 Mapping between financial indices correlation and log10 (trade) for Germany (DEU) and France (FRA), for four snapshots over time, two before the crisis (2005 and 2006) and two into the crisis period (2008 and 2009). For DEU the best fit lines having slopes: **a** 1.35 ± 0.27 for 2005, **b** 1.35 ± 0.26 for 2006, **c** 1.51 ± 0.41 for 2008 and **d** 1.25 ± 0.31 for 2009. For FRA the best fit line having slope, **e** 2.00 ± 0.23 for 2005, **f** 1.97 ± 0.28 for 2006, **g** 1.80 ± 0.47 for 2008 and **h** 1.65 ± 0.35 for 2009

the analysis for France (FRA). We have considered four snapshots over time, two before the crisis (2005 and 2006) and two into the crisis period (2008 and 2009). For DEU the best fit lines having slopes: (a) 1.35 ± 0.27 for 2005, (b) 1.35 ± 0.26 for 2006,(c) 1.51 ± 0.41 for 2008 and (d) 1.25 ± 0.31 for 2009. For FRA the best fit lines having slopes: (e) 2.00 ± 0.23 for 2005, (f) 1.97 ± 0.28 for 2006, (g) 1.80 ± 0.47 for 2008 and (h) 1.65 ± 0.35 for 2009. One interesting feature is that during the crisis period, many countries become much more correlated and hence create a cluster, most notably in the case of Germany in periods 2008–09. However, in all the cases, it seems to be a clear positive correlation between pairwise trade flow and index correlation and this relationship is seemingly robust with respect to the occurrence of the crisis period. Three countries viz. Latvia (LVA), Romania (ROU) and Slovakia (SVK) seems to be far less correlated than the rest of the countries in the sample. However, removal of them does not affect the direction of the relationship.

Mapping Between Economic Complexity Index (ECI) and Financial Indices

To find out the production characteristics of large economies, Economic Complexity Index (ECI) is a holistic measure proposed by Hidalgo and Hausmann [15] in 2009. The goal of this index is to explain an economy as a whole rather than the sum of its parts. To see the mapping between equity and ECI, we regress the normalized EVC's of financial indices and ECI of 51 countries across the globe during 2007–2010, as shown in Fig. 9.5. Equity and ECI are sharing a positive relationship among themselves. Also the evolution of three variables: per capita GDP, ECI and EVC's of financial indices during 2002–2014 is shown in Fig. 9.6.

This finding is not very surprising as there are two fundamental relationships. One, typically larger (and more developed) countries have higher complexity index. Two, there is a strong relationship between return centrality and size (we explore it below in more details). Combining the two, we see that ECI could also have positive correlation with return centrality.

Estimation Results Controlling for Variations Across Countries

All analyses done so far were essentially correlation study without controlling for other country-specific characteristics. Here, we present a sequence of regression tables done across years with control variables in place (2001–09; see Tables: 1–9 in arXiv:1805.06829). We have used foreign direct investment, total credit as a percentage of GDP, trade openness (total trade/GDP), size variables (GDP and GDP per capita) as control variables. As can be seen, the relationship is not robust to inclusion of aggregate size (i.e. GDP). We have discussed this issue below in details.

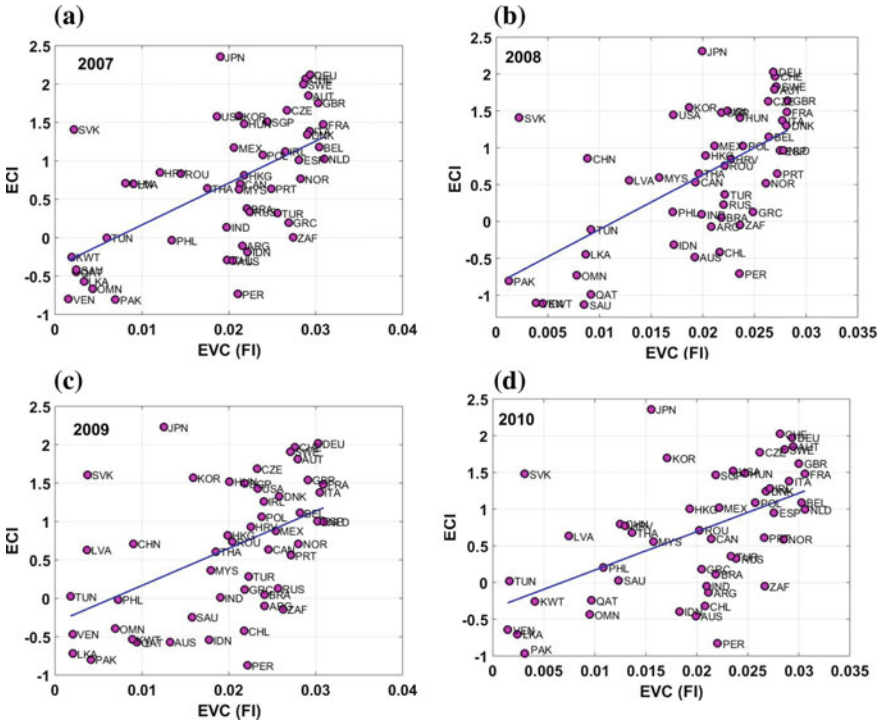


Fig. 9.5 Mapping between the EVC's of financial indices and economic complexity index (ECI), with the best fit lines having a slopes: **a** 54.5 ± 10.7 for 2007, **b** 73.9 ± 13.9 for 2008, **c** 48.5 ± 12.3 for 2009 and **d** 52.4 ± 11.7 for 2010

Next, we have constructed an instrumental variable based on geographic centrality of the countries. The assumption we make is that geographic centrality should be orthogonal to size, but related to trade centrality (because of gravity equation; see below). The results are presented in Table 10 in arXiv:1805.06829 (2 stage least square estimation) and Table 11 in arXiv:1805.06829 (limited information maximum likelihood estimation). As can be seen, the sign is preserved and in the expected direction but the relationship is not statistically significant at 5%. This is somewhat problematic as it indicates that the instrument is not very good for this test.

Finally, we put all data into one balanced panel structure and find panel estimates without incorporating the control variables (unfortunately all data are not available). In this case, as can be expected, the relationship prevails (see Tables 12 and 13 in arXiv:1805.06829). Hausman test confirms that a random effect model is more appropriate here. We checked if there is any relationship between EVC from trade and EVC from return across time (rather than across countries, as we have discussed above). In particular, it would be of interest to see if there is any strong indication of *Granger causality*. The results are presented in Tables 14–31 in arXiv:1805.06829. We see that there is no systematic relationship across these variables over time (we

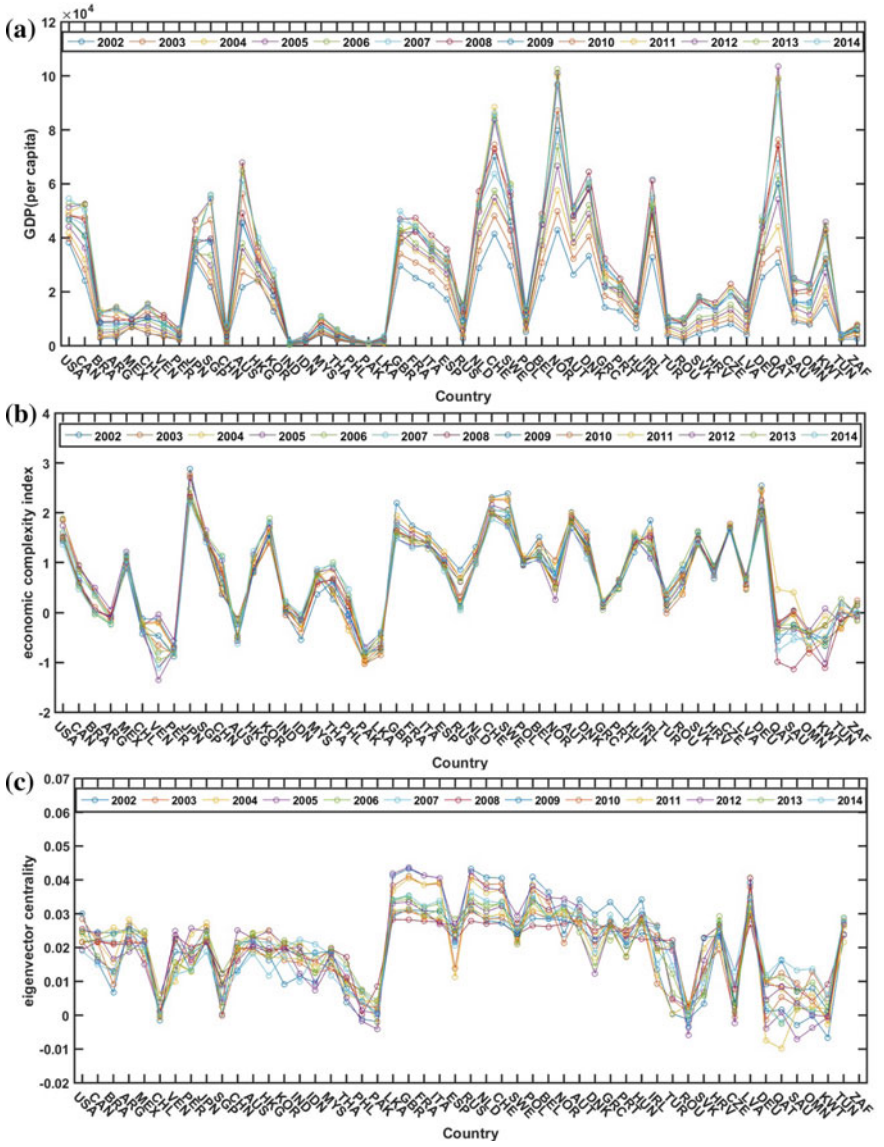


Fig. 9.6 Time evolution of **a** GDP per capita, **b** economic complexity index and **c** eigenvector centrality of market indices of different countries across the globe for the period 2002–2014

have included two lags for all estimations). Note that the time length is very small (9 years). Hence, we cannot infer much from the *VAR* analysis.

Economic Interpretation and Econometric Issues

We have shown that there is a mapping between the networks of real and nominal variables. It is important to stress that this establishes the novelty of the present approach over and above the basic findings of Sharma et al. [2]. The main statement of Sharma et al. [2] is that centrality in the financial market is related to the size effect. In the present case, the same still holds true and that can be explained easily through gravity equation of trade, which states that the trade volume (T_{ij}) between two countries is approximately proportional to the product of the size of the countries (Y_i and Y_j) and inversely proportional to their distance (d_{ij}). This can be stated as

$$T_{ij} \sim \frac{Y_i \times Y_j}{d_{ij}}. \tag{9.5}$$

This implies that the EVC from the trade matrix is highly correlated with size itself. This, in turn, implies that the relationship we find between the financial network and the trade network, may actually be a manifestation of the centrality-size relationship, similar to the finding in Sharma et al. [2].

This raises a fundamental question about the nature of the relationship. Is it centrality-centrality or centrality-size? We cannot provide a complete answer to that. There are three points that need to be considered. First, centrality-size identification is an extremely difficult exercise as typically these two variables are highly correlated. Second, to characterize spill-over effects, network structures are useful whereas the size effect is not. Finally, the relationship between EVC of financial network and the trade network, is not monotonic. The linear fit captures the positive relationship. But a non-linear fit shows that the effect of higher centrality in trade diminishes after a steep initial increase. Thus, the multi-layered network view (with EVC-EVC as opposed to EVC-size relationship) is important to recognize the non-monotonic behavior.

Snapshot of the World Stock Market

Figure 9.7 shows the minimum spanning tree (MST) of 51 market indices obtained from the Pearson cross-correlation matrix across the globe during the period 2013–2014. The nodes in the tree represent the market indices of the corresponding countries and the links between the nodes represent the relative distances of the distance matrix, $d = \sqrt{2(1-\rho)}$, where ρ represents the correlation matrix. Thus, the minimum spanning tree reveals the structure of the global market indices and provides simple visualization about the patterns of links between different markets, similar to what was observed by Wang et al. [7]. The MST indicates that geographic proximity plays big role in shaping up the correlation structure across markets. This feature has been noted and documented by other researchers as well [16]. One can con-

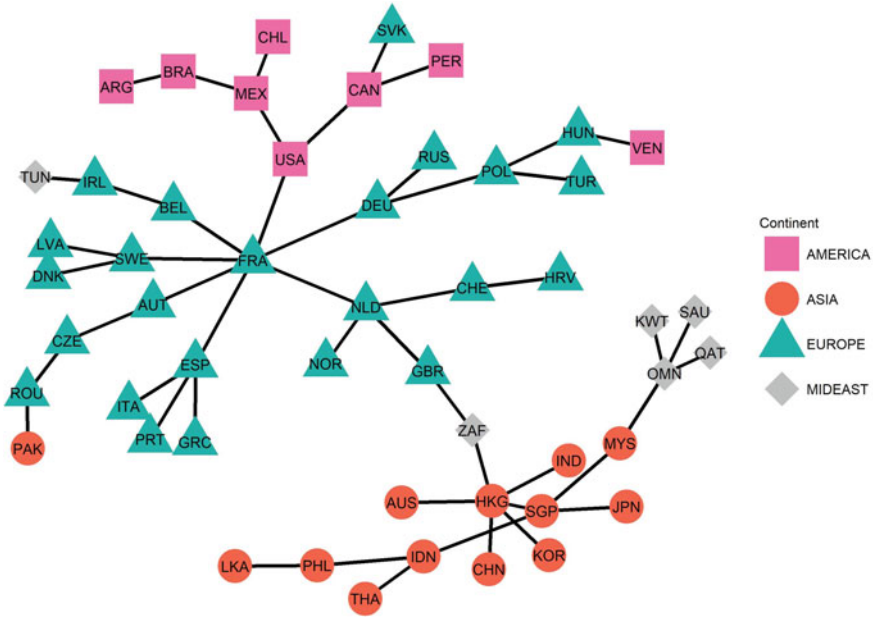


Fig. 9.7 Minimum spanning tree of 51 market indices across the globe during the period 2013–2014. The MST shows 6 African markets (gray diamonds), 13 Asian markets (orange circles), 24 European markets (green triangles) and 8 Latin America markets (magenta squares)

jecture that the main factor behind this observation is that financial markets react very quickly to news and hence, any bout of volatility in a market will be transmitted to another market when that opens. For example, Tokyo stock exchange opens before London stock exchange. Hence, it is conceivable that there would be volatility spillover from Tokyo to London. Although this qualitative explanation is intuitive, it remains unclear how to understand the underlying mechanism quantitatively. On a similar vein, it does not clearly explain the structure of the MST either.

Summary

In this chapter, we have demonstrated using multi-layered networks, the existence of an empirical linkage between the dynamics of the financial network constructed from the market indices and the macroeconomic networks constructed from macroeconomic variables such as trade, foreign direct investments, etc., for several countries across the globe. The time scales of the dynamics of the financial variables and the macroeconomic variables are orders of magnitude different, which makes the empirical linkage even more interesting and significant. Also, we found that there exist in the respective networks, core-periphery structures (determined through eigenvector

centrality measures) that are composed of similar sets of countries—a result that may be related through the 'gravity model' of the country-level macroeconomic networks. Thus, from a multi-lateral openness perspective, we showed that for individual countries, larger trade connectivity is positively associated with higher financial return correlations. We have specifically studied the two countries: Germany (DEU) and France (FRA), with respect to the other European countries. This revealed that mapping between the trade and financial indices correlation is quite robust across several years.

Furthermore, we showed that the Economic Complexity Index and the equity markets have a positive relationship among themselves, as is the case for Gross Domestic Product; the time evolution of the three variables have interesting periodicities and correlation patterns. For certain countries the dispersions in the variables are rather pronounced than in other countries. To reveal the structure and dynamics of the global market indices, we have also studied the minimum spanning tree, which indicated that the geographical proximity does play an important role in the correlation structure across different markets. Perhaps the time-lagged correlation studies would reveal further the lead-lag structure of the markets.

As noted by many researchers, network approach illuminates several interesting facets of the structure of global economy. However, standard econometric techniques show that all superficial observations are not necessarily robust. In particular, whenever one wants to move from correlations to causality, one has to use extra caution. In the end, we note a simple point. Many proposed empirical relationships in the econophysics literature fail the test for robustness (both in economic and statistical sense). Usage of econometrics combined with simple economic intuitions could remedy the problem to a large extent.

Acknowledgements ASC acknowledges the support by the institute grant (R&P), IIM Ahmedabad. AC and KS acknowledge the support by grant number BT/BI/03/004/2003(C) of Govt. of India, Ministry of Science and Technology, Department of Biotechnology, Bioinformatics division, DST-PURSE of the Jawaharlal Nehru University, and University of Potential Excellence-II grant (Project ID-47) of the Jawaharlal Nehru University, New Delhi. KS acknowledges the University Grants Commission (Ministry of Human Resource Development, Govt. of India) for her senior research fellowship.

References

1. Lee, Y., Amaral, L.A.N., Canning, D., Meyer, M., Stanley, H.E.: Universal features in the growth dynamics of complex organizations. Phys. Rev. Lett. **81**(15), 3275 (1998)
2. Sharma, K., Gopalakrishnan, B., Chakrabarti, A.S., Chakraborti, A.: Financial fluctuations anchored to economic fundamentals: a mesoscopic network approach. Sci. Rep. **7**(1), 8055 (2017)
3. Lee, K.M., Goh, K.I.: Strength of weak layers in cascading failures on multiplex networks: case of the international trade network. Sci. Rep. **6**, 26,346 (2016)
4. Lee, K.M., Min, B., Goh, K.I.: Towards real-world complexity: an introduction to multiplex networks. Eur. Phys. J. B **88**(2), 48 (2015)

5. Duenas, M., Fagiolo, G.: Modeling the international-trade network: a gravity approach. J. Econ. Interact. Coord. **8**, 155–178 (2013)
6. Squartini, T., Fagiolo, G., Garlaschelli, D.: Null models of economic networks: the case of the world trade web. J. Econ. Interact. Coord. **8**, 75–107 (2013)
7. Wang, G.J., Xie, C., Stanley, H.E.: Correlation structure and evolution of world stock markets: evidence from pearson and partial correlation-based networks. Comput. Econ. **51**(3), 607–635 (2018)
8. Qadan, M., Yagil, J.: International co-movements of real and financial economic variables. Appl. Econ. **47**(31), 3347–3366 (2015)
9. Bookstaber, R., Kenett, D.Y.: Looking deeper, seeing more: a multilayer map of the financial system (2016). https://financialresearch.gov/briefs/2016/07/14/multilayer-map
10. Thompson Reuters Eikon database (2016). Accessed 9th November 2016. https://customers.thomsonreuters.com/eikon/index.html
11. Economic complexity index database. Atlas of economic complexity (2017). http://atlas.cid.harvard.edu/rankings/country/
12. GDP (per capita by country) database. Knoema world data atlas (2017). https://knoema.com/pjeqzh/gdp-per-capita-by-country-1980-2014
13. Lee, C.S.: The linkage between the FDI and trade of China, Japan and Korea: the Korean perspective. In: Seoul: Korean Institute for International Economic Policy. Paper Prepared for Presentation at the DRC/NIRA/KIEP Symposium on Strengthening Economic Cooperation in Northeast Asia: Facilitating Investment Between China, Japan and Korea Held in Beijing, vol. 29, p. 2002 (2002)
14. Liu, L., Graham, E.: The relationship between trade and foreign investment: empirical results for Taiwan and South Korea. Working paper series (Institute for International Economics (U.S.)) (1998). https://books.google.co.in/books?id=rIAUnQAACAAJ
15. Hidalgo, C.A., Hausmann, R.: The building blocks of economic complexity. Proc. Natl. Acad. Sci. **106**(26), 10570–10575 (2009)
16. Curme, C., Stanley, H.E., Vodenska, I.: Coupled network approach to predictability of financial market returns and news sentiments. Int. J. Theor. Appl. Financ. **18**(07), 1550043 (2015)

Chapter 10
Evolution and Dynamics of the Currency Network

Pradeep Bhadola and Nivedita Deo

Abstract We study the statistical and spectral properties of the foreign exchange of 21 different currencies from January 4, 1999 to March 30, 2018. The correlation matrix is calculated for different periods with a rolling window method and the properties are studied for each window. The basic statistics on the correlation matrix shows that the currencies are more and more correlated with times. The distribution of the correlation matrix was very asymmetric with non zero skewness which shows a fat tail behavior for the initial years but approach Gaussian distribution for the later time. The spectral properties of the correlation matrices for each window when compare with the properties of the correlation matrix formed for the complete period and with analytical results for Wishart matrices shows that the distribution is different for the windows comprising the calm and the crisis period. The study of the number of eigenvalues which are outside the random matrix bounds for each window on both sides of spectrum reveals that for the crisis period, the number of eigenvalues outside the lower bound increases as compared to the calm period. This increase in the number of eigenvalues on the lower side of the spectrum for a window also implies a crisis in the near future. The lower end of the spectra contains more information than the higher side as revealed by the entropic measures on the eigenvalues. This entropic measure shows that the eigenvectors on the lower side are more informative and localized. In this work, the analysis of individual eigenvector captures the evolution of interaction among different currencies with time. The analysis shows that the set of most interacting currencies that are active during the calm period and the crisis period are different. The currencies which was dominating in the calm period suddenly lose all weight and new set of currencies become active at the onset and during the crisis. The largest eigenvector of the correlation matrix can separate currencies based on their geographical location.

P. Bhadola
The Institute for Fundamental Study, Naresuan University, Phitsanulok, Thailand
e-mail: bhadola.pradeep@gmail.com

N. Deo (✉)
Department of Physics and Astrophysics, University of Delhi, Delhi 110007, India
e-mail: ndeo@physics.du.ac.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_10

Introduction

The foreign exchange (FX) are the market which are global, decentralized, over-the-counter which includes all aspects of exchanging, buying or selling currencies determining the foreign exchange rates. The financial markets are deemed as one of the most complex systems that captures human financial activity on a global scale with a very high trading volume leading to high liquidity. The FX market directly or indirectly affects all other financial exchanges or markets therefore peeks the interest of academic community to study the structure, statistical properties and topology of the FX network. The main reasons to study the structure and nature of the world foreign network is first, the world foreign exchange (currency) market is considered to be the largest financial market, which according to the Triennial Central Bank Survey show a daily average trading for over $5.09 trillion per day in April 2016 [1] extending over all countries. Second, the inability to express the absolute price of a given currency i.e. the absence of a reference frame, makes it more complex than any other financial system. In FX, one has to represents a currency in terms of the other base currency. Since internal dynamics of the base currency greatly depends on multiple factors such as political or social changes, economy, inflation as well as and sensitive to events in any part of the world, the local events can have a global effects in the FX network. With the digital age, there is an enormous growth of the electronically recorded financial data. Even with these huge data sets together with the modern day high-throughput computational methods, the understanding of the complex nature, structure, interaction, dynamics and behavior of the foreign exchange rate remains a challenge. One of the important features of the financial systems is the existence of the correlations between different financial commodities or agents. The study of the cross correlations at different time scales between the financial data is been widely study [2–4] and are used for portfolio optimization or asset risk management [5, 6].

System, Data and Nomenclature

For the analysis, the data set comprises of the daily FX rates for 21 different currencies from January 4, 1999 to March 30, 2018 which spans a period of over 18 years. The currency is denoted according to the ISO 4217 standards using three letter code. The list of the countries and their currencies used for the analysis is shown in Table 10.1 The FX rates are expressed in term of a base currency, which in the current case is United States Dollar (USD). The base currency serves as a frame of reference for all other currencies. The dynamics of other currencies by using the same base currency is equivalent to study the dynamics from the perspective of the base currency (USD). In other words, the evolution of the all other currencies are studied in the frame in which the base currency is at rest.

Table 10.1 Countries and their respective currencies used for the analysis

Country	Currency	Country	Currency	Country	Currency
Brazil	BRL	South Korea	KRW	Switzerland	CHF
Canada	CAD	Mexico	MXN	Taiwan	TWD
China	CNY	Norway	NOK	Thailand	THB
Denmark	DKK	Sweden	SEK	Australia	AUD
Hong Kong	HKD	South Africa	ZAR	Euro	EUR
India	INR	Singapore	SGD	New Zealand	NZD
Japan	JPY	Sri Lanka	LKR	United Kingdom	GBP

The daily FX exchange rates data was first preprocessed for filtering to remove any numerical artifacts. Let $S_i(t)$ is daily FX rate of a currency i on day t expressed in terms of USD. The logarithmic returns $R_i(t)$ of the currency i on day t is defined as

$$R_i(t) = ln(S_i(t)) - ln(S_i(t - \Delta)) \tag{10.1}$$

where $\Delta t = 1$ day. The normalized logarithmic returns is then given by

$$r_i(t) = \frac{R_i(t) - < R_i >}{\sigma_i} \tag{10.2}$$

where $< R_i >$ is the time average of the returns over the time period and σ_i is the standard deviation of $R_i(t)$ defined as $\sigma_i = < R_i^2 > - < R_i >^2$.

Correlation Coefficients

The Pearson correlation coefficients are used to estimate the correlation between different FX rates. The correlation coefficient between currency i and j is given by

$$C_{i,j} = < r_i(t) r_j(t) > . \tag{10.3}$$

The correlation coefficients are obtained, such that $-1 \leq C_{i,j} \leq 1$ where $C_{i,j} = 1$ represents perfect correlation and $C_{i,j} = -1$ represents perfect anti-correlation. The correlation matrix is a $N \times N$ symmetric matrix where $N = 21$ in this case. These correlation are correlation as viewed in the frame of base currency USD.

To check the evolution of the FX exchange rates the correlation matrix is calculated with a rolling window of size 250 days with a shift of 50 days. Our data spans the period from January 4, 1999 to March 30, 2018, which results in a 97 windows. Various statistical and spectral properties of correlation matrices for each window is studied and compared.

Statistics Correlation Matrix

Before analyzing the spectral properties of the correlation matrices, we investigate the statistical properties of the correlation matrix of each window along with the probability density function (PDF) of the independent elements of the correlation matrix i.e. for C_{ij} for $i < j$. This results in the $N(N-1)/2 = 210$ elements. Figure 10.1 shows the PDFs of the correlation coefficients C_{ij}, calculated for the complete interval from 1999–2018 and the correlation coefficients C_{ij}, calculated for the each window. Where the window is represented by the starting date of the window.

From Fig. 10.1, it is evident that the PDFs of the correlation coefficients for the complete dataset (1999–2018) is a non symmetrical distribution with a positive mean (0.289). This distribution (differs from the Gaussian distribution) having a a right-skewed distribution with skewness 0.65. The distribution is right tailed and with a kurtosis of 3.10 (for Gaussian the kurtosis is 3.0). From Fig. 10.1 we can conclude that for the FX exchange rates the positive cross-correlations are more common than negative cross-correlations. With window (time) the PDFs for of the correlation coefficients shifts towards right (the larger positive correlations) and amount of the negative correlations decreases significantly with window (time).

For each window, we calculated and study the descriptive statistics (i.e., the mean, standard deviation, skewness, and kurtosis) of the cross-correlation coefficients $C_{ij}; i > j$, which is shown in Fig. 10.2. For first window (which starts at 04–01–1999), the mean of the correlation coefficient is low (0.10), which start to increase with rolling windows (time). This implies that with time, the FX currency exchange network becomes more and more correlated. An interesting observation in plot of skewness with window Fig. 10.2, is that during the crisis period there is a decrease in the magnitude of skewness. The skewness which was positive for the period far from the crisis (2008), changes sign and becomes negative just before and during the crisis. The absolute value of skewness decreases with window form 1.96 in 1999 to 0.40 in 2017, the same trend is seen in kurtosis which also shows a

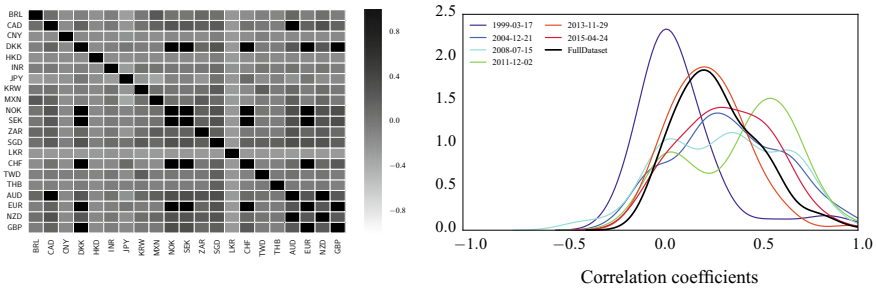


Fig. 10.1 First figure shows the color map of the correlation coefficients between the FX rates of different currencies for the complete interval from 1999–2018. The second figure shows PDFs of C_{ij} for the each window, where the date indicates the start of the window

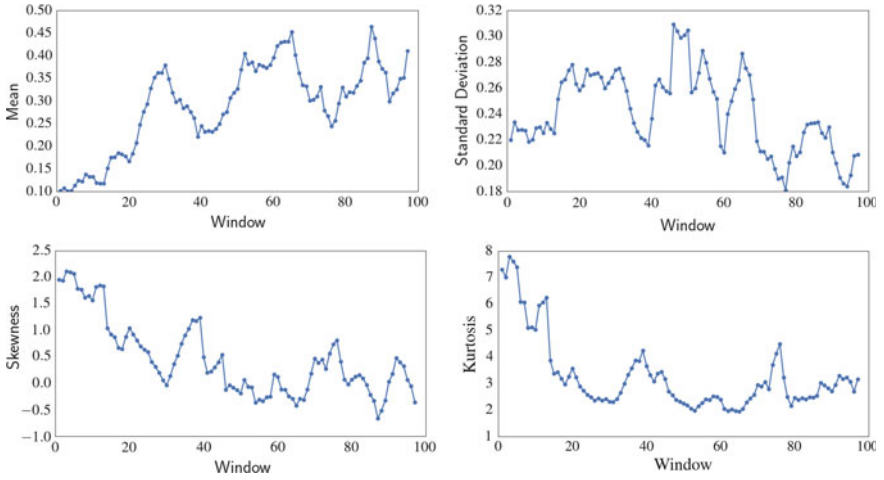


Fig. 10.2 First figure descriptive statistics (mean, standard deviation, skewness and kurtosis) of the cross correlation coefficients $C_{ij}; i > j$ for each window

decrease in value with time (window), in 1999 the kurtosis was 7.8 which decreases to 3.18 in 2017.

For the PDFs and the basic statistics we can conclude that the FX currency network becomes more and more positively correlated with time. Also, as seen with the skewness, the distribution of the cross correlation is becoming more and more symmetric and Gaussian like with the passage of time as the skewness approaching close to zero and the kurtosis approaching close to 3 which are the standard for the Gaussian distribution.

Spectral Properties

The FX exchange correlation structure, in the reference of the base currency USD, can be described by the eigen spectra of the correlation matrix C. For every correlation matrix C, (correlation matrix corresponding to each window), the complete set of the eigenvalues λ_i and eigenvectors v_i, are determined from the eigenvalue equation $C\mathbf{v_i} = \lambda_i \mathbf{v_i}$ where $i = 1 \ldots N$. These eigenvalues are arranged in an ascending order such that $\lambda_1 \leq \lambda_2 \leq \lambda_3 \cdots \leq \lambda_N$. The spectral properties of the correlation matrix for each window is studied.

In the current work, we first compare the eigenvalue distribution of correlation matrix corresponding to each window with the correlation matrix for the complete data set. This may result in the useful insights on how the dynamics for a small scale (window) differs from the over all system dynamics (full data set).

To distinguish noise from the information contained in the system, we constructed a null model by randomly shuffling the FX exchange rates data for each currency. We find that the results from the random shuffling are numerically equivalent to the analytical results for the Wishart matrices. The eigen-spectra of Wishart matrices is well studied [7], where the density function for the eigenvalues $P(\lambda)$ is defined as

$$P(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}. \tag{10.4}$$

with $Q = \frac{N}{L} \geq 1$. Where N the number of currencies and L is the number of days for which the exchange rates are observed (used) and $\sigma = 1$ the standard deviation. Equation (10.4) is known as Marcenko–Pastur distribution where the the upper and lower bounds for the eigenvalue λ are defined as

$$\lambda_\pm = \sigma^2\left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}\right) \tag{10.5}$$

We use the above analytical results, to established the bounds for noise in the eigen-spectrum of the system. For the full data set $N = 21$, number of currencies and the exchange rate data is taken for $L = 4835$ days. This gives $\lambda_+^c = 1.14$ and $\lambda_-^c = 0.87$ where c in the superscript implies that this is for the complete data set. For each window, the number of days are fixed to 250, therefore $L = 250$, the size of the window and $N = 21$ is the number of currencies used for the analysis giving $\lambda_+ = 1.66$ and $\lambda_- = 0.50$.

For each window, we estimate the number of eigenvalues outside the RMT bounds on each side of the eigenvalue spectra ($\lambda \leq \lambda_-$) and ($\lambda \geq \lambda_+$) and is plotted. Figure 10.3 shows that the number of eigenvalues outside the RMT upper bound is very less, (two in most cases but the second largest eigenvalue is very close to λ_+). On the other hand, the number of eigenvalues on the lower side of the spectra for which $\lambda \leq \lambda_-$ are higher in number. On an average 10 out of 21, eigenvalues are outside the lower RMT bounds, which are nearly 50% of the total eigenvalues. This indicates that most of the information about the correlation and interactions between different currencies is located on the lower side of the spectrum.

On of the interesting observation is the number of eigenvalues smaller than the lower RMT bounds, increases at the time of the stress in the global economy. Just before the 2008 crisis, the number of eigenvalues outside the lower bound increases from 9 to 11, which further increase to 13 during the 2008–2009 crisis. After the crisis, the number of eigenvalues outside the lower RMT bounds drops. But further increase in the number (13 eigenvalues $\leq \lambda_-$) is observed for the window corresponding to the period 2011–09–20 to 2012–09–17, this period corresponds to European sovereign debt crisis and the United States debt ceiling crisis [8, 9]. In Fig. 10.3, this was followed a decrease in the number of eigenvalues $\leq \lambda_-$, and again during the 2015–16 Chinese stock market turbulence following slowdown in China and its currency devaluation [9], we observe a increase in number of eigenvalues $\leq \lambda_-$ (again the

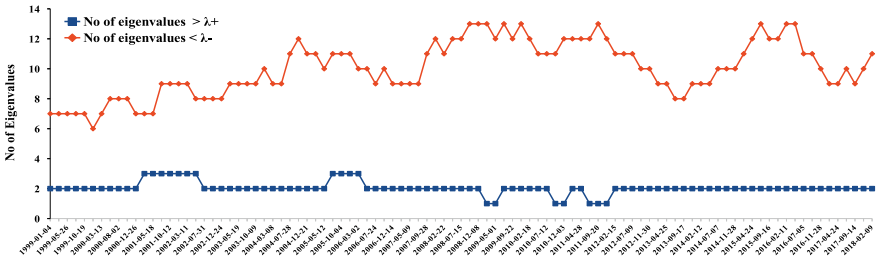
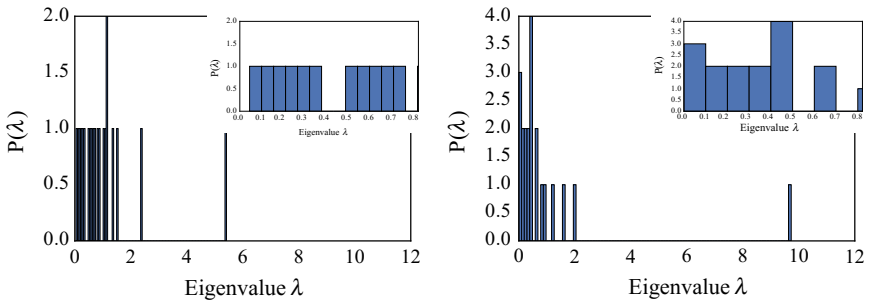


Fig. 10.3 Number of eigenvalues outside the RMT bound on each side of the spectra for every window, where the date indicates the start of the window

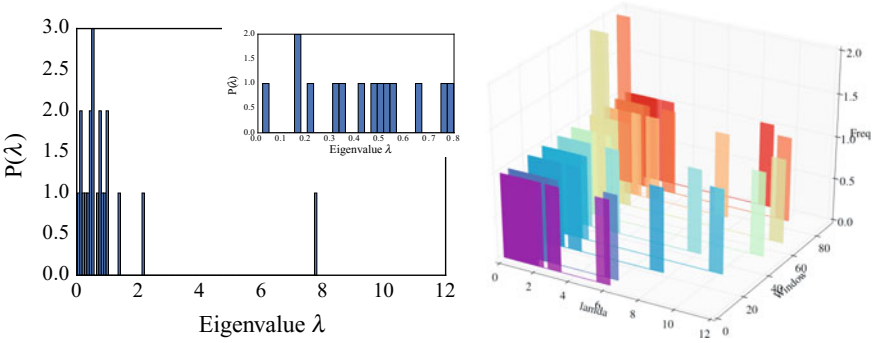


(a) Window 6. (b) Window 50.

Fig. 10.4 Eigenvalue distributions for different windows. Insets show eigenvalues outside the lower bound

number is 13). The increase in number of eigenvalues on the lower side of spectra that are less than the lower RMT bounds correlates very well with the time of the financial crisis.

Figures 10.4 and 10.5 shows the eigenvalue distribution for different windows and for the complete period (1999–2018). For Fig. 10.4a which is the calm period (1999–12–30 to 2000–12–22), the magnitude of the largest eigenvalue is 5.4 which is very less compared to the period of high financial stress (crisis) window 50 (2008–09–24 to 2009–09–21, with largest λ equal to 9.6) as shown in Fig. 10.4b. We observe that for the less financial stress the largest eigenvalue of correlation matrix for the FX exchange is lower as compared to the period of high financial stress. The largest eigenvalue of the correlation matrix of FX exchange rates for the complete period is 7.8. We study the time evolution of each eigenvalue which is outside the RMT bound. We observe the the sum of the eigenvalues on the lower side of spectra outside the lower RMT bound is opposite to the dynamics of largest eigenvalue as shown in Fig. 10.6.



(a) For full period (1989−2018). (b) For different windows.

Fig. 10.5 Eigenvalue distributions. Insets show eigenvalues outside the lower bound

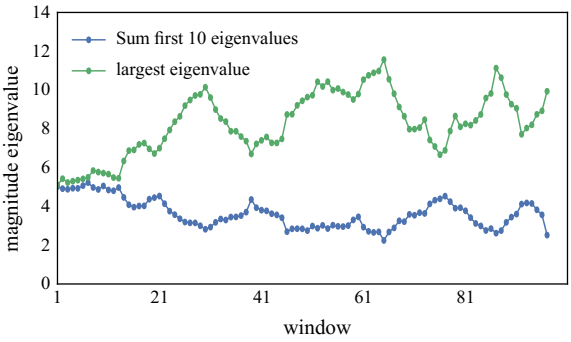


Fig. 10.6 Dynamics of the largest and the sum first few eigenvalues

Information Content of Eigenvectors

The information content of each eigenvector is estimated by estimating the Shannon's entropy. The entropy of an eigenvector v_i is defined as

$$H_i = -\sum_{j=1}^{N} u_i(j) log_L(u_i(j)), \tag{10.6}$$

where N is the total number of currencies used for the analysis (number of eigenvector components) and $u_i(j) = (v_i(j))^2$ is the square of the jth component of the ith normalized vector v_i. The entropy estimates also helps to calculate the localization of the eigenvector. Eigenvector with low entropy should be highly localized. In the current analysis, we find that the eigenvector corresponding to the small eigenvalues are very informative as well as highly localized as compared to the eigenvector corresponding to the large eigenvalues. These eigenvectors are further used to estimate the strong interacting pair of currencies. For many systems especially for correlation

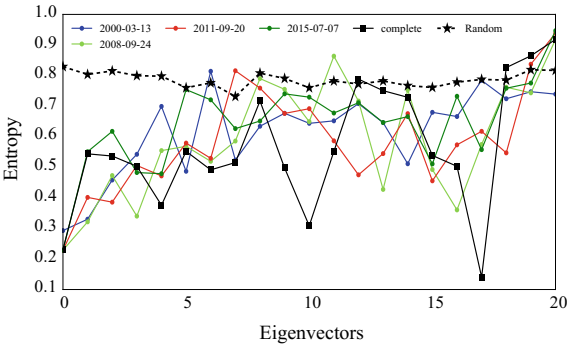


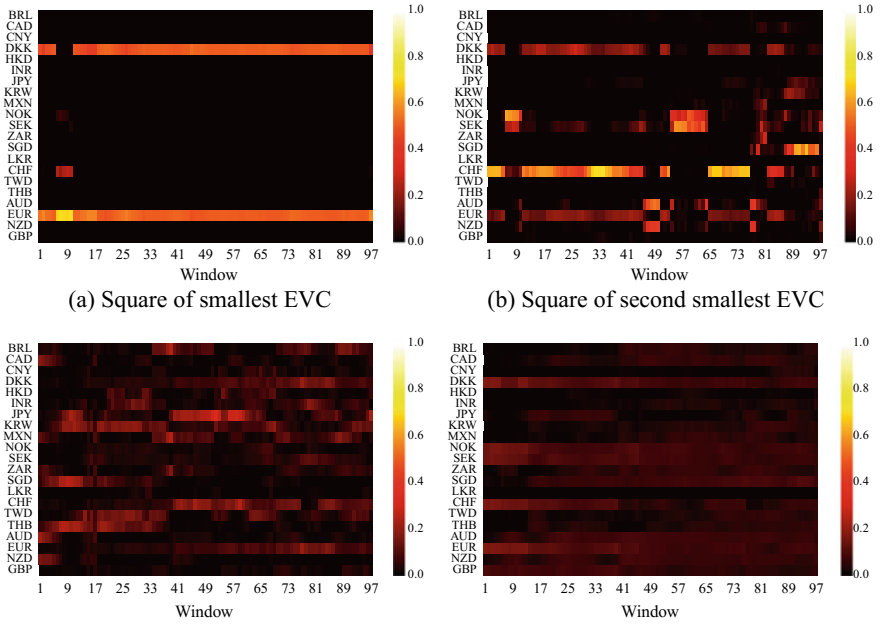
Fig. 10.7 Entropy of eigenvector for different windows, where the date indicates the start of the window

matrices between position in a protein family, it is established that the eigenvector corresponding to the small eigenvalues can identify important positions and interaction responsible for the formation of motifs and sectors [10]. In the Markowitzs theory of optimal portfolios [11], the least risky portfolios corresponds to the lowest eigenvalues and the corresponding eigenvectors. In recent work [12], which involves the analysis of the correlation structure of global financial indices, it was established that the lower side of the eigenvalue spectrum is more informative and localized which is able to capture most of the system dynamics.

Figure 10.7 compares the eigenvector entropy with for different windows with the random shuffled system and eigenvector entropy for the complete period. The eigenvectors corresponding to the large eigenvalues have entropy values very close to the random system (black dashed line). But there is a clear distinction of the entropy of the small eigenvectors with the random system. The fist few eigenvectors have very small entropy (≈ 0.25) for all windows. These eigenvectors are highly localized and gives the set of highly interacting currencies. Analyzing each eigenvector independently reveals clusters of currencies with very close ties.

Eigenvector Components

The analysis of the eigenvector corresponding to the eigenvalues outside the RMT bounds should contain the information present within the system. We use the square of the eigenvector components, to determine their contribution towards that a given eigenvalue. As it is already been discussed in the previous section, that the low eigenvector are highly localized as compared to the high eigenvector components (EVC). Figure 10.8 shows the variation of the square of the component for each currency with window, for the two lowest and two largest eigenvectors. From the Fig. 10.8a it is clear that the lowest eigenvector is highly localized over all windows and there are only a few currencies that are contributing to it. Most of the time (window), there are only two contributing currencies (Euro (EUR) and Danish Krone



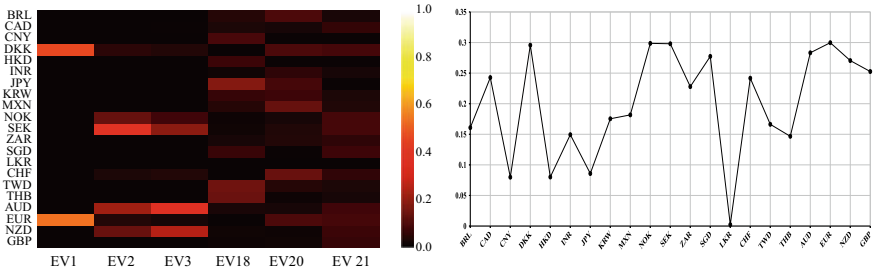
(a) Square of smallest EVC (b) Square of second smallest EVC

(c) Square of second largest EVC (d) Square of largest EVC

Fig. 10.8 The square of the eigenvector components (EVC) for different windows

(DKK)) but only for a short period from window number 6–10 (period from 1999–12–30 to 2001–10–11) Swiss Franc (CHF) have high interactions with Euro. This implies that there exist strong interaction between the Euro (EUR) and Danish Krone (DKK).

For the second smallest EVC Fig. 10.8b, we find the contribution depends on the time period of observation (window). The second smallest eigenvector shows a drastic change in the contributing currencies for the period of financial stress. The early 2000s recession represented by window number 6–10 (period from 1999–12–30 to 2001–10–11), contributing components changes from Euro (EUR), Danish Krone (DKK) and Swiss Franc (CHF) to Norwegian Krone (NOK), Swedish Krona (SEK), Australian Dollar (AUD) and New Zealand Dollar (NZD). A similar change is observed for during the global 2008 financial crisis (window 46–50, period from 2007–12–10 to 2009–07–10), European sovereign debt crisis and the United States debt ceiling crisis a period from 2010–02–18 to 2012–09–17 (window 57–65). The three currencies (EUR, DKK and CHF) dominates the second smallest eigenvector forming very close ties during the calm period but at the onset and during crisis, their interaction cease to exist and a new interaction between NOK, SEK, AUD and NZD is formed. These four currencies dominates forming very strong interactions during the crisis. During the Chinese stock market turbulence captured by window 79–85, there are other currencies such as SGD, ZOR, MXN, JPY, KRW that have



(a) Square of EVC (b) Largest EVC

Fig. 10.9 Eigenvector components (EVC) for the complete time period 1999–2017

contribution. The largest and the second largest eigenvector for the windows are not localized and have contribution from all the currencies and the contribution changes with time (window) Fig. 10.8c, d.

We compare the eigenvector components obtained for each window with the eigenvector components for the complete time period (1999–2018). Figure 10.9a shows the distribution of the square of the smallest and largest three eigenvector. The small eigenvectors are localized and more informative where as the larger eigenvectors are non localized and plagued with noise. The smallest eigenvector overall shows the same behavior as the shown by the smallest eigenvector for each window where the only contribution are from Euro (EUR) and Danish Krone (DKK). For the second smallest eigenvector, there are a total of 7 contributing currencies Euro (EUR), Danish Krone (DKK), Swiss Franc (CHF), Norwegian Krone (NOK), Swedish Krona (SEK), Australian Dollar (AUD) and New Zealand Dollar (NZD). These currencies are same as obtained analyzing the second smallest eigenvector for all windows Fig. 10.8b, but there are divided into two set one set (EUR), DKK and CHF) dominates during the calm period where the other set (NOK, SEK, AUD and NZD) which is more active during the crisis period.

Analyzing the components of the largest eigenvector Fig. 10.9b, we found that the currencies can be separated by their geographical location based on the range of the value of their contribution. If we choose components by the magnitude, and we define two groups, first group which has magnitude of component greater than 0.2 and the other group with magnitude less than or equal to 0.2. The details of the currencies and their geographical location are given in Table 10.2. The Asian and the Latin American countries form a one group which corresponds to the components less than 0.2 Fig. 10.9b, where as the other group is formed by the countries whose component greater than 0.2. These are the countries belonging to Europe, Australia, one each from North America, Africa and Asia.

Table 10.2 Currencies with respective countries separated based on the magnitude of the largest eigenvector of correlation matrix for the complete dataset

EVC <= 0.2			EVC > 0.2		
Currency	County	Continent	Currency	County	Continent
BRL	Brazil	Latin America	CAD	Canada	North America
CNY	China	Asia	DKK	Denmark	Europe
HKD	Hong Kong	Asia	NOK	Norway	Europe
INR	India	Asia	SEK	Sweden	Europe
JPY	Japan	Asia	ZAR	South Africa	Africa
KRW	South Korea	Asia	SGD	Singapore	Asia
MXN	Mexico	Latin America	CHF	Switzerland	Europe
LKR	Sri Lanka	Asia	AUD	Australia	Australia
TWD	Taiwan	Asia	EUR	Euro	Europe
THB	Thailand	Asia	NZD	New Zealand	Australia

Conclusions

In this work, we try to study and understand the relation and interaction between the foreign exchanges rates for a period from January 1999 to March 2018. All the foreign exchange rates are expressed in terms of a base currency which for the current case is USD. All the dynamics and properties studied in this paper will be in the reference frame where the USD is at rest. The evolution of FX rates is studied using a rolling window of size 250 days with a shift of 50 days. The statistics and the correlation between different currencies is calculated and studied. The statistics on the correlation matrix reveals that with time the currencies are getting more and more correlated. At the start of the period (1999) the mean of the correlation matrix was very less as well was very highly non symmetric (non zero skewness) and high Kurtosis but with time the distribution of correlation coefficients try to approach Gaussian distribution by approaching zero skewness 0 and Kurtosis (three) standard for Gaussian distribution. The spectral properties of the correlation matrices are studied for each window and then compared with the correlation matrix formed from the complete data set and with the analytical results for Wishart matrices. The distribution of eigenvalues reveals the distribution is different for the calm and the crisis period. The number of eigenvalues for each window which are outside the random matrix bounds are higher on the lower sides. For the period of the crisis the number of eigenvalues outside the lower bound increases as compared to the calm period. This may be to incorporate the addition information generated during the crisis. We propose that the if there is increase in the number of eigenvalues outside the rmt bound then that may indicate a crisis in the near future. The information content and localization of each eigenvector is estimated by the using an entropic measure. This measures shows that the eigenvalues on the lower side of the spectra are more localized as well as informative as compared

to the eigenvalue on the higher side of the spectra. The analysis of the individual eigenvectors gives information about the interaction between different currencies. We observe in the second smallest eigenvector, at each crisis period, the contribution and interaction among the currencies changes. The currencies which was dominating in the calm period suddenly lose all their contribution during the crisis and a new group of interaction between currencies become active at the onset and during the crisis. The components of largest eigenvector of the correlation matrix for the complete period can separate the currencies based on their geographical location based on the magnitude of the components.

Acknowledgements We acknowledge the Department of Science and Technology (DST), India, (SERB-DST No- EMR/2016/006536) for financial support.

References

1. Triennial Central Bank Survey of foreign exchange and OTC derivatives markets (2016). https://www.bis.org/publ/rpfx16.htm
2. Kenett, D.Y., Huang, X., Vodenska, I., Havlin, S., Stanley, H.E.: Partial correlation analysis: applications for financial markets. Quant. Financ. **15**(4), 569–578 (2015)
3. Podobnik, B., Stanley, H.E.: Detrended cross-correlation analysis: a new method for analysing two non-stationary time series. Phys. Rev. Lett. **100**, 084102 (2008)
4. Conlon, T., Ruskin, H.J., Crane, M.: Random matrix theory and fund of funds portfolio optimisation. Phys. A **382**(2), 565–576 (2007)
5. Kenett, Dror Y., Shapira, Yoash, Madi, Asaf, Bransburg-Zabary, Sharron, Gur-Gershgoren, Gitit, Ben-Jacob, Eshel: Dynamics of stock market correlations. AUCO Czech Econ. Rev. **4**(3), 330–341 (2010)
6. Kenett, Dror Y., Preis, Tobias, Gur-Gershgoren, Gitit, Ben-Jacob, Eshel: Quantifying meta-correlations in financial markets. Eur. Lett. **99**(3), 38001 (2012)
7. Bowick M. J., Brezin E.: Phys. Lett. B **268**, 21 (1991); Feinberg J., Zee A.J.: Stat. Phys. **87**, 473 (1997)
8. Jayech, Selma: The contagion channels of July–August-2011 stock market crash: a DAG-copula based approach. Eur. J. Oper. Res. **249**(2), 631–646 (2016)
9. Clements, Adam, Hurn, Stan, Shi, Shuping: An empirical investigation of herding in the US stock market. Econ. Model. **67**, 184–192 (2017)
10. Bhadola, Pradeep, Deo, Nivedita: Targeting functional motifs of a protein family. Phys. Rev. E **94**(4), 042409 (2016)
11. Elton E.J., Gruber, M.J.: Modern Portfolio Theory and Investment Analysis. Wiley, New York (1995); Markowitz, H.: Portfolio Selection: Efficient Diversification of Investments. Wiley, New York (1959). See also: Bouchaud J.P., Potters, M.: Theory of Financial Risk. Alea-Saclay, Eyrolles, Paris (1997) (in French)
12. Pradeep Bhadola, Nivedita Deo.: Extreme eigenvector analysis of global financial correlation matrices. In: Econophysics and Sociophysics: Recent Progress and Future Directions, pp. 59–69, Springer, Cham (2017)

Chapter 11
Some Statistical Problems with High Dimensional Financial data

Arnab Chakrabarti and Rituparna Sen

Abstract For high dimensional data some of the standard statistical techniques do not work well. So modification or further development of statistical methods are necessary. In this paper we explore these modifications. We start with important problem of estimating high dimensional covariance matrix. Then we explore some of the important statistical techniques such as high dimensional regression, principal component analysis, multiple testing problems and classification. We describe some of the fast algorithms that can be readily applied in practice.

Introduction

A high degree of interdependence among modern financial systems, such as firms or banks, is captured through modeling by a network $G(V, E)$, where each node in V represents a financial institution and an edge in E stands for dependence between two such institutions. The edges are determined by calculating the correlation coefficient between asset prices of pairs of financial institutions. If the sample pairwise correlation coefficient is greater than some predefined threshold then an edge is formed between corresponding nodes. This network model can be useful to answer important questions on the financial market, such as determining clusters or sectors in the market, uncovering possibility of portfolio diversification or investigating the degree distribution [4, 29]. See Fig. 11.1 for illustration of one such network. Using correlation coefficients to construct the economic or financial network has a serious drawback. If one is interested in direct dependence of two financial institutions the high observed correlation may be due to the effect of other institutions. Therefore a more appropriate measure to investigate direct dependence is partial correlation. Correlation and partial correlation coefficients are related to covariance and inverse

A. Chakrabarti · R. Sen (✉)
Indian Statistical Institute, Chennai, Tamil Nadu, India
e-mail: rsen@isichennai.res.in

A. Chakrabarti
e-mail: arnab@isichennai.res.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_11

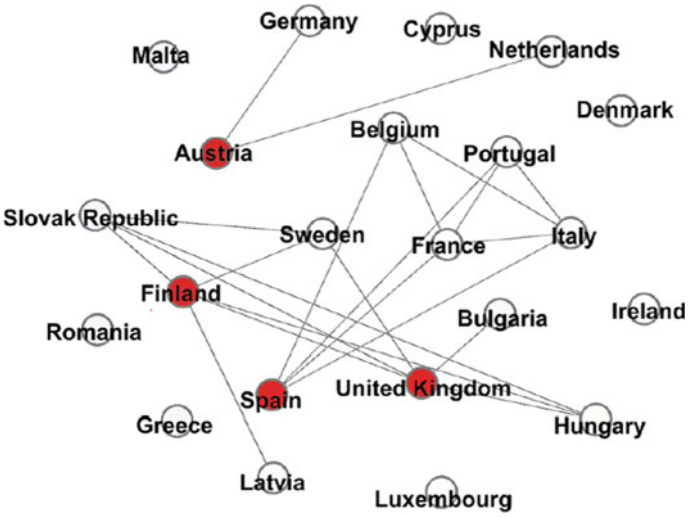


Fig. 11.1 Network topology of European economies in post-euro period as described in [21]

covariance matrix respectively. Therefore in order to have meaningful inference on the complex system of financial network, estimation of covariance matrix accurately is of utmost importance. In this paper we investigate how inference based on covariance matrix for high dimensional data can be problematic and how to solve the problem.

The rest of the paper is organized as follows. In section "Distribution of Eigenvalues" we discuss the distribution of eigenvalues of covariance matrix. In section "Covariance Matrix Estimator" the problem and possible solution of covariance matrix estimation is discussed. Section "Precision Matrix Estimator" deals with estimation of precision matrix. Sections "Multiple Hypothesis Testing Problem and False Discovery Rate" and "High Dimensional Regression" deals with multiple testing procedure and high dimensional regression problem respectively. We discuss high dimensional principal component analysis and several classification algorithms in sections "Principal Components" and "Classification".

Distribution of Eigenvalues

Eigenvalues of Covariance Matrix

In multivariate statistical theory, the sample covariance matrix is the most common and undisputed estimator because it is unbiased and has good large sample properties with growing number of observations when number of variables is fixed. But if the ratio of the number of variables (p) and the number of observations (n) is large, then sample covariance does not behave as expected. It can be showen that if p grows at the

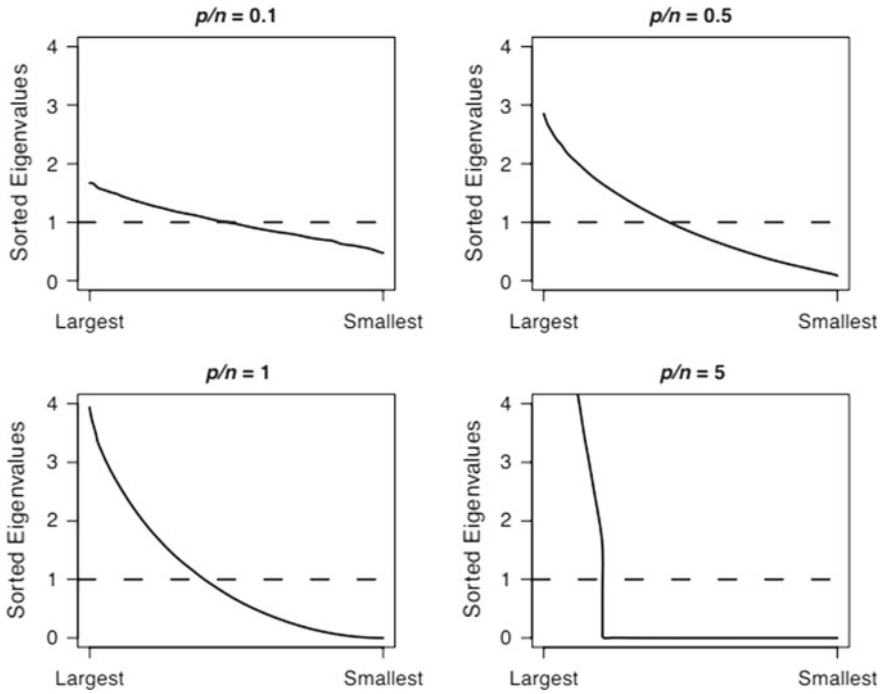


Fig. 11.2 Plot of true (dotted line) and sample (solid line) eigenvalues

same rate as n i.e. $p/n \to y > 0$ the sample covariance matrix becomes inconsistent and therefore can not be relied upon [28]. In Fig. 11.2 eigenvalues of population covariance matrix and sample covariance matrix are plotted for different values of p and n where the population covariance matrix is the identity matrix. It is evident that the true and sample spectra differ a lot as the ratio p/n grows. So for high dimensional data ($p/n \to y > 0$) there is a need to find an improved estimator.

Even though the sample eigenvalues are not consistent anymore, the limiting distribution of eigenvalues of the sample covariance matrix and the connection it has with the limiting eigenvalue distribution of population covariance matrix are of importance. Determining the limiting spectral distribution can be very useful to test the underlying assumption of the model. In this section we will very briefly discuss some results of random matrix theory that answers this kind of questions. Throughout we will denote the ratio p/n as y_n.

Marchenko Pastur Law and Tracy Widom Law

Suppose that $\{x_{ij}\}$ are iid Gaussian variables with variance σ^2. If $p/n \to y > 0$ then the empirical spectral distribution (distribution function of eigenvalues) of sample covariance matrix $S_n$converges almost surely to the distribution F with the density

$$f(x) = \frac{1}{2\pi\sigma^2 yx}\sqrt{(b-x)(x-a)}I(a \le x \le b)$$

if $y < 1$, where $a = a(y) = \sigma^2(1-\sqrt{y})^2$ and $b = b(y) = \sigma^2(1+\sqrt{y})^2$. If $y > 1$ It will take additional positive mass $1 - \frac{1}{y}$ at 0.

σ is called the scale parameter. The distribution is known as Marchenko-Pastur distribution.

If $p/n \to 0$ then empirical spectral distribution of $W_n = \sqrt{\frac{n}{p}(S_n - \sigma^2 I)}$ converges almost surely to the *semicircle law* with density:

$$f(x) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - x^2}I(|x| \le 2\sigma)$$

Although the sample eigenvalues are not consistent estimator, the limiting spectral distribution is related to the population covariance matrix in a particular way.

Also if $p \to \infty$ and $n \to \infty$ such that $\frac{p}{n} \to y > 0$, then $\frac{\lambda_1 - \mu_{np}}{\sigma_{np}} \xrightarrow{\mathcal{L}} W_1$ where λ_1 is the largest eigenvalue of sample covariance $\mu_{np} = \left(\sqrt{n} + \sqrt{p}\right)^2$ and $\sigma_{np} = \left(\sqrt{n} + \sqrt{p}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}}\right)^{\frac{1}{3}}$ and W_1 is Tracy-Widom Law distribution.

Covariance Matrix Estimator

Stein's Approach

We see from Fig. 11.2 that the sample eigenvalues can differ a lot from the population eigenvalues. Thus shrinking the eigenvalues to a central value is a reasonable approach to take. Such an estimator was proposed by Stein [26] and takes the following form:

$$\hat{\Sigma} = \hat{\Sigma}(S) = P\psi(\Lambda)P'$$

where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $\psi(\Lambda)$ is also a diagonal matrix. If $\psi(\lambda_i) = \lambda_i \forall i$ then $\hat{\Sigma}$ is the usual estimator S. In this approach the eigen vectors are kept as it is, but the eigenvalues are shrink towards a central value. As the eigen vectors are not altered or regularized this estimator is called rotation equivariant covariance estimator. To come up with a choice of ψ, we can use entropy loss function

$$L = \text{tr}(\hat{\Sigma}\Sigma^{-1}) - \log(\hat{\Sigma}\Sigma^{-1}) - p$$

or Frobeneous loss function

$$L_2 = \text{tr}(\hat{\Sigma}\Sigma^{-1} - I)^2.$$

Under entropy risk ($= E_\Sigma(L)$), we have $\psi(\lambda_i) = \frac{\lambda_i n}{\alpha_i}$ where

$$\alpha_i = \left(n - p + 1 + 2\lambda_i \Sigma_{i \neq j} \frac{1}{\lambda_i - \lambda_j} \right).$$

The only problem with this estimator is that some of the essential properties of eigenvalues, like monotonicity and nonnegativity, are not guaranteed. Some modifications can be adopted in order to force the estimator to satisfy those conditions (see [18, 20]). An algorithm was proposed to avoid such undesirable situations by pooling adjacent estimators together in [27]. In this algorithm first the negative α_i's are pooled together with previous values until it becomes positive and then to keep the monotonicity intact, the estimates ($\psi's$) are pooled together pairwise.

Ledoit-Wolf Type Estimator

As an alternative to the above mentioned method, the empirical Bayes estimator can also be used to shrink the eigenvalues of sample covariance matrix. Reference [14] proposed to estimate Σ by

$$\hat{\Sigma} = \frac{np - 2n - 2}{n^2 p} \tilde{\alpha} I + \frac{n}{n+1} S,$$

where $\tilde{\alpha} = (\det(S))^{1/p}$. This estimator is a linear combination of S and I which is reasonable because although S is unbiased, it is highly unstable for high dimensional data and αI has very little variability with possibly high bias. Therefore a more general form of estimator would be

$$\hat{\Sigma} = \alpha_1 T + \alpha_2 S$$

where T is a positive definite matrix and α_1 (shrinkage intensity parameter), α_2 can be determined by minimising the loss function. For example [17] used

$$L(\hat{\Sigma}, \Sigma) = \frac{1}{p} \text{tr}(\hat{\Sigma} - \Sigma)^2$$

to obtain a consistent estimator with $T = I$. A trade off between bias and variance is achieved through the value of shrinkage parameter. In Fig. 11.3, bias, variance and MSE are plotted against shrinkage parameter. The optimum value of shrinkage intensity is that for which MSE is minimum.

It can be shown that if there exists k_1 and k_2 independent of n such that $p/n \leq k_1$ and $\frac{1}{p} \sum_{i=1}^{p} \text{E}[Y_i]^8 \leq k_2$ where Y_i is the ith element of any row of the matrix of principal components of X and if

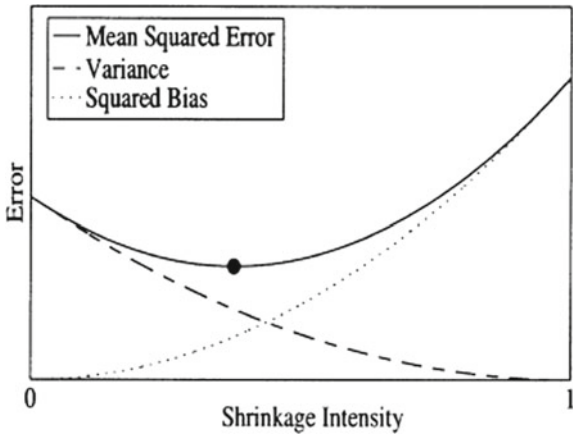


Fig. 11.3 Plot for Error versus shrinkage intensity [17]

$$\lim_{n\to\infty} \frac{p^2}{n^2} \times \frac{\sum_{(i,j,k,l)\in Q_n} (\mathrm{Cov}[Y_i Y_j, Y_k Y_l])^2}{\mathrm{cardinality}(Q_n)} = 0$$

where Q_n denotes the set of all quadruples made of four distinct integers between 1 and p, then the following estimator S_n^* (a convex combination of I and S) is consistent for Σ, see [17]:

$$S_n^* = \frac{b_n^2}{d_n^2} m_n I_n + \frac{d_n^2 - b_n^2}{d_n^2} S_n$$

where $X_{k.}$ is the kth row of X and

$$\begin{aligned} m_n &= \frac{1}{p} \mathrm{tr}(S_n^{'} I_n) \\ d_n^2 &= \|S_n - m_n I_n\|^2 \\ b_n^2 &= \min\left(d_n^2, \frac{1}{n^2} \sum_{k=1}^{n} \|X_{k.}^{'} X_{k.} - S_n\|^2 \right). \end{aligned}$$

The first condition clearly deals with the interplay between sample size, dimension and moments whereas the second one deals with the dependence structure. For $p/n \to 0$ the last condition for dependence structure can be trivially verified as a consequence of the assumption on moments. This estimate is also computationally easy to work with. In fact as S is still unbiased estimator one possible way to reduce the variance is to use bootstrap-dependent techniques like bagging. But that is far more computationally demanding compared to this method.

With the additional assumption that $\mathrm{var}\left(\frac{\sum_{i=1}^{p} Y_i^2}{p} \right)$ is bounded as $n \to \infty$, [17] showed that

$$lim_{n\to\infty}\left[E||S_n - \Sigma_n||^2 - \frac{p}{n}\left(m_n^2 + var\left(\frac{\sum_{i=1}^{p} Y_i^2}{p}\right)\right)\right] = 0$$

This result implies that expected loss of sample covariance matrix, although bounded, does not usually vanish. Therefore consistency of usual sample covariance matrix is achieved only when $\frac{p}{n} \to 0$ or $m_n^2 + \text{var}\left(\frac{\sum_{i=1}^{p} Y_i^2}{p}\right) \to 0$. In the latter case most of the random variables are asymptotically degenerate. The difference between these two cases is that in the first, the number of variables is very less compared to n and in the latter $O(n)$ degenerate variables are augmented with the earlier lot. Both of these essentially indicate sparsity.

A more general target matrix T can be used instead I. For example, under Gaussian distribution, if $T = {}^{\text{tr}(S)}/_{p}I$, $\alpha_1 = \lambda$ (intensity parameter) and $\alpha_2 = 1 - \lambda$, then optimal shrinkage intensity is

$$min\left(\frac{\sum_{i=1}^{p} ||x_i x_i' - S||_F^2}{n^2[tr(S^2) - tr^2(S)/p]}, 1\right)$$

which implies that the shrinkage estimator is a function of the sufficient statistics S and therefore can be further improved upon by using Rao-Blackwell theorem [7]. The resulting estimator becomes $\lambda_{RBLW}T + (1 - \lambda_{RBLW})S$ where

$$\lambda_{RBLW} = \frac{\frac{n-2}{n}tr(S) + tr^2(S)}{(n+2)[tr(S^2) - \frac{tr^2(S)}{p}]}$$

If we take $T = \text{Diag}(S)$, that is, the diagonal elements of S then the optimal intensity that minimises $E[\|\hat{\Sigma} - \Sigma\|^2]$ can be estimated as

$$\frac{\frac{1}{n}(\hat{a}_2 + p\hat{a}_{12}) - \frac{2}{n}\hat{a}_{2*}}{\frac{n+1}{n}\hat{a}_2 + \frac{p}{n}\hat{a}_1^2 - \frac{n+2}{n}\hat{a}_{2*}}$$

where $\hat{a}_1 = {}^{1}/_{p}\text{tr}(S)$, $\hat{a}_2 = \frac{n^2}{(n-1)(n+2)}\frac{1}{p}\left[\text{tr}S^2 - \frac{1}{n}(\text{tr}S)^2\right]$, $\hat{a}_2^* = \frac{n}{n+2}\text{tr}(T^2)/p$ as shown in[11]. Reference [25] chose the shrinkage parameter to be

$$\lambda* = \frac{\sum_{i=1}^{p} \hat{\text{var}}(s_i) - \hat{\text{cov}}(t_i, s_i) - \hat{\text{Bias}}(s_i)(t_i - s_i)}{\sum_{i=1}^{p}(t_i - s_i)^2}.$$

Along with conventional target matrix (I) they used five other target matrices summarised in the following Table 11.1.

Table 11.1 Conventional target matrix (I) used along with five other target matrices

Target A: "diagonal, unit variance"	Target B: "diagonal, common variance"
0 estimated parameters	1 estimated parameter: v
$t_{ij} = \begin{cases} 1 & if\ i = j \\ 0 & if\ i \neq j \end{cases}$	$t_{ij} = \begin{cases} v = avg(s_{ii}) & if\ i = j \\ 0 & if\ i \neq j \end{cases}$
$\hat{\lambda}^* = \frac{\sum_{i \neq j} \hat{var}(s_{ij}) + \sum_{i=j} \hat{var}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii}-1)^2}$	$\hat{\lambda}^* = \frac{\sum_{i \neq j} \hat{var}(s_{ij}) + \sum_i \hat{var}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii}-v)^2}$
Target C: "common (co)variance"	**Target D: "diagonal, unequal variance "**
2 estimated parameters: v,c	p estimated parameters: s_{ii}
$t_{ij} = \begin{cases} v = avg(s_{ii}) & if\ i = j \\ c = avg(s_{ij}) & if\ i \neq j \end{cases}$	$t_{ij} = \begin{cases} v = s_{ii} & if\ i = j \\ 0 & if\ i \neq j \end{cases}$
$\hat{\lambda}^* = \frac{\sum_{i \neq j} \hat{var}(s_{ij}) + \sum_{i=j} \hat{var}(s_{ii})}{\sum_{i \neq j} (s_{ij}-c)^2 + \sum_i (s_{ij}-v)^2}$	$\hat{\lambda}^* = \frac{\sum_{i \neq j} \hat{var}(s_{ij})}{\sum_{i \neq j} s_{ij}^2}$
Target E: "perfect positive correlation"	**Target F: "constant correlation"**
p estimated parameters: s_{ij}	$p+1$ estimated parameters, s_{ii},$\bar{r}$
$t_{ij} = \begin{cases} s_{ii} & if\ i = j \\ \sqrt{s_{ij} s_{ji}} & if\ i \neq j \end{cases}$	$t_{ij} = \begin{cases} s_{ii} & if\ i = j \\ \bar{r}\sqrt{s_{ij} s_{ji}} & if\ i \neq j \end{cases}$
$f_{ij} = \frac{1}{2}\{\sqrt{\frac{s_{jj}}{s_{ii}}} \hat{Cov}(s_{ii}, s_{ij}) + \sqrt{\frac{s_{ii}}{s_{jj}}} \hat{Cov}(s_{jj}, s_{ij})\}$	$f_{ij} = \frac{1}{2}\{\sqrt{\frac{s_{jj}}{s_{ii}}} \hat{Cov}(s_{ii}, s_{ij}) + \sqrt{\frac{s_{ii}}{s_{jj}}} \hat{Cov}(s_{jj}, s_{ij})\}$
$\hat{\lambda}^* = \frac{\sum_{i \neq j} \hat{Var}(s_{ij}) - f_{ij}}{\sum_{i \neq j} (s_{ij} - \sqrt{s_{ii} s_{jj}})^2}$	$\hat{\lambda}^* = \frac{\sum_{i \neq j} \hat{Var}(s_{ij}) - \bar{r} f_{ij}}{\sum_{i \neq j} (s_{ij} - \bar{r}\sqrt{s_{ii} s_{jj}})^2}$

Element-Wise Regularization

Under the assumption of sparsity, some element-wise regularization methods can be used. In contrast to [17] type of estimator, where only the eigenvalues were shrunk, here both eigenvalues and vectors are regularised. We will first discuss popular methods like Banding and Tapering which assume some order between the variables and as a result, the estimator is not invariant under permutation of variables. So this is useful for time-dependent data.

Banding

The idea behind banding is that the variables are ordered in such a way that elements of the covariance matrix, further away from the main diagonal, are negligible. An $l-$banded covariance matrix is defined as $B(S_l) = [s_{ij} I(|i-j| \leq l)]$, where $S = [s_{ij}]$ is the $p \times p$ sample covariance matrix and l $(\leq p)$ is band length, determined through cross validation. One can question- which kind of population covariance matrix can be well approximated by Banded sample covariance matrix. Intuitively, such a matrix should have decaying entries as one moves away from the main diagonal. Reference [3] showed that the population covariance can be well approximated uniformly over the following class of matrices: $\{\Sigma : \max_j \sum_i |\sigma_{ij}| I(i-j \geq k) \leq C.k^{-\alpha}$, and $0 < \varepsilon \leq \lambda_{\min}(\Sigma) < \lambda_{\max}(\Sigma) \leq \varepsilon^{-1}\}$, where C is

a constant and α captures the rate of decay of the entries σ_{ij} as i goes away from j. Although p is large, if $\log(p)$ is very small compared to n, that is, $\frac{\log(p)}{n} \to 0$, then such a Σ can be well-approximated by accurately chosen band length and the error in approximation depends on $\log(p)/n$ and α. Same result holds also for the precision matrix. Banded covariance estimation procedure does not guarantee positive definiteness.

Tapering

Tapering the covariance matrix is another possible way and it can preserve positive definiteness. $T(S) = S \circ T$ is a tapered estimator where S is the sample covariance matrix, T is the tapering matrix and '$\circ$' denotes the Hadamard product (element-wise product). Properties of Hadamard product suggest that $T(S)$ is positive definite if T is so. The banded covariance matrix is a special case of this with $T = ((1_{[|i-j|\leq l)}))$, which is not positive definite.

Thresholding

The most widely applicable element-wise regularization method is defined through Thresholding Operator. The regularized estimator is $T_\lambda(S) = ((s_{ij} I(s_{ij}) > \lambda)))$, where $S = ((s_{ij}))$ is the sample covariance matrix and $\lambda > 0$ is the threshold parameter. λ can be determined through cross validation. Although it is much simpler than other methods, like penalized lasso, it has one problem. The estimator preserves symmetry but not positive definiteness. With Gaussian assumption, consistency of this estimator can be shown uniformly over a class $\{\Sigma : \sigma_{ii} \leq C, \sum_{j=1}^{p} |\sigma_{ij}|^q \leq s_0(p), \forall i\}$ with $0 \leq q \leq 1$, $^{\log(p)}/_n = o(1)$ and $\lambda = M\sqrt{\frac{log(p)}{n}}$ for sufficiently large M [3]. For $q = 0$ the condition $\sum_{j=1}^{p} |\sigma_{ij}|^q \leq s_0(p)$ reduces to $\sum_{j=1}^{p} I(\sigma_{ij} \neq 0) \leq s_0(p)$. The rate of convergence is dependent on the dimension (p), sample size (n) and s_0, the determining factor of the number of nonzero elements in Σ. Similar result can be shown for precision matrix. For non-Gaussian case, we need some moment conditions in order to achieve consistency result [3].

The result goes through for a larger class of thresholding operators. One such is called generalized thresholding operators with the following three properties:

1. $|s_\lambda(x)| \leq |x|$ (shrinkage)
2. $s_\lambda(x) = 0$ for $|x| \leq \lambda$(thresholding)
3. $|s_\lambda(x) - x| \leq \lambda$(constraint on amount of shrinkage)

Apart from being consistent under suitable conditions discussed earlier, if variance of each variable is bounded then this operator is also "sparsistent" i.e. able to identify true zero entries of population covariance matrix with probability tending to one.

For both thresholding and generalized thresholding operators λ is fixed for all entries of the matrix. An adaptive threshold estimator can be developed [5] to have different parameters for different entries where

$$\lambda_{ij} \propto \sqrt{\frac{log(p)}{n} \hat{var}(Y_i - \mu_i)(Y_j - \mu_j)}$$

Approximate Factor Model

Sometimes the assumption of sparsity is too much to demand. For such situations estimation methods of a larger class of covariance matrices is required. A simple extension is possible to the class of matrices that can be decomposed into sum of low rank and sparse matrix: $\Sigma = FF^T + \Psi$, where F is low rank and Ψ is sparse matrix. Due to similarity with Factor model where Ψ is diagonal, this model is called approximate factor model. To estimate Σ, one can decompose S similarly as, $S = \sum_{i=1}^{q} \hat{\lambda}_i \hat{e}_i \hat{e}_i^T + R$, where the first part involves the first q principal components and the second part is residual. As R is sparse we can now use thresholding/adaptive thresholding operators to estimate it [6].

Positive Definiteness

Sometimes positive definiteness of the estimator is required in order to be used in classification or covariance regularised regression. As we have discussed thresholding estimators do not guarantee positive definite estimator. In the following section we will describe a couple of methods to achieve that. One possible way is to replace the negative eigenvalues in eigen decomposition of $\hat{\Sigma}$ by zero. But this manipulation destroys the sparse nature of the covariance matrix. An alternative way is necessary that will ensure sparsity and at the same time will produce positive definite output. Let us denote sample correlation matrix by R matrix $M \succ 0$ if it is symmetric and positive definite ($M \succeq 0$ for positive semi definite) and $M_{j,-j} = M_{-j,j} = j$th column of symmetric matrix M with it's jth element removed. $M_{-j,-j}$ =matrix formed after removing jth column and jth row of M and M^+ is the diagonal matrix with the same diagonal elements as M. Define $M^- = M - M^+$. Then a desirable positive definite estimator is

$$\hat{\Sigma}_\lambda = (S^+)^{\frac{1}{2}} \hat{\Theta}_\lambda (S^+)^{\frac{1}{2}}$$

where $S^+ = diag(S)$ and estimated correlation matrix is

$$\hat{\Theta} = \text{argmin}_{\Theta \succ 0} \|\Theta - R\|_F^2/2 - \tau log|\Theta| + \lambda|\Theta_-|$$

with $\lambda$and $\tau > 0$ respectively being tuning parameter and a fixed small value. The log-determinant term in the optimization function ensures positive definiteness. Regularizing the correlation matrix leads to faster convergence rate bound and scale invariance of the estimator. Under suitable and reasonable conditions this estimator is consistent [23]. For fast computation the following algorithm has been developed.

- Input Q- a symmetric matrix with positive diagonals, λ, τ and initialise $(\Sigma_0,\ \Omega_0)$ with $\Omega_0 > 0$. Follow steps $1-3$ for $j = 1, 2, \ldots, p$ and repeat till convergence.
  Step1: $\sigma_{jj}^{(k+1)} = q_{jj} + \tau\omega_{jj}^{(k)}$ and solve the lasso penalized regression:

$$\Sigma_{j,-j}^{(k+1)} = \text{argmin}_\beta \frac{1}{2}\beta^T\left(I + \frac{\tau}{\sigma_{jj}^{k+1}}\Omega_{-j,-j}^{(k)}\right)\beta - \beta^T Q_{-j,j} + \lambda\|\beta\|_1$$

  Step2: $\Omega_{j,-j}^{(k+1)} = -\Omega_{-j,-j}^{(k)}\Sigma_{j,-j}^{(k+1)}/\sigma_{jj}^{(k+1)}$.
  Step3: Compute $\omega_{jj}^{(k+1)} = \left(1 - \Sigma_{j,-j}^{(k+1)}\Omega_{j,-j}^{(k+1)}\right)\Big/\sigma_{jj}^{(k+1)}$.

An alternative estimator has been proposed based on *alternating direction method* [31]. If we want a positive semi definite matrix then the usual objective function along with l_1 penalty term should be optimized with an additional constraint for positive semi-definiteness:

$$\Sigma^+ = \text{argmin}_{\Sigma\succeq 0}\|\Sigma - S\|_F^2/2 + \lambda|\Sigma|_1.$$

For positive definite matrix we can replace the constraint $\Sigma \succeq 0$ with $\Sigma \succ \varepsilon I$ for very small $\varepsilon > 0$. Introducing a new variable Θ, we can write the same as

$$(\hat{\Theta}_+, \hat{\Sigma}_+) = \text{argmin}_{\Theta,\Sigma}\|\Sigma - S\|_F^2/2 + \lambda|\Sigma|_1 :\ \Sigma = \Theta,\ \Theta \succeq \varepsilon I.$$

Now it is enough to minimize its augmented Lagrangian function for some given penalty parameter μ:

$$L(\Theta,\ \Sigma;\ \Lambda) = \|\Sigma - S\|_F^2/2 + \lambda|\Sigma|_1 - < \Lambda,\ \Theta - \Sigma > + {}^{\|\Theta-\Sigma\|_F^2}/_{2\mu},$$

where Λ is the Lagrange multiplier. This can be achieved through the following algorithm ($\mathbf{S}$ being Soft Thresholding Operator):

- Input μ, Σ^0, Λ^0.
- Iterative alternating direction augmented lagrangian step: for the ith iteration:
  1. Solve $\Theta^{i+1} = (\Sigma^i + \mu\Lambda^i)_+$
  2. Solve $\Sigma^{i+1} = \{\mathbf{S}(\mu(S - \Lambda^i) + \Theta^{i+1}; \mu\lambda)\}/(1 + \mu)$
  3. Update $\Lambda^{i+1} = \Lambda^i - (\Theta^{i+1} - \Sigma^{i+1})/\mu$.
- Repeat the above cycle till convergence.

Precision Matrix Estimator

In some situations instead of covariance matrix, the precision matrix (Σ^{-1}) needs to be calculated. One example of such situation is financial network model using partial correlation coefficients because the sample estimate of partial correlation between

two nodes is $\hat{\rho}_{ij} = -\hat{\omega}_{ij}/\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}$, where $\hat{\omega}_{ij} = (\hat{\Sigma}^{-1})_{ij}$. Of course $\hat{\Sigma}^{-1}$can be calculated from $\hat{\Sigma}$ but that inversion involves $O(p^3)$ operations. For high dimensional data it is computationally expensive. On the other hand, if it is reasonable to assume sparsity of the precision matrix, that is, most of the off-diagonal elements of the precision matrix are zeros, then we can directly estimate the precision matrix. Although the correlation for most of the pairs of financial institutions would not be zero, the partial correlations can be. So this assumption of sparsity would not be a departure from reality in many practical situations. In such cases starting from a fully connected graph we can proceed in a *backward stepwise* fashion, by removing the least significant edges. Instead of such sequential testing procedure, some multiple testing strategy, for example, controlling for false discovery rate, can also be adopted. We discuss this in detail in section "Multiple Hypothesis Testing Problem and False Discovery Rate". After determining which off-diagonal entries of precision matrix are zeros (by either sequential or multiple testing procedure), maximum likelihood estimates of nonzero entries can be found by solving a convex optimization problem: maximizing the concentrated likelihood subject to the constraint that a subset of entries of precision matrix equal to zero [8, 22].

Alternatively, under Gaussian assumption a penalized likelihood approach can be employed. If $Y_1, ..., Y_p \sim N_p(0, \Sigma)$, the likelihood function is

$$L(\Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} exp\left(-\frac{1}{2}\sum_{i=1}^{n} Y_i' \Sigma^{-1} Y_i \right).$$

The penalized likelihood $l(\Sigma^{-1}) = log|\Sigma^{-1}| - tr(S\Sigma^{-1}) - \lambda\|\Sigma^{-1}\|_1$, with penalty parameter $\lambda > 0$, can be used to obtain a sparse solution [32]. The fastest algorithm to obtain the solution is called graphical lasso [12], described as follows:

1. Denote $\Theta = \Sigma^{-1}$. Start with a matrix W that can be used as a proxy of Σ. The choice recommended in Friedman et. al is $W = S + \lambda I$.
2. Repeat till convergence for $j = 1, 2, \ldots, p$:
   a. Partition matrix W in two parts, jth row and column, and the matrix $W_{11}-$ composed by the remaining elements. After eliminating the jth element, the remaining part of jth column ($p - 1$dimensional) is denoted as w_{12} and similarly the row is denoted as w_{21}. Similarly, define S_{11}, s_{12}, s_{21}, s_{22} for S matrix. (For $j = p$, the partition will look like: $W = \begin{pmatrix} W_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$ and $S = \begin{pmatrix} S_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$).
   b. Solve the estimating equations

$$W_{11}\beta - s_{12} + \lambda.Sign(\beta) = 0,$$

   using cyclical coordinate-descent algorithm to obtain $\hat{\beta}$.
   c. Update $w_{12} = W_{11}\hat{\beta}$.

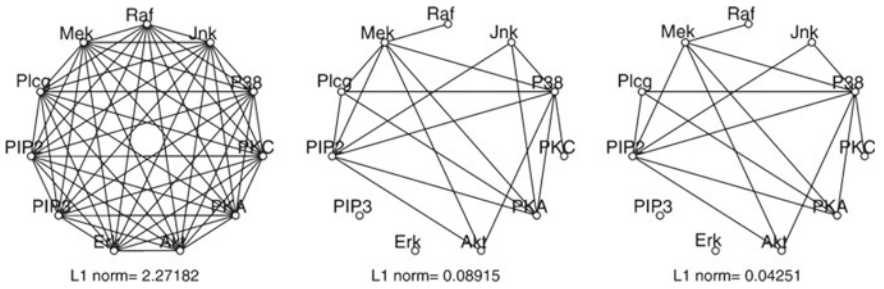


Fig. 11.4 Resulting networks given by graphical lasso algorithm for different values of penalty parameter lambda [12]

3. In the final cycle, for each j, solve for $\hat{\Theta}_{12} = -\hat{\beta}\hat{\Theta}_{22}$, with $\hat{\Theta}_{22}^{-1} = w_{22} - w_{12}'\hat{\beta}$. Stacking up $(\hat{\Theta}_{12}, \hat{\Theta}_{21})$ will give the jth column of Θ.

Figure 11.4 shows undirected graph from Cell-signalling data obtained through graphical lasso with different penalty parameters [12].

Multiple Hypothesis Testing Problem and False Discovery Rate

We can encounter large scale hypothesis testing problems in many practical situations. For example, in Section "Precision Matrix Estimator", we discussed to remove edge from a fully connected graph, we need to perform $p(p-1)/2$ testing problems- $H_{ij} : \rho_{ij} = 0$ vs $K_{ij} : \rho_{ij} \neq 0$. A detailed review can be found in [9].

Suppose we have N independent hypothesis $H_1, H_2, \ldots, H_N$ to test. In such situations it is important to control not only the type I error of individual hypothesis tests but also the overall (or combined) error rate. It is due to the fact that the probability of atleast one true hypothesis would be rejected becomes large: $1 - (1-\alpha)^N$, where α being the level of significance, generally taken as 0.05 or 0.01. The conventional way to resolve this problem is by controlling the familywise error rate (FWER)- $P(\cup_{i=1}^{N} H_{0i}$ is rejected when it is true). One example of such is Bonferroni correction. The problem with this procedure is that it is overly conservative and as a consequence the power of the test will be small. A much more liberal and efficient method for high dimension has been proposed by Benjamini and Hochberg [1]. In Fig. 11.5, out of N hypothesis tests in N_0 cases null hypothesis is true and in N_1cases null hypothesis is false. According the decision rule, in R out of N cases null hypothesis is rejected. Clearly R is observed but N_0 or $N_1$are not. The following algorithm controls the expected false discovery proportion:

1. The test statistic of $H_1, H_2, \ldots, H_N$ yield $p$values $p_1, \ldots, p_N$.
2. Order the p values $p_{(1)}, \ldots, p_{(N)}$.

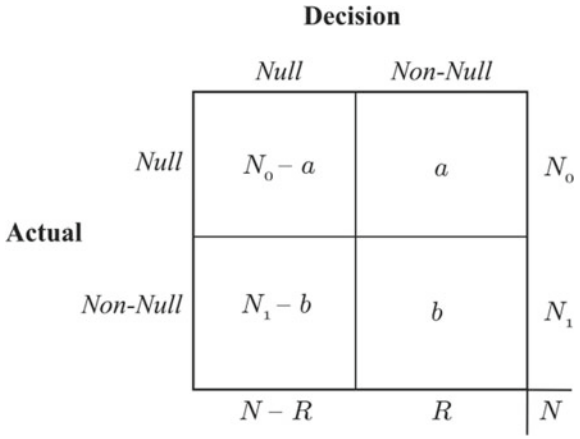


Fig. 11.5 False discovery rate a/R [10]

3. Rank the hypothesis $H_1, H_2, \ldots, H_N$ according to the $p$values.
4. Find largest j, say j^*, such that $p_j \leq \frac{j}{N}\alpha$.
5. Reject the top j^*tests as significant.

It can be shown that if the p values are independent of each other then the rule based on the algorithm controls the expected false discovery proportion by α, more precisely, $E(a/R) \leq \frac{N_0}{N}\alpha \leq \alpha$.

High Dimensional Regression

In financial econometrics one can often encounter multiple regression analysis problem. A large number of predictors implies large number of parameters to be estimated which reduces the degrees of freedom. As a result prediction error will be increased. So in high dimensional regression regularization is an essential tool.

In this section we will briefly discuss the multivariate regression problem with q responses and p predictors, which requires estimation of pq parameters in the regression coefficient matrix. Suppose the matrix of regressors, responses and coefficient matrix are X, Y and B respectively. As we know $\hat{B}_{OLS} = (X'X)^{-1}X'Y$ (under multivariate normality, this is also the maximum likelihood estimator) with pq parameters. Estimated covariance matrix (with $q(q+1)/2$ parameters) of Y is $\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})'(Y - X\hat{B})$. When p and q are large then both these estimators exhibits poor statistical properties. So here again, shrinkage and regularization of $\hat{B}$ would help to obtain a better estimator. It can be achieved through Reduced Rank Regression which attempts to solve a constrained least square problem:

$$\hat{B}_r = \underset{B:\ \mathrm{rank}(B)=r\leq \min(p,q)}{\operatorname{argmin}} tr[(Y - XB)'(Y - XB)]$$

The solution of this constrained optimization is $\hat{B}_r = (X'X)^{-1}X'YHH'$ where $H = (h_1, ..., h_r)$ with h_k being normalized eigenvector corresponding to the kth largest eigenvalue of the matrix $Y'X(X'X)^{-1}X'Y$. Choice of r is important because this is the parameter that balances between bias and variance of prediction.

Alternatively, a regularized estimator can be obtained by introducing a nonnegative penalty function in the optimization problem:

$$\hat{B} = \text{argmin}_B\{tr[(Y - XB)'(Y - XB)] + \lambda C(B)\}$$

when C is a scalar function and λ is nonnegative quantity. Most common choices of $C(B)$ are l_p norms. $C(B) = \sum_{j,k} |b_{jk}|$ leads to lasso estimate where as $C(B) = \sum_{j,k} b_{jk}^2$ amount to ridge regression. $C(B) = \alpha \sum_{j,k} |b_{jk}| + {}^{(1-\alpha)}/_2 \sum_{j,k} b_{jk}^2$ for $\alpha \in (0, 1)$ and $C(B) = \sum_{j,k} |b_{jk}|^\gamma$ for $\gamma \in [1, 2]$ are called elastic net and bridge regression respectively. Grouped lasso with $C(B) = \sum_{i=1}^{p} (b_{j1}^2 + \cdots + b_{jq}^2)^{0.5}$ imposes group-wise penalty on the rows of B, which may lead to exclusion of some predictors for all the responses.

All the above mentioned methods regularize the matrix B only while leaving Σaside. Although a little more complicated, it is sometimes appropriate to regularize both B and Σ^{-1}. One way to do that is to adding separate lasso penalty on B and Σ^{-1}in the negative log likelihood:

$$l(B, \Sigma) = \text{tr}\left[\frac{1}{n}(Y - XB)\Sigma^{-1}(Y - XB)\right] - \log|\Sigma^{-1}| + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{j,k} |b_{j,k}|$$

where λ_1 and λ_2 are as usual tuning parameter, $B = ((b_{jk}))$ and $\Sigma^{-1} = \Omega = ((\omega_{j',j}))$. This optimization problem is not convex but biconvex. Note that solving the above mentioned optimization problem for Ω with B fixed at B_0 reduced to the optimization problem:

$$\hat{\Omega}(B_0) = \text{argmin}\left\{tr(\hat{\Sigma}\Omega) - log|\Omega| + \lambda_1 \sum_{i \neq j} |\omega_{ij}|\right\}$$

where $\hat{\Sigma} = \frac{1}{n}(Y - XB_0)'(Y - XB_0)$. If we fix Ω at a non-negative definite Ω_0 it will lead to

$$\hat{B}(\Omega_0) = \text{argmin}_B\left\{tr\left[\frac{1}{n}(Y - XB_0)\Omega(Y - XB_0)'\right] + \lambda_2 \sum_{j,k} |b_{j,k}|\right\}$$

It can be shown that the original problem can be solved by using the following algorithm prescribed by [24]-

- Fix λ_1 and λ_2, initialize $\hat{B}^{(0)} = 0$ and $\hat{\Omega}^{(0)} = \hat{\Omega}(\hat{B}^{(0)})$.
  - Step 1: Compute $\hat{B}^{(m+1)} = \hat{B}(\hat{\Omega}^{(m)})$ by solving

$$\text{argmin}_B \left\{ tr\left(\frac{1}{n}(Y - XB)\Omega(Y - XB)\right) + \lambda_2 \sum_j \sum_k |b_{jk}| \right\}$$

by coordinate descent algorithm.

- Step 2: Compute $\hat{\Omega}^{(m+1)} = \hat{\Omega}(\hat{B}^{(m+1)})$ by solving

$$\text{argmin} \left\{ tr(\hat{\Sigma}\Omega) - log|\Omega| + \lambda_1 \sum_{i \neq j} |\omega_{ij}| \right\}$$

by Graphical lasso algorithm.

- Step 3: If $\sum_{i,j} |\hat{b}_{ij}^{(m+1)} - \hat{b}_{ij}^{(m)}| < \varepsilon \sum_{i,j} \hat{b}_{ij}^R$ where $((\hat{b}_{ij}^R))$ is the Ridge estimator of B.

Principal Components

In many high dimensional studies estimates of principal component loadings are inconsistent and the eigenvectors consists of too many entries to interpret. In such situation regularising the eigenvectors along with eigenvalues would be preferable. So it is desirable to have loading vectors with only a few non-zero entries. The simplest way to achieve that is through a procedure called SCoTLASS [16]. In this approach a lasso penalty is to be imposed on the PC loadings. So the first PC loading can be obtained by solving the optimization problem:

$$\text{maximize}_v vX'Xv \text{ subject to } ||v||_2^2 \leq 1,\ ||v||_1 \leq c$$

The next PC can be obtained by imposing extra constraint of orthogonality. Note that this is not a minimization problem and so can be difficult to solve. However the above problem is equivalent to the following one:

$$\text{maximize}_{u,v} u'Xv \text{ subject to } ||v||_1 \leq c,\ ||v||_2^2 \leq 1,\ ||u||_2^2 \leq 1$$

The equivalence between the two can be easily verified by using Cauchy Schwartz inequality to $u'Xv$ and noting that equality will be achieved for $u = \frac{X'v}{||X'v||_2}$. The optimization problem can be solved by the following algorithm [30]

- Initialize v to have l_2 norm 1.
- Iterate until convergence
  (a) $u \leftarrow \frac{Xv}{||Xv||_2}$
  (b) $v \leftarrow \frac{s(X'u,\Delta)}{||s(X'u,\Delta)||_2}$, where S is a soft thresholding operator, and $\Delta = 0$ if the computed v satisfies $||v||_1 \leq c$; otherwise $\Delta > 0$ with $||v||_1 = c$.

Classification

Suppose there are n independent observations of training data $(\mathbf{X}_i,\ Y_i)$, $i = 1(1)n$, coming from an unknown distribution. Here Y_i denotes the class of the ith observation and therefore can take values $\{1, 2, 3, \ldots, K\}$ if there are K classes. $\mathbf{X}_i$, generally a vector of dimension p, is the feature vector for the ith observation. Given a new observation $\mathbf{X}$, the task is to determine the class, the observation belongs to. In other words we have to determine a function from the feature space to $\{1, 2, \ldots, K\}$. One very widely used class of classifiers is distance based classifiers. It assigns an observation to a class k, if the observation is closer to class k on average compared to other classes i.e. $k = \text{argmin}_i \text{dist}(\mathbf{X}, \mu_i)$, where μ_i is the center of the feature space of class i. As an example if there are two classes and the feature distribution for the first class is $\mathbf{X} \sim N(\mu_1, \Sigma)$ and for the second class is $\mathbf{X} \sim N(\mu_2, \Sigma)$ then under the assumption that both the classes have equal prior probabilities, the most widely used distance measure is called Mahalanobis's distance

$$dist(X, \mu_k) = \sqrt{(X - \mu_k)\Sigma^{-1}(X - \mu_k)}, k = 1, 2.$$

So class 1 is to be chosen when

$$\sqrt{(X - \mu_1)\Sigma^{-1}(X - \mu_1)} \leq \sqrt{(X - \mu_2)\Sigma^{-1}(X - \mu_2)}$$

This technique is called Fisher Discriminant Analysis. For high dimensional data Fisher Discriminant Analysis does not perform well because it involves accurate estimation of precision matrix [2]. In the following section we will discuss some high dimensional classification methods.

Naive Bayes Classifier

Suppose we classify the observation, with feature $\mathbf{x}$, by some predetermined function δ i.e. $\delta(\mathbf{x}) \in \{1, 2, \ldots, K\}$. Now to judge the accuracy we need to consider some Loss function. A most intuitive loss function is the zero-one loss: $L(\delta(\mathbf{x}), Y) = I(\delta(\mathbf{x}) \neq Y)$, where $I(.)$ is the indicator function. Risk of δ is the expected loss-$E(L(\delta(\mathbf{x}), Y)) = 1 - P(Y = \delta(\mathbf{x})|\mathbf{X} = \mathbf{x})$. The optimal classifier, minimizing the risk, is $g(\mathbf{x}) = \arg\max_k P(Y = k|\mathbf{X} = \mathbf{x})$. If π be the prior probability of an observation being in class k then by Bayes Theorem $P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\pi(k)P(\mathbf{X}=\mathbf{x}|Y=k)}{\sum \pi(k)P(\mathbf{X}=\mathbf{x}|Y=k)}$. So $g(\mathbf{X}) = \arg\max_k \pi(k)P(\mathbf{X} = \mathbf{x}|Y = k)$. This is called Bayes classifier. When $\mathbf{X}$ is high dimensional $P(\mathbf{X}|Y)$ is practically impossible to estimate. The Naive Bayes Classifier works by assuming conditional independence: $P(\mathbf{X} = \mathbf{x}|Y = k) = \prod_i P(X_i = x_i|Y = k)$ where X_j is the jth component of $\mathbf{X}$. Naive Bayes classifier is used in practice even when the conditional independent assumption is not valid. In case of the previous example, under some conditions Naive Bayes classifier outper-

forms Fisher Discriminant function as long as dimensionality p does not grow faster than sample size n.

Centroid Rule and k-Nearest-Neighbour Rule

The centroid rule classifies an observation to kth class if its distance to the centroid of kth class is less than that to the centroid of any other class. The merit of this method is illustrated for $K = 2$. Suppose $n_1$and n_2 are fixed and $p \to \infty$ and within each class observations are iid. The observation of two classes can be denoted by $Z_1 = (Z_{11}, Z_{12}, \ldots, Z_{1p})$ and $Z_2 = (Z_{21}, \ldots, Z_{2p})$ respectively. With the assumption that as $p \to \infty$, $\frac{1}{p}\sum_{i=1}^{p} var(Z_{1i}) \to \sigma^2$, $\frac{1}{p}\sum_{i=1}^{p} var(Z_{2i}) \to \tau^2, \sigma^2/n_1 > \tau^2/n_2$ and $\frac{1}{p}\sum_{i=1}^{p}[E(Z_{1i}^2) - E(Z_{2i}^2)] \to \kappa^2$ a new observation is correctly classified with probability converging to 1 as $p \to \infty$ if $\kappa^2 \geq {}^{\sigma^2}/_{n_1} - {}^{\tau^2}/_{n_2}$[15].

The k-Nearest Neighbour rule determines the class of a new observations by help of k nearest data points from the training data. The new observation is to be assigned to the class closest to

$$g(\mathbf{X}) = \frac{1}{k} \sum_{i \mathbf{X}_i \in N_k(\mathbf{X})} Y_i$$

where $N_k(\mathbf{x})$ is the set k nearest points around $\mathbf{x}$.

Support Vector Machine

In Bayes classifier, as we discussed, one tries to minimize $\sum_i I(g(X) \neq Y)$, with respect to $g(.)$. But it is difficult to work with as indicator function is neither smooth nor convex. So one can think of using a convex loss function. Support vector machine (SVM) claims to resolve that problem. Suppose for binary classification problem, Y takes -1 and 1 to denote two classes. The SVM replaces zero-one loss by convex hinge loss $H(x) = [1 - x]_+$ where $[u]_+ = max\{0, u\}$, the positive part of u. The SVM tries to minimize $\sum_i H(Y_i g(\mathbf{X}_i)) + \lambda J(g)$ with respect to g. Here λ is a tunning parameter and J is a complexity penalty of g. If the minimizer is $\hat{g}$ then the SVM classifier is taken to be $sign(\hat{g})$. $J(.)$ can be taken as L_2 penalty.

It can be shown that the function minimizing $E(H(Yg(\mathbf{X})) + \lambda J(g))$ is exactly $sign\big(P(Y = +1|\mathbf{X}{=}\mathbf{x}) - \frac{1}{2}\big)$ [19]. In fact instead of working with $P(Y|\mathbf{X} = \mathbf{x})$ as in Bayes classifier SVM directly tries to estimate the decision boundary $\{\mathbf{x} : P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{2}\}$.

AdaBoost

Among the recently developed methodologies one of the most important is Boosting. It is a method that combines a number of 'weak' classifiers to form a powerful 'committee'. AdaBoost is the most commonly used boosting algorithm. An interesting result of this algorithm is that it is immune to over-fitting i.e. the test error decreases consistently as more and more classifiers are added. Suppose we have two classes represented as -1 and 1 and denoted by y. We have n training data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We want to produce a committee $F(x) = \sum_{i=1}^{M} c_m f_m(x)$ where f_m is a weak classifier (with values either 1 or -1) that predicts better than random guess. The initial classifiers f_m are trained on an weighted version of training sample giving more weight to the cases that are currently misclassified. The ultimate prediction will be based on $sign(F(x))$. This following algorithm is called discrete AdaBoost (as the initial classifier can take only two discrete values $+1$ and -1) [13].

Discrete AdaBoost Algorithm:

1. Start with weights $w_i = 1/n$ for $i = 1(1)n$
2. Repeat for $m = 1(1)M$:
   a. Fit the classifier $f_m(x) \in \{-1, 1\}$ with weights w_i on training data.
   b. Compute error $e_m = E_w[I_{(y \neq f_m(x))}]$, where E_w is the expectation over the training data with weights $w = (w_1, w_2, \ldots, w_n)$ and $I(.)$ is an indicator function.
   c. $c_m = \log(\frac{1-e_m}{e_m})$.
   d. Set $w_i \leftarrow w_i \exp[c_m I_{(y_i \neq f_m(x_i))}]$ for $i = 1, 2, \ldots, n$ and then renormalize to get $\sum_i w_i = 1$.
3. Final classifier: $sign(\sum_{m=1}^{M} c_m f_m(x))$

The base classifier $(f_m(.))$ of Discrete AdaBoost algorithm is binary. It can be generalized further to obtain a modification over discrete AdaBoost algorithm.

Real AdaBoost Algorithm:

1. Start with weights $w_i = 1/n$ for $i = 1(1)n$.
2. Repeat for $m = 1(1)M$:
   a. Fit the classifier to obtain the class probability estimate $p_m(x) = \hat{P}_w(y = 1|x) \in [0, 1]$, with weights $w = (w_1, w_2, \ldots, w_n)$ on the training data.
   b. Set $f_m(x) \leftarrow \frac{1}{2}\log(p_m(x)/(1 - p_m(x))) \in \mathbf{R}$.
   c. Set $w_i \leftarrow w_i \exp[-y_i f_m(x_i)]$, $i = 1, 2, .., n$ and renormalize so that $\sum_i w_i = 1$.
3. Finale classifier $sign[\sum_m f_m(x)]$.

It can be shown that AdaBoost method of classification is equivalent to fitting an additive logistic regression model in a forward stagewise manner [13].

Concluding Remarks

With the availability of high dimensional economic and financial data many classical statistical methods do not perform well. We have discussed some of the commonly encountered problems related to inference for high dimensional financial data. In many of the approaches significant improvement can be achieved by bias variance trade off. Some feasible solutions to the problems and some efficient algorithms are discussed. It is to be noted that there are many other challenges related to high dimensional data. Some solutions have been proposed based on simulation studies without desired theoretical justifications.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodological) 289–300 (1995)
2. Bickel, P.J., Levina, E., et al.: Some theory for fisher's linear discriminant function, naive bayes', and some alternatives when there are many more variables than observations. Bernoulli **10**(6), 989–1010 (2004)
3. Bickel, P.J., Levina, E., et al.: Covariance regularization by thresholding. Ann. Stat. **36**(6), 2577–2604 (2008)
4. Boginski, V., Butenko, S., Pardalos, P.M.: Statistical analysis of financial networks. Comput. Stat. Data Anal. **48**(2), 431–443 (2005)
5. Cai, T., Liu, W.: Adaptive thresholding for sparse covariance matrix estimation. J. Am. Stat. Assoc. **106**(494), 672–684 (2011)
6. Chamberlain, G., Rothschild, M.: Arbitrage, factor structure, and mean-variance analysis on large asset markets (1982)
7. Chen, Y., Wiesel, A., Hero, A.O.: Shrinkage estimation of high dimensional covariance matrices. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP, pp. 2937–2940. IEEE (2009)
8. Dempster, A.P.: Covariance selection. Biometrics 157–175 (1972)
9. Drton, M., Perlman, M.D.: Multiple testing and error control in gaussian graphical model selection. Stat. Sci. 430–449 (2007)
10. Efron, B.: Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, vol. 1. Cambridge University Press, Cambridge (2012)
11. Fisher, T.J., Sun, X.: Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Comput. Stat. Data Anal. **55**(5), 1909–1918 (2011)
12. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)
13. Friedman, J., Hastie, T., Tibshirani, R., et al.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. **28**(2), 337–407 (2000)
14. Haff, L.: Empirical bayes estimation of the multivariate normal covariance matrix. Ann. Stat. 586–597 (1980)
15. Hall, P., Marron, J.S., Neeman, A.: Geometric representation of high dimension, low sample size data. J. R. Stat. Soc.: Ser. B (Statistical Methodology) **67**(3), 427–444 (2005)
16. Jolliffe, I.T., Trendafilov, N.T., Uddin, M.: A modified principal component technique based on the lasso. J. Comput. Graph. Stat. **12**(3), 531–547 (2003)
17. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. **88**(2), 365–411 (2004)

18. Lin, S.P.: A monte carlo comparison of four estimators for a covariance matrix. Multivar. Anal. **6**, 411–429 (1985)
19. Lin, Y.: Support vector machines and the bayes rule in classification. Data Min. Knowl. Discov. **6**(3), 259–275 (2002)
20. Naul, B., Rajaratnam, B., Vincenzi, D.: The role of the isotonizing algorithm in stein's covariance matrix estimator. Comput. Stat. **31**(4), 1453–1476 (2016)
21. Papadimitriou, T., Gogas, P., Sarantitis, G.A.: European business cycle synchronization: a complex network perspective. In: Network Models in Economics and Finance, pp. 265–275. Springer, Berlin (2014)
22. Pourahmadi, M.: High-Dimensional Covariance Estimation: with High-dimensional Data, vol. 882. Wiley, New York (2013)
23. Rothman, A.J.: Positive definite estimators of large covariance matrices. Biometrika **99**(3), 733–740 (2012)
24. Rothman, A.J., Levina, E., Zhu, J.: Sparse multivariate regression with covariance estimation. J. Comput. Graph. Stat. **19**(4), 947–962 (2010)
25. Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. **4**(1) (2005)
26. Stein, C.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University Stanford, US (1956)
27. Stein, C.: Estimation of a covariance matrix, rietz lecture. In: 39th Annual Meeting IMS, Atlanta, GA (1975)
28. Stein, C., et al.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. Math. Stat. Probab. **1**, 197–206 (1956)
29. Vandewalle, N., Brisbois, F., Tordoir, X., et al.: Non-random topology of stock markets. Quant. Financ. **1**(3), 372–374 (2001)
30. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**(3), 515–534 (2009)
31. Xue, L., Ma, S., Zou, H.: Positive-definite 1-penalized estimation of large covariance matrices. J. Am. Stat. Assoc. **107**(500), 1480–1491 (2012)
32. Yuan, M., Lin, Y.: Model selection and estimation in the gaussian graphical model. Biometrika **94**(1), 19–35 (2007)

Chapter 12
Modeling Nelson–Siegel Yield Curve Using Bayesian Approach

Sourish Das

Abstract Yield curve modeling is an essential problem in finance. In this work, we explore the use of Bayesian statistical methods in conjunction with Nelson–Siegel model. We present the hierarchical Bayesian model for the parameters of the Nelson–Siegel yield function. We implement the MAP estimates via BFGS algorithm in `rstan`. The Bayesian analysis relies on the Monte Carlo simulation method. We perform the Hamiltonian Monte Carlo (HMC), using the `rstan` package. As a by-product of the HMC, we can simulate the Monte Carlo price of a Bond, and it helps us to identify if the bond is over-valued or under-valued. We demonstrate the process with an experiment and US Treasury's yield curve data. One of the interesting observation of the experiment is that there is a strong negative correlation between the price and long-term effect of yield. However, the relationship between the short-term interest rate effect and the value of the bond is weakly positive. This is because posterior analysis shows that the short-term effect and the long-term effect are negatively correlated.

Introduction

In financial applications, accurate yield curve modeling is of vital importance. Investors follow the bond market carefully, as it is an excellent predictor of future economic activity and levels of inflation, which affect prices of goods, stocks and real estate. The 'yield curve' is a curve showing the interest rates across different maturity spans (1 month, one year, five years, etc.) for a similar debt contract. The curve illustrates the relationship between the interest rate's level (or cost of borrowing) and the time to maturity, known as the 'term.' It determines the interest rate pattern, which you can use to discount the cash flows appropriately. The yield curve is a crucial representation of the state of the bond market. The short-term and long-term rates

S. Das (✉)
Chennai Mathematical Institute, India and Commonwealth Rutherford Fellow, University of Southampton, Southampton, UK
e-mail: sourish@cmi.ac.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_12

are usually different and short-term is lower than the long-term rates. The long-term rates are higher since the risk is more in long-term debt. The price of long-term bond fluctuates more with interest rate changes. The 'term structure' tells us, at a given time, how the yield depends on maturity. The most important factor in the analysis of the fixed-income asset is the yield curve. Any analysis of the fixed-income attribution requires evaluating how changes in the curve are estimated, and its impact on the performance of a portfolio. Some form of mathematical modeling of the yield curve is necessary, as it explains the curve's movement to extrapolate.

The slope of the yield curve is an essential indicator of short-term interest rates and is followed closely by investors [15]. As a result, this has been the center of significant research effort. Several statistical methods and tools, commonly used in econometrics and finance, are implemented to model the yield curve (see for example, [5, 9, 14, 20]). The [14] introduces a parametrically parsimonious model for yield curves that has the ability to represent the shapes generally associated with yield curves; monotonic, humped and $\mathscr{S}$-shaped.

Bayesian inference was applied to Dynamic Nelson–Siegel Model with Stochastic Volatility which models the conditional heteroscadasticity [11]. Bayesian inference for the stochastic volatility Nelson–Siegel (SVNS) model was introduced by [16]. This models the stochastic volatility in the underlying yield factors. Bayesian extensions to Diebold-Li term structure model involve the use of a more flexible parametric form for the yield curve [13]. It allows all the parameters to vary in time using a structure of latent factors, and the addition of a stochastic volatility structure to control the presence of conditional heteroskedasticity observed in the interest rates.

The Nelson–Selgel class of functions that produces the standard yield curve shapes associated with solutions to differential equations. If a differential equation produces the spot rates, then forward rates, being forecasts, will be the solution to the equations. Hence, the expectations theory of the term structure of the interest rates motivate investigating the Nelson–Siegel class. For example, if the immediate forward rate at maturity τ, denoted $r(\tau)$, is given by the solution to a second-order differential equation, with real and unequal roots, then we have

$$r(\tau) = \beta_0 + \beta_1 . \exp(-\tau/\lambda) + \beta_2 . [(\tau/\lambda) . \exp(-\tau/\lambda)].$$

The yield to maturity on a bond, denoted by $\mu(\tau)$, is average of the forward rates

$$\mu(\tau) = \frac{1}{\tau} \int_0^\tau r(\tau) d\tau,$$

the resulting function is popularly known as the Nelson–Siegel function [14], which has the form

$$\mu(\tau) = \beta_0 + (\beta_1 + \beta_2) \left\{ \frac{1 - \exp(-\tau/\lambda)}{\tau/\lambda} \right\} - \beta_2 \exp \left\{ -\frac{\tau}{\lambda} \right\}, \tag{12.1}$$

where

- β_0 is known as the long-run interest rate levels,
- β_1 is the short-term effect,
- β_2 is the midterm effect,
- λ is the decay factor.

The small value of λ leads to slow decay and can better fit the curve at longer maturities. Several literature [5, 9, 14, 20] reports that the model explains more than 90% variations in yield curve. The movement of the parameters through time reflects the change in the monetary policy of Federal Reserve and hence the economic activity. The high correlation indicates the ability of the fitted curves to predict the price of long term US Treasury bond. *Estimations and statistical inference about the parameters are extremely important, as each parameter space* $\theta = (\beta_0, \beta_1, \beta_2, \lambda)$ *of the model* (12.1), *has its own economic interpretation*. In this chapter, we discuss a Bayesian approach for the estimation of the yield curve and further inference.

Why Bayesian Method?

The Bayesian methods provides a consistent way of combining the prior information with data, within the decision theoretical framework. We can include past information about a parameter or hypothesis and form a prior distribution for the future analysis. When new observations become available, the previous posterior distribution can be used as the prior distribution. This inferences logically follow from the Bayes? theorem. The Bayesian analysis presents inferences that are conditional on the data and are exact, without dependence on asymptotic approximation. When the sample size is small, the inference proceeds in the same way, as if one had a large sample. The Bayesian analysis can estimate any functions of parameters directly, without using the 'plug-in' method.

In Bayesian inference, probability represents the degree of belief. In frequentist statistics, the probability means the relative frequency of an event. Therefore the frequentist method cannot assign the probability to a hypothesis (which is a belief), because a hypothesis is not an event characterized by a frequency. Instead, frequentist statistics can only calculate the probability of obtaining the data of an event, assuming a hypothesis is true. Therefore the Bayesian inference can calculate the probability that a hypothesis is valid, which is usually what the researchers want to know. By contrast, the frequentist statistics calculate the p-value, which is the probability of the more extreme data to obtain under the assumption that the null hypothesis is true. This probability, $\mathbb{P}$(data|null hypothesis is true), usually does not equal the probability that the null hypothesis is surely true.

Frequentist statistics has only one well-defined hypothesis—the null hypothesis and alternate hypothesis is simply defined as the 'null hypothesis' is wrong. However, the Bayesian method can have multiple well-defined hypotheses. The frequentist methods transform the data into a test statistics and the p-value, then com-

pares this value to an arbitrarily determined cutoff value and employs the decision-making approach to judge the significance, for example, reject the null hypothesis (or not reject) based on whether $p < \alpha$, where $0 < \alpha < 1$. The best way to do frequentist analysis would be to determine the sample size n before you start collecting the data. In Bayesian methods, because the probabilities represent the degrees of belief, it allows more nuanced and sophisticated analyses. We can calculate likelihoods and posterior probabilities for multiple hypotheses. We can enter data as we collect them; then update the degrees of belief so that we worry less about the arbitrary cut-off values. It makes sense to choose a hypothesis with maximum posterior probability, out of multiple hypotheses and not to worry about the arbitrary single value of significance. We can also do useful and straightforward analysis such as marginalizing over nuisance parameters, calculating likelihood ratio, or Bayes factor etc. In Bayesian methods, probability calculation follows the axiomatic foundation of the probability theory (e.g., the sum and product rules of the probability). By contrast, inferential frequentist method uses a collection of different test procedures that are not necessarily obtained from a coherent, consistent basis.

Having said that one should be aware of some possible disadvantages with the Bayesian methods. The prior distributions are often difficult to justify and can be a significant source of inaccuracy. There can be too many hypotheses, which may lead to the low posterior probability of each hypothesis, making the analysis sensitive to the choice of the prior distribution. The analytical solutions can be difficult to derive. Analytical evaluation of posterior inference can be intensive; but we can bypass this by using the state of the art Monte Carlo methods.

Bayesian Approach to Modeling

Bayesian approach to the statistical modeling follows three steps. First, we define the likelihood model, also known as the data model in some machine learning literature. In the second step, we describe the prior distributions, and the third step follows to obtain the posterior distribution model via Bayes theorem. Once we get the posterior model, all the Bayesian statistical inferences and predictions can be carried out based on the posterior model.

Prior Distribution

The prior probability distribution of an unknown parameter is the probability distribution that would express analysts beliefs about the quantity before any evidence is taken into consideration. We can develop a prior distribution, using a number of techniques [4] describe below.

1. We can determine a prior distribution, from past data, if historical data exits.
2. We can elicit prior distribution, from the subjective assessment of an experienced expert in the domain. For example if an expert believe that long-term interest rate

will never be more than 4%, then that can be used to define the prior distribution for β_0.
3. When no information is available, we can create an uninformative prior distribution to reflect a balance among outcomes.
4. The prior distribution can also be chosen according to some objective principle, such as the maximizing entropy for given constraints. For examples: the Jeffreys prior or Bernardo's reference prior. These priors are often known as the objective.
5. If the family of conjugate priors exists, then considering a prior from that family simplifies the further calculation.

Prior for Interest Rate Levels

In the Nelson–Siegel model as described in (12.1), the three interest rate parameters are: (i) β_0 is the long-run interest rate levels, (ii) β_1 represents short-term interest rate and (iii) β_2 represents medium-term interest rate. It is very rare that interest rate is negative. In fact, many argue if interest rate becomes negative then financial system collapse. So all practical purposes, we can assume that interest rates are positive and we can assume a prior probability distribution with its support only on the positive side of the real line. For example we can assume the inverse-gamma probability distribution over $\{\beta_0, \beta_1, \beta_2\}$. The fourth parameter of the model is the decay parameter λ, and it is natural to assign a prior distribution, the support of which is positive. Question is what would be a practical parameter value of the inverse gamma prior distribution? One choice could be the `Inverge-Gamma(a=1,b=1)`. The reason for such choice is if $\texttt{a} \leq 1$, then the moments of the `Inverge-Gamma` distribution does not exist. If one does not have the idea about the mean and variance of the parameters, then such choice of prior could be used. Having said that one could check the $\mathbb{P}[0 \leq \beta \leq 30] \approx 0.97$, that is the prior belief; and there is 97% chance that interest rates are below 30%. Such kind of probabilistic statement conveys a vague idea about the possible values of interest rate. We assumes that in the prior distributions the parameters are exchangeable and the there is no dependence among the parameters. So the first prior distribution we consider is

$$\pi(\beta_0, \beta_1, \beta_2, \lambda, \sigma) = \pi(\beta_0)\pi(\beta_1)\pi(\beta_2)\pi(\lambda)\pi(\sigma),$$

where $\beta_0, \beta_1, \beta_2, \lambda, \sigma \sim$ `Inverge-Gamma(a=1,b=1)`.

In the history of finance, the negative yield is though rare, it has occurred. Therefore it would be wise to consider an alternative model, which allows negative effect on yield. Therefore we consider `Normal`$(\mu_\beta, \sigma_\beta)$ prior distribution over the interest rate and `Inverge-Gamma(a=1,b=1)` over the σ_β, which models the scaling effect of the interest rate. Such kind of model is also known as the hierarchical Bayes model. The detail of the second model presented as Model 2 in the Table 12.1. Besides, we considered a slight variation of the Model 2 and named it as the Model 3. In the Model 3, we considered `Inverge-Gamma(a=0.1,b=0.1)` over the $\{\lambda, \sigma, \sigma_\beta\}$.

Table 12.1 Three different prior distributions for Nelson–Siegel Model. Note second and third model allow negative effect over yield on prior distributions

	Description
Model 1	$(\beta_0, \beta_1, \beta_2, \lambda, \sigma) \sim$ `Inverge-Gamma(a=1,b=1)`
Model 2	$(\beta_0, \beta_1, \beta_2) \sim$ `Gaussian`$(0, \sigma_\beta)$ $(\lambda, \sigma, \sigma_\beta) \sim$ `Inverge-Gamma(a=1,b=1)`
Model 3	$(\beta_0, \beta_1, \beta_2) \sim$ `Gaussian`$(0, \sigma_\beta)$ $(\lambda, \sigma, \sigma_\beta) \sim$ `Inverge-Gamma(a=0.1,b=0.1)`

Likelihood Function

Now we discuss one of the most important concept of Statistics, known as the 'likelihood'. Note that both frequestist and Bayesian statistics agrees that there has to be a data model or likelihood function to do any statistical inference. In order to understand the concept of the likelihood function, we consider a simple example.

Example

Suppose y is the number insurance claims that follow Poisson distribution

$$\mathbb{P}(Y = y) = e^{-\lambda}\frac{\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots; 0 < \lambda < \infty,$$

and in the dataset we have only one data points and that is $Y = 5$. We don't have any idea about λ. Different values of λ will result into different values of $\mathbb{P}[Y = 5]$. Note that $\mathbb{P}[Y = 5|\lambda = 3]$ denotes the probability of $Y = 5$, when $\lambda = 3$. Here we compare the probabilities for different values of λ,

$$\begin{aligned}
\mathbb{P}[Y = 5|\lambda = 3] &= e^{-3}\frac{3^5}{5!} \approx 0.101,\\
\mathbb{P}[Y = 5|\lambda = 4] &= e^{-4}\frac{4^5}{5!} \approx 0.156,\\
\mathbb{P}[Y = 5|\lambda = 5] &= e^{-5}\frac{5^5}{5!} \approx 0.175,\\
\mathbb{P}[Y = 5|\lambda = 6] &= e^{-6}\frac{6^5}{5!} \approx 0.161,\\
\mathbb{P}[Y = 5|\lambda = 7] &= e^{-7}\frac{7^5}{5!} \approx 0.128.
\end{aligned}$$

We can conclude that $\lambda = 5$ justifies the data almost 75% better than $\lambda = 3$. That is because if the value of λ is close to 5 then the the likelihood of the seeing the data 'Y = 5' is much higher than when $\lambda = 3, 4, 6$ or 7.

The model for the given data is presented as a function of the unknown parameter λ, is called likelihood function. The likelihood function can be presented as

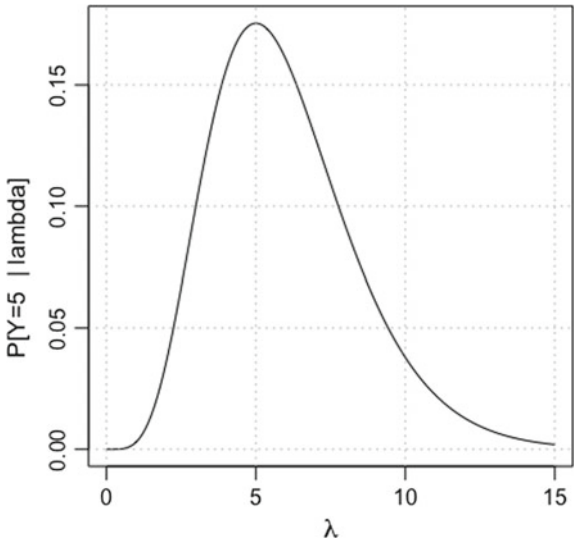


Fig. 12.1 The plot shows the most likely value of an unknown parameter λ, which generates the data $Y = 5$. The y-axis represents the likeliness of seeing the data $Y = 5$ for different possible values of λ. The curve is known as the likelihood curve

$$l(\lambda|Y = 5) = p(5|\lambda).$$

However, in reality we typically have the multiple observations like $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$. Then of course we have to look into the joint probability models of $\boldsymbol{y}$. The likelihood function for such model would be

$$l(\theta|y_1, y_2, \ldots, y_n) = p(Y_1 = y_1, \ldots, Y_n = y_n|\theta),$$

where θ is parameters of the model and θ could be scalar or vector, depending on the model (Fig. 12.1).

Likelihood Function for Nelson–Siegel Model

The Nelson–Siegel function is believed to be the model which explain the behaviour of the yield curve rate. Now it is expected there will be some random shock or unexplained error in the observed rate. Hence the expected data model would be

$$y_i(\tau_j) = \mu_i(\tau_j) + e_{ij}, \tag{12.2}$$

where

$$\mu_i(\tau_j) = \beta_0 + (\beta_1 + \beta_2)\left\{\frac{1 - \exp(-\tau_j/\lambda)}{\tau_j/\lambda}\right\} - \beta_2 \exp\left\{-\frac{\tau_j}{\lambda}\right\},$$

$j = 1, 2, \ldots, m$, $i = 1, 2, \ldots, n$, and $e_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$. Note that in the data model (12.2), it is the error or unexplained part which is stochastic or random. The assumption of the *independent and identically distributed* (aka. *i.i.d*) error provides us to assume that each observations y_{ij} independently follow

$$y_i(\tau_j) \overset{indep}{\sim} N\big(\mu_i(\tau_j), \sigma^2\big).$$

The likelihood function of the Nelson–Siegel function can be modeled as

$$l(\mathscr{D}|\theta) = \prod_{i=1}^{n}\prod_{j=1}^{m} p\big(\mu_i(\tau_j), \sigma^2\big), \tag{12.3}$$

where $\theta = (\beta_0, \beta_1, \beta_2, \lambda, \sigma^2)$ is the parameter vector needs to be estimated and $p(.)$ is the probability density function (pdf) of the Gaussian distribution, $\mathscr{D} = \{y_{11}, \ldots, y_{nm}, \tau_1, \ldots, \tau_m\}$ is the data or evidence.

Posterior Distribution

The posterior probability distribution of unknown parameters, conditional on the data obtained from an experiment or survey. The "Posterior," in this context, means after taking into account the relevant data related to the particular study. The posterior probability of the parameters θ given the data $\mathscr{D}$ is denoted as $p(\theta|\mathscr{D})$. On the contrary, the likelihood function is the probability of the evidence given the parameters is denoted as the $l(\mathscr{D}|\theta)$. The concepts are related via Bayes theorem as

$$\begin{aligned} p(\theta|\mathscr{D}) &= \frac{l(\mathscr{D}|\theta)\ p(\theta)}{\int_{\Theta} l(\mathscr{D}|\theta)\ p(\theta)\ d\theta} \\ &= \frac{l(\mathscr{D}|\theta)\ p(\theta)}{p(\mathscr{D})}. \end{aligned} \tag{12.4}$$

There are two points to note.

- The denominator of the (12.4), is free of θ. Therefore, often the posterior model is presented as proportion of likelihood times prior, i.e.,

$$p(\theta|\mathscr{D}) \propto l(\mathscr{D}|\theta)\ p(\theta).$$

- The integration of the in the denominator of (12.4) is high-dimensional integration problem. For example, in the Model 2 of the Table 12.1, there are six parameters in the θ, i.e., $\theta = \{\beta_0, \beta_1, \beta_2, \lambda, \sigma, \sigma_\beta\}$. So the integration will be a six-dimensional integration problem.
- Since having an analytical solution of the six-dimensional integration problem is almost impossible; we resort to Monte Carlo simulation methods.

Posterior Inference

In Bayesian methodology, the posterior model for the parameter θ, contains all the information. However, that is too much information to process. Hence we look for the summary statistics of the posterior probability distribution. The Bayesian estimation methods consider the measurements of central tendency of the posterior distribution as the representative value of the parameter. The measures are:

- **Posterior Median**: For one-dimensional problems, an unique median exists for the real valued parameters. The posterior median is also known as the robust estimator. If $\int_{\mathbb{R}} p(\theta \mid \mathscr{D}) < \infty$, then posterior median $\tilde{\theta}$ is

$$\mathbb{P}(\theta \leq \tilde{\theta} \mid \mathscr{D}) = \int_{-\infty}^{\tilde{\theta}} p(\theta \mid \mathscr{D})\, d\theta = \frac{1}{2}. \tag{12.5}$$

  The posterior median is Bayes estimator under squared error loss function [2].
- **Posterior Mean**: If there exists a finite mean for the posterior distribution, then we can consider the posterior mean as the estimate of the parameter, i.e.,

$$\hat{\theta} = \mathbb{E}(\theta \mid \mathscr{D}) = \int_{\Theta} \theta p(\theta \mid \mathscr{D})\, d\theta. \tag{12.6}$$

  The posterior mean is Bayes estimator under squared error loss function [2].
- **Posterior Mode**: The mode of the posterior distribution, also known as the maximum a posteriori probability (MAP) estimate,

$$\bar{\theta} = \arg\max_{\Theta} p(\theta \mid \mathscr{D}). \tag{12.7}$$

  The posterior mode is Bayes estimator under Kullback–Leibler type loss function [8].

Note that the posterior mean and the posterior median is an integration problem and the posterior mode is an optimization problem.

Posterior Analysis of US Treasury Yield Data

Here we present the Bayesian posterior analysis of the Nelson–Siegel model (12.1), with three different prior distribution models, presented in the Table 12.1. We worked all computational issues using the `rstan` package in `R` statistical software. In this analysis, considered only six days of data (from May 01, 2018 to May 08, 2018) presented in the Table 12.2. *Note that the purpose of this toy analysis is to demonstrate how the Bayesian analysis works*!

Table 12.2 US treasur yield curve rate ove first six business days of May 2018. [22]

Date	1 Mo	3 Mo	6 Mo	1 Yr	2 Yr	3 Yr	5 Yr	7 Yr	10 Yr	20 Yr	30 Yr
05/01/18	1.68	1.85	2.05	2.26	2.50	2.66	2.82	2.93	2.97	3.03	3.13
05/02/18	1.69	1.84	2.03	2.24	2.49	2.64	2.80	2.92	2.97	3.04	3.14
05/03/18	1.68	1.84	2.02	2.24	2.49	2.62	2.78	2.90	2.94	3.02	3.12
05/04/18	1.67	1.84	2.03	2.24	2.51	2.63	2.78	2.90	2.95	3.02	3.12
05/07/18	1.69	1.86	2.05	2.25	2.49	2.64	2.78	2.90	2.95	3.02	3.12
05/08/18	1.69	1.87	2.05	2.26	2.51	2.66	2.81	2.93	2.97	3.04	3.13

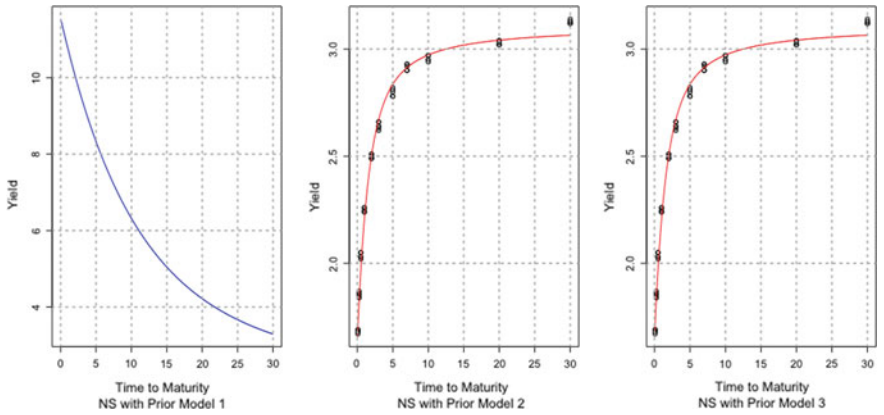


Fig. 12.2 Fitted Nelson Siegel yield curve with the three prior models described in Table 12.1. The fitted yield curve with the prior model 1 is unrealistic, indicates undesirable prior distribution. The fitted yield curve with prior model 2 and 3 show an excellent fit to data

The Fig. 12.2, exhibit the fitted Nelson Siegel yield curve with the three different prior distributions illustrated in Table 12.1. The fitted yield curve with the prior model 1 is unrealistic, indicates an unsatisfactory prior distribution. The fitted yield curve with prior model 2 and 3 show an excellent fit to data. The MAP estimates for the Nelson–Siegel parameters is presented in the Table 12.3, with the three prior distributions in Table 12.1. In case of the model 1, the long-run effect β_0 for the prior model 1, is less than the medium-term effect β_2, makes the prior model 1 an undesirable. The similar MAP estimates of the Nelson–Siegel parameters for model 2 and 3 indicates robust posterior analysis, in spite of differences in the prior parameters.

As we see the model 1 is really undesirable, hence we drop this model from the further discussion and we only focus our discussion on the hierarchical model considered in model 2 and 3. **Having said that one must note that the Bayesian methodology is not a magic bullet. A poorly chosen prior distribution may lead to an undesirable model**. The Monte Carlo estimates of the posterior mean, standard deviation, median, 2.5 and 97% quantile of the Nelson–Siegel parameters under the prior distribution model 2 and 3, under Table 12.4.

Table 12.3 The MAP estimates for the Nelson–Siegel parameters, with the three prior distributions in Table 12.1. Note that the long-run effect β_0 for the prior model 1, is less than the medium-term effect β_2, makes the prior model 1 an undesirable. The similar MAP estimates of the Nelson–Siegel parameters for model 2 and 3 indicates robust posterior analysis, in spite of differences in the prior parameters

	β_0	β_1	β_2	λ	σ	σ_β
Model 1	1.639	0.255	4.831	9.052	0.143	–
Model 2	3.111	−1.440	−0.016	0.950	0.043	1.636
Model 3	3.111	−1.440	−0.012	0.954	0.036	1.705

Table 12.4 Monte Carlo estimates of the posterior mean, standard deviation, median, 2.5 and 97% quantile of the Nelson–Siegel parameters under model 2 and 3

Parameters	Model	Mean	sd	2.5%	Median	97.5%
β_0	m2	3.11	0.01	3.08	3.11	3.14
	m3	3.11	0.01	3.09	3.11	3.13
β_1	m2	−1.44	0.02	−1.47	−1.44	−1.40
	m3	−1.44	0.01	−1.47	−1.44	−1.41
β_2	m2	0.00	0.16	−0.33	0.00	0.32
	m3	0.01	0.14	−0.27	0.01	0.27
λ	m2	0.97	0.12	0.76	0.96	1.23
	m3	0.97	0.10	0.79	0.97	1.18
σ	m2	0.05	0.00	0.04	0.05	0.06
	m3	0.04	0.00	0.03	0.04	0.05
σ_β	m2	2.32	1.18	1.11	2.02	5.43
	m3	2.70	1.81	1.13	2.21	7.36
lp	m2	155.84	1.82	151.57	156.18	158.32
	m3	177.40	1.80	173.18	177.72	179.87

We present the US Treasury yield curve data in the Fig. 12.3, and present the MAP estimate of the Nelson–Siegel parameters in the Fig. 12.4. We present the scatter plot of the daily MAP values of β_0 and β_1 of Nelson–Siegel Model in the Fig. 12.5. The plot indicates a negative relationship between the long and short-term effect. However, we assume independence among all the parameters in the prior distribution.

Dynamic Nelson–Siegel Model

The Dynamic Nelson–Siegel (DNS) model [5, 9, 14] for yield curve can be presented as

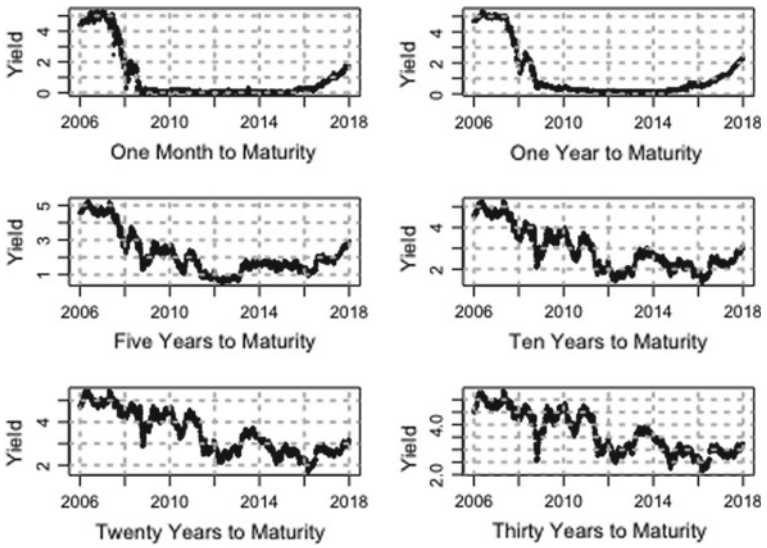


Fig. 12.3 US Tresury's yield rate (from 02-Oct-2006 to 08-May-2018) presented in six panels. Subtitle of each panel state the yield rate corresponding to the 'time to maturity'. The x-axis present the years and y-axis represent the yield rate

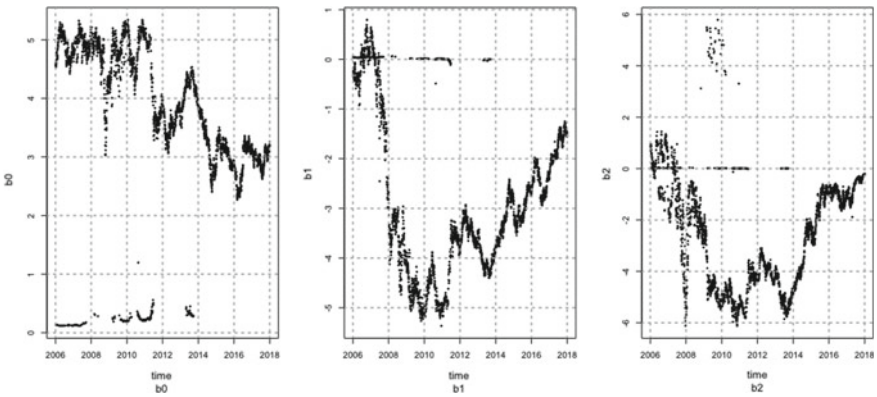


Fig. 12.4 The MAP estimates of the Nelson–Siegel Parameters from the US treasury yield data [22] and presented in the Fig. 12.3

$$y_t(\tau_j) = \beta_{1t} + \beta_{2t}\left(\frac{1-\exp\{-\tau_j/\lambda\}}{\tau_j/\lambda}\right) + \beta_{3t}\left(\frac{1-\exp\{-\tau_j/\lambda\}}{\tau_j/\lambda} - \exp\{-\tau_j/\lambda\}\right) + \varepsilon_t(\tau_j),$$

$$\varepsilon_t(\tau_j) \sim N(0, \sigma_\varepsilon^2),$$

$$\beta_{it} = \theta_{0i} + \theta_{1i}\beta_{i,t-1} + \eta_i, \quad i = 1, 2, 3, \quad \eta_i \sim N(0, \sigma_\eta^2), \quad t = 1, 2, \ldots, T, \quad j = 1, 2, \ldots, m,$$

where $y_t(\tau)$ is the yield for maturity τ (in months) at time t. The three factors β_{1t}, β_{2t} and β_{3t} are denoted as level, slope and curvature of slope respec-

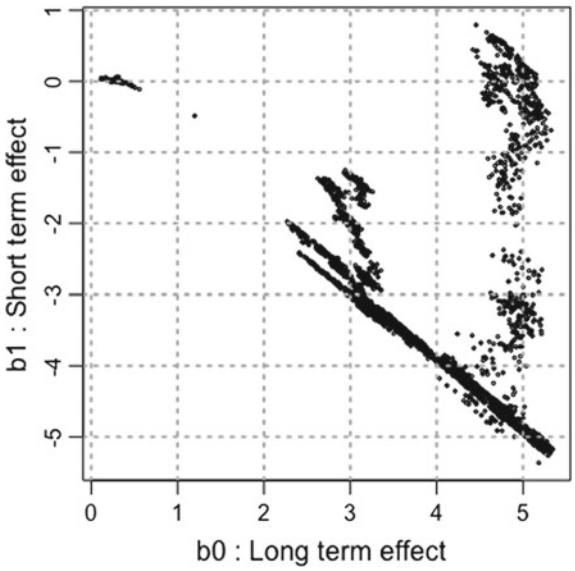


Fig. 12.5 Scatter plot of daily MAP values of β_0 and β_1 of Nelson–Siegel Model. The plot indicates a negative relation between the two parameters. However, we assume independence among all the parameters in the prior distribution

tively. Parameter λ controls exponentially decaying rate of the loadings for the slope and curvature. The goodness-of-fit of the yield curve is not very sensitive to the specific choice of λ [14]. Therefore [5] chose λ to be known. In practice, λ can be determined through grid-search method. There are eight static parameters $\boldsymbol{\theta} = (\theta_{01}, \theta_{02}, \theta_{03}, \theta_{11}, \theta_{12}, \theta_{13}, \sigma_\varepsilon^2, \sigma_\eta^2)$ in the model. In matrix notation the DNS model can be presented as

$$\boldsymbol{\beta}_t = \theta_0 + \boldsymbol{Z}\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \tag{12.8}$$

$$\boldsymbol{y}_t = \boldsymbol{\phi}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \tag{12.9}$$

where $\boldsymbol{y}_t = \begin{pmatrix} y_t(\tau_1) \\ y_t(\tau_2) \\ \vdots \\ y_t(\tau_m) \end{pmatrix}_{m\times 1}$, $\boldsymbol{\phi} = \begin{pmatrix} 1 & f_1(\tau_1) & f_2(\tau_1) \\ 1 & f_1(\tau_2) & f_2(\tau_2) \\ \vdots & \vdots & \vdots \\ 1 & f_1(\tau_m) & f_2(\tau_m) \end{pmatrix}_{m\times 3}$, $\boldsymbol{\beta}_t = \begin{pmatrix} \beta_{0t} \\ \beta_{1t} \\ \beta_{2t} \end{pmatrix}_{3\times 1}$, $\boldsymbol{\varepsilon}_t = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}_{m\times 1}$, such that $f_1(\tau_j) = \left(\frac{1-\exp\{-\tau_j/\lambda\}}{\tau_j/\lambda}\right)$ and $f_2(\tau_j) = \left(\frac{1-\exp\{-\tau_j/\lambda\}}{\tau_j/\lambda} - \exp\{-\tau_j/\lambda\}\right)$, $j = 1, 2, \ldots, m$, $\theta_0 = \begin{pmatrix} \theta_{01} \\ \theta_{02} \\ \theta_{03} \end{pmatrix}$ and $\boldsymbol{Z} = \begin{pmatrix} \theta_{11} & 0 & 0 \\ 0 & \theta_{12} & 0 \\ 0 & 0 & \theta_{13} \end{pmatrix}$. Note that $\boldsymbol{\varepsilon}_t \sim \boldsymbol{N}_m(0, \sigma_\varepsilon^2 \boldsymbol{I}_m)$ and $\boldsymbol{\eta}_t \sim \boldsymbol{N}_3(0, \sigma_\eta^2 \boldsymbol{I}_3)$. Note that (12.8) is *system equation* and

(12.9) is *observation equation*. If available, we can use the generalized linear models (GLM) to incorporate any additional predictor variable, see [6, 7].

Relation Between DNS Model and Kalman Filter

The term "Kalman filter" refers to recursive procedure for inference. A beautiful tutorial paper on the same was written by [12]. The key notion here is that given the data $\boldsymbol{Y}_t = (\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_1)$ inference about $\boldsymbol{\beta}_t$ and prediction about $\boldsymbol{y}_{t+1}$ can be carried via Bayes theorem, which can be expressed as

$$\mathbb{P}(\boldsymbol{\beta}_t|\boldsymbol{Y}_t) \propto \mathbb{P}(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{Y}_{t-1}) \times \mathbb{P}(\boldsymbol{\beta}_t|\boldsymbol{Y}_{t-1}). \tag{12.10}$$

Note that the expression on the left of Eq. (12.10) is the *posterior distribution* of $\boldsymbol{\beta}$ at time t, whereas the first and second expression on the left side of (12.10) is the *likelihood* and *prior distribution* of $\boldsymbol{\beta}$, respectively. At $t-1$, our knowledge about $\boldsymbol{\beta}_{t-1}$ is incorporated in the probability statement for $\boldsymbol{\beta}_{t-1}$:

$$(\boldsymbol{\beta}_{t-1}|\boldsymbol{Y}_{t-1}) \sim \boldsymbol{N}_3(\hat{\boldsymbol{\beta}}_{t-1}, \Sigma_{t-1}), \tag{12.11}$$

where $\hat{\boldsymbol{\beta}}_{t-1}$ and Σ_{t-1} are the expectation and the variance of $(\boldsymbol{\beta}_{t-1}|\boldsymbol{Y}_{t-1})$. In effect, (12.11) is the posterior distribution of $\boldsymbol{\beta}_{t-1}$. We now look forward to time t in two steps.

1. prior to observing $\boldsymbol{y}_t$,
2. posterior or after observing $\boldsymbol{y}_t$, and
3. inference about $\boldsymbol{y}_t^*$ at maturity τ^*.

Step 1: Prior to observing $\boldsymbol{y}_t$, our best choice for $\boldsymbol{\beta}_t$ is governed by the system Eq. (12.8) and is given as $\theta_0 + \boldsymbol{Z}\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t$. Since $\boldsymbol{\beta}_{t-1}$ is describe in (12.11), therefore

$$(\boldsymbol{\beta}_t|Y_{t-1}) \sim \boldsymbol{N}_3(\theta_0 + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1}\ ,\ R_t = \boldsymbol{Z}\Sigma_{t-1}\boldsymbol{Z}^T + \sigma_\eta^2 \boldsymbol{I}_3) \tag{12.12}$$

is the prior distribution of $\boldsymbol{\beta}$ at time t. In obtaining (12.12) we use the result for any constant B,

$$X \sim \boldsymbol{N}(\mu, \Sigma) \implies a + BX \sim \boldsymbol{N}(a + B\mu, B\Sigma B^T).$$

Step 2: On observing $\boldsymbol{y}_t$, our objective is to obtain the posterior $\boldsymbol{\beta}_t$ using (12.10). However, to do this, we need the likelihood $\mathscr{L}(\boldsymbol{\beta}_t|\boldsymbol{Y}_t)$, or equivalently $\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{Y}_{t-1})$. Let e_t is the error in predicting $\boldsymbol{y}_t$ from previous time point $t-1$; thus

$$\boldsymbol{e}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t = \boldsymbol{y}_t - \boldsymbol{\phi}\theta_0 - \boldsymbol{\phi}\boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1}. \tag{12.13}$$

Since, $\boldsymbol{\phi}$, $\boldsymbol{Z}$, θ_0 and $\hat{\boldsymbol{\beta}}_{t-1}$ are known, observing $\boldsymbol{y}_t$ is equivalent to observing $\boldsymbol{e}_t$. Therefore (12.10) can be expressed as:

$$\mathbb{P}(\boldsymbol{\beta}_t|\boldsymbol{y}_t, \boldsymbol{Y}_{t-1}) = \mathbb{P}(\boldsymbol{\beta}_t|\boldsymbol{e}_t, \boldsymbol{Y}_{t-1}) \propto \mathbb{P}(\boldsymbol{e}_t|\boldsymbol{\beta}_t, \boldsymbol{Y}_{t-1}) \times \mathbb{P}(\boldsymbol{\beta}_t|\boldsymbol{Y}_{t-1}),$$

where $\mathbb{P}(\boldsymbol{e}_t|\boldsymbol{\beta}_t, \boldsymbol{Y}_{t-1})$ is the likelihood. Using the fact that $\boldsymbol{y}_t = \boldsymbol{\phi}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t$, (12.13) can be expressed as $\boldsymbol{e}_t = \boldsymbol{\phi}(\boldsymbol{\beta}_t - \theta_0 - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1}) + \boldsymbol{\varepsilon}_t$, so that $\mathbb{E}(\boldsymbol{e}_t|\boldsymbol{\beta}_t, \boldsymbol{Y}_{t-1}) = \boldsymbol{\phi}(\boldsymbol{\beta}_t - \theta_0 - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1})$. Since, $\boldsymbol{\varepsilon}_t \sim \boldsymbol{N}_m(0, \sigma_\varepsilon^2 \boldsymbol{I}_m)$, it follows the likelihood as

$$(\boldsymbol{e}_t \mid \boldsymbol{\beta}_t, \boldsymbol{Y}_{t-1}) \sim \boldsymbol{N}_m(\boldsymbol{\phi}(\boldsymbol{\beta}_t - \theta_0 - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1}) \ , \ \sigma_\varepsilon^2 \boldsymbol{I}_m). \tag{12.14}$$

Now in order to find the posterior, we use the standard result of the Gaussian distribution ([1], pp. 28–30). If $X_1 \sim N(\mu_1, \Sigma_{11})$ and

$$(X_2|X_1 = x_1) \sim N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \ , \ \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}), \tag{12.15}$$

then

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \boldsymbol{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]. \tag{12.16}$$

In our case, lets consider $X_1 \iff \boldsymbol{\beta}_t$ and $X_2 \iff \boldsymbol{e}_t$. Since $(\boldsymbol{\beta}_t|Y_{t-1}) \sim \boldsymbol{N}_3(\theta_0 + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1} \ , \ R_t)$, we note that

$$\mu_1 \iff \theta_0 + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1} \text{ and } \Sigma_{11} \iff R_t.$$

If in (12.15), we replace X_1,X_2, μ_1 and Σ_{11} by $\boldsymbol{\beta}_t, \boldsymbol{e}_t, \theta_0 + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1}$ and R_t respectively and compare the result (12.14), then

$$\mu_2 + \Sigma_{21}R_t^{-1}(\boldsymbol{\beta}_t - \theta_0 - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1}) \iff \boldsymbol{\phi}(\boldsymbol{\beta}_t - \theta_0 - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1}),$$

so that $\mu_2 \iff \boldsymbol{0}$ and $\Sigma_{21} \iff \boldsymbol{\phi}R_t$; following the same method

$$\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \Sigma_{22} - \boldsymbol{\phi}R_t\boldsymbol{\phi}^T \iff \sigma_\varepsilon^2 \boldsymbol{I}_m,$$

so that $\Sigma_{22} \iff \boldsymbol{\phi}R_t\boldsymbol{\phi}^T + \sigma_\varepsilon^2 \boldsymbol{I}_m$. Under the result (12.15) and (12.16) the joint distribution of $\boldsymbol{\beta}_t$ and $\boldsymbol{e}_t$, given $\boldsymbol{Y}_{t-1}$ can be described as

$$\begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{e}_t \end{pmatrix} \Bigg| \boldsymbol{Y}_{t-1} \sim \boldsymbol{N}\left[\begin{pmatrix} \theta_0 + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} R_t & R_t^T\boldsymbol{\phi}^T \\ \boldsymbol{\phi}R_t & \boldsymbol{\phi}R_t\boldsymbol{\phi}^T + \sigma_\varepsilon^2 \boldsymbol{I}_m \end{pmatrix}\right].$$

So we have the posterior distribution of $\boldsymbol{\beta}_t$ at time point t is

$$(\boldsymbol{\beta}_t|\boldsymbol{Y}_t) = (\boldsymbol{\beta}_t|\boldsymbol{e}_t, \boldsymbol{Y}_{t-1}) \sim \boldsymbol{N}(\hat{\boldsymbol{\beta}}_t, \Sigma_t),$$

where

$$\hat{\boldsymbol{\beta}}_t = \mathbb{E}(\boldsymbol{\beta}_t|\boldsymbol{Y}_t) = \theta_0 + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1} + R_t^T\boldsymbol{\phi}^T[\boldsymbol{\phi}R_t\boldsymbol{\phi}^T + \sigma_\varepsilon^2\boldsymbol{I}_m]^{-1}\boldsymbol{e}_t, \quad (12.17)$$

and

$$\Sigma_t = R_t - R_t\boldsymbol{\phi}^T[\boldsymbol{\phi}R_t\boldsymbol{\phi}^T + \sigma_\varepsilon^2\boldsymbol{I}_m]^{-1}\boldsymbol{\phi}R_t.$$

Step 3: Now in order to predict yield at a new maturity point(s) τ^* we can simply plug-in $\hat{\boldsymbol{\beta}}_t$ in observation Eq. (12.9), i.e.,

$$\hat{y}_t(\tau^*) = \boldsymbol{\phi}(\tau^*)\hat{\boldsymbol{\beta}}_t. \quad (12.18)$$

Gaussian Process Prior

The Gaussian process prior for DNS is presented by [18]. This can be accomplished very easily by introducing a random component in observation Eq. (12.9). The modified observation equation is

$$\boldsymbol{y}_t = \boldsymbol{\phi}\boldsymbol{\beta}_t + W_t(\tau) + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{y}_t$, $\boldsymbol{\phi}$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\varepsilon}_t$ are defined as in (12.9) and $W_t(\tau) \sim \boldsymbol{N}_m(\boldsymbol{0}, \mathbf{K})$, where $\mathbf{K} = \rho(\tau, \tau')$. Following the structure of the GP model [17], at time point t is

$$\begin{aligned} f_t &\sim \boldsymbol{N}_m\left(\boldsymbol{\phi}\boldsymbol{\beta}_t, \mathbf{K}\right), \quad \boldsymbol{\varepsilon}_t \sim \boldsymbol{N}_m(0, \sigma_\varepsilon^2\boldsymbol{I}_m) \\ \boldsymbol{y}_t &\sim \boldsymbol{N}_m(\boldsymbol{\phi}\boldsymbol{\beta}_t, \mathbf{K} + \sigma_\varepsilon^2\boldsymbol{I}_m). \end{aligned} \quad (12.19)$$

We consider the same system equation as (12.8). Since the system equation is same, therefore step 1 and 2 for DNS with the Gaussian process prior would be same as in the section "Relation Between DNS Model and Kalman Filter".

Marginal Likelihood

It will be useful to compute the probability that DNS with a given set of parameters (prior distribution, transition and observation models) would produce an observed signal. This probability is known as the 'marginal likelihood' because it integrates out the hidden state variables $\boldsymbol{\beta}_t$, so it can be computed using only the observed data $\boldsymbol{y}_t$. The marginal likelihood is useful to estimate different static parameter choices using Bayesian computation technique.

It is easy to estimate the marginal likelihood as a side effect of the recursive filtering calculation. By the chain rule, the likelihood can be factored as the product of the probability of each observation given previous observations,

$$p(\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{t=0}^{T} p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \ldots, \boldsymbol{y}_0, \boldsymbol{\theta})$$

and because the Kalman filter describes a Markov process, all relevant information from previous observations is contained in the current state $(\boldsymbol{\beta}_t|\boldsymbol{Y}_{t-1})$. Note that $\boldsymbol{\theta}$ is the static parameter(s). Thus the marginal likelihood is given by

$$\begin{aligned}
p(\boldsymbol{Y}_T|\boldsymbol{\theta}) &= \prod_{t=0}^{T} p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}) d\boldsymbol{y}_t \\
&= \prod_{t=0}^{T} \int p(\boldsymbol{y}_t|\boldsymbol{\beta}_t) p(\boldsymbol{\beta}_t|\boldsymbol{Y}_{t-1}) d\boldsymbol{\beta}_t \\
&\quad \text{consider the likelihood (12.19) and prior at time t (12.12)} \\
&= \prod_{t=0}^{T} \int \boldsymbol{N}_m(\boldsymbol{y}_t; \boldsymbol{\phi}\boldsymbol{\beta}_t, \tilde{\mathbf{K}}) \boldsymbol{N}_3(\boldsymbol{\beta}_t; \hat{\boldsymbol{\beta}}_{t|t-1}\ ,\ R_t) d\boldsymbol{\beta}_t \\
&\quad \text{where } \tilde{\mathbf{K}} = \mathbf{K} + \sigma_\varepsilon^2 \boldsymbol{I}_m \text{ and } \hat{\boldsymbol{\beta}}_{t|t-1} = \theta_0 + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{t-1} \\
&= \prod_{t=0}^{T} \boldsymbol{N}_m(\boldsymbol{y}_t; \boldsymbol{\phi}\hat{\boldsymbol{\beta}}_{t|t-1}, \tilde{\mathbf{K}} + \boldsymbol{\phi} R_t \boldsymbol{\phi}^T), \\
&= \prod_{t=0}^{T} \boldsymbol{N}_m(\boldsymbol{y}_t; \boldsymbol{\phi}\hat{\boldsymbol{\beta}}_{t|t-1}, \boldsymbol{S}_t), \ \text{where} \boldsymbol{S}_t = \tilde{\mathbf{K}} + \boldsymbol{\phi} R_t \boldsymbol{\phi}^T
\end{aligned}$$

i.e., product of multivariate normal densities. This can easily be calculated as a simple recursive update. However, to avoid numeric underflow, it is usually desirable to estimate the log marginal likelihood $l = \log p(\boldsymbol{Y}_T|\boldsymbol{\theta})$. We can do it via recursive update

$$l^{(t)} = l^{(t-1)} - \frac{1}{2}\Big\{ \ln|\boldsymbol{S}_t| + m \ln 2\pi + (\boldsymbol{y}_t - \boldsymbol{\phi}\hat{\boldsymbol{\beta}}_{t|t-1}) \boldsymbol{S}_t^{-1} (\boldsymbol{y}_t - \boldsymbol{\phi}\hat{\boldsymbol{\beta}}_{t|t-1})^T \Big\}.$$

Note that computation of $\boldsymbol{S}_t^{-1}$ involves the complexity of $O(n^3)$, where n is the number of data point. For fast GP regression, see [19].

Computational Issues

In general, it is impossible to obtain explicit analytical form for MAP (12.7), posterior mean (12.6), or posterior median (12.5). This implies that we have to resort to numerical methods, such as the Monte Carlo, or optimization subroutine. We implement the optimization for MAP estimates using the BFGS method. This method has

the time complexity of $O(p^2)$ per iteration where p is the number of parameters. The order of convergence for BFGS method is super-linear.

We implement the posterior mean, posterior median, via the Hamiltonian Monte Carlo (HMC) algorithm for hierarchical models [3, 10], using the `rstan` software [21]. The `rstan` can also implement the BFGS optimization method.

Monte Carlo Pricing of Bond

An interesting by-product of the Monte Carlo method is it helps us to estimate the theoretical price of a bond. We know the price of a bond is a non-linear function of the yield curve. That is

$$P = f\big(Y(\tau, \theta)\big), \tag{12.20}$$

where P is the price of the bond and $Y(\tau)$ is the yield curve modeled by the Nelson–Siegel function (12.1). Suppose $\{\theta_1^*, \theta_2^*, \ldots, \theta_M^*\}$ are the Monte Carlo simulation of the θ in (12.1). Then we can plug-in each of the θ_i^* in the pricing Eq. (12.20) and we can get the Monte Carlo price $\{P_i^* \mid i = 1, 2, \ldots, M\}$, where M is the simulation size. Now we can estimate the posterior mean, median and $100 \times (1 - \alpha)\%$ confidence interval for the price of the bond. If P_l is the lower bound and P_u is the upper bound of the interval, and if the 'traded price' is below the P_l then that will indicate that the bond is undervalued. Similarly, if the 'traded price' is above the P_u, then that will indicate that the bond is overvalued.

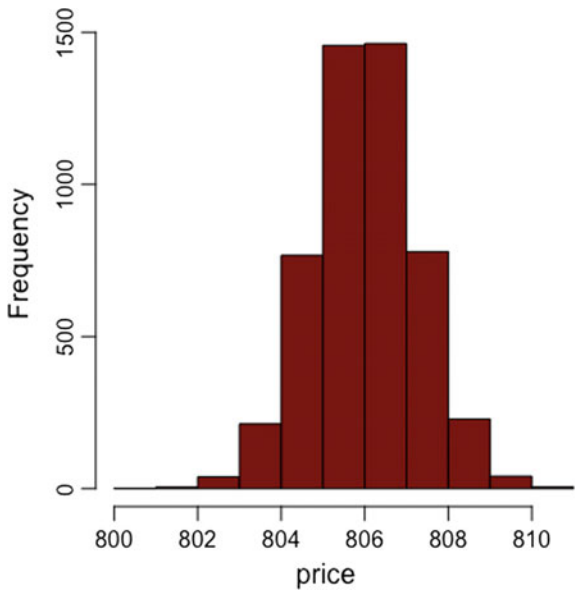


Fig. 12.6 Histogram of 5000 Monte Carlo price of the bond

Table 12.5 Posterior summary of bond price

Posterior mean	Posterior median	95% Posterior confidence interval
806.01	806.01	(803.59, 808.46)

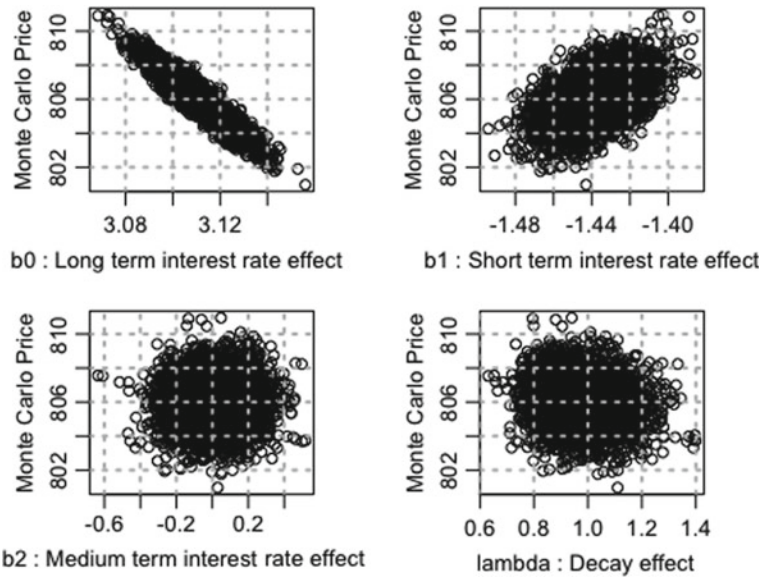


Fig. 12.7 Relationship between the Monte Carlo price of the bond and the parameters of the Nelson–Siegel model. We can see the strong negative correlation between the price and long-term effect of yield, i.e., β_0 and a weak positive correlation between short-term interest rate effect β_1 and the value of the bond

Experiment: We demonstrate the concept with a simple experiment. Suppose we have a bond which will mature in 15 years, pays coupon twice a year at 4% annual rate with a par value of $ 1000. If we have yield data as presented in Table 12.2; what would be the Bayesian price of the bond on May 9th, 2018?

We considered the prior model 3 presented in the Table 12.1 for this task. The parameter values of the Nelson–Siegel model were simulated from the posterior model using the No-U turn sampler of the HMC algorithm via `rstan` package. Then we calculate the 5000 Monte Carlo price of the bond and present histogram of the 5000 simulated price in the Fig. 12.6. We present the posterior summary of the bond price in the Table 12.5. The expected price is $ 806.01, and the traded price should stay within ($ 803.59, $ 808.46). If the traded price goes below $ 803.59, then we can consider the bond to be undervalued; while if the traded price goes above $ 808.46 then we can consider the bond to be over-priced. The Fig. 12.7 presents an exciting relationship between the Monte Carlo price of the bond and parameters of the Nelson–Siegel function. We can see the strong negative correlation between the price and long-term effect of yield, i.e., β_0 and a weak positive correlation between short-term interest rate effect and the value of the bond.

Conclusion

In this work, we present the hierarchical Bayesian methodology to model the Nelson–Siegel yield curve model. We demonstrate that ad-hoc choice of prior may lead to undesirable results. However, the proposed the hierarchical Bayesian method is much more robust and deliver the desired effect. We used BFGS algorithm in `rstan` for the MAP estimates of the Nelson–Siegel's parameters. We also implemented full Bayesian analysis using the HMC algorithm available in `rstan` package. As a by-product of the HMC, we simulate the Monte Carlo price of a Bond, and it helps us to identify if the bond is over-valued or under-valued. We demonstrate the process with example and US treasury's yield curve data. One interesting finding is that there is a strong negative correlation between the price and long-term effect of yield, i.e., β_0. However, the relationship between the short-term interest rate effect and the value of the bond is weakly positive. This is phenomenon is observed because the posterior analysis shows an inverse relationship between the long-term and the short-term effect of the Nelson–Siegel model.

References

1. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. Wiley, New York (1984)
2. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1985)
3. Betancourt, M., Girolami, M.: Hamiltonian Monte Carlo for hierarchical models. Technical report (2013). https://arxiv.org/abs/1312.0906
4. Carlin, B.P., Louis, T.A.: Bayesian Methods for Data Analysis, 3rd edn. CRC Press, New York (2008)
5. Chen, Y., Niu, L.: Adaptive dynamic nelson-siegel term structure model with applications. J. Econ. **180**(1), 98–115 (2014)
6. Das, S., Dey, D.: On bayesian analysis of generalized linear models using jacobian technique. Am. Stat. **60**, (2006). https://doi.org/10.1198/000313006X128150
7. Das, S., Dey, D.: On dynamic generalized linear models with applications. Methodol. Comput. Appl. Probab. **15**, (2013). https://doi.org/10.1007/s11009-011-9255-6
8. Das, S., Dey, D.K.: On bayesian inference for generalized multivariate gamma distribution. Stat. Probab. Lett. **80**, 1492–1499 (2010)
9. Diebold, F., Li, C.: Forecasting the term structure of government bond yields. J. Econ. **130**(1), 337–364 (2006)
10. Hoffman, M.D., Gelman, A.: The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. **15**, 1593–1623 (2014)
11. Joao, F.C., Laurin., M.P., Marcelo, S.P.: "Bayesian Inference Applied to Dynamic Nelson-Siegel Model with Stochastic Volatility" Brazilian Review of Econometrics, Sociedade Brasileira de Econometria - SBE, **30**(1), (2010)
12. Meinhold, R.J., Singpurwalla, N.D.: Understanding the kalman filter. Am. Stat. **37**(2), 123–127 (1983)
13. Mrcio, P.L., Luiz, K.H.: Bayesian extensions to diebold-li term structure model. Int. Rev. Financ. Anal. **19**, 342–350 (2010)
14. Nelson, C.R., Siegel, A.F.: Parsimonious modeling of yield curve. J. Bus. **60**(4), 473–489 (1987)
15. Nielsen, B.: Bond yield curve holds predictive powers treasury rates (2017). http://www.investopedia.com/articles/economics/08/yield-curve.asp

16. Nikolaus, H., Fuyu, Y.: Bayesian inference in a stochastic volatility Nelson–Siegel model. Comput. Stat. Data Anal. **56**, 3774–3792 (2012)
17. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge (2005)
18. Sambasivan, R., Das, S.: A statistical machine learning approach to yield curve forecasting. IEEE proceedings ICCIDS-2017 (2017)
19. Sambasivan, R., Das, S.: Fast gaussian process regression for big data. Big Data Res. (2018). https://doi.org/10.1016/j.bdr.2018.06.002
20. Spencer Hays, H.S., Huang, J.Z.: Functional dynamic factor models with applications to yield curve forecasting. Ann. Appl. Stat. **6**(3), 870–894 (2012)
21. Team., S.D.: Stan modeling language users guide and reference manual (2016). http://mc-stan.org/documentation/
22. Us treasury yield curve rates. https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=yield. Accessed 08-May-2018

Chapter 13
Pareto Efficiency, Inequality and Distribution Neutral Fiscal Policy—An Overview

Sugata Marjit, Anjan Mukherji and Sandip Sarkar

Abstract A structure of taxes and transfers that keep the income distribution unchanged even after positive or negative shocks to an economy, is referred as a Distribution Neutral Fiscal Policy. Marjit and Sarkar (Distribution neutral welfare ranking-extending pareto principle, 2017, [14]) referred this as a Strong Pareto Superior (SPS) allocation which improves the standard Pareto criterion by keeping the degree of inequality, not the absolute level of income, intact. In this paper we show the existence of a SPS allocation in a general equilibrium framework, and we provide a brief survey of distribution neutral fiscal policies existing in the literature. We also provide an empirical illustration with Indian Human Development Survey data.

Introduction

Marjit [15] has argued that researchers must look for robust results in economic theory that hold across space and time. This is likely to bring economics closer to physical sciences. In this paper we review one such robust result also referred as the "*Distribution Neutral Fiscal Policy*" (DNFP). By DNFP we mean a structure of taxes and transfers that keep the income distribution unchanged even after positive or negative shocks to an economy. The idea of distribution neutrality evolved in different areas of economics in different forms. For example, in a classic article [3] uses this approach in the context of poverty decomposition analysis. Very recently,

S. Marjit (✉)
Reserve Bank of India Professor of Industrial Economics, Centre for Studies in Social Sciences, Calcutta (CSSSC), R1, B. P. Township, Kolkata, India
e-mail: marjit@gmail.com

A. Mukherji
Professor Emeritus, Centre for Economic Studies and Planning, Jawaharlal Nehru University, New Delhi, India
e-mail: anjan.mukherji@gmail.com

S. Sarkar
CTRPFP, Centre for Studies in Social Sciences, Calcutta, R1, B. P. Township, Kolkata, India
e-mail: sandip.isi.08@gmail.com

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_13

Marjit et al. [14] generalized the notion of Pareto efficiency by focusing on Pareto efficient allocations that do not aggravates inequality. In this paper we discuss these results.

The Pareto ranking or Pareto efficiency is a topic economists are exposed to very early in their career. In particular the basic welfare comparison between two social situations starts with the ranking in terms of a principle Pareto had talked about in the nineteenth century. If we compare two social situations A and B, we say B is Pareto superior to A iff everyone is as well of in B as in A and at least one is strictly better off in situation B compared to A. This comparison is done in terms of utility or welfare levels individuals enjoy in A and B. In the theory of social welfare this has been a widely discussed topic with seminal contributions from [1, 4, 19, 20, 23] and recent treatments include [2, 12] etc.

Pareto's principle provides a nice way to compare situations when some gain and some lose by considering whether transfer from gainers to losers can lead to a new distribution in B such that B turns out to be Pareto superior to A, the initial welfare distribution. It is obvious that if sum of utilities increases in B relative to A, then whatever be the actual distribution in B, a transfer mechanism will always exist such that transfer-induced redistribution will make B Pareto superior to A. The great example is how gains from international trade can be redistributed in favor of those who lose from trade such that everyone gains due to trade. Overall gains from trade lead to a higher level of welfare under ideal conditions and therefore one can show that under free trade eventually nobody may lose as gainers *'bribe'* the losers. But whatever it is Pareto ranking definitely does not address the inequality issue. There will be situations where B will be Pareto superior to A, but inequality in B can be much greater than A. The purpose of this survey paper is to extend the basic principle of Pareto's welfare ranking subjecting it to a stricter condition that keeps the degree of inequality intact between A and B after transfer from gainers to losers, but at the same time guaranteeing that everyone gains in the end.

In this paper we also provide the existence and uniqueness of SPS in a general equilibrium framework. Considering the case with only two agents we show that an SPS allocation would always exist on the contract curve i.e. in the set of Pareto efficient allocations. Further, we also show that this point is unique. We also provide an empirical illustration with Indian Human Development Survey data.

The paper is organized as follows. In section "Inequality Measurement" we present a brief discussion on the measurement of income inequality. In section "Strong Pareto Superiority" we present a discussion on Strong Pareto Superiority allocations. In the next section we discuss other Distribution Neutral Fiscal policies that exists in the literature. In section "Empirical Illustration" we provide an empirical illustration with Indian data. The paper is concluded in section "Conclusion".

Inequality Measurement

In this section, we provide a discussion on the measurement of income inequality. We begin with some well-known axioms or postulates that any income inequality measure must satisfy. We then discuss some standard inequality measures proposed in the literature. We also provide a discussion on the Lorenz curve that stands as a basic tool for inequality ordering.

Inequality index is a scalar measure of interpersonal income differences within a given population. The class of inequality measures can be classified in two broad types, namely the relative and the absolute inequality measures. The relative inequality measures satisfy the scale invariance property, i.e., the inequality measure should remain unchanged if we multiply income of all individuals by a positive scalar. The absolute inequality measures are translation invariant, i.e., these inequality measures should remain unchanged if we change the origin by a positive scalar. has referred the relative and absolute the inequality measures as the rightists and leftists inequality measures, respectively.

The leftist inequality measures assigns a higher weight to the bottom of the income distribution. This can be illustrated as follows: consider the following income distribution $Y = \{1, 100\}$. Now following the relative inequality measure inequality should remain unchanged if we multiply the income of both individuals by a positive scalar. If we choose this scalar as 10, the new distribution becomes $Y^* = 10, 1000$. If we observe Y and Y^* closely, the poorer individual has gained only 9, whereas the richer individual has gained 900. The relative inequality remains unchanged despite the fact that absolute income differences between individuals increase. Notwithstanding this problem, the relative inequality index is widely used. This is because it is simple and because income inequality measures do not depend on the choice of the units. That is income inequality measured in dollars or pounds will exhibit same values.

We now discuss the following axioms that both absolute and relative inequality measures are expected to satisfy. We begin with the transfer axiom.

Transfer Axiom: A progressive transfer of income is defined as a transfer of income from a person to anyone who has a lower income so that the donor does not become poorer than the recipient. A regressive transfer is defined as a transfer from a person to anybody with a higher income, keeping their relative positions unchanged. This axiom requires that inequality should decline and increase as a result of progressive and regressive transfer, respectively. If the income distribution is ordered either in an ascending or in a descending order, then the transfers cannot change the rank orders of the individuals. Hence, these transfers are sometimes referred as rank preserving transfers.

Symmetry: This axiom requires that an income measure should not distinguish individuals by anything other than their incomes. This axiom is also referred as an anonymity axiom.

Normalization axiom: This axiom states that if for a society income of all individuals are same then there is no inequality. Eventually any inequality index that satisfies this axiom should take the value zero. This axiom always ensures that an inequality measure takes non-negative values.

Population Replication Invariance: This axiom states that inequality measurement should be invariant for replication of incomes. This implies that any inequality measure satisfying this axiom should be invariant between $Y = (2, 3)$ and $Y^* = (2, 3, 2, 3, 2, 3)$. This is because Y^* is obtained following three-fold replication of incomes of Y. Following this axiom, we can compare two income distributions with different population size. For example, consider $Y = (2, 3)$ and $X = (1, 2, 3)$. If the inequality index is population invariant, then we can compare $Y^* = (2, 3, 2, 3, 2, 3)$ and $X^* = (1, 2, 3, 1, 2, 3)$.

A well known relative inequality measure that satisfies all the axioms listed above is the Gini coefficient. Variance is an absolute inequality measure that also satisfies all the axioms discussed so far.

Researchers often expressed interest on whether different inequality indices can rank alternative distributions of income in the same way. One may consider the Lorenz curve in order to address this issue. Before we address this issue let us formally define Lorenz curve. Assume that the income distribution $Y_t = (y_1, y_2, \ldots, y_n)$ is designed in an ascending order. Lorenz curve represents the share of the total income enjoyed by the bottom p% of the population. The Lorenz curve is defined as the plot $L(Y_t, k/n) =$ against p where $p = k/n$. Note that 0% of the population enjoys 0% of the total income. Further, 100% of the population possesses the entire income. Hence, the curve starts from the south-west corner with coordinates (0,0) of the of unit square and terminates at the diametrically opposite north-east corner with coordinates (1,1). In the case of perfect equality, Lorenz curve coincides with the diagonal line of perfect equality, which is also referred to as the egalitarian line. This follows from the fact that if everybody has equal income then every p% of the population enjoys p% of the total income. In all other cases the curve will lie below the egalitarian line. If there is complete inequality, i.e., in a situation where only one person has positive income and all other persons have zero income, the curve will be L shaped. That is the curve will run through the horizontal axis until we reach the richest person where it will rise perpendicularly. In the context of Lorenz curve one important concept is that of Lorenz dominance. Any income distribution Yis said to be Lorenz dominant income distribution X if the Lorenz curve of Y lies strictly above X for at least one point and not below X at any of the point. In this context, we can also refer that inequality in Yis less than that of X, for all relative inequality measures that satisfies Symmetry, Transfer, and Population Replication Invariance. These class of measures is also referred as Lorenz consistent inequality measures.

Strong Pareto Superiority

Pareto superiority (PS) is defined as the situation where no one loses from the initial to the final period but at least one individual gains. However, PS allocation may aggravate inequality. We thus introduce "*Strong Pareto Superiority*" (SPS). By SPS we mean a situation where the utility of all the individuals increases and the inequality (either absolute or relative) also remains same, compared to that of the initial distribution. In order to derive such a SPS allocation we first assume that their exists a social planner who taxes a subgroup of the population and distributes the collected tax to the rest of the population. Note that as we move through this paper we show that in general the SPS allocations and their associated tax transfer vector will be different for the relative and absolute inequality measures. From here onwards we refer the SPS allocation's that preserves relative and absolute inequality as RSPS and ASPS, respectively.

In order to illustrate the existence of SPS we consider a two-period economy with only two individuals observed in both the period. Let the distribution in the first period be denoted by the vector $Y_0 = (y_1, y_2)$ and that of the second period as $Y_1 = (g_1 y_1, g_2 y_2)$, where g_1 and g_2 denotes the growth rate of first and second individual respectively. Let $g_1 y_1 > g_2 y_2$. In order to keep the degree of relative inequality we must tax the first individual such that the following condition is satisfied:

$$\frac{y_1}{y_2} = \frac{g_1 y_1 - T}{g_2 y_2 + T} \tag{13.1}$$

We can solve T from the above equation in the following fashion

$$T = \frac{y_1 y_2 (y_1 - y_2)}{g_1 y_1 + g_2 y_2} \tag{13.2}$$

Note that the income profiles $y = (y_1, y_2)$ and $\hat{y} = (\hat{y}_1 = g_1 y_1 - T, \hat{y}_2 = g_2 y_2 + T)$ have same level of inequality, following any measure that is relative in nature. Further, $\hat{y}_1 > y_1$ and $\hat{y}_2 > y_2$, holds if the society enjoys positive growth (i.e., $g_1 y_1 + g_2 y_2 > y_1 + y_2$). Hence $\hat{y} = (\hat{y}_1, \hat{y}_2)$ qualifies as a relative-SPS allocation. Similarly we can design the absolute-SPS allocation, where T has to be solved from the following absolute inequality preserving condition: $y_1 - y_2 = (g_1 y_1 - T) - (g_2 y_2 - T)$. Note that following the results from [6, 13, 14] it can be easily established that SPS always exists and is unique. The existence of an SPS can also follow if we do the same exercise in the utility space. We use the well known box-diagram (Fig. 13.1) to highlight the welfare implication of the result.

Along the contract curve AB all points are Pareto superior to M. All such points are non-comparable in the sense of Pareto. This is because one can move from one point to the other (within AB) only by redistribution. Thus one is made better off and the other is worse off. However, our discussion so far ensures that among the non-comparable Pareto points on AB there exists a unique point which is distribution neutral. The formal proof is provided in the appendix.

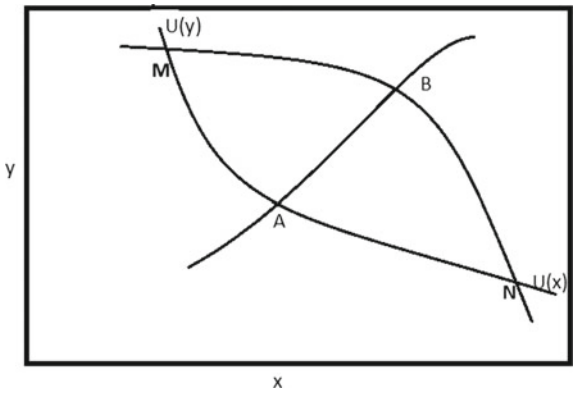


Fig. 13.1 Edgeworth Box Digram reflecting existence and uniqueness of SPS

So far our discussion has been limited only two the case where the number of individuals is 2. Consider the number of individuals being n. Let $Y_0 = (y_1, y_2, \ldots, y_n)$ be the initial income distribution. $Y_1 = (g_1.y_1, g_2.y_2, \ldots, g_n.y_n)$ as the final income distribution. Thus g_i ($g_i \in \Re$) denotes the growth rate of income of the i^{th} individual $\forall i \in \{1, 2, \ldots, n\}$. The SPS allocation that have same level of relative inequality as in Y_0 is given by

$$RSPS = (y_1^{RSPS}, y_2^{RSPS}, \ldots, y_n^{RSPS}) \tag{13.3}$$

where $y_i^{RSPS} = y_i.\left(\frac{\sum_{i=1}^{n} g_i.y_i}{\sum_{i=1}^{n} y_i}\right)$.

Note that if there is growth in the economy (i.e., $\sum_{i=1}^{n} g_i.y_i > \sum_{i=1}^{n} y_i$), then every individual is better off in the distribution RSPS and relative inequality also remains the same. The tax transfers vector following which one can derive the RSPS allocation form the final distribution is given by the following vector:

$$t_{RSPS} = (t_1, t_2, \ldots, t_n) \tag{13.4}$$

where $t_i = g_i.y_i - RSPS_i$. Note that if $t_i > 0$ then the individual has to pay tax. One the other hand if $t_i < 0$ then he enjoys transfer.

One can also preserve absolute inequality and make every one better off considering the following distribution.

$$ASPS = (y_1^{ASPS}, y_2^{ASPS}, \ldots, y_n^{ASPS}) \tag{13.5}$$

where $y_i^{ASPS} = y_i + \left(\sum_{i=1}^{n} g_i.y_i - \sum_{i=1}^{n} y_i\right)$.

$$t_{RSPS} = (\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_n) \tag{13.6}$$

where $\tilde{t}_i = g_i.y_i - ASPS_i$. The individuals pays tax (receives transfer) if $\tilde{t}_i > 0$ ($\tilde{t}_i < 0$).

Related Literature

The notion of distribution neutral fiscal policies exists in different forms in the literature. In this section we provide a discussion on some of the papers in this direction. We begin with a discussion of the poverty decomposition analysis introduced in [3]. The authors introduced a methodology that decomposes poverty changes into growth and redistribution components.[1] The methodology is applicable only if the poverty measure is a function of poverty line (z) mean income (μ_t) and the underlying Lorenz curve (L_t). That is the poverty measure can be written as $P_t = P(z/\mu_t, L_t)$. Here t denotes the time point. Note that the level of poverty may change due to a change in the mean income μ_t relative to the poverty line, or due to a change in relative inequalities i.e. L_t. Let L_r denotes the Lorenz curve at any given reference frame.[2] Following [3], a change in poverty from time t to $t + n$ can be decomposed as

$$P_{t+n} - P_t = G(t, t+n; r) - D(t, t+n; r) + R(t, t+n; r) \tag{13.7}$$

where $G(t, t+n; r)$, $D(t, t+n; r)$ and $R(t, t+n; r)$ denotes the growth component, redistribution component and the residual, respectively.

The growth component (i.e., G()) refers to the the change in poverty due to a change in the mean while holding the Lorenz curve constant at the reference frame (i.e., L_r). Mathematically this can be written as:

$$G(t, t+n; r) = P(z/\mu_{t+n}, L_r) - P(z/\mu_t, L_r) \tag{13.8}$$

Note that if the Lorenz curve of the initial distribution is set to be the reference frame (i.e., $L_r = L_t$) then the component $G(t, t+n; r)$ can also be interpreted as the change of poverty from a SPS allocation.

The redistribution component is the change in poverty due to a change in the Lorenz curve while keeping the mean income constant at μ_r.

$$D(t, t+n; r) = P(z/\mu_r, L_{t+n}) - P(z/\mu_r, L_t) \tag{13.9}$$

Poverty reduction is an important fiscal policy, and the total reduction of poverty following a distribution that do not aggravates inequality might also be an important tool for policy analysis. In the next section we shed some further lights on this issue using data from India. Note that presence of the residual term is often considered as

[1] Readers interested in this topics are referred to [7–9, 21].

[2] This can also be interpreted as the desired level of inequality the policy maker is anticipating.

a major criticism of the [3] poverty decomposition methodology. This problem has been addressed in [9, 21].

Poverty decomposition analysis discussed is also a widely used in the pro-poor growth literature. The notion "*pro poor growth*" may be defined in two different senses. In general growth is said to be pro poor in an absolute sense, if it raises income of the poor and consequently poverty reduces (see [11]). Following [10], growth is labeled as "pro-poor" in a relative sense, if it raises the incomes of poor proportionately more than that of the non poor. For further references readers interested are referred to [5, 16–18, 22].

Empirical Illustration

In this section we provide an application of distribution neutral fiscal policies. We show how poverty reduction which is viewed as an important fiscal policy is effected because of presence of inequality. Or in other words we are interested in the additional poverty reduction if the final distribution being replaced by the SPS allocation. This can also be viewed as an application of the [3] methodology that has been discussed earlier.

For this exercise we use the India Human Development Survey (IHDS) which is a nationally representative, multi-topic survey of households in India. The survey consists of two rounds. The first round was completed in 2004–2005. The second round was conducted in 2011–2012. This is an unbalanced panel data. That is, in the second round most households in 2004–2005 were interviewed again. From here onwards we denote IHDS data for 2004–2005 and 2011–2012 as IHDS1 and IHDS2, respectively. Throughout this exercise, we consider IHDS1 as the initial time point and IHDS2 as the final time point. We use monthly per-capita expenditure as a proxy of income.[3]

Before we move through the formal application of the distributional fiscal policies we discuss how poverty and inequality has changed from 2004–2005 to 2011–2012. Throughout this exercise we focus our analysis on three mutually exclusive and exhaustive subgroups of India, namely rural, urban and the slum areas. Following Table 13.1 it is readily observable that poverty has decreed substantially from period 2004–2005 to 2011–2012. However, following the same table it is also readily observable that inequality figures has also increased substantially. Clearly following the discussions in the previous sections the poverty reduction would have been more, had the degree of inequality remaining same as that of the initial distribution.

In Table 13.1 we present the actual poverty reduction. That is we simply compute the difference between poverty 2004–2005 to that of 2011–2012. In the same table we

[3]Note that per-capita income data is available in this survey. However, income data have standard problems in the sense that people often misreports their income. Furthermore, the poverty line in India is usually constructed using the per-capita consumption figures. Thus using such figures for income data might be incomplete.

Table 13.1 Summary statistics

	2004–2005			2011–2012		
	Rural	Urban	Slum	Rural	Urban	Slum
Real and nominal mean						
Mean	694.23	1792.67	1159.49	2881.88	828.02	2019.67
Inequality						
Gini	0.36	0.38	0.37	0.39	0.32	0.33
Atkinson (a = 0.5)	0.11	0.12	0.12	0.13	0.08	0.09
Generalized entropy (e = 0)	0.22	0.24	0.23	0.25	0.16	0.18
Generalized entropy (e = 1)	0.25	0.29	0.26	0.31	0.17	0.21
Generalized entropy (e = −1)	0.24	0.27	0.27	0.29	0.18	0.20
Poverty rates						
Head count ratio	0.37	0.18	0.24	0.10	0.38	0.18
Poverty gap	0.10	0.04	0.06	0.02	0.10	0.04
Poverty reduction						
Head count ratio						
Net poverty reduction				−0.188	−0.140	−0.207
SPS poverty reduction				−0.216	−0.151	−0.225
Poverty gap						
Net poverty reduction				−0.062	−0.039	−0.061
SPS poverty reduction				−0.070	−0.043	−0.065
Squared poverty gap						
Net poverty reduction				−0.026	−0.015	−0.025
SPS poverty reduction				−0.030	−0.016	−0.027

Notes

[1] Author's calculations based on monthly per capita expenditure from IHDS data

[2] Poverty figures are obtained considering poverty line as suggested by the Tendulkar Committee report for 2004–05. The poverty line for 2011–12 is obtained by inflating the 2004–05 line using CPIAL and CPIIW for rural and urban India respectively. Thus the poverty line for rural and urban India for 20111–12 is $\frac{611}{340} \times 477 = 803$ and $\frac{195}{112} \times 579 = 1007$ (ignoring the decimals), respectively. Real Mean is also obtained following the same procedure

[3] Poverty differences:

Net Poverty Reduction = (poverty rate in 2011–12)–(poverty rate in 2004–05)

SPS Poverty Reduction = (poverty rate for the SPS distribution)-(poverty rate in 2004–05)

also compute the SPS poverty reduction. This is the difference between the poverty figures of the initial distribution and SPS distribution. It is readily observable that the absolute value of the SPS poverty reduction rates is much higher than that of the actual rates. For example, the net poverty reduction for rural India is 18.8% where as the same for an SPS distribution is 21.6%. Thus moving from the final to the SPS distribution would have reduced additional 3% poverty. Similar result also holds for the rest of the cases.

Conclusion

In this paper, we provide a survey on distributional neutral fiscal policy and related literature highlighting the ideas of our new research in this area. We introduce the notion of "*Strongly Pareto Superior*" allocation or SPS. In order to focus on inequality-neutral or distribution neutral Pareto superior allocation we first discuss SPS allocation introduced in [14] which guarantees higher individual welfare keeping the degree of inequality same as before. Also [13] provide an analysis in the context of international trade.

It is possible to generate a SPS allocation by taxing a subgroup of population and redistributing the collected tax to the rest of the population. The only condition required is that there should be growth in the society. The construction of SPS is different when relative and absolute inequality is preserved. The SPS allocation preserving the relative inequality is obtained by redistribution of the aggregate gains among the individuals proportional to their utilities of the initial distribution. On the other hand, the SPS allocation which preserves absolute inequality is obtained by equally distributing the aggregate gains among all the individuals. SPS is a general condition and whenever there is growth in the society one can generate both relative and absolute SPS uniquely. This paper also provides the existence and uniqueness of SPS in a general equilibrium framework. Considering the case with only two agents we show that SPS allocation is a unique point in the edgeworth box and also lies in the locus of the contract curve. Major contributions of this approach is that it provides a new interpretation of Pareto criterion, an interpretation that should have been done long ago. We propose that Pareto criterion IS also about inequality or distribution, a point totally missed and misinterpreted so far.[4] Additionally we rigorously prove that a unique Pareto efficient allocation is DIFFERENT from all non-comparable Pareto allocations on the well known contract curve.

We also provide an empirical illustration on distribution neutral fiscal policies. We show that poverty reduction following a SPS distribution is actually 3% higher

[4] An Wikipedia entry argues that "*It would be incorrect to treat Pareto efficiency as equivalent to societal optimization, as the latter is a normative concept that is a matter of interpretation that typically would account for the consequence of degrees of inequality of distribution.*" (https://en.wikipedia.org/wiki/Pareto_efficiency).

than that of the actual poverty reduction. The exercise is somewhat similar to the poverty decomposition analysis introduced in [3].

The paper is consistent with the concern raised in [15] regarding robust results in economics. Independent of the country or context we focus, we can always find out a Pareto efficient allocation that will keep the distribution intact. Our idea convincingly proves the point that Pareto criterion could be refined to include distribution and would have wider applications in the areas of growth, trade and public policy.

Acknowledgements Sugata Marjit is indebted to the participants of the conference titled "*The Economy as a Complex System IV: Can economics be a physical science?*" arranged by the Institute of Mathematical Sciences (IMSc). We are also grateful to the seminar participants at IMF, Washington D.C. in July 2017, for helpful comments. We would also like to acknowledge the helpful comments by Sanjeev Gupta of the Fiscal Affairs Department of the IMF. The usual dissimilar applies.

Mathematical Appendix

Let $\{W^i\}$: endowment vectors $i = 1, 2, \ldots, n; \sum_i W^i = W$.

Utility functions, real valued, assumed to be continuous, increasing, strictly quasiconcave : $\{\bar{U}^i\}$ The consumption possibility set is $\Re^n_+$ the non-negative orthant. The set of feasible allocation vectors $\mathfrak{F} = \{\{Y^i\} : Y^i \geq 0, \sum_i Y^i \leq W\}$; $\{W^i\} \in \mathfrak{F}$; $\mathfrak{U} = \{\{U^i\} : \exists \{Y^i\} \in \mathfrak{F}, U^i(Y^i) = \bar{U}^i \forall i\}$ is the set of feasible utilities. Notice $\{U^i(W^i)\} \in \mathfrak{U}$.

Statement 1 $\mathfrak{U}$ is a nonempty compact subset of $\Re^n_+$

Proof: The above follows given the continuity of U^i over a compact set $\mathfrak{F}$.

Let $U^i(W^i) = \hat{U}^i \forall i$; there is a $P^* = (P_1^*, P_2^*, \ldots, P_n^*)$ which is a competitive equilibrium i.e., X^{i*} solves the problem $max U^i(.)$ subject to $P^* x \leq P^*.W^i \forall i$ and $\sum_i X^{i*} = W$. Let $U^i(X^{i*}) = U^{i*}$.

Define $\mathfrak{U}^{\mathfrak{P}} = \{\{\bar{U}^i\} : \{\bar{U}^i\} \in \mathfrak{U}, \quad \exists \{U^i\} \in \mathfrak{U} \text{ such that } , U^i > \bar{U}^i \forall i\}$: Pareto Frontier. The First Fundamental Theorem assures us that $\{U^{i*}\} \in \mathfrak{U}^{\mathfrak{P}}$.

Consider next $\bar{\theta} < \hat{\theta}$; from the property of the supremum, there is $\tilde{\theta} > \bar{\theta}$ such that $U(\tilde{\theta}) \in \mathfrak{U}$ i.e., there is $\tilde{Y}^i$ such that $\{\tilde{Y}\} \in \mathfrak{F}$ and $U^i(\tilde{Y}^i) = \tilde{\theta} U^i(W^i) \forall i$. Now there must be a scalar $1 > \alpha_i \geq 0$ such that $U^i(\alpha_i \tilde{Y}^i) = \bar{\theta}.U^i(W^i)$ since by shrinking the scalar α_i we can make the left hand side go to zero (U is increasing), whereas for $\alpha_i = 1$, the left hand side is greater; also then we can claim $\{\alpha_i \tilde{Y}^i\} \in \mathfrak{F}$ since

$$\alpha_i \tilde{Y}^i \geq 0$$

and

$$\sum_i \alpha_i \tilde{Y}^i \leq max_i \alpha_i \sum_i \tilde{Y}^i \leq W$$

□

Statement 3 $U(\hat{\theta}) \in \mathfrak{V}$

Thus one may conclude that $\{\hat{Y}^i\}$ is a Pareto optimal configuration with the property that $U^i(\hat{Y}^i) = \hat{\theta}U^i(W^i)\forall i$. And since $\hat{\theta}$ is uniquely determined, so is the configuration $\{\hat{Y}^i\}$.

References

1. Arrow, K.J.: Social Choice and Individual Values, vol. 12. Yale University Press (1963)
2. Cornes, R., Sandler, T.: Pareto-improving redistribution and pure public goods. Ger. Econ. Rev. **1**(2), 169–186 (2000)
3. Datt, G., Ravallion, M.: Growth and redistribution components of changes in poverty measures: A decomposition with applications to brazil and india in the 1980s. J. Dev. Econ. **38**(2), 275–295 (1992)
4. De Scitovszky, T.: A note on welfare propositions in economics. Rev. Econ. Stud. **9**(1), 77–88 (1941)
5. Duclos, J.Y.: What is pro-poor? Soc. Choice Welf. **58**, 32–37 (2009)
6. Gupta, S., Marjit, S., Sarkar, S.: An application of distribution-neutral fiscal policy. IMF Working Paper No. 18/12 (2018)
7. Jain, L.R., Tendulkar, S.D.: Role of growth and distribution in the observed change in headcount ratio measure of poverty: a decomposition exercise for india. Indian Econ. Rev. 165–205 (1990)
8. Kakwani, N., Subbarao, K.: Rural poverty and its alleviation in india. Econ. Polit. Wkly. A2–A16 (1990)
9. Kakwani, N.: On measuring growth and inequality components of poverty with application to thailand. J. Quant. Econ. **16**(1), 67–80 (2000)
10. Kakwani, N., Pernia, E.: What is pro-poor growth? Asian Dev. Rev. **16**, 1–22 (2000)
11. Kraay, A.: When is growth pro-poor ? evidence from a panel of countries. J. Dev. Econ. **80**, 198–222 (2006)
12. Mandler, M.: Simple pareto-improving policies. J. Econ. Theory **84**(1), 120–133 (1999)
13. Marjit, S., Sarkar, S., Yang, L.: Gains from trade, inequality and distribution-neutral fiscal policy. mimeo-CSSSC (2018)
14. Marjit, S., Sarkar, S.: Distribution neutral welfare ranking-extending pareto principle. CESifo Working Paper No. 6397 (2017)
15. Marjit, S.: On economics as a physical science. Eur. Phys. J. Spec. Top. **225**(17–18), 3269–3273 (2016)
16. Nssah, E.: A unified framework for pro-poor growth analysis. Econ. Lett. **89**, 216–221 (2005)
17. Osmani, S.: Defining pro-poor growth. One Pager Number 9, vol. 9. International Poverty Center, Brazil (2005)
18. Ravallion, M., Chen, S.: Measuring pro-poor growth. Econ. Lett. **78**(1), 93–99 (2003)
19. Samuelson, P.A.: Aspects of public expenditure theories. Rev. Econ. Stat. 332–338 (1958)
20. Sen, A.: Collective Choice and Social Welfare (1970)
21. Shorrocks, A.F.: Decomposition procedures for distributional analysis: a unified framework based on the shapley value. J. Econ. Inequal. **11**(1), 99–126 (2013)
22. Son, H.H.: A note on pro-poor growth. Econ. Lett. **82**, 307–314 (2004)
23. Stiglitz, J.E.: Pareto efficient and optimal taxation and the new new welfare economics. Handb. Public Econ. **2**, 991–1042 (1987)

Chapter 14
Tracking Efficiency of the Indian Iron and Steel Industry

Aparna Sawhney and Piyali Majumder

Abstract Recycling of steel scrap has been one of the key drivers of improving energy efficiency in steel manufacturing worldwide. Energy intensity of Indian steel plants is higher than the world average, and the strategy of scrap recycling to enhance energy efficiency has gained policy momentum. We track the energy intensity of Indian steel plants during the period 1999–2014, to determine whether scrap-use provided energy-saving benefits. We consider energy intensity as a function of various plant characteristics, while controlling for plant heterogeneity and industry sub-group (by 5-digit National Industrial Classification). We find that energy-intensity of Indian steel plants has declined significantly over the years, and more so for privately-owned steel plants, but the use of scrap in the production process has not helped reduce energy consumption. Indeed scrap users have lower energy-efficiency that may be driven by poor quality of raw materials which our analysis is unable to capture.

Introduction

Enhancing energy efficiency with economic growth has increasingly gained priority in India as the country grapples with twin challenges of economic poverty and energy poverty. The institutional framework to reduce energy intensity of the economy is provided by the Energy Conservation Act of 2001, whereby standards, regulations and norms have been implemented. The Bureau of Energy Efficiency, established in 2002, under the Act, facilitates the implementation of different initiatives for energy conservation and efficiency. In order to promote energy conservation in the industrial sector, the Bureau is implementing the National Mission for Enhanced Energy Efficiency (NMEEE) under the National Action Plan on Climate Change 2008, which

A. Sawhney (✉) · P. Majumder
Centre for International Trade and Development, Jawaharlal Nehru University, New Delhi 110067, India
e-mail: aparnasawhney@yahoo.com

P. Majumder
e-mail: piyalimajumder88@yahoo.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_14

aims to reduce the specific energy consumption of the most energy-intensive industries. In particular, the NMEEE identified nine energy-intensive sectors of the country for target energy efficiency norms to achieve a low carbon path for the economy, including iron and steel, a key infrastructure development industry.

India is one of the leading producers of steel in the world—third largest producing country after China and Japan, and the growth prospects continue to be high. However, energy consumption of Indian iron and steel plants are much higher than steel plants abroad—according to the Ministry of Steel the integrated steel plants require 6–6.5 Giga calories per tonne of crude steel compared to 4.5–5 in steel plants abroad. The higher energy intensity of Indian steel production is attributed to obsolete technologies (added problem of retrofitting modern technologies in old plants), old operating practices and poor quality of raw materials.

While steel production is energy-intensive, modern energy management systems recycling steel scrap have reduced energy intensity of steel production in the world. Recycling scrap helps in conservation of energy as remelting of scrap requires much less energy than production of iron or steel from iron ore. Steel is infinitely recyclable and the recycling of steel accounts for significant energy and raw material savings: estimated to be over 1,400 kg of iron ore, 740 kg of coal and 120 kg of limestone saved for every 1,000 kg of steel scrap converted into new steel [1]. In India, the re-use of iron and steel scrap serves as a critical input substituting for ore in basic metal manufacturing and as well as other large manufacturing sectors like metal casting, machinery equipment, etc. The consumption of scrap is mainly reported by Induction Furnace and Electric Arc Furnace units, integrated steel plants and alloy steel and foundry industries in India. The consumption of iron and scrap by remelting also reduces the burden on land fill disposal facilities and prevents the accumulation of abandoned steel products in the environment [2]. Since recycling of steel scrap offers significant energy saving, it is pertinent to understand whether recycling strategy has enhanced energy-efficiency in the Indian iron and steel industry during the last two decades. Our analysis examines the energy efficiency and overall productivity performance of Indian iron and steel plants, distinguishing the use of scrap as an input in the production process.

Energy Efficiency of Iron and Steel Industry in the Post-Liberalization Period

Following deregulation and de-licencing in 1991–92, the iron and steel private sector has grown in leaps and bounds in India and now accounts for more than three-quarters of total production of crude steel and finished steel products in the country [3]. During the last decade, several initiatives were undertaken to improve energy efficiency in the industry. For example, a UNDP project during 2004–13 facilitated low carbon technologies in 34 Indian steel re-rolling mills to bring down energy consumption and reduce GHG emissions by 25–50%. This was followed by a replication project

during 2013–16 covering 300 mini steel mills, including 5 Induction furnace units. An analysis of the Indian iron and steel sector had found that in the decades prior to liberalization, productivity had been declining with poor energy efficiency, but following deregulation the trend began to reverse in the latter half of 1990s [4]. The study observed that technological change in Indian iron and steel industry has been biased towards energy and material-use, although it has been improving since late 1990s. More recent estimates confirm that energy efficiency in the industry has been improving since 1998 [5]. A decomposition of the energy efficiency improvement in the iron and steel industry during 1991–2005 showed that the decline in energy intensity was due to technical change, which more than offset an adverse effect from structural change (towards more energy-intensive products) [6].

Recycling Scrap—An Energy Conservation Strategy

Recycling scrap conserves energy, as remelting of scrap requires less energy than production of virgin metal from ore. The production of secondary steel, utilizing scrap, uses 74% less energy than the production of steel from iron ore [7]. The use of recycled scrap serves to substitute for virgin ore as well as conserve energy in the production process. One can thus expect to see higher factor productivity (including energy) in scrap-recycling plants compared to non-scrap steel plants.

The consumption of scrap is mainly reported by Induction Furnace and Electric Arc Furnace units, integrated steel plants and alloy steel and foundry industries in India. It is understood that integrated steel plants are more resource efficient since they use better technology compared to the smaller plants manufacturing other steel products. Indeed, product and process mix has significant impact on energy performance.

The question of interest that arises, is whether plants in the Indian iron and steel industry using recycled scrap have experienced gains through lower energy use. So here we consider the energy efficiency of plants in the Indian iron and steel industry, while distinguishing between the scrap users and non-scrap users. Since existing studies have recorded an improvement in the energy efficiency in the iron and steel industry in the post-liberalization period, the energy saving from scrap usage should be observable after controlling for this trend as well as other plant characteristics that impact energy consumption of plants. It is established that specific energy consumption among the steel plants varies due to different processes, quality of material, types of products produced by the plants. Thus, when tracking the energy-efficiency of production we control for these factors, as well as other features. We use the measure of autonomous energy efficiency of plants, which is measured by the total energy used per unit of output.

Model and Estimation Results

In order to track the energy efficiency of the iron and steel plants, we use a simple log linear specification of the physical energy intensity of output of the plant. The energy intensity of the plant is a function of the technology, scale of operation, products produced, ownership, and location characteristics of the plants. Our measure of physical energy intensity is given by real energy used per unit of real output, measured in MJ per constant thousand Rs. In order to calculate the energy used in production, we aggregate the energy in the fuels reported by each plant—including electricity and coal. The value of total output produced by a plant is deflated by the wholesale price index of that industry for that year, in order to obtain the value of output in constant money value (base year 2004–05). We use manufacturing plant level data from the Annual Survey of Industries, to analyze the energy-intensity of iron and steel plants for the period 1998–99 through 2013–14. We deflate all money values (in Indian rupees) by the appropriate price indices. For example, our measure of physical capital is calculated as the value of plant and machinery deflated by price index of plant and machinery.

While we consider plants in the basic iron and steel industry, we distinguish them by their industry sub-group based on the main products produced by them (since these are multi-product firms). We utilize the most dis-aggregated industrial classification available, namely the 5-digit National Industrial Classification 2008 (concorded for earlier years beginning 1998). In our analysis we have seven industry sub-groups under the basic iron and steel manufacturing, including manufacture of steel ingot, ferro alloys, direct reduced iron, hot-rolled and cold-rolled products, rail track material, and other basic iron-steel products. Since integrated iron and steel plants (produced pig iron as well as steel products) are structurally very different in nature to other plants, we analyze the energy intensity behavior of plants by dropping them from one sample. Our econometric model of the energy-intensity of plants is as follows:

$$log(E_{it}) = T + log(k_{it}) + log(s_{it}) + O_{it} + C_{it} + R_i + \varepsilon_{it} \tag{14.1}$$

Where E denotes the energy intensity of plant i in year t, T is the time trend of energy intensity of plant i during the 16-year period, $k$denotes the capital labour ratio of plant i in year t, s denotes the scale of operation of the plant, O denotes the ownership status (1 for private, 0 for public) of plant i in year t, C is the scrap use of plant i in year t; R is the location dummy (30 Indian states) of the plant, ε is the error term (controlling for the product group classification of plant i in year t).

We de-trend the energy intensity of plants over the 16-year period in order to isolate the impact of scrap use on the energy consumption per unit output. The nature of technology used in the plant is captured by the capital-labour ratio (k), while the scale or size of the plant is measured by the employment. Ownership of the firm, whether it is private or public, is a dichotomous variable taking on value 1 for private sector and 0 for public sector units. We also incorporate state fixed effects, to account for time-

invariant location factors that affect plant efficiency. We use two measures of scrap recycled in production, C: first, as a dichotomous variable (1 when a plant uses scrap, 0 if scrap is not used); and second, as a continuous variable measured by the value share of scrap in total raw material used in production. Our multivariate regression controls for plant heterogeneity, distinguishing between plants belonging to separate product groups at the 5-digit NIC codes, and estimating with heteroskedastic robust standard errors. The results are summarized in the Table. Column (1) and (3) show the regression estimation for all plants in the manufacturing of basic iron and steel industry; while columns (2) and (4) show the regression estimations for the sample without the integrated steel plants. It should be noted here that the sample of plants in the regressions (2) and (4) are on average smaller than in the integrated steel plants and use far less capital-intensive technological processes.

The results indicate that energy intensity of production in the iron and steel industry has indeed been declining over time, especially considering the integrated steel plants—since the time trend is significant and negative in regressions (1) and (3), but less so in the sample without the integrated steel plants regression (2). It is important to note that the negative time trend observed for the sample with large integrated steel plants re-confirms the observation in the existing literature that energy efficiency has improved in the industry since the late 1990s. However, energy efficiency improvement is not readily observed in the smaller steel plants—as the magnitude of the trend coefficient in regression (2) is lower and is less significant compared to regression (1). With the alternative measure of scrap usage, in regression (4), the negative time trend for the smaller steel plants becomes insignificant. The positive significant coefficients of capital-labour ratio and plant size shows that the energy intensity is higher in more capital-intensive larger plants. After controlling for plant characteristics and location, we find that energy intensity in the plants that recycled scrap is significantly higher than those which did not (regressions 1 and 2) or those which used less scrap in the material mix (regressions 3 and 4). Alternatively, we can examine the total factor productivity of steel plants to measure efficiency of plants recycling scrap versus non-scrap users (summarized in the Appendix). An analysis of total factor productivity gives us the same qualitative result, where scrap-using plants are found to have lower factor productivity than non-scrap users (discussed in the appendix).

Concluding Observations

While recycling has been identified as one of the key drivers of improvement in energy efficiency in the iron and steel industry, we do not find support for this phenomenon for Indian iron and steel manufacturing industry. Our analysis suggests that scrap recycling in steel plants in India are less efficient in total factor use as well as in energy use. The lower energy-efficiency of the plants using scrap may not be energy saving due to the other factors, like poor quality of raw material used that our analysis has not been able to capture. This has important implications for the strategy to encourage

growth in the secondary steel production in India, as envisioned in the National Steel Policy 2017. Going forward, the government intends to encourage scrap-based steel manufacturing in order to save energy. However, promoting scrap-use may not be energy-efficient unless the overall factor productivity of the scrap-using plants is also improved.

Appendix: Total Factor Productivity of Plants in Basic Iron and Steel Manufacturing

Efficiency in production is widely measured through factor productivity, which measures the output per factor use. As a robustness check of our result obtained in energy intensity, we also estimated the total factor productivity across the plants in basic iron and steel manufacturing during 1998–99 to 2013–14, using the same data from the Annual Survey of Industries of India (Table 14.1). Utilizing a standard production function specification given by the Eq. 14.2, we estimate the plant productivity as the residual of this functional relationship.

$$Y_{it} = A_{it} K_{it}^{\beta_k} L_{it}^{\beta_l} M_{it}^{\beta_m} \tag{14.2}$$

K, L and M are the capital, labour and material employed by the firm. A is the efficiency parameter of the firm which is unobserved to the researchers. If we take a

Table 14.1 Regression results for iron and steel plants, 1998–99 to 2013–14. (Dependent variable: log of energy intensity of plant in MJ/'000 constant Rs)

	(1) all plants	(2) non-integrated plants	(3) all plants	(4) non-integrated plants
Year	−0.009[a] [0.003]	−0.006[b] [0.003]	−0.006[c] [0.003]	–0.002 [0.003]
Log(capital labour ratio)	0.052[a] [0.008]	0.047[a] [0.008]	0.068[a] [0.008]	0.064[a] [0.008]
Ownership	−0.613[a] [0.115]	−0.608[a] [0.124]	−0.535[a] [0.113]	−0.534[a] [0.121]
Log(plant size)	0.033[a] [0.009]	0.018[b] [0.009]	0.070[a] [0.009]	0.057[a] [0.009]
Scrap use dummy	1.033[a] [0.026]	1.055[a] [0.026]		
Scrap usage			1.352[a] [0.032]	1.370[a] [0.032]
Observations	16,464	15,800	16,413	15,754
R-squared	0.174	0.175	0.174	0.175
State fixed effects	Yes	Yes	Yes	Yes
F-test	109.7	110.5	111.7	112.2
Prob>F	0.00	0.00	0.00	0.00

All regressions control for heterogeneity across plants in the different product groups (at the 5-digit NIC classification)

Robust standard errors in parentheses [a]p<0.01, [b]p<0.05, [c]p<0.1

log transformation of the above equation and add an error term:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \beta_m m_{it} + \varepsilon_{it} \tag{14.3}$$

where, $y = log(Y), k = log(K), l = log(L), m = log(M)$, $\beta_0 + \varepsilon_{it} = log(A_{it})$ β_0 is the mean efficiency level across firms and over time. ε_{it} is the time and producer specific deviation from that mean which can be further decomposed into an observable and unpredictable part. We empirically estimated the total factor productivity A_{it} as a residual of the production function represented by Eq. 14.3 using the Levinsohn Petrin Methodology [8]. An important problem while estimating productivity is the correlation between observable productivity shocks and input levels. For a positive (negative) productivity shock a firm will try to make certain adjustments in the input to expand (reduce) the output. Therefore, if the firm has prior knowledge about the productivity shock, then at the time of input decision endogeneity may arise and the input quantities are partly determined by prior beliefs of productivity. In Levinsohn Petrin methodology firm's intermediate material inputs or fuels are used as a proxy to control the effect of unobservable productivity shocks thereby correcting for the simultaneity bias under the assumption that demand for intermediate inputs is a monotonic function of productivity. Our sample for productivity estimation is an unbalanced panel of 6811 firms, with 17,836 observations, for the period 1998–99 to 2013–14. On comparing the total factor productivity between scrap vs. non-scrap users, we observed that scrap-using firms exhibited lower total factor productivity as compared to the non-scrap using firms. Our result suggests that using scrap as an intermediate raw material did not have productivity enhancing effect across firms in the iron and steel industry.

References

1. Fact Sheet, World Steel Association (2018). https://www.worldsteel.org/
2. Indian minerals yearbook 2014 (Part II: metals and alloys): iron and steel and scrap, Indian Bureau of Mines, Ministry of Mines, Government of India, pp. 9–19 (2016)
3. Report of the working group on steel industry for the twelfth five year plan (2012–2017), Ministry of Steel, Government of India (2011)
4. Schumacher, K., Sathaye, J.: Indias iron and steel industry: productivity, energy efficiency and carbon emissions, environmental energy technologies division, Ernest Orlando Lawrence Berkeley National Laboratory, LBNL-41844 (1998)
5. Dasgupta, S., Roy J.: Analysing energy intensity trends and decoupling of growth from energy-use in Indian manufacturing industries during 1973–74 to 2011–2012. Energy Effic. 1–19 (2016)
6. Reddy, B.S., Ray, B.K.: Decomposition of energy consumption and energy intensity in Indian manufacturing industries. Energy Sustain. Dev. **14**, 35–47 (2010)
7. Energy information administration, recycling is the primary energy efficiency technology for aluminium and steel manufacturing, US Energy Information Administration, 9 May 2014. http://www.eia.gov/todayinenergy/detail.php?id=16211
8. The estimates obtained from Levinsohn Petrin methodology are unbiased and efficient (compared to the estimates ordinary least squares or fixed effects methods) as it corrects for the endogeneity bias involved in the estimation of the productivity of firm

Part II
Sociophysics

Chapter 15
Social Integration in a Diverse Society: Social Complexity Models of the Link Between Segregation and Opinion Polarization

Andreas Flache

Abstract There is increasing societal and scholarly interest in understanding how social integration can be maintained in a diverse society. This paper takes a model of the relation between opinion polarization and ethnic segregation as an example for social complexity. Many argue that segregation between different groups in society fosters opinion polarization. Earlier modeling work has supported this theoretically. Here, a simple model is presented that generates the opposite prediction based on the assumption that influence can be assimilative or repulsive, depending on the discrepancy between interacting individuals. It is discussed that these opposite results from similar models point to the need for more empirical research into micro-level assumptions and the micro-to-macro transformation in models of opinion dynamics in a diverse society.

Introduction

Migration both within and between countries has strongly increased in recent decades [11]. For many Western societies this comes with more ethnic and cultural diversity of their population. Other societies, like India, know high levels of diversity already for many centuries. Ethnic and cultural diversity have many benefits, for example in terms of a broader pool of talent or more creativity in diverse teams in organizations [16]. But diversity also constitutes a challenge for societies. Often diversity is associated with high levels of segregation between different groups [9], or with differences between groups in attitudes on fundamental issues such as civil rights of homosexuals, legalization of abortion, or gender equality [38].

There are no easy answers to the question under which conditions diversity can endanger societal integration and foster instead persistent disagreement or even polarization. Polarization can be described as the tendency of a population to fall apart

A. Flache (✉)
Department of Sociology/ICS, University of Groningen, Groningen, The Netherlands
e-mail: a.flache@rug.nl
URL: https://flache.gmw.rug.nl/

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_15

into a small number of subgroups with large agreement within and disagreement between them [6]. Whether and under which conditions polarization arises is notoriously hard to predict. The evolution of the distribution of opinions in a society results from numerous simultaneous interactions between individuals both within their own cultural subgroups as well as across intergroup boundaries. Some of these interactions may drive groups apart, others may foster consensus. For example, research in the tradition of contact theory emphasizes that intergroup contacts reduce prejudice and promote agreement [1, 14, 40], but negative interactions at the individual level can also result in deeper divisions between groups [43, 44]. Polarization can thus be an outcome that results from the interactions of multiple individuals who neither expect nor intend to bring it about. One reason is the possibility that small changes in opinion distributions can have large unexpected consequences, for instance when disagreement emerging between some individuals in a local region of a network spreads and then quickly becomes amplified by social contagion [7, 20]. Identifying the conditions and mechanisms under which social influence dynamics in a diverse population result in polarization is therefore a major scientific issue with a long tradition of vivid debate [34].

In search for tools to tackle the inherent complexity of collective opinion dynamics, researchers used in recent decades increasingly agent-based computational modeling [5, 22, 28]. While this has greatly helped to understand the complex interplay of individual-level social interactions with macro-level outcomes, it also highlighted that the outcomes of opinion dynamics can sensitively depend on the exact assumptions researchers make about the process of social influence at the micro-level. In this paper, I will illustrate this with a model of the relationship between segregation and polarization in a diverse society. According to many, segregation between different subgroups is one of the important reasons for persistent disagreement between groups. Segregation is the separation of different groups, for example between residential areas of a city [9], or between different clusters in a social network [12, 36]. Segregation can reduce the extent of intergroup contact [8] and thus exacerbate prejudice. A further problem is that segregation can create 'bubbles' within which only like-minded people meet and interact. As former U.S. president Obama pointed out in his farewell address, in such a bubble we are "surrounded by people who look like us and share the same political outlook and never challenge our assumptions", such that "we become so secure ... that we start accepting only information, whether it's true or not, that fits our opinions" [39].

The argument that segregation fosters polarization seems compelling, but social complexity models showed how different equally plausible micro-level theories of social influence can generate radically different implications. A number of formal models is consistent with Obama's intuition. Building on persuasive argument theory [37, 46], models proposed by Mäs and coauthors [31, 33] assume that agents with more similar opinions are more likely to persuade each other to strengthen their already prevailing opinion tendency. Simulations demonstrated how then opinions in different subgroups can be pushed towards opposing poles of an opinion spectrum if agents prefer interacting with similar others, based on the principle of homophily [35]. Similar dynamics have also been derived from models of "biased assimilation"

[4, 13] in which agents are assumed to put more weight on those influences in the process of assimilation that are in line with their current opinion.

Models of persuasive arguments and biased assimilation suggest that segregation fosters opinion polarization. Building on earlier work [17], I will show in this paper that radically different conclusions can be drawn from another class of models. I follow a number of studies [2, 3, 15, 18, 20, 26, 29, 30, 41] which incorporated into models of social influence the assumption that influence can not only be assimilative, reducing opinion differences, but also repulsive. When influence is repulsive, individuals strive to be dissimilar to people they dislike, accentuating disagreement with others. But this only happens when those others are perceived as being too discrepant, otherwise influence is assimilative. This combination of assimilative and repulsive influence is suggested by theories of fundamental psychological processes in the formation of attitudes, like Heider's balance theory [23] or Festinger's theory of cognitive dissonance [19]. In a number of formal models elaborating this idea, it has in particular been assumed that perceived discrepancy not only arises from disagreement in opinions between individuals, but also from 'demographic' differences representing fixed characteristics like gender, race or ethnicity [17, 21, 30].

I will demonstrate in what follows that a model combining assimilative and repulsive influence implies that more segregation reduces opinion polarization between groups. The model will be presented in section "Modelling the Link Between Segregation and Opinion Polarization", results are described in section "Results". Section "Discussion and Conclusion" concludes with a more general reflection on the role of social complexity models for our understanding of social integration in a diverse society.

Modelling the Link Between Segregation and Opinion Polarization

First, the micro-level assumptions about social influence are introduced in section "Microlevel Assumptions About Social Influence". Second, the model of spatial network segregation is described in section "Modeling the Spatial Structure: Local Interaction and Segregation".

Microlevel Assumptions About Social Influence

The model contains a population of N individuals i who are throughout members of either group 0 or group 1, indicated by group membership $g_i \in \{0, 1\}$. For simplicity, I assume that both subgroups are always equally large. Every individual i adopts at every time point t an opinion o_{it}, with $0 \leq o_{it} \leq 1$. Following [17, 20], individuals are connected in a static interaction network (see section "Modeling the Spatial Structure:

Local Interaction and Segregation" for details) and can only interact with network neighbors.

Dynamics of the model unfold in consecutive discrete time steps t. In every time step, a pair of two different network neighbors i and j is selected at random with equal probability. All individuals k who are not involved in an interaction at time point t do not change their opinions, thus $o_{k,t+1} = o_{kt}$. If i and j interact, then both can modify their current opinions to move closer towards or away from the opinion of the interaction partner as given by Eqs. (15.1) and (15.2).

$$o_{i,t+1} = o_{it} + \Delta o_{it} = o_{it} + \mu w_{ijt}\,(o_{jt} - o_{it}) \tag{15.1}$$

$$o_{j,t+1} = o_{jt} + \Delta o_{jt} = o_{jt} + \mu w_{jit}\,(o_{it} - o_{jt}) \tag{15.2}$$

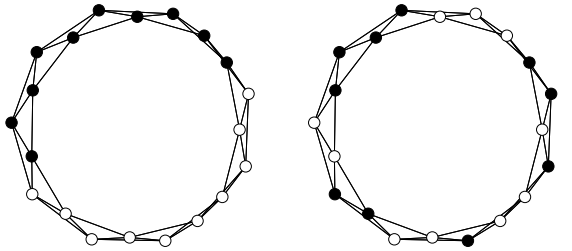
The parameter μ ($0 < \mu \leq 0.5$) in Eqs. (15.1) and (15.2) defines the rate of opinion change and will be kept at $\mu = 0.5$ in the present paper. The influence weights w_{ijt} and w_{jit} in Eqs. (15.1) and (15.2) represent the direction and magnitude of the influence of i on j and j on i, respectively. Weights are constrained by $-1 \leq w_{ij} \leq 1$. A positive weight w_{km} entails assimilative influence (k moving her opinion closer towards m's opinion), whereas a negative weight imposes repulsive influence (k moving her opinion away from m's opinion). A zero weight implies no change, reflecting indifference towards the source of influence. In this basic form, Eqs. (15.1) and (15.2) allow interactions to push the opinion outside of the opinion interval $[\,0,\,1\,]$ if weights are negative. In this case, the resulting opinion is truncated to the interval boundary that was crossed by the opinion shift. In some models that combine assimilation and repulsive influence, opinions are constrained with smoother functions [20, 21, 25], but this seems to have little effect on the main model dynamics.

The link between diversity, disagreement and social influence is implemented as follows. The influence weight w_{ijt} expresses the similarity that i experiences at time point t between herself and j. More precisely, the influence weight declines in the current level of disagreement $|o_{jt} - o_{it}|$, and is reduced if i and j belong to different groups. Equation (15.3) formalizes the computation of influence weights.

$$w_{ijt} = 1 - 2\left(\beta_O |o_{jt} - o_{it}| + \beta_D |g_j - g_i|\right). \tag{15.3}$$

Equation (15.3) shows that influence becomes repulsive when the discrepancy $\beta_O |o_{jt} - o_{it}| + \beta_D |g_j - g_i|$ exceeds 0.5, half of the theoretical maximum of 1. The parameters β_O and β_D in Eq. (15.3) scale the relative impact that respectively, *opinion* disagreement and *demographic* differences have on discrepancy. For convenience, I impose the constraint $\beta_O + \beta_D = 1$.

The model assumed here uses a simple linear transformation of discrepancy into influence weights w_{ijt}. Some studies have adopted a non-linear weight function in an otherwise similar framework [26, 32], but did only consider disagreement in opinions. Future work should combine a non-linear weight function with both disagreement and intergroup differences to explore possible new implications.



(a) segregated network (b) desegregated network

Fig. 15.1 Example for result of desegregation algorithm with desegregation rate $d = 0.5$ (right), starting from initially maximally segregated ring lattice (left). $N = 20$, range of interaction $= 2$

Modeling the Spatial Structure: Local Interaction and Segregation

The key condition of interest in the simulation experiments is segregation between groups. Modeling local interaction in a simple way, I employ a ring-lattice network in which all agents have the same number of local network neighbors to the left and to the right, called range of interaction r. Figure 15.1a shows the baseline condition of maximal segregation between the two groups for $r = 2$, $N = 20$, and two equally large subgroups. In both subgroups, only 4 out of 10 agents have any outgroup-neighbor among their 4 network neighbors. Of those 4 agents, half have 2 outgroup neighbors and the other half has only 1.

The degree of segregation is manipulated as follows.[1] Starting from a maximally segregated network (see Fig. 15.1a), a subset of N_s distinct agents from group 0 is randomly chosen for relocation. N_s is given by the desegregation rate d, ($0 \leq d \leq 0.5$), rounded to the integer nearest to $d\ ^N/_2$. For every chosen agent of group 0, a unique randomly selected agent from group 1 is picked. In all the selected pairs thus formed, network positions of the group 0 agents are swapped with those of the group 1 agents. Figure 15.1b shows an example for a network generated with a desegregation rate of $d = 0.5$.

To quantify segregation, a segregation measure S is computed. S indicates the fraction of same-group neighbors among all network neighbors of an agent, averaged over all agents and divided by the theoretically possible maximal fraction of ingroup neighbors, given N and r. The exact value of S given d varies randomly, depending on which pairs of agents were selected for a position swap.

The relation between desegregation rate d and average segregation S is non-linear. The closer the desegregation rate comes to 0.5, the less impact further increase has on the segregation level S. To account for this, the average value of S per level of d will be used to show how segregation affects polarization in the experiments that follow, whereas segregation will be manipulated with stepwise variation in the desegregation rate d.

[1] All computations, simulations and graphics in this paper were produced with Wolfram Mathematica©Version 11.2.

Results

In Section "Design and Measures", design and outcome measures of the computational experiments are described. Two experiments are conducted in both of which segregation S is manipulated. In the first experiment it is assumed that there is no initial group-specific disagreement in opinions. In the second experiment, a moderate group-specific disagreement is introduced. Results of experiment 1 and experiment 2 are described in sections "Experiment 1" and "Experiment 2", respectively. Section "Robustness Tests" is devoted to a brief description of some robustness tests.

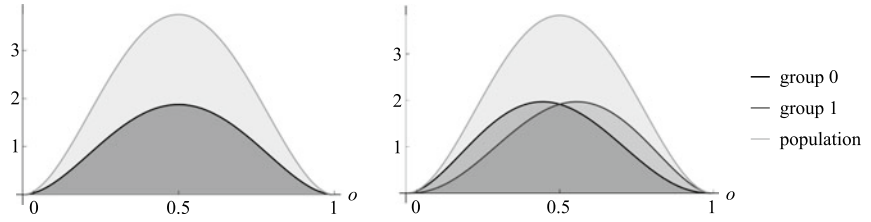
Design and Measures

The following simple baseline scenario was used in both experiments. Population size was set to $N = 100$ with 50 members in groups 0 and 1, respectively. The network was a circular ring lattice with interaction range $r = 5$. The relative impact of demographic dissimilarity on the influence weight w_{ijt} was set to $\beta_D = 1/3$. With this value, polarization between groups is possible but not trivial.

In both experiments, the desegregation rate d was varied from 0 to 0.5 in steps of 0.025, over 21 different levels. This resulted in variation of the average segregation measure S between $S = 1$ at $d = 0$ and $S \approx 0.522$ at $d = 0.5$. Except for $d \geq 0.45$ the 95% confidence intervals of mean S were non-overlapping for consecutive levels of d in a sample of 500 independent realizations per condition. For every level of d, 500 independent realizations of the simulation model were conducted, each running for 1000 $N = 100,000$ time steps. This was more than enough for all conditions in experiment 1 and 2 to reach stable outcomes.

It is an important question whether opinion polarization between groups can arise even if these groups have no systematic disagreement prior to interaction. For this reason, I drew in experiment 1 initial opinions randomly from the same Beta distribution $Beta(3, 3)$ for both groups, shown in Fig. 15.2a. This distribution has expected mean value of 0.5 and a standard-deviation of about 0.189. For culturally salient issues it is, however, more plausible that different groups also have different initial opinion tendencies. To model this, initial opinions were in experiment 2 randomly drawn from two symmetric Beta distributions $Beta(3, 3.5)$ and $Beta(3.5, 3)$ for groups 0 and 1, respectively, as shown in Fig. 15.2b. Mean opinions were about 0.462 for group 0 and 0.538 for group 1. Initial opinions in both groups had the same expected standard deviation of approximately 0.182.

The key outcome of interest in the simulation experiments was the degree of polarization both within the population as a whole and between the two groups. To assess between-group polarization, I measured the absolute value of the difference between the mean opinions in both groups, $P_t^g = |\overline{o_{t,g=1}} - \overline{o_{t,g=0}}|$. If this difference is close to one, this is a clear sign of strong between-group polarization. A low difference between the mean opinions of the groups, however, does not necessarily



(a) Experiment 1: both groups same initial distribution

(b) Experiment 2: moderate initial disagreement between groups

Fig. 15.2 Initial opinion distributions

show that there is no polarization at all. The population can also fall apart into two opposed factions that both contain members of both groups. To distinguish this form of 'population polarization' from between-group polarization, population polarization P_t^p at time point t is computed as the variance of all pairwise opinion distances in the population (adapted from [20]), as given by Eq. (15.4).

$$P_t^p = \frac{4}{N^2} \sum_{ij}^{i=N,j=N} \left(|o_{jt} - o_{it}| - \overline{|o_{kt} - o_{mt}|} \right)^2 . \tag{15.4}$$

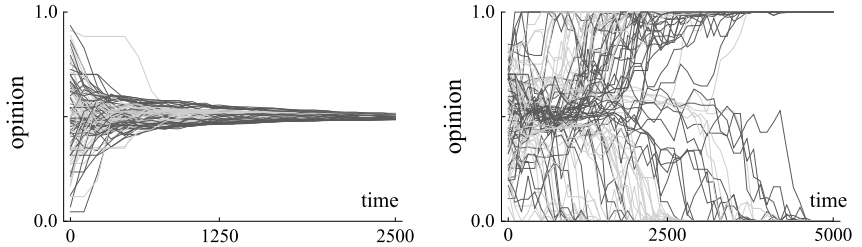
In Eq. (15.4), $\overline{|o_{kt} - o_{mt}|}$ denotes the average opinion distance across all pairs (km) in the population. The minimum level of polarization ($P = 0$) obtains when all pairwise distances are zero, corresponding to full consensus in the population. P^p obtains its maximal value of 1 if the population is split into two equally large factions with maximal mutual disagreement and full agreement within each of the factions.

Experiment 1

I begin with showing the dynamics for two prototypical runs. Figure 15.3a shows a run with maximal segregation $S = 1$, Fig. 15.3b displays a run with minimal segregation $S \approx 0.522$.

Figure 15.3 reveals remarkable differences between the two runs. In the maximally segregated population (Fig. 15.3a), members of both groups were quickly drawn to almost perfect population-wide consensus on an opinion at approximately 0.5. After 100,000 time steps, the standard deviation of opinions declined to practically zero.[2] This outcome occurred in about 90% of all runs in this condition. In the maximally desegregated population (Fig. 15.3b), the result was strikingly different.

[2]Perfect consensus is only obtained in the time limit. The simulation program computed a standard deviation of about 2.31 10^{-9} after 100,000 time steps for this run.



(a) Maximally segregated network ($S = 1$) (b) Maximally desegregated network ($S \approx 0.522$)

Fig. 15.3 Change opinions in single runs without initial difference between group means ($N = 100$, $\beta_D = 1/3$, $r = 5$). Dark: group 0, Light: group 1

The population was split almost perfectly into two opposing camps after 5000 time steps. Population polarization reached in this condition a level of $P_t^p \geq 0.99$ in 97% of all runs at $t = 100{,}000$. Yet, the emergent camps were not perfectly divided between groups. On average, between-group polarization was about $P_t^g = 0.497$ in the final state.

The strong difference between the two scenarios can be explained as follows. With high segregation, only few agents had neighbors who belong to another group. If interacting agents i and j belong to the same group, it is highly unlikely that their initial opinion disagreement is large enough to trigger repulsive influence ($w_{ij} < 0$). This happens only when their disagreement exceeds $|o_{jt} - o_{it}| = 0.75$. However, with the initial distribution of $Beta(3, 3)$ this was practically impossible.[3] Thus, within both groups influence was overwhelmingly assimilative, pulling all agents towards the mean value of the initial distribution (0.5). Only those few agents who were located on the interface between groups had outgroup-neighbors. With outgroup-neighbors, disagreement only needed to exceed $|o_{jt} - o_{it}| = 0.5$ to trigger repulsive influence. In a randomly chosen pair of neighbors from different groups, this happens at the outset with a probability of about 0.056. The few events of repulsive influence that occurred pushed agents to move away from each other towards the extremes of the opinion space. However, in most cases they were pulled back towards less extreme opinions in subsequent interactions with moderate ingroup members. This explains why in this condition about 90% of all runs ended in consensus. Yet, in the remaining approximately 10% of runs, interactions on the interface of groups became repulsive, driving agents on opposite sides of the boundary towards opposite extremes in the opinion space. Consensus within groups remained high at the same time. As a consequence, opinions of the two groups were driven apart. These runs ended in almost perfect between-group polarization.

In a highly desegregated population outcomes were different. With $S \approx 0.522$ agents had on average about 50% outgroup-neighbors. Likely, on at least some places in the network neighboring agents disagreed enough to develop repulsive influence. As a consequence, they increasingly shifted opinions away from each other, towards

[3]The probability was about 0.00155.

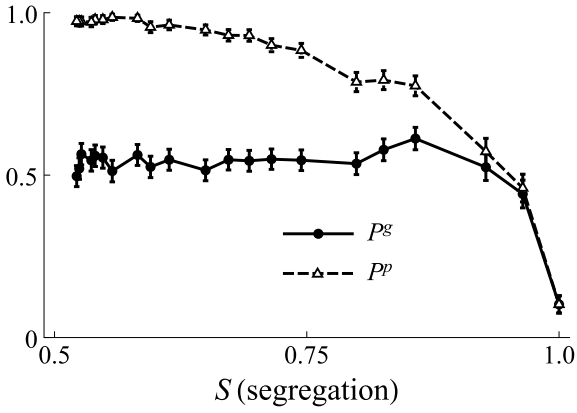


Fig. 15.4 Experiment 1: Effect of segregation on polarization measures. Bars indicate 0.95 CI around mean values

opposing extremes on the opinion scale. The dynamic of increasing differentiation between neighboring agents occurred simultaneously in different local regions of the network, because groups were well-mixed in this condition. This explains why population polarization reached its maximum of $P^p = 1.0$ here. At the same time, members of the two groups differentiated in different ways from each other in different local regions of the network. Thus, within both groups, members moved to both extreme ends of the opinion spectrum. Within the same group different poles were adopted at different places in the network. This was the reason why between-group polarization fell far below its theoretical maximum, with about $P^g = 0.5$ on average in the final state.

Figure 15.4 reports the results of experiment 1 for the entire range of segregation levels that were inspected. More precisely, the figure shows how the level of segregation S affected between-group polarization P^g and population polarization P^p in the final state, averaged across 500 realizations per condition. In line with the explanation given above, less segregation was on the whole associated with more population polarization. Also between-group polarization is on the whole higher in desegregated networks than in highly segregated ones. However, Fig. 15.4 also shows a non-linear association. When segregation increased from its minimum of $S \approx 0.522$, average between-group polarization remained fairly constant up to about $S \approx 0.8$, then increased to its peak-level at $S \approx 0.85$, to finally drop to a minimum of $P^g = 0.102$ in maximally segregated networks.

Figure 15.5 helps explaining the non-linearity identified by Fig. 15.4. Figure 15.5 shows the effect of segregation on the proportion of three types of outcomes in the final state, consensus ($P^p \leq 0.01$), group-split ($P^g \geq 0.99$) and population-split ($P^p \geq 0.99$). The share of runs with population-split and with consensus were largely mirror images of each other in this experiment. The more runs generated population-split, the less runs ended in consensus. In other words, lower segregation increasingly drove populations into a polarized state. But this state did not need to be group-split. In the region between about $S = 0.8$ and $S = 0.9$ population-split decreased with

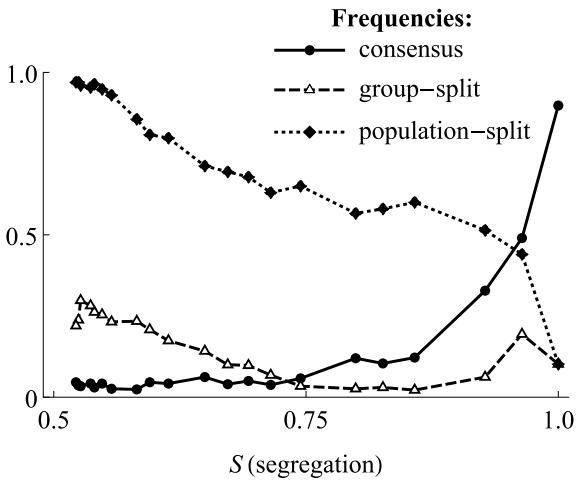


Fig. 15.5 Experiment 1: Effect of segregation on proportions of three outcome types

more segregation, while group-split increased. Above $S = 0.9$ both lines move again in the same direction.

The reason for the difference between group-split and population-split is the spatial coherence of groups under high segregation. This can be best understood by traversing the change of outcome proportions from right to left in Fig. 15.5, starting from a maximally segregated population. As the figure shows, moderate amounts of 'mixing' induce more polarization. This is due to more between-group interactions. But moderate mixing does not yet disrupt the spatial connectedness of groups, which therefore can still develop internal consensus. Thus, population polarization largely was found to be between-group polarization between $S = 0.9$ and $S = 1$. Once the segregation level was reduced further below $S = 0.9$, more interactions across group boundaries fueled more polarization, while consensus within groups was disrupted at the same time, due to more disconnectedness within groups. This explains the simultaneous decline of group-split and increase of population polarization when segregation moves downward from $S = 0.9$ to $S = 0.8$. Only when segregation levels further declined, even more individuals were spread across the network so that again most had at least some members of their own group in their local network, allowing for more within-group coordination in the process of population polarization. As a consequence, group-split and population-split moved again in the same direction when segregation levels were lower than about $S = 0.8$. However, the low levels of group-split between about $S = 0.75$ and $S = 0.9$ do not show that there was no systematic disagreement between groups at all. With an interaction range of $r = 5$, most individuals are connected with ingroup peers at all levels of segregation. Thus, some degree of coordination remains, explaining that on average between-group disagreement never fell below about 0.5, as shown by Fig. 15.4. Moreover, declining levels of group-split between $S = 0.5$ and $S = 0.8$ did not show up in declining average between-group polarization P^g, because also the proportion of runs in consensus declined in favor of more runs with medium-levels of between-group polarization.

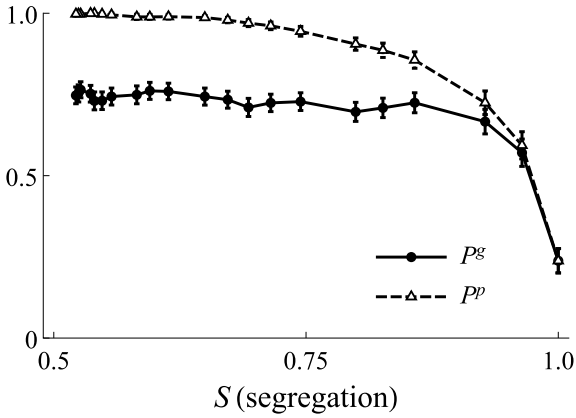


Fig. 15.6 Experiment 2: Effect of segregation on polarization measures. Bars indicate 0.95 CI around mean values

Experiment 2

Experiment 1 showed that considerable levels of between-group polarization came about in an unsegregated population, even when there was no systematic initial disagreement between groups. Experiment 2 tested whether the effect of segregation remained the same when mean opinions of the two groups differed from the outset.

Figures 15.6 and 15.7 report results of the ceteris-paribus replication of experiment 1, the only difference being that initial opinions were randomly drawn from the Beta-distributions shown in Fig. 15.2b. Comparison of Fig. 15.6 with the corresponding result for experiment 1 shows that on average between-group polarization in the final state was considerably higher across all levels of segregation. While in experiment 2 between-group polarization declined from about $P^g = 0.75$ to $P^g = 0.25$ between the lowest and the highest level of segregation, this decline happened at a lower level ($P^g \approx 0.5$ to $P^g \approx 0.1$) in experiment 1. Also population polarization was consistently higher in experiment 2, but this difference was less pronounced. A further noteworthy difference was that there was no longer a discernible non-monotonous effect of segregation on between-group polarization.

Figure 15.7 further confirms these differences and helps to explain them. The share of runs ending in group-split was slightly but consistently above the levels found in experiment 1, while the share of runs ending in consensus was slightly but consistently below this level. Population-split clearly was at a considerably higher level. The most striking qualitative difference was that group-split did no longer increase when small amounts of mixing were added to a maximally segregated network, but instead started to drop immediately. This illustrates the most important explanation for the differences between the experiments. In experiment 2, initial between-group differences were high enough to trigger mutual distancing on the interface between groups. Thus about 25% of runs were ending in group-split in the maximally segregated networks. While reducing segregation from this point fueled more population polarization - like in experiment 1 - it also blurred the boundaries between the groups

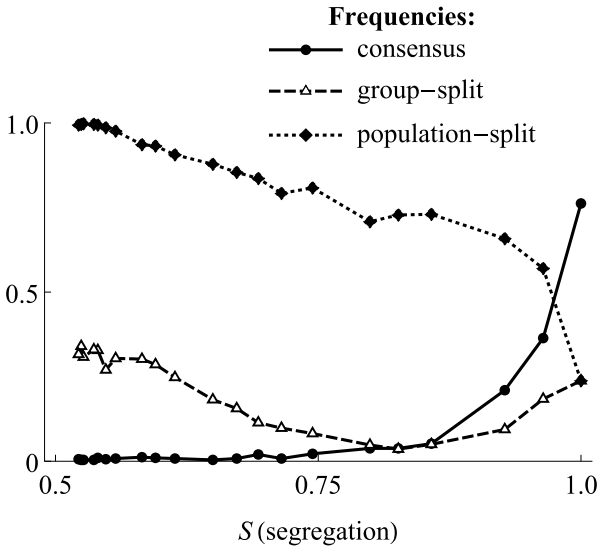


Fig. 15.7 Experiment 2: Effect of segregation on proportions of three outcome types

at the global level. Locally, agents from different groups were even more prone to end up on opposing sides of the spectrum than in experiment 1, but this was globally less coordinated than in the maximally segregated networks. This explains why starting from a maximally segregated network, mixing groups immediately decreased group-split in experiment 2, unlike it did in experiment 1.

Robustness Tests

The results of experiment 1 and experiment 2 rest on a number of assumptions about the model and the specific scenario. A full exploration of the robustness of results against meaningful variations is impossible in the space of this paper. As a start, I conducted two main robustness tests. First, a ceteris-paribus replication of both experiments was conducted with the range of interaction reduced from $r = 5$ to $r = 1$. A smaller range of interaction greatly reduces the interface between groups in highly segregated populations and inhibits the spreading of locally emergent extreme opinions in the network. The robustness test showed that this did not change the main qualitative effects of segregation found in experiments 1 and 2. More precisely, replicating experiment 1 with $r = 1$, it was found that increasing segregation from its minimal level first slightly increased, then reduced between-group polarization, while population polarization was reduced monotonously. Similarly, replicating experiment 2 it was found that there was no more non-monotonicity under higher initial between-group disagreement, but more segregation still reduced polarization

both in the population and between groups. However, it should be noted that $N1000$ time steps were not enough to obtain stable outcomes in all conditions with $r = 1$. This is especially not the case in highly segregated networks, where polarization that starts on the interface between groups can take a long time to spread and pull all group members into opposing camps in the sparse network with $r = 1$.

The second robustness test was to reduce the relative impact of demographic group differences on discrepancy in the influence process. A relatively lower value of β_D can be expected to reduce the overall potential for polarization, because individuals from different groups need more disagreement to develop a mutually negative relationship. To assess this, a ceteris-paribus replication of experiments 1 and 2 was conducted, setting $\beta_D = {}^1/_4$ (vs. $\beta_D = {}^1/_3$ in the baseline condition). As expected, both forms of polarization declined. Most importantly, segregation still reduced polarization, where the difference between segregated and desegregated networks was actually larger than for $\beta_D = {}^1/_3$ across both experiments.

Discussion and Conclusion

Intuitive reasoning as well as a number of formal models of opinion dynamics suggest that cultural diversity can under certain conditions be a threat to societal consensus, despite all its undoubted benefits. It has been argued that polarization between groups in a diverse society may be particularly likely when the society is highly segregated, echoing concerns raised by former U.S. president Barack Obama and results obtained with formal models of socially complex opinion dynamics. In these models, interactions between like-minded people can make them more and more convinced of their prevailing opinion tendencies, resulting in opinions that are increasingly extreme and different from those outside of their segregated world [13, 31, 33]. I presented in this paper a formal model drawing on social-psychological theories of cognitive balance that points to the opposite conclusion. Building on earlier work [17, 20, 21, 26, 29], this model combines assimilative with repulsive social influence, assuming that mutual disagreement between interacting agents is particularly likely to become accentuated and extreme when they interact with members of other groups that are separated from them by socially salient boundaries.

The point of my paper is not to show that the one or the other line of modeling is right or wrong about the link between segregation and polarization. What I would like to demonstrate is that formal modeling of socially complex dynamics can help us to better understand counter-intuitive and often unanticipated consequences of simple and familiar principles of social interaction. Principles such as influence, repulsion, persuasion, homophily or xenophobia are well known from our daily lives and from research conducted by social scientists. However, their possible implications at the societal level are often less well understood. An important contribution of social complexity models is that they can focus attention of empirical researchers on testing those assumptions in models that can be particularly critical for key social outcomes, such as polarization between groups.

The two lines of modeling work discussed in this chapter serve as an example in case. The obvious contradiction between their implications has motivated researchers in recent years to conduct systematic behavioral experiments. While this endeavor is still in progress, some of this work speaks to the models presented here. For example, while several empirical studies point to some evidence for repulsive influence in experimental and field settings [24, 27], recent experimental research tested repulsive influence more systematically in a controlled lab setting and found no support [10, 45]. This suggests that repulsive influence may be less easily triggered in social interactions than most formals models assume. At the same time, experimental tests have been conducted that lend some support to models of argument persuasion [31]. However, it would be premature to therefore entirely discard the possibility that segregation may sometimes preclude polarization. Experimental tests hitherto could not capture situations of strong between-group antagonism nor could they observe groups with mutually strongly exclusive social identities, conditions that appear to be plausible candidates for triggering repulsive influence that may drive groups apart.

Social complexity models have revealed important challenges for our scientific understanding of polarization. In line with calls from recent reviews of the field [22, 42], I believe that for tackling these challenges, we need to move forward towards a deeper connection of formal models with empirical insights from behavioral experiments and field research in the social sciences. The potential threat from polarization in diverse societies is an issue important enough to merit this effort.

Acknowledgements An earlier version of this paper has been presented at the Econophys-2017 & APEC-2017 Conference, held in November 2017 at the Jawaharlal Nehru University and Delhi University, New Delhi, India. The author wishes to thank the participants and especially the organizers of the conference, as well as the editors of this volume, for creating a pleasant and intellectually stimulating environment for this work.

References

1. Allport, G.W.: The Nature of Prejudice. Addison-Wesley, Cambridge (1954)
2. Altafini, C.: Consensus problems on networks with antagonistic interactions. IEEE Trans. Automat. Contr. **58**(4), 935–946 (2013). https://doi.org/10.1109/TAC.2012.2224251
3. Baldassarri, D., Bearman, P.: Dynamics of political polarization. Am. Sociol. Rev. **72**(5), 784–811 (2007)
4. Banisch, S., Olbrich, E.: Opinion polarization by learning from social feedback. J. Math. Sociol., (2018). https://doi.org/10.1080/0022250X.2018.1517761
5. Bianchi, F., Squazzoni, F.: Agent-based models in sociology. Wiley Interdiscip. Rev. Comput. Stat. **7**(4), 284–306 (2015). https://doi.org/10.1002/wics.1356
6. Bramson, A., Grim, P., Singer, D.J., Fisher, S., Berger, W., Sack, G., Flocken, C.: Disambiguation of social polarization concepts and measures. J. Math. Sociol. **40**(2), 80–111 (2016). https://doi.org/10.1080/0022250X.2016.1147443
7. Centola, D.: The social origins of networks and diffusion. Am. J. Sociol. **120**(5), 1295–1338 (2015). https://doi.org/10.1086/681275
8. Christ, O., Schmid, K., Lolliot, S., Swart, H., Stolle, D., Tausch, N., Al Ramiah, A., Wagner, U., Vertovec, S., Hewstone, M.: Contextual effect of positive intergroup contact on outgroup

prejudice. Proc. Natl. Acad. Sci. **111**(11), 3996–4000 (2014). https://doi.org/10.1073/pnas.1320901111
9. Clark, W.A.V., Fossett, M.: Understanding the social context of the Schelling segregation model. Proc. Natl. Acad. Sci. U. S. A. **105**(11), 4109–4114 (2008)
10. Clemm von Hohenberg, B., Maes, M., Pradelski, B.: Micro influence and macro dynamics of opinion formation (2017). https://ssrn.com/abstract=2974413
11. Collier, P.: Exodus. Immigration and Multiculturalism in the 21st Century. Penguin, London (2013)
12. Currarini, S., Jackson, M.O., Pin, P.: Identifying the roles of race-based choice and chance in high school friendship network formation. Proc. Natl. Acad. Sci. U. S. A. **107**(11), 4857–4861 (2010)
13. Dandekar, P., Goel, A., Lee, D.T.: Biased assimilation, homophily, and the dynamics of polarization. Proc. Natl. Acad. Sci. U. S. A. **110**(15), 5791–5796 (2013)
14. Dovidio, J.F., Love, A., Schellhaas, F.M.H., Hewstone, M.: Reducing intergroup bias through intergroup contact: twenty years of progress and future directions. Group Process. Intergroup Relat. **20**(5), 606–620 (2017). https://doi.org/10.1177/1368430217712052
15. Eger, S.: Opinion dynamics and wisdom under out-group discrimination. Math. Soc. Sci. **80**, 97–107 (2016). https://doi.org/10.1016/j.mathsocsci.2016.02.005
16. Ellemers, N., Rink, F.: Diversity in work groups. Curr. Opin. Psychol. **11**, 49–53 (2016). https://doi.org/10.1016/j.copsyc.2016.06.001
17. Feliciani, T., Flache, A., Tolsma, J.: How, when and where can spatial segregation induce opinion polarization? Two competing models. JASSS **20**(2), (2017). https://doi.org/10.18564/jasss.3419. http://jasss.soc.surrey.ac.uk/20/2/6.html
18. Fent, T., Groeber, P., Schweitzer, F.: Coexistence of social norms based on in- and out-group interactions. Adv. Complex Syst. **10**, 271–286 (2007). https://doi.org/10.1142/S0219525907000970
19. Festinger, L.: A Theory of Cognitive Dissonance. Stanford University Press, Stanford (1957)
20. Flache, A., Macy, M.W.: Small worlds and cultural polarization. J. Math. Sociol. **35**(1–3), 146–176 (2011). https://doi.org/10.1080/0022250X.2010.532261
21. Flache, A., Mäs, M.: How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams. Comput. Math. Organ. Theory **14**(1), 23–51 (2008). https://doi.org/10.1080/0022250X.2010.532261
22. Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., Lorenz, J.: Models of social influence: towards the next frontiers. JASSS **20**(4), (2017). https://doi.org/10.18564/jasss.3521. http://jasss.soc.surrey.ac.uk/20/4/2.html
23. Heider, F.: Attitudes and cognitive organization. J. Psychol. **21**(1), 107–112 (1946)
24. Hovland, C.I., Harvey, O.J., Sherif, M.: Assimilation and contrast effects in reactions to communication and attitude change. J. Abnorm. Soc. Psychol. **55**(2), 244–252 (1957). https://doi.org/10.1037/h0048480
25. Huet, S., Deffuant, G.: Openness leads to opinion stability and narrowness to volatility. Adv. Complex Syst. **13**(3), 405–423 (2010). https://doi.org/10.1142/S0219525910002633
26. Jager, W., Amblard, F.: Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. Comput. Math. Organ. Theory **10**(4), 295–303 (2005)
27. Liu, C.C., Srivastava, S.B.: Pulling closer and moving apart: interaction, identity, and influence in the U.S. Senate, 1973–2009. Am. Sociol. Rev. **80**(1), 192–217 (2015). https://doi.org/10.1177/0003122414564182
28. Macy, M.W., Flache, A.: Social dynamics from the bottom up: agent-based models of social interaction. In: Bearman, P., Hedström, P. (eds.) Oxford Handbook of Analytical Sociology, pp. 245–268. Oxford University Press, Oxford (2009)
29. Macy, M.W., Kitts, J., Flache, A., Benard, S.: Polarization and dynamic networks. a hopfield model of emergent structure. In: Breiger, R., Carley, K., Pattison, P. (eds.) Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, pp. 162–173. The National Academies Press, Washington (2003)

30. Mark, N.P.: Culture and competition: homophily and distancing explanations for cultural niches. Am. Sociol. Rev. **68**(3), 319–345 (2003)
31. Mäs, M., Flache, A.: Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. PLoS One **8**(11), (2013). https://doi.org/10.1371/journal.pone.0074516
32. Mäs, M., Flache, A., Kitts, J.A.: Cultural integration and differentiation in groups and organizations (2014). https://doi.org/10.1007/978-3-319-01952-9_5
33. Mäs, M., Flache, A., Takács, K., Jehn, K.A.: In the short term we divide, in the long term we unite: demographic crisscrossing and the effects of faultlines on subgroup polarization. Organ. Sci. **24**(3), 716–736 (2013). https://doi.org/10.1287/orsc.1120.0767
34. Mason, W.A., Conrey, F.R., Smith, E.R.: Situating social influence processes: dynamic, multidirectional flows of influence within social networks. Personal. Soc. Psychol. Rev. **11**(3), 279–300 (2007). https://doi.org/10.1177/1088868307301032
35. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Annu. Rev. Sociol. **27**(1), 415–444 (2001)
36. Moody, J.: Race, school integration, and friendship segregation in America. Am. J. Sociol. **107**(3), 679–716 (2001). https://doi.org/10.1086/338954
37. Myers, D.G.: Polarizing effects of social interaction. In: Brandstätter, H., Davis, J.H., Stocker-Kreichgauer, G. (eds.) Group Decision Making, pp. 125–161. Academic Press, London (1982)
38. Norris, P., Inglehart, R.F.: Muslim integration into western cultures: between origins and destinations. Polit. Stud. **60**(2), 228–251 (2012). https://doi.org/10.1111/j.1467-9248.2012.00951.x
39. Obama, B.: Farewell address (2017). Retrieved 06 07 2018. https://obamawhitehouse.archives.gov/Farewell
40. Pettigrew, T.F., Tropp, L.R.: A meta-analytic test of intergroup contact theory. J. Pers. Soc. Psychol. **90**, 751–783 (2006)
41. Proskurnikov, A.V., Matveev, A.S., Cao, M.: Opinion dynamics in social networks with hostile camps: consensus versus polarization. IEEE Trans. Automat. Contr. **61**(6), 1524–1536 (2016). https://doi.org/10.1109/TAC.2015.2471655
42. Sobkowicz, P.: Modelling opinion formation with physics tools: call for closer link with reality. J. Artif. Soc. Soc. Simul. **12**(1), 11 (2009). http://jasss.soc.surrey.ac.uk/12/1/11.html
43. Stark, T.H., Flache, A., Veenstra, R.: Generalization of positive and negative attitudes toward individuals to outgroup attitudes. Personal. Soc. Psychol. Bull. **39**(5), 608–622 (2013). https://doi.org/10.1177/0146167213480890
44. Stark, T.H., Mäs, M., Flache, A.: Liking and disliking minority-group classmates: explaining the mixed findings for the influence of ethnic classroom composition on interethnic attitudes. Soc. Sci. Res. **50**, 164–176 (2015). https://doi.org/10.1016/j.ssresearch.2014.11.008
45. Takács, K., Flache, A., Mäs, M.: Discrepancy and disliking do not induce negative opinion shifts. PLoS One **11**(6), e0157,948 (2016). https://doi.org/10.1371/journal.pone.0157948
46. Vinokur, A., Burnstein, E.: Depolarization of attitudes in groups. J. Pers. Soc. Psychol. **36**(8), 872–885 (1978)

Chapter 16
Competitive Novel Dual Rumour Diffusion Model

Utkarsh Niranjan, Anurag Singh and Ramesh Kumar Agrawal

Abstract Rumors have been present in society for very long. In modern days with the better penetration of online social media technologies, we have been able to share anything with anybody. Rumors have become an undesirable but built-in feature of social networking technologies. In this work, we represent a model of dual rumor propagation in population. Our model is an extension of the basic SIR model with six states. We present a detailed numerical analysis of our model to show the impact of various parameters on the density of nodes in different states. When two rumors are competing in the population the rumor with high spreading rates wins the race. In our model, we also present a study of the impact of individuals biasness toward one type of rumor. This biasness arises if a rumor is originating from a popular and credible source. For a relatively high stifling rate with respect to spreading rate, we find that a large fraction of population remains ignorant of rumors.

Introduction

A mammoth volume of individuals is online spreading all sort of information. According to the Facebook newsroom, 1.47 billion active daily users and 2.23 billion active monthly users access their Facebook accounts to perform social networking activities. Official YouTube google blog states that 1.9 billion logged-in users come to YouTube every month to create and/or enjoy videos. Similarly, WhatsApp also

U. Niranjan · R. K. Agrawal (✉)
School of Computer and Systems Sciences, Jawaharlal Nehru University,
New Delhi 110067, India
e-mail: rka@gmail.com

U. Niranjan
e-mail: utkarshkumarniranjan@gmail.com

A. Singh
Department of Computer Science and Engineering, National Institute of Technology Delhi,
New Delhi 110040, India
e-mail: name@email.address

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_16

provides messenger services to more than 1 billion users daily. With the help of online social media networking technologies, People are able to share information with any number of individuals simultaneously. In summary, we can say that online social networking is a great technology. It provides easy access to information to the people. Every new technology comes with its own set of challenges.

One of the challenges of social networking is the reliability and truth value of the available information. Information available online finds its way to you through various sources. The identity of these sources is sometimes unknown. A widely disseminated opinion or information from the discernible source is called rumor. These rumors are present in every section of society. Rumors are present in educational campuses, political arenas, movie industry, and product market. Any sector you name it rumors are present there. Rumors have great power to influence individuals opinion about things they are connected. Rumors can make a new product a phenomenon overnight. Rumors can change the outcome of an election. Rumors can also tarnish the image of any public figure in a few hours.

Information, rumors, malicious computer programs (viruses, worms etc.), and disease-causing microorganisms (viruses, bacteria, fungi etc.) disseminate in similar fashion on any given contact network [7, 8, 12]. This similarity in the dissemination process stimulated researchers to use already well-developed science of human epidemiology in the field of rumor spreading mechanism and computer epidemics. Susceptible-Infected-Removed/Recovered (SIR) model [3] is one of the early models of human epidemiology. SIR model was proposed by Kermack, McKendrick in 1927. In this model whole population of individuals is divided into three groups namely susceptible, infected and removed/recovered. An individual in the susceptible state is free from contagion but can acquire contagion from any of infected neighbor. The infected state represents the group of individuals who are carrying in the contagion and also able to spread it to their neighbors. Individuals in the recovered state are free from infection and immune to the infection. The Initial model of rumor spreading proposed by Daley and Kendall [1] was also inspired by SIR model. This model is also known as the DK model. Another model of rumor spreading was proposed by Maki and Thompson [6] in 1973. Both of these models assume homogeneous mixing of the population. A variant of Susceptible-Exposed-Infected-Removed (SEIR) model for rumor propagation is presented by Xia et al. [10]. In this model, they consider hesitating mechanism with the help of exposed (E) state. An individual in the exposed state is hesitating to spread the rumor. They also presented simulation on the small world and BA networks. Agent-based model for rumor spreading analysis is presented by Serrano et al. [9]. In this study, they used two twitter rumor datasets. Their model depends on the hypothesis that a recovered user will not influence any neighbor in the network to change their state as recovered. Zhao et al. [11] proposed an extension of the SIR model for rumor propagation based on Propagation force. They defined propagation force as a fuzzy variable. They also introduced a fuzzy reproduction number. Li and Ma [4] studied the role of governments policies and individuals sensitivity toward rumor with SIR and SIS model.

They also present a study on Facebook and POK social networks. In their analysis, they find out that increasing punishment for spreading false rumor and individual sensitivity towards rumors can control the spreading rumor. Hu et al. [2] presented a rumor spreading model with wisemen present in the network. Wisemen act as a sink for the rumor. In their work authors presents a system of differential equations. They also presented stability analysis of Rumor-free and Rumor equilibrium using Routh–Hurwitz criteria. An extended version of SIR model is proposed by Liu et al. [5] for competitive information diffusion. They called their model Susceptible-Hesitated-Infected-Removed (SHIR). In this model, two competing pieces of information are spreading in the network. They also gave a detailed analysis of their model. Zhuang et al. [13] proposed an information diffusion model for dual information propagation in the network. They considered possibilities of cooperation and competition between the dual information on any topic. In this model Zhaung et al. also presented an analysis of real-world data from weibo.com. Another challenge with the rumors present in social networking is that often we find two or more rumors spreading online about the same topic. These rumors can be either competing or cooperation piece of information. Real issues arise with competing rumors. It is very difficult to decide which rumor is true. Very few previous works [5, 13] focuses on the competing rumor spreading modeling. In the SHIR model authors assume the diffusion of two competing information. In this model, they assume that a node in the ignorant state can acquire both the information simultaneously and move to a hesitator state. Hesitators are the nodes which are influenced by both information and also spreading both information. In a real-world scenario getting two information at the same instance of time is almost impossible so this transmission is not very realistic. In the Zhuangs [13] model there is no state for the nodes influenced by both the pieces of information. They assumed that spreading rate is same for both the piece of information.

In our work, we propose a more accurate model for competing rumors/information. In our work, we present a model with six states. We consider different spreading rates for different rumor. In our model, no ignorant node can directly become hesitater/vacillator. Stiflers of different rumors are considered to be different in our model. Information coming from a famous and credible source is considered to be truer than information coming from an ordinary source. Individuals in a population has a bias towards such rumor. In our work, we also model this biasness. No other previous study considers biasness of the population towards one type of information. We present a detailed numerical analysis of our model comprising the impact of different parameters.

Remaining sections of the paper are organized as follows. Section “Model” explains the model. Section “Numerical Simulation Experiments” presents the numerical and result analysis. In section “Conclusion” we present the conclusion of our paper.

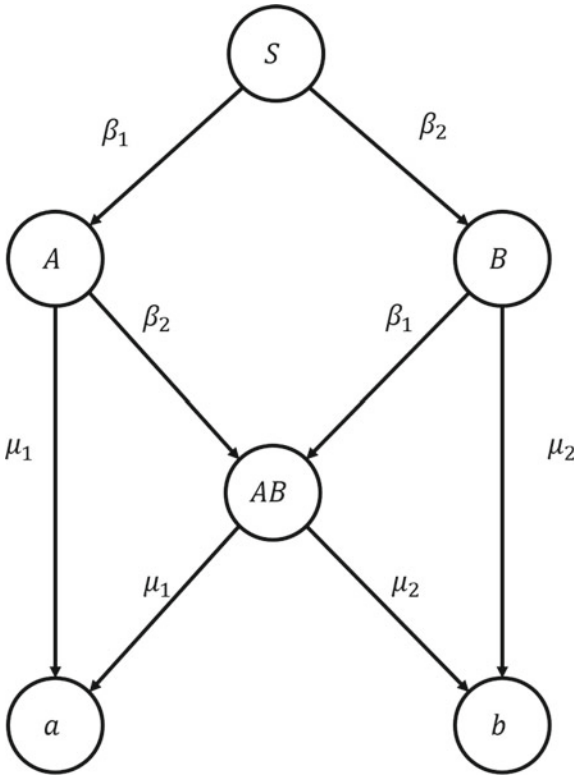


Fig. 16.1 Block diagram of model

Model

In this paper, we propose a model of dual information diffusion. In this model, there are six states. A node can exist in any of the six states at any point of time. These sates are shown in Fig. 16.1 and detail description of states is given below.

State Description

There are six states in the proposed model namely S, A, B, AB, a and b. Nodes in state S are susceptible/ignorant and are unaware of any information. The population density of susceptible/ignorant nodes at time t is denoted by $S_S(t)$. A node in state A is capable of information A and node in state B is capable of spreading information B. $S_A(t)$ and $S_B(t)$ denote the population densities in state A and state B respectively at time t. A vacillator node in state AB is capable of spreading both information A and B and $S_{AB}(t)$ represents the population density in state AB at time t. $S_a(t)$ and $S_b(t)$ denote the population of striflers of information A and information B at time t.

Table 16.1 Different notation and their meaning

Notation	Definition
N	Total number of nodes in the system
$S_S(t)$	Population density in state S, or ignorant nodes
$S_A(t)$	Population density in state A, or spreaders of Information A
$S_B(t)$	Population density in state B, or spreaders of Information B
$S_{AB}(t)$	Population density in state AB, or spreaders of both information A and B
$S_a(t)$	Population density in state a, or stiflers of Information A
$S_b(t)$	Population density in state b, or stiflers of Information A
α_1, α_2	Frequencies with which a vacillator node spreads information A and Information B respectively
β_1, β_2	Spreading rates for information A and Information B respectively
μ_1, μ_2	Stifling rates for information A and Information B respectively

The sum of the nodes in all these state is N at any time, i.e. $S_S + S_A + S_B + S_{AB} + S_a + S_b = N$, where N is population of the system the total. Different symbols used and their definition have been given in Table 16.1.

State Transition Dynamics

From state S,a node can have following transitions,

- When a an ignorant node in S state communicates with a node in state A or AB, it can move to state A with the rate β_1.
- When a an ignorant node in S state communicates with a node in state B or AB, it can move to state B with the rate β_2.

From state A, a node can have following transitions,

- When a spreader node in state A communicates with a node in state AB or B, it can move to state AB with the rate β_2.
- When a spreader node in state A communicates with a node in the state a or A, it can move to state a with the rate μ_1.

From state B, a node can have following transitions

- When a spreader node in state B communicates with a node in state AB or A, it can move to state AB with the rate β_1.
- When a spreader node in state B communicates with a node in state B or b, it can move to state b with the rate μ_2.

From state AB, a node can have following transitions

- When a spreader node in state AB communicates with a node in state A or a, it can move to state a with the rate μ_1.
- When a spreader node in state AB communicates with a node in state B or b, it can move to state b with the rate μ_2.

α_1 and α_2 are the frequencies with which a vacillator (state AB) node spreads information A and B respectively. The average probability of receiving information A at any node

$$R_A = \beta_1 \left(S_A(t) + \frac{\alpha_1}{\alpha_1 + \alpha_2} S_{AB}(t) \right) \tag{16.1}$$

The average probability of receiving information B at any node

$$R_B = \beta_2 \left(S_B(t) + \frac{\alpha_2}{\alpha_1 + \alpha_2} S_{AB}(t) \right) \tag{16.2}$$

The average probability of becoming stifler information A at any node

$$R_a = \mu_1 \left(S_a(t) + S_A(t) \right) \tag{16.3}$$

The average probability of becoming stifler information B at any node

$$R_b = \mu_2 \left(S_b(t) + S_B(t) \right) \tag{16.4}$$

With the help of state transition dynamics and probabilities defined in Eqs. (16.1–16.4), we can establish the following mathematical model.

$$\frac{dS_S(t)}{dt} = -S_S(t) \left(k\beta_1 \left(S_A(t) + \frac{\alpha_1}{\alpha_1 + \alpha_2} S_{AB}(t) \right) + k\beta_2 \left(S_B(t) + \frac{\alpha_2}{\alpha_1 + \alpha_2} S_{AB}(t) \right) \right) \tag{16.5}$$

$$\frac{dS_A(t)}{dt} = S_S(t) k\beta_1 \left(S_A(t) + \tfrac{\alpha_1}{\alpha_1+\alpha_2} S_{AB}(t) \right) - S_A(t) k\beta_2 \left(S_B(t) + \tfrac{\alpha_2}{\alpha_1+\alpha_2} S_{AB}(t) \right) - S_A(t) k\mu_1 \left(S_a(t) + S_A(t) \right) \tag{16.6}$$

$$\frac{dS_B(t)}{dt} = S_S(t) k\beta_2 \left(S_B(t) + \tfrac{\alpha_2}{\alpha_1+\alpha_2} S_{AB}(t) \right) - S_B(t) k\beta_1 \left(S_A(t) + \tfrac{\alpha_1}{\alpha_1+\alpha_2} S_{AB}(t) \right) - S_B(t) k\mu_2 \left(S_b(t) + S_B(t) \right) \tag{16.7}$$

$$\frac{dS_{AB}(t)}{dt} = S_A(t) k\beta_2 \left(S_B(t) + \tfrac{\alpha_2}{\alpha_1+\alpha_2} S_{AB}(t) \right) + S_B(t) k\beta_1 \left(S_A(t) + \tfrac{\alpha_1}{\alpha_1+\alpha_2} S_{AB}(t) \right) - S_{AB}(t) \left(k\mu_1 \left(S_a(t) + S_A(t) \right) + k\mu_2 \left(S_b(t) + S_B(t) \right) \right) \tag{16.8}$$

$$\frac{dS_a(t)}{dt} = \left(S_A(t) + S_{AB}(t) \right) k\mu_1 \left(S_a(t) + S_A(t) \right) \tag{16.9}$$

$$\frac{dS_b(t)}{dt} = \left(S_B(t) + S_{AB}(t) \right) k\mu_2 \left(S_b(t) + S_B(t) \right) \tag{16.10}$$

Where k is the average degree of the network.

Numerical Simulation Experiments

To examine the impact of different parameters in the model on the process of competitive Information/rumor diffusion, we perform numerical analysis of our model equation (16.5–16.10). Initial conditions for numerical solutions are given below

$$S_S(0) = N - 2,\ S_A(0) = 1,\ S_B(0) = 1,\ S_{AB}(0) = 0,\ S_a(0) = 0,\ S_b(0) = 0. \tag{16.11}$$

For simplicity, we have taken values of all parameters to be same. We also present visual results in the form of plots for mathematical analysis. Densities of nodes in different states are plotted in all plots. In Fig. 16.2 we present a plot for the numerical solutions and simulation on a homogeneous network. In Fig. 16.2a we have overlapping curves for the population densities of spreaders and stiflers in both the states. These overlapping curves are present in results because our model is symmetric and we have taken the same values for parameters. Figure 16.2a, b are not in proper agreement because numerical simulation executes in continuous time and network simulation runs in discrete time steps. We are using random numbers to compare probability so we are not getting overlapping curves in Fig. 16.2b, like numerical solutions.

Single Information Diffusion and Reproduction of SIR Model

In Fig. 16.3 we represent the population dynamics when only single rumor/information is spreading. When only single information is spreading, parameters

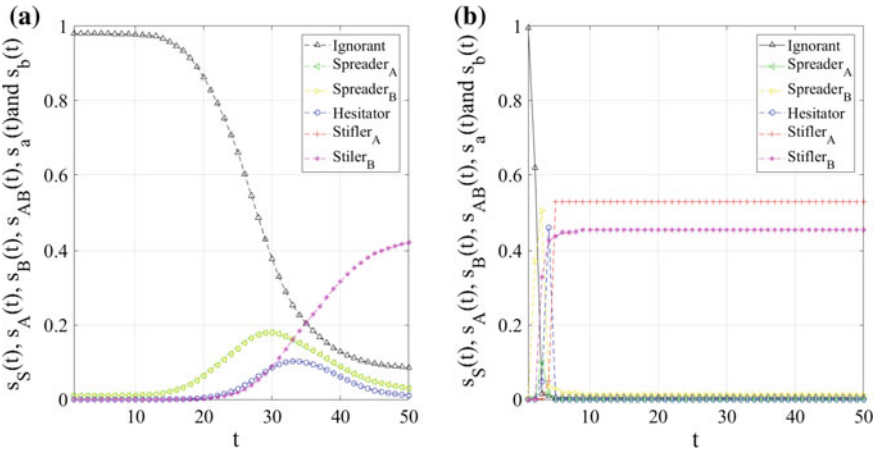


Fig. 16.2 The population density of nodes in different states **a** Numerical Solution **b** Homogeneous network simulation

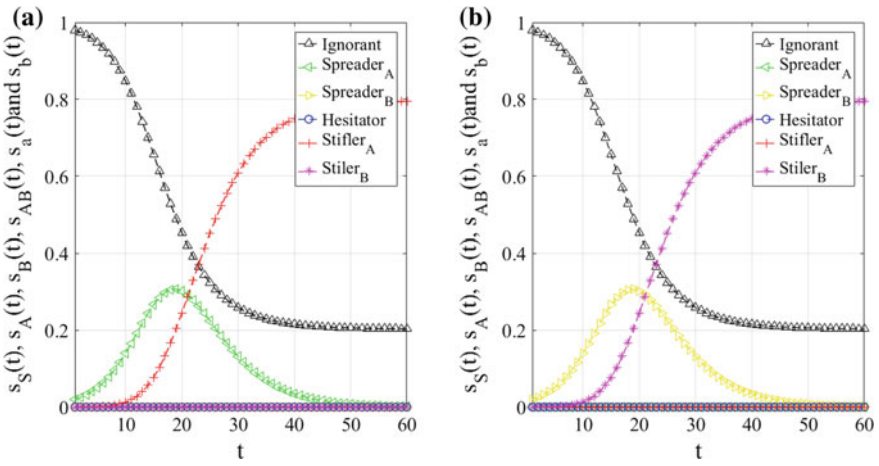


Fig. 16.3 The population density of nodes in different states **a** with Information A only **b** with information B only

related to the other information are zero i.e. if only information A is spreading in the network then we consider parameters related to Information B to be zero ($\alpha_2 = 0, \beta_2 = 0$ and $\mu_2 = 0$). So initial condition for a numerical solution in the presence of information A are

$$S_S(0) = N - 1, S_A(0) = 1, S_B(0) = 0, S_{AB}(0) = 0, S_a(0) = 0, S_b(0) = 0. \tag{16.12}$$

Similarly, we can have initial conditions when only information B is spreading in population by exchanging values of $S_A(0)$ and $S_B(0)$. Figure 16.3a presents population dynamics in the presence of information A only and Fig. 16.3b presents the population dynamics in the presence of information B only. Model represented by Eqs. (16.5–16.10) is an extension of the basic SIR model. So in the presence of only single information, we are able to reproduce the results of the basic SIR model.

Effect of Spreading Rates

In Fig. 16.4 we have plotted the effect of spreading rate on the density of stifler nodes. In Fig. 16.4a we vary the value of β_1 only and in Fig. 16.4b we vary the values of β_2 only. In both cases, all other parameters are kept fixed. We find that with the increase in the value of spreading rate the prevalence of information increase in the population. Since our model is symmetric so we have symmetric plots in Fig. 16.4a, b. In Fig. 16.5 we present the covariation of β_1 and β_2 on the penetration of rumors in the population. Population density of spreaders of both information are plotted in Fig. 16.5a–c for a different pair of β_1 and β_2 values. Similarly, we present the population density of Stiflers in Fig. 16.5d–f. The difference in the value

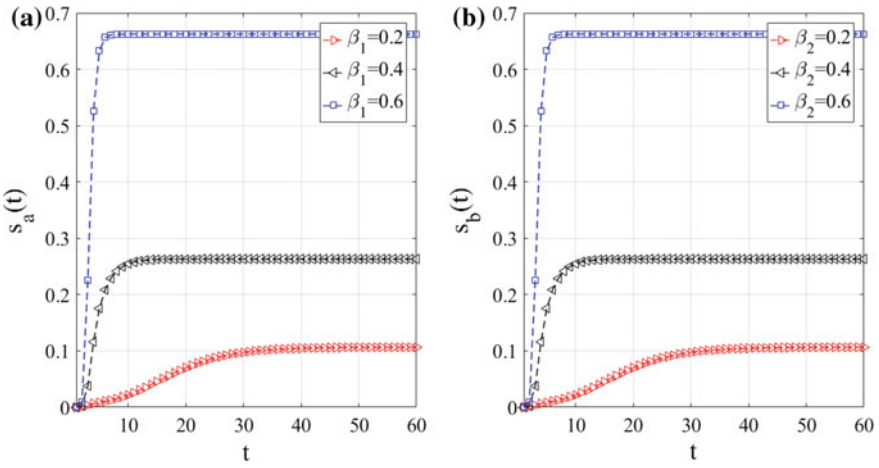


Fig. 16.4 The population density of nodes in Stifler state **a** for different values of β_1 **b** for different values of β_2

of β_1 and β_2 is more in Fig. 16.5a, c. In these figures, we can observe that the density of spreaders for the smaller β value persists for a longer period of time as compared to the density of spreaders for the larger β value, i.e. in Fig. 16.5a density of spreaders of information A ($\beta_1 = 0.3$) persist for a longer duration ($t \approx 40$) as compared to density of spreader of information B($\beta_2 = 0.5$). Similarly in Fig. 16.5c density of spreaders of information B ($\beta_2 = 0.5$) persists for longer duration ($t \approx 60$) than density of spreader of information A ($\beta_1 = 0.9$). Such behavior is present due to the fact that the average transition probability (equations 16.3 and 16.4) from spreaders to stiflers is function of stiflers density. So when the density of spreader is more the probability of transmission to stiflers is also more, which results in quick transmission to stiflers from spreaders. From Fig. 16.5d, f we can clearly notice that the density of stiflers is high for the information with higher spreading rate. In Fig. 16.5d, $\beta_1 = 0.3$ and $\beta_2 = 0.5$ so we have higher density of stiflers of information B. In Fig. 16.5f, $\beta_1 = 0.9$ and $\beta_2 = 0.5$, so we have a higher density of stiflers of information A. Therefor a rumor with higher spreading rate wins the race and become more prevailing with a high density of followers.

Effect of Population Biasness

An information either originating from some popular sources like some famous movie actors, players, politicians etc. or citing some credible source such as BCC, NASA, UNICEF etc. is considered to be truer than the information originating from some ordinary source. Such information creates a bias in population towards the information. In our model, we try to analyze this behavior with the help of parameter α. α_1 and α_2 are the frequencies with which a vacillator node spreads information A

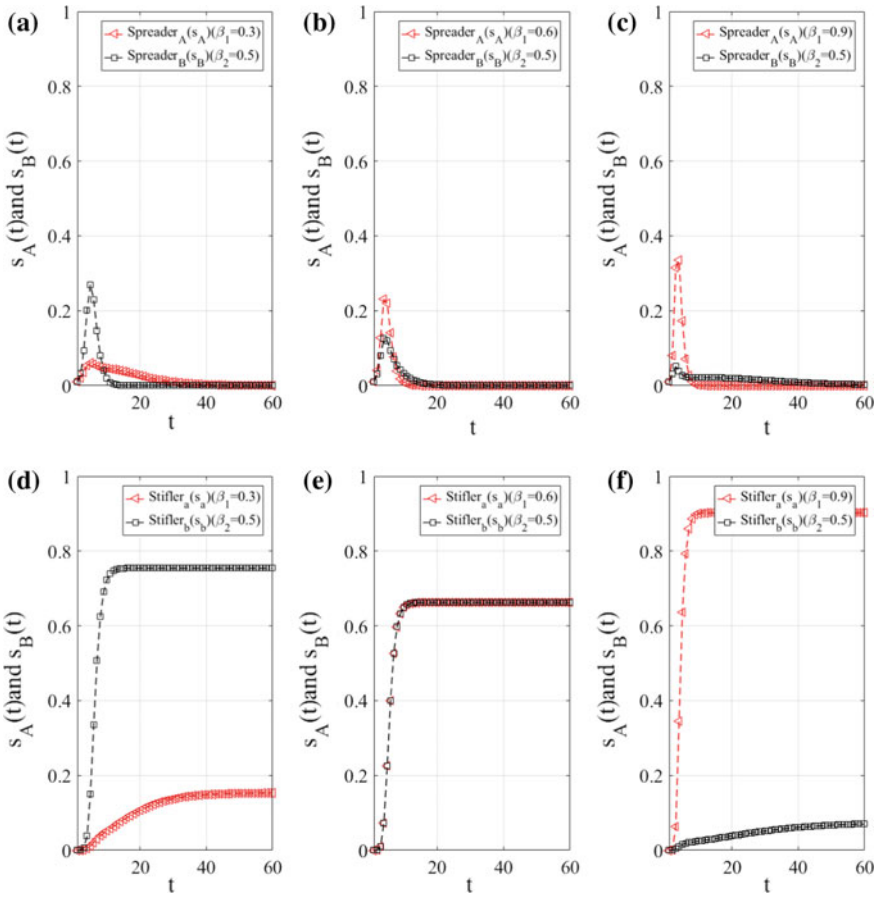


Fig. 16.5 Covariation of spreading rate **a, b, c** density of nodes in spreader state **d, e, f** density of nodes in stifler state

and B respectively. A higher value of α_1 than α_2 implies that more individuals in a population are supporting information A. From Fig. 16.6 it is clear that information with higher α value has more penetration in population. Which is very intuitive. This is the reason why big brands uses famous celebrities to endorse their products.

Effect of High Stifling Rates

Figure 16.7 analyzes the effect of relatively high stifling rates. If the values of stifling rates (μ_1, μ_2) are high as compared to the spreading rates (β_1, β_2) a large density of population remains in the ignorant state. This can be observed form Fig. 16.7. This behavior occurs because with the higher value of stifling rate, spreaders quickly become stiflers and a less population of the spreader is available to spread the information.

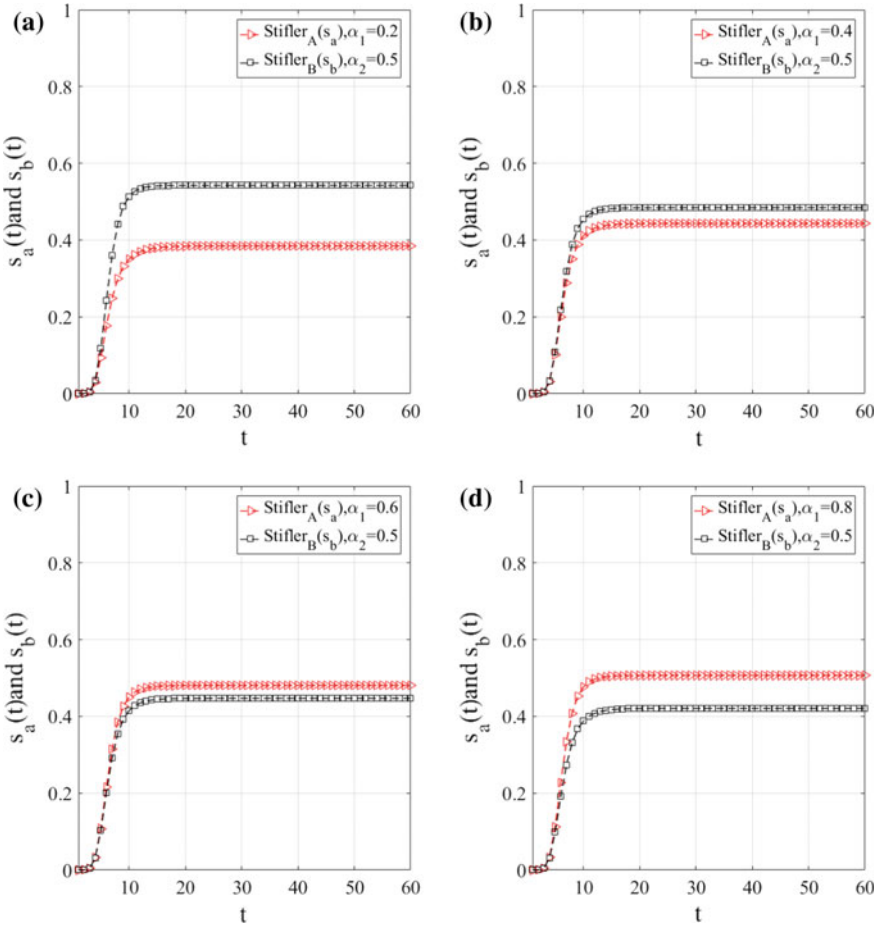


Fig. 16.6 Effect of biasness on the stiflers density of nodes

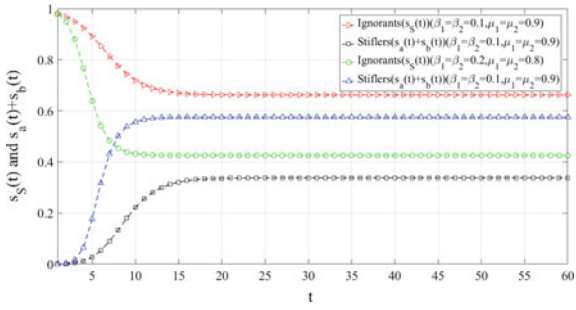


Fig. 16.7 Effect of relatively high stifling rate as compared with spreading rates

Conclusion

In this work, we present a new model for competitive dual rumor diffusion. In our model, we present six different states namely S, A, B, AB, a and b. We have performed a detailed numerical analysis for the impact of various parameters of models. We noted that to gain more penetration for a rumor relatively high spreading rate is required. In the presence of only one rumor in the population, we have successfully reproduces SIR model like behavior in our model. Which strengthen the fact that our model is an extension of the basic SIR model. We have also analyzed our model when there is individuals' bias towards one type of rumor. It is noted that the rumor with bias has more coverage in the population. If the stifling rate is relatively high as compared with spreading rate we find that a large fraction of population remains ignorant of rumors.

Acknowledgements First and third author are thankful to University Grant Commission (UGC), India and Department of Science and Technology- Promotion of University Research and Scientific Excellence (DST-PURSE), Government of India.

References

1. Daley, D.J., Kendall, D.G.: Stochastic rumours. IMA J. Appl. Math. **1**(1), 42–55 (1965)
2. Hu, Y., Pan, Q., Hou, W., He, M.: Rumor spreading model considering the proportion of wisemen in the crowd. Phys. A: Stat. Mech. Appl. **505**, 1084–1094 (2018)
3. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. Proc. R. Soc. Lond. A **138**(834), 55–83 (1932)
4. Li, D., Ma, J.: How the governments punishment and individuals sensitivity affect the rumor spreading in online social networks. Phys. A: Stat. Mech. Appl. **469**, 284–292 (2017)
5. Liu, Y., Diao, S.M., Zhu, Y.X., Liu, Q.: Shir competitive information diffusion model for online social media. Phys. A: Stat. Mech. Appl. **461**, 543–553 (2016)
6. Maki, D.P., Thompson, M.: Mathematical models and applications: with emphasis on the social life, and management sciences. Technical report (1973)
7. Murray, W.H.: The application of epidemiology to computer viruses. Comput. Secur. **7**(2), 139–145 (1988)
8. Newman, M.E.: Spread of epidemic disease on networks. Phys. Rev. E **66**(1), 016,128 (2002)
9. Serrano, E., Iglesias, C.Á., Garijo, M.: A novel agent-based rumor spreading model in twitter. In: Proceedings of the 24th International Conference on World Wide Web, pp. 811–814. ACM (2015)
10. Xia, L.L., Jiang, G.P., Song, B., Song, Y.R.: Rumor spreading model considering hesitating mechanism in complex social networks. Phys. A: Stat. Mech. Appl. **437**, 295–303 (2015)
11. Zhao, Z.j., Liu, Y.m., Wang, K.x.: An analysis of rumor propagation based on propagation force. Physica A: Statistical Mechanics and its Applications **443**, 263–271 (2016)
12. Zhou, J., Xiao, G., Cheong, S.A., Fu, X., Wong, L., Ma, S., Cheng, T.H.: Epidemic reemergence in adaptive complex networks. Phys. Rev. E **85**(3), 036,107 (2012)
13. Zhuang, Y.B., Chen, J., Li, Zh: Modeling the cooperative and competitive contagions in online social networks. Phys. A: Stat. Mech. Appl. **484**, 141–151 (2017)

Chapter 17
Dynamical Evolution of Anti-social Phenomena: A Data Science Approach

Syed Shariq Husain and Kiran Sharma

Abstract Human interactions can be either positive or negative, giving rise to different complex social or anti-social phenomena. The dynamics of these interactions often lead to certain spatio-temporal patterns and complex networks, which can be interesting to a wide range of researchers—from social scientists to data scientists. Here, we use the publicly available data for a range of anti-social and political events like ethnic conflicts, human right violations and terrorist attacks across the globe. We aggregate these anti-social events over time, and study the temporal evolution of these events. We present here the results of several time-series analyses like recurrence intervals, Hurst R/S analysis, etc., that reveal the long memory of these time-series. Further, we filter the data country-wise, and study the time-series of these anti-social events within the individual countries. We find that the time-series of these events have interesting statistical regularities and correlations. Using multi-dimensional scaling technique, the countries are then grouped together in terms of the co-movements with respect to temporal growths of these anti-social events. The data science approaches to studying these anti-social phenomena may provide a deeper understanding about their formations and spreading. The results can help in framing public policies and creating strategies that can check their spread and inhibit these anti-social phenomena.

Introduction

Humans prefer to form groups and act collectively. These groups have evolved from simple settlements in ancient times to huge nations in modern times; defined by multiple causes like languages, common heritage, geographical boundaries, and even

S. S. Husain · K. Sharma (✉)
School of Computational and Integrative Sciences,
Jawaharlal Nehru University, New Delhi 110067, India
e-mail: kiransharma1187@gmail.com

S. S. Husain
e-mail: shariq.iitk@gmail.com

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_17

ideologies. Often human cooperation has been the motivating force behind the rapid progress of man. This cooperation [1] has extended from blood relatives to totally unrelated individuals. Contrarily, evolution has been responsible for drawing distinctions among themselves in their bids for the "survival of the fittest". The segregation [2, 3] can be seen in various forms of race, caste, class, religion, political ideology, etc. The assortment of positive and negative aspects of human social behavior makes it extremely complex and convoluted with multiple parameters playing crucial roles. Thus, it is extremely difficult to assess and model the complexity of human social behavior, ranging from bonding, co-operation, support to greed, jealousy, conflict, aggression, coup, war, etc.

Entire world has seen time and again different forms of conflicts, aggression, war, and terrorism, which have plagued mankind from antiquity. Anti-social phenomenon notably possesses very different characteristics than normal social behavior. The interactions among anti-social agents are very low and the occurrences of events tend to be independent of each other. A conflict is an activity which takes place between conscious (not necessarily rational) beings when their interests are mutually inconsistent with each other. A conflict is usually associated with violent activities. The human society has been riddled with conflicts. The first known conflict, a case of inter-group violence, was in eastern Africa around 10,000 years ago as an attempt to seize resources - territory, women, food stored in pots, which resulted in the killing of over two dozen prehistoric men, women, and children [4]. However, as there has been more progress in civilization, humans have become more materialistic and self-centered, and gone beyond competition for tangible resources; they have adopted causes like religion, racial superiority, etc., as pretexts for killing others.

Many people, including Karl Marx and Friedrich Engels, have proposed theories of social conflicts. Apart from social scientists, physicists and data scientists have recently tried to perform in-depth studies and provide mathematical models, statistical and time series analysis of the empirical data and tried to propose potential solutions to the menaces of terrorism, conflicts and other social phenomena, leading to the development of the field of sociophysics [5–8]. Sociophysics is marked by the belief that large-scale statistical measurement of social variables reveals underlying relational patterns that can be explained by theories and laws found in natural sciences, and physics in particular.

In this chapter, we focus on the data dependent statistical analyses of three major anti-social phenomena, viz., ethnic conflicts (EC), human right violations (HR), and terrorism (GTD) [9–11]. An ethnic conflict is a conflict between two or more contending ethnic groups where each group fights for its position within the society on the basis of ethnicity, derived from common descent, culture, language and sometimes, even a common identity. Similarly, a human right violation is said to occur when the basic fundamental rights of a person or a group of persons are infringed upon. Both these anti-social phenomena are based on the conflicts between one or more contesting parties (two sets of actors). However, terrorism differs from these conflicts in the sense that the casualties occurring in a terrorist event are direct or indirect targets of the terrorist groups (sources). The aim of terrorism is not limited to eliminating the target group or destruction of resources, rather it is specifically

carried out to send out a psychological message to the adversary [12]. In other words, unlike in ethnic conflicts and human rights violations, the terrorist attacks are carried out to send across a message to the opponent [13].

Here, we use the publicly available data from: (a) GDELT database [14, 15], which has news reports in media consisting of records of a wide range of socio-economic and political events, viz. ethnic conflicts and human rights violations, over a long period of time, and (b) GTD project [16, 17], which has recorded the terrorist attack incidents that occurred in the last half-century across the globe. We aggregate these anti-social events over time, and study the temporal evolution of these events. We present here the results of several analyses like recurrence intervals, Hurst R/S analysis, etc., that reveal the long memory of these time series [18]. Further, we filter the data country-wise, and study the correlations of these anti-social events within the individual countries. Using the multi-dimensional scaling, we cluster the countries together in terms of the co-movements with respect to temporal growths of these anti-social events. The time series of these events reveal interesting statistical regularities and correlations.

The article is organized as follows. Section "Data description, Methodology and Results" describes the data description, methodology and results in detail. Section "Concluding Remarks" contains the concluding remarks.

Data Description, Methodology and Results

Data Description

We have used the Global Database of Events, Language, and Tone (GDELT) [14, 15] which is an open source database hosted and managed by GDELT project through *Google Cloud*. GDELT monitors the world's news media from nearly every corner of every country in print, broadcast, and web formats, in over 100 languages, every moment of every day. The GDELT project is a real-time open database, where the human society is seen through the eyes of the world's news media, reaching deeply into local events, reaction, discourse, and emotions of the most remote corners of the world. The entire GDELT event database is available and can be extracted using *Google BigQuery*. We filtered all events related to ethnic conflicts (EC) and human rights violations (HR) happening around the world spanning over a large time scale. We procured 45, 942 events for EC and 48, 295 for HR for a 15 year period, 2001–2015.

We have also analysed the data on terrorism events. For the analysis we have used the Global Terrorism Database (GTD) which is an open-source database provides a detailed account of terrorist events around the world from 1970–2017 [16, 17]. The event database is hosted by the National Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland. We procured 72, 521 events for the same 15 year period, 2001–2015.

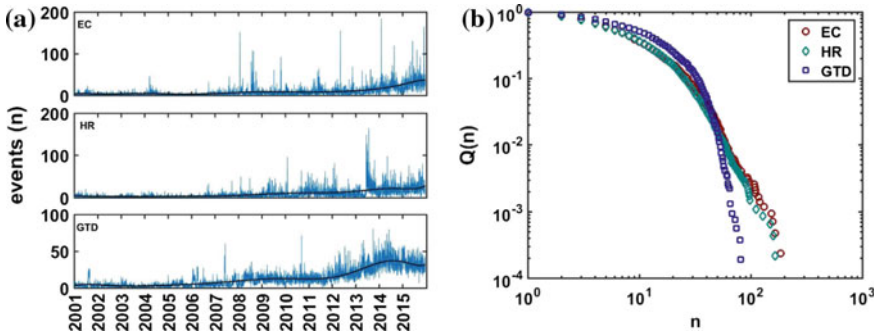


Fig. 17.1 **a** Plots for the time evolution of the number of events n reported daily for EC, HR and GTD during the period 2001–2015 with trends (black solid curves). **b** The complementary cumulative density function (CCDF) $Q(n)$ that n or more events are reported on a particular day. The data seems to fit well to a stretched exponential of the form (exp$[-an^b]$) for EC (red circles), HR (green diamonds) and GTD (blue squares), with exponents given in Table 17.2

The list of all the countries analysed, containing names along with their three letter ISO codes, is given in Table 17.1.

Methodology and Results

The GDELT and GTD data sets contain detailed information about the anti-social events, viz. ethnic conflicts (EC), human rights violations (HR) and terrorist attacks (GTD), on the scale of a day. Our overarching aim is to observe nature of the memory of each of the time-series (EC, HR and GTD) and the cross-correlations among them, within a country. Further, we would like to group the countries together on the basis of their long-term evolution trends and correlations. First, we study simple statistics of auto-correlations, Hurst R/S analysis and recurrence intervals distribution, of the detrended time series. Later, we study the co-movements of the countries on the events spaces using the multidimensional scaling technique.

We have considered the data for the period 2001–2015, and generated daily time series of EC, HR and GTD, as shown in Fig. 17.1a. The black curves show the long time trends, which imply that the time-series are not stationary. To see the spread of the events, we computed the complementary cumulative density function (CCDF) of the events: For the probability density function (PDF) $P(n)$ as n reported events per day, the cumulative density function (CDF) is $F(n) = P(N \leq n)$; then, the CCDF is $Q(n) = 1 - F(n)$, such that it estimates the probability of the events are above a particular level n, $P(N > n)$. As the empirical PDFs are often too noisy (specially toward the tails) to be relied upon for statistics, it is known that integrating a signal improves its "Signal-to-Noise ratio". So, we plot the CCDF as it reduces the noise content and makes the information contained by the signal clearer.

Table 17.1 List of countries and their 3-letter ISO codes

S. No.	Code	Country	S. No.	Code	Country	S. No.	Code	Country
1	**AFG**	Afghanistan	36	**GEO**	Georgia	71	**NLD**	Netherlands
2	**AGO**	Angola	37	**GHA**	Ghana	72	**NOR**	Norway
3	**ALB**	Albania	38	**GMB**	Gambia	73	**NPL**	Nepal
4	**ARG**	Argentina	39	**GRC**	Greece	74	**PAK**	Pakistan
5	**ARM**	Armenia	40	**HKG**	Hong Kong	75	**PER**	Peru
6	**AUS**	Australia	41	**HRV**	Croatia	76	**PHL**	Philippines
7	**AZE**	Azerbaijan	42	**HTI**	Haiti	77	**POL**	Poland
8	**BDI**	Burundi	43	**HUN**	Hungary	78	**PRK**	North Korea
9	**BEL**	Belgium	44	**IDN**	Indonesia	79	**RUS**	Russia
10	**BGD**	Bangladesh	45	**IND**	India	80	**RWA**	Rwanda
11	**BGR**	Bulgaria	46	**IRL**	Ireland	81	**SAU**	Saudi Arabia
12	**BIH**	Bosnia-Herzegovina	47	**IRN**	Iran	82	**SDN**	Sudan
13	**BLR**	Belarus	48	**IRQ**	Iraq	83	**SEN**	Senegal
14	**BRA**	Brazil	49	**ISR**	Israel	84	**SLE**	Sierra Leone
15	**BTN**	Bhutan	50	**ITA**	Italy	85	**SLV**	El Salvador
16	**CAN**	Canada	51	**JOR**	Jordan	86	**SOM**	Somalia
17	**CHE**	Switzerland	52	**JPN**	Japan	87	**SRB**	Serbia
18	**CHL**	Chile	53	**KEN**	Kenya	88	**SWE**	Sweden
19	**CHN**	China	54	**KGZ**	Kyrgyzstan	89	**SYR**	Syria
20	**CIV**	Cote D'ivoire	55	**KHM**	Cambodia	90	**TCD**	Chad
21	**COG**	Democratic Republic of the Congo	56	**KOR**	South Korea	91	**THA**	Thailand
22	**COL**	Colombia	57	**KWT**	Kuwait	92	**TUN**	Tunisia
23	**CUB**	Cuba	58	**LBN**	Lebanon	93	**TUR**	Turkey
24	**CYP**	Cyprus	59	**LBR**	Liberia	94	**TWN**	Taiwan
25	**CZE**	Czech Republic	60	**LBY**	Libya	95	**UGA**	Uganda
26	**DEU**	Germany	61	**LKA**	Sri Lanka	96	**UKR**	Ukraine
27	**DNK**	Denmark	62	**LVA**	Latvia	97	**USA**	United States
28	**DZA**	Algeria	63	**MDA**	Moldova	98	**UZB**	Uzbekistan
29	**ECU**	Ecuador	64	**MEX**	Mexico	99	**VEN**	Venezuela
30	**EGY**	Egypt	65	**MKD**	Macedonia	100	**VNM**	Vietnam
31	**ESP**	Spain	66	**MMR**	Myanmar	101	**YEM**	Yemen

(continued)

Table 17.1 (continued)

S.No.	Code	Country	S.No.	Code	Country	S.No.	Code	Country
32	**ETH**	Ethiopia	67	**MNP**	Northern Mariana Islands	102	**ZAF**	South Africa
33	**FJI**	Fiji	68	**MYS**	Malaysia	103	**ZMB**	Zambia
34	**FRA**	France	69	**NAM**	Namibia	104	**ZWE**	Zimbabwe
35	**GBR**	United Kingdom	70	**NGA**	Nigeria			

Table 17.2 Exponent values for events of EC, HR, and GTD

Series	Exponents	
	a	b
EC	0.50 ± 0.01	2.49 ± 0.01
HR	0.76 ± 0.01	2.91 ± 0.01
GTD	0.48 ± 0.01	3.17 ± 0.01

As the number of news entries n per day is a stochastic variable, we often see bursts of activities for all the three anti-social phenomena: EC, HR and GTD. Due to large inter-day fluctuations in the number of reports and the bursty nature, the CCDF shows a broad distribution. Figure17.1b shows the plots for the CCDF $Q(n)$ that n or more events are reported on a particular day, for the three time-series. Each of the curves is well-fitted by a stretched exponential of the form, $(\exp[-an^b])$ with exponents given in Table 17.2.

The auto-correlation is the correlation of a signal with a time-delayed copy of itself, as a function of delay or lag. In simple words, it is the similarity between observations as a function of the time lag between them, which can be used for finding repeating patterns or periodicity obscured by noise. The Hurst exponent is a popular measure of long-term memory in a time series, which relates to the auto-correlations of the same and the rate at which these auto-correlations decrease as the time-lag between the pair of values increases.

Extreme events are rare in natural as well as social phenomena, but it is essential to study their properties as the consequences of extreme events are often enormous [19–21]. As researchers, we are often interested in the question that how long would we have to wait for extreme events of a certain magnitude to recur. We thus fix a threshold $X^{(q)}$ and consider only the events of magnitude higher than $X^{(q)}$, where q denotes the quantile. We define the *recurrence interval* as the time interval between two consecutive extreme events:

$$R_t = \begin{cases} \text{NA}, & X(t) < X^{(q)} \\ \inf\left\{\tau > 0 \mid X(t+\tau) \geq X^{(q)}\right\}, & X(t) \geq X^{(q)} \end{cases}, \tag{17.1}$$

where $X(t)$ is an event occurring at time t, $X^{(q)}$ is the threshold, and τ is a time lag.

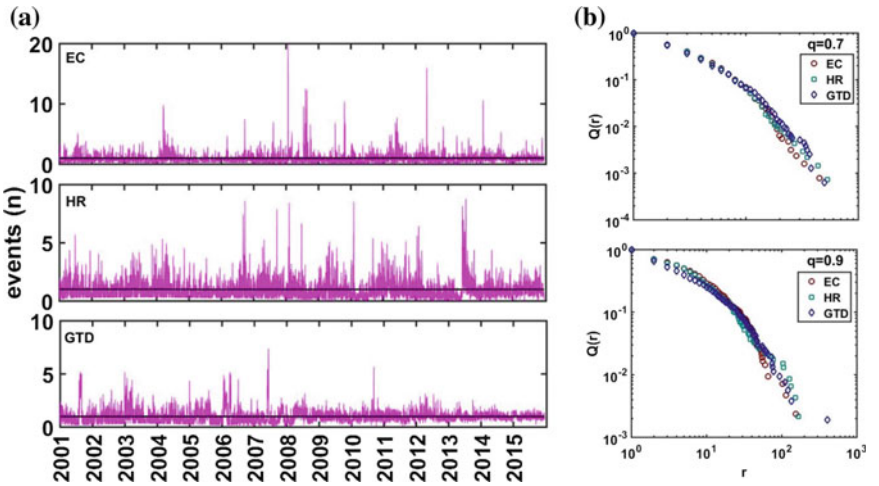


Fig. 17.2 **a** Detrended time series for EC, HR and GTD with a flat trend (black solid line). **b** shows the plots of complementary cumulative density function $Q(r)$ that r events recurred at quantiles $q = 0.7$ and $q = 0.9$. The data for quantiles $q = 0.7$ and $q = 0.9$ have been found to fit well to stretched exponentials, $(a \exp[-bn^c])$ with parameters given in Tables 17.3 and 17.4

Table 17.3 Parameter values for recurrence CCDF at $q = 0.7$

Series	Parameters		
	a	b	c
EC	3.53 ± 0.37	1.27 ± 0.10	0.49 ± 0.03
HR	3.69 ± 0.46	1.31 ± 0.12	0.49 ± 0.03
GTD	6.95 ± 0.71	1.94 ± 0.10	0.38 ± 0.01

A real data series usually exhibit non-stationarity of various forms such as "seasonal effects", "trends", etc. Though it is very difficult to completely eliminate non-stationarity, its effect can be reduced by introducing some corrective measures. Each type of non-stationarity requires a different type of correction. Here, none of the original time series of events is stationary, as can be seen in Fig. 17.1, where the black dashed line shows the inherent trend of the time series. We calculate the trend with a polynomial of degree 10, and then divide the original signal by the computed trend, resulting in a detrended series. Figure 17.2a shows the events time series after detrending it. The black dashed line, which shows trend of this detrended time series, is thus flat. The detrended events time series can be assumed to be weakly stationary.

We hence analyse the CCDF of the recurrence time intervals on the detrended time series (see Fig. 17.2a) to quantify the extreme events or duration of recurrences of an event. Figure 17.2b shows the plots of CCDF $Q(r)$ that r events recurred at quantiles $q = 0.7$ and $q = 0.9$. The data for quantiles $q = 0.7$ and $q = 0.9$ seem to fit well to stretched exponentials, $(a \exp[-bn^c])$, with parameters given in Tables 17.3 and 17.4. It should be mentioned that as the quantile q increases, the distribution becomes

Table 17.4 Parameter values for recurrence CCDF at $q = 0.9$

Series	Parameters		
	a	b	c
EC	1.67 ± 0.09	0.54 ± 0.04	0.50 ± 0.02
HR	1.69 ± 0.10	0.56 ± 0.05	0.51 ± 0.02
GTD	4.72 ± 0.72	1.58 ± 0.15	0.28 ± 0.02

fatter, i.e., lower recurrence time intervals occur less frequently. This observation is quite obviously explained by the fact that at higher values of q there are fewer extreme events and they are spread apart.

It is often not possible to comprehend certain effects using empirical data. Thus, the results obtained by analyses of empirical data generally need to be compared against standard benchmarks. In such situations, artificial data can be simulated according to required specifications and the simulated data can then serve as reliable benchmarks. Therefore, we first use Gaussian noises (white and fractional) to understand certain effects and use them as benchmarks for comparing the empirical statistics.

Gaussian noise is a statistical noise having a probability density function equal to that of the Normal (or Gaussian) distribution; a special case is the white Gaussian noise (wGn) or Brownian motion, in which the increments (values at any pair of times) are identically distributed and statistically independent (and hence uncorrelated). Thus, it has no auto-correlation for positive lags, and an exponentially decreasing recurrence interval distribution. We illustrate a white Gaussian noise in Fig. 17.3a.

A fractional Brownian motion is a generalization of Brownian motion. The main difference between fractional Brownian motion and regular Brownian motion is that the increments in Brownian motion are independent, whereas increments for fractional Brownian motion are not. A fractional Gaussian noise (fGn) with Hurst exponent $0 \leq H \leq 1$, is defined as a continuous-time Gaussian process $B_H(t)$ on $[0, T]$, which starts at zero, has expectation zero for all t in $[0, T]$, and has a covariance function. Mathematically,

$$\forall(t, s) \in \mathbb{R}^2_+,$$
$$\mathbb{E}[B_H(t)] = 0 \tag{17.2}$$
$$\mathbb{E}[B_H(t)B_H(s)] = \frac{|t|^{2H} + |s|^{2H} - |t - s|^{2H}}{2}. \tag{17.3}$$

The auto-correlation function (ACF) of a fractional Gaussian noise with Hurst exponent H is given by:

$$ACF(\tau) \rightarrow \frac{|\ \tau + 1\ |^{2H} + |\ \tau - 1\ |^{2H} - 2\ |\ \tau\ |^{2H}}{2}. \tag{17.4}$$

For a stationary process with auto-correlations decaying $ACF(\tau) \sim \tau^{-\gamma}$ (long-memory processes), it can be shown mathematically $\gamma = 2 - 2H$ [22].

Figure 17.3b shows the ACF and PDF of recurrence time intervals for fractional Gaussian noise with Hurst exponent $H = 0.8$. If the underlying time series has auto-correlation, then the extreme events are auto-correlated as well. Evidently the presence of the auto-correlation renders the probability density function of the recurrence time intervals to be a stretched exponential, instead of a pure exponential as observed in the case of white Gaussian noise.

The Hurst exponent is a useful statistical method for inferring the properties of a time series. There are various methods to calculate Hurst exponent, which measures the existence of trend or 'persistence' or long-range memory present in the time series. We used the rescaled range (R/S) method to compute the Hurst exponent [23]. The rescaled range (R/S) method is calculated for a time series, $X_1, X_2, \ldots, X_T$, as follows [24]: We first break the long time series with T data points, into shorter windows of n data points, such that there are $m = T/n$ windows. For each of the m windows of size n, we have the partial time series $X_1, X_2, \ldots, X_n$, for which we calculate the rescaled range:

1. Calculate the mean $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. Create a mean adjusted series $Y_i = X_i - \mu$ for $i = 1, 2, \ldots, n$.
3. Calculate the cumulative deviate series $Z_t = \sum_{i=1}^{t} Y_i$ for $t = 1, 2, \ldots, n$.
4. Compute the range $R(n) = \max(Z_1, Z_2, \ldots, Z_n) - \min(Z_1, Z_2, \ldots, Z_n)$.
5. Compute the standard deviation $S(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}$, where μ is the mean for the partial time series $X_1, X_2, \ldots, X_n$.
6. Calculate the rescaled range $R(n)/S(n)$ and average over all the partial time series of length n.

The Hurst exponent is estimated by fitting the power law: $\mathbb{E}[R(n)/S(n)] = Cn^H$ to the empirical data, where C is a constant. This can be done by plotting $\log[R(n)/S(n)]$ as a function of $\log n$, and fitting a straight line. The value of the slope gives the Hurst exponent H, such that

- A value in the range $0 \leq H < 0.5$ indicates a time series with 'anti-persistent' behavior,
- a value in the range $0.5 < H \leq 1$ indicates a time series with long-term positive auto-correlation ('persistent' behavior),
- a value of $H = 0$ indicates a pink noise,
- a value of $H = 0.5$ indicates a completely uncorrelated series (Brownian motion).

Figure 17.4 shows the auto-correlation of the detrended time series and Hurst exponent based on R/S analysis having exponent for (a) EC 0.75 ± 0.01, (b) HR 0.78 ± 0.01 and (c) GTD 0.82 ± 0.01. The auto-correlation for GTD is decaying exponentially. As the value of the exponent is greater than 0.5 so the time series shows the persistence behavior for all EC, HR and GTD.

Next, we study the co-movements among the different countries across the globe. Figure 17.5a shows the time series of n events (EC, HR and GTD) during the period of

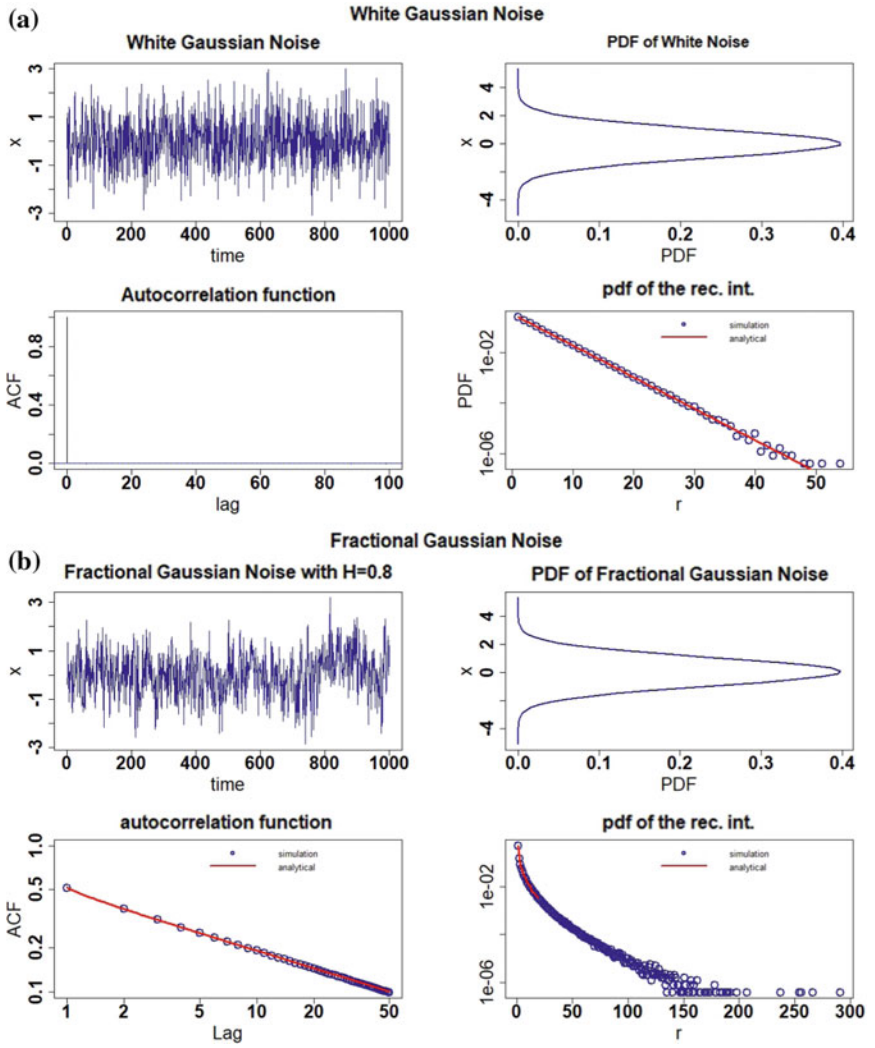


Fig. 17.3 **a** Time series for white Gaussian noise, probability density function for white noise (independent and identically distributed variables), auto-correlation function of time series, probability density function of recurrence time intervals at $q = 0.75$. The time series was generated using the *rnorm()* function in R-software for statistical computing. **b** Time series for fractional Gaussian noise with Hurst index $H = 0.8$, probability density function of the fractional Gaussian noise (dependent and identically distributed variables), auto-correlation function of the time series, probability density function for recurrence intervals at $q = 0.75$. The time series was generated using the *simFGN0()* function in R-software for statistical computing

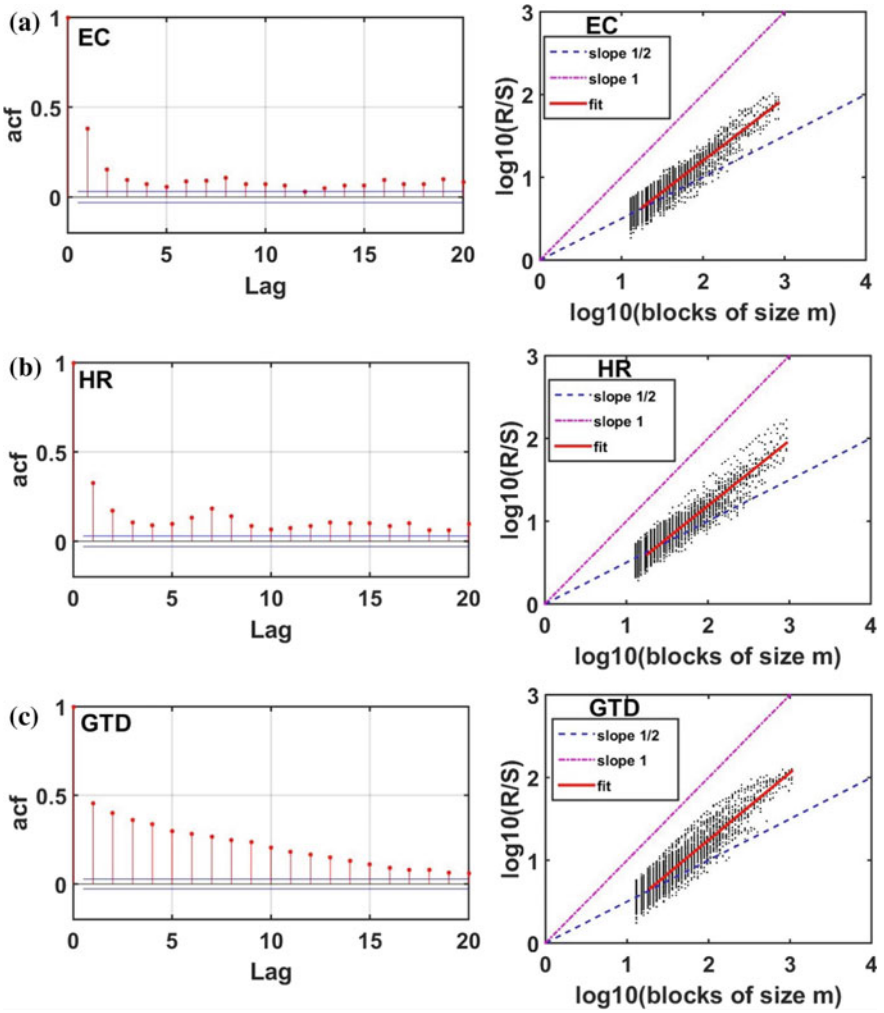


Fig. 17.4 Plot for the auto-correlation of detrended time series and Hurst exponent based on R/S analysis having exponent for **a** EC 0.75 ± 0.01, **b** HR 0.78 ± 0.01 and **c** GTD 0.82 ± 0.01. As the value of the exponent is greater than 0.5 so the time series shows the persistence behavior for all EC, HR and GTD

2001–2015 for a few countries chosen arbitrarily. We take N countries and aggregate the events over a year, producing $T = 15$ data points for the period 2001–2015. To build the correlation matrices, we define the equal-time Pearson cross-correlation coefficient for the time series of the number of events per year c_i as

$$\rho_{ij}(\tau) = \frac{\langle c_i c_j \rangle - \langle c_i \rangle \langle c_j \rangle}{\sigma_i \sigma_j}. \tag{17.5}$$

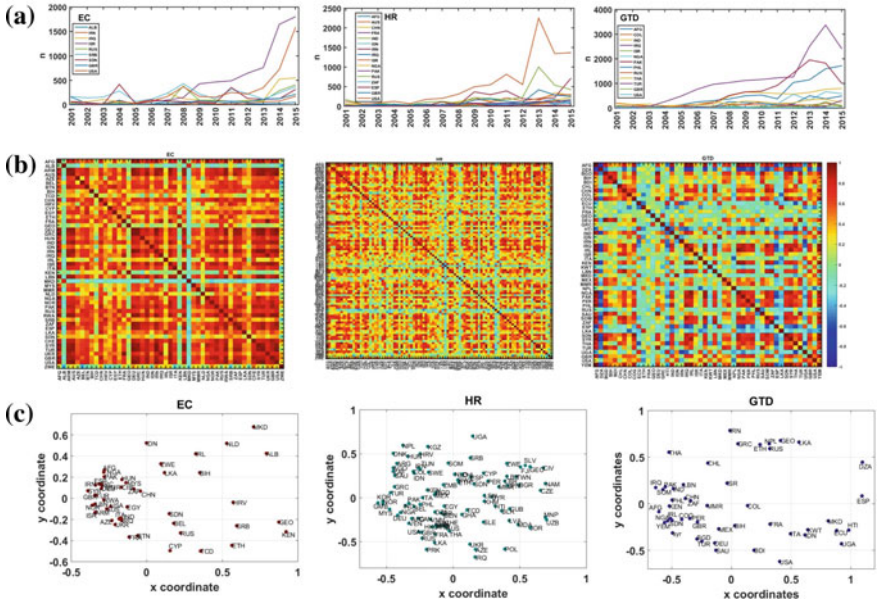


Fig. 17.5 **a** Time series plots for number of events n of different countries for EC, HR and GTD during the period of 2001–15. The list of country names can be seen in Table 17.1. **b** Correlation matrices for EC, HR and GTD. As number of countries N are more than length of time series T, i.e. $N >> T$, the power mapping technique with a distortion $\varepsilon = 0.6$ is applied to the correlation matrix to suppress the noise. **c** 2D MDS plots for EC, HR and GTD during the period 2001–2015. The MDS plots show the co-movement of the countries: similar countries are grouped together and dissimilar ones placed far apart

where $\sigma_i = \sqrt{\langle c_i^2 \rangle - \langle c_i \rangle^2}$ is the standard deviation of $c_i, i, j = 1, \ldots, N$, and $\langle \cdots \rangle$ denotes average over the time period τ. The elements ρ_{ij} are restricted to the domain $-1 \leq \rho_{ij} \leq 1$, where $\rho_{ij} = 1$ signifies perfect correlations, $\rho_{ij} = -1$ perfect anti-correlations, and $\rho_{ij} = 0$ corresponds to uncorrelated pairs.

It is difficult to estimate the exact correlation among N time series, each of length T, as spurious correlations or 'noise' are present in the finite time series (see Ref. [25]). The quality of the estimation of true correlation in a matrix strongly depends upon the ratio of the length of the time series T and the number of time series N, $Q = T/N$. Correlation matrices are less noisy for higher value of Q. As $N > T$, the corresponding cross-correlation matrices are also singular with $N - T + 1$ zero eigenvalues, which leads to poor eigenvalue statistics. Thus, we use the power map technique [25–27] to break the degeneracy of eigenvalues at zero and suppress the noise. In this method, a non-linear distortion is given to each cross-correlation coefficient (ρ_{ij}) of the correlation matrix $\boldsymbol{\rho}$ by: $\rho_{ij} \rightarrow (\text{sign } \rho_{ij})|\rho_{ij}|^{1+\varepsilon}$, where ε is the distortion parameter; here we used $\varepsilon = 0.6$ (see Refs. [25, 27] for choice of the parameter).

Figure 17.5b shows the correlation matrices (after using the power mapping method), computed over different time series across the countries by using Eq. 17.5. The correlation matrix for EC shows more correlations (colored red) among the countries as compared to anti-correlations (colored blue). The correlation matrices for HR and GTD look very different. In order to visualize the correlations, we apply the multidimensional scaling (MDS) technique. First, we transform the correlation matrix $\boldsymbol{\rho}$ into distance matrix $\mathbf{D}$, as

$$d_{ij} = \sqrt{2(1 - \rho_{ij})}, \tag{17.6}$$

such that $2 \geq d_{ij} \geq 0$. After transforming the correlation matrix into distance matrix, we generate the MDS map. The MDS algorithm is used to display the structure of similarity in terms of distances, as a geometrical map where each country corresponds to a set of coordinates in the multidimensional space. MDS arranges different countries in this space according to the strength of the pairwise distances between them. Two similarly behaving countries are represented by two points that are close to each other, and two dissimilarly behaving countries are placed far apart in the map. In general, we choose the embedding dimension to be 2, so that we are able to plot the coordinates in the form of a map. It may be noted that coordinates are not necessarily unique, as we can arbitrarily translate and rotate them, as long as such transformations leave the distances unaffected. Figure 17.5c shows the 2D MDS plots for EC, HR and GTD based on the similarities/ distances among them.

At the end, we also calculated the correlation among the different time series for individual countries. The correlations are computed among EC-HR, EC-GTD and HR-GTD. Few countries like ESP, IDN, ITA and RUS have low correlations among EC-HR, whereas ESP, FRA, ITA, LKA and RUS show anti-correlations for EC-GTD; countries like ESP, FRA, GRC, IDN, ITA, LKA show anti-correlations for HR-GTD. For further details, see Table 17.5. It must be noted that these are just linear correlations, and causal relations cannot be inferred.

Concluding Remarks

In this paper, our goal was to do the time series analysis and apply data science approaches to the study of the daily anti-social events like ethnic conflicts (EC), human right violations (HR) and terrorist attacks (GTD). As the time series were non-stationary, so we made them stationary by detrending them. We computed the recurrence interval distribution of events and made attempts to relate it with its auto-correlation function. Then we computed the Hurst exponent using the rescaled range (R/S) analyses, which gives the information about whether long memory is present or not. Further, our interest was to study the co-movements of the countries in the respective events spaces. To visualize the co-movements, we computed the cross-correlations among different countries, transformed the correlations into distances and then projected the distances into 2D multidimensional scaling maps.

Table 17.5 Cross-correlation among events across countries

S.No.	Country	EC-HR	EC-GTD	HR-GTD	S.No.	Country	EC-HR	EC-GTD	HR-GTD
1	**AFG**	0.72	0.80	0.84	11	**IRQ**	0.61	0.73	0.85
2	**CHN**	0.81	0.64	0.89	12	**ISR**	0.90	0.43	0.33
3	**DEU**	0.65	0.95	0.49	13	**ITA**	0.21	−0.05	−0.09
4	**ESP**	0.47	−0.52	−0.28	14	**LKA**	0.57	−0.22	−0.39
5	**FRA**	0.91	−0.08	−0.10	15	**NGA**	0.73	0.77	0.77
6	**GBR**	0.81	0.72	0.84	16	**PAK**	0.83	0.65	0.82
7	**GRC**	0.91	0.18	−0.01	17	**RUS**	0.29	−0.07	0.04
8	**IDN**	0.42	0.69	−0.10	18	**TUR**	0.97	0.95	0.90
9	**IND**	0.94	0.85	0.89	19	**USA**	0.61	0.38	0.11
10	**IRN**	0.80	0.11	0.45	20	**ZAF**	0.96	0.21	0.17

Acknowledgements The authors would like to thank Anirban Chakraborti, Vishwas Kukreti, Arun S. Patel and Hirdesh K. Pharasi for critical discussions and inputs. KS acknowledges the University Grants Commission (Ministry of Human Resource Development, Govt. of India) for her senior research fellowship. SSH and KS acknowledge the support by University of Potential Excellence-II grant (Project ID-47) of JNU, New Delhi, and the DST-PURSE grant given to JNU by the Department of Science and Technology, Government of India.

References

1. Perc, M., Jordan, J.J., Rand, D.G., Wang, Z., Boccaletti, S., Szolnoki, A.: Statistical physics of human cooperation. Phys. Rep. **687**, 1–51 (2017)
2. Schelling, T.: Models of segregation. Am. Econ. Rev. **59**(2), 488–93 (1969)
3. Schelling, T.: Dynamic models of segregation. J. Math. Sociol. **1**, 143–186 (1971)
4. Lahr, M.M., Rivera, F., Power, R.K., Mounier, A., Copsey, B., Crivellaro, F., Edung, J.E., Fernandez, J.M.M., Kiarie, C., Lawrence, J., Leakey, A., Mbua, E., Miller, H., Muigai, A., Mukhongo, D.M., Van Baelen, A., Wood, R., Schwenninger, J.L., Grn, R., Achyuthan, H., Wilshaw, A., Foley, R.A.: Inter-group violence among early holocene hunter-gatherers of west Turkana, Kenya. Nature **529**, 394–398 (2016)
5. Abergel, F., Aoyama, H., Chakrabarti, B.K., Chakraborti, A., Deo, N., Raina, D., Vodenska, I.: Econophysics and Sociophysics: Recent Progress and Future Directions. Springer, Berlin (2017)
6. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. **81**(2), 591 (2009)
7. Chakrabarti, B.K., Chakraborti, A., Chatterjee, A.: Econophysics and Sociophysics: Trends and Perspectives. Wiley-VCH, Germany (2006)
8. Sen, P., Chakrabarti, B.K.: Sociophysics: An Introduction. Oxford University Press, Oxford (2014)
9. Clauset, A., Young, M., Gleditsch, K.S.: On the frequency of severe terrorist events. J. Confl. Resolut. **51**(1), 58–87 (2007)

10. Sharma, K., Sehgal, G., Gupta, B., Sharma, G., Chatterjee, A., Chakraborti, A., Shroff, G.: A complex network analysis of ethnic conflcts and human rights violations. Sci. Rep. **7**(1), 8283 (2017)
11. Husain, S.S., Sharma, K., Kukreti, V., Chakraborti, A.: Identifying the global terror hubs and vulnerable motifs using complex network dynamics (2018). arXiv:1802.01147
12. Richardson, L.: The Roots of Terrorism. Routledge, London (2013)
13. Cutter, S.L., Richardson, D.B., Wilbanks, T.J.: The Geographical Dimensions of Terrorism. Routledge, London (2014)
14. GDELT - data format codebook v 1.03, as on 25 Aug 2013 (2016). http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf
15. The global database of events, language and tone (GDELT) (2016). www.gdeltproject.org/
16. Global terrorism database (GTD)-codebook: inclusion criteria and variables, as on 3 Jan 2013 (2018). https://www.start.umd.edu/gtd/downloads/codebook.pdf
17. Global terrorism database (GTD) (2018). https://www.start.umd.edu/gtd/contact/
18. Tilak, G.: Studies of the recurrence-time interval distribution in financial time-series data at low and high frequencies (2012)
19. Chicheportiche, R., Chakraborti, A.: Copulas and time series with long-ranged dependencies. Phys. Rev. E **89**, 042,117 (2014)
20. Chicheportiche, R., Chakraborti, A.: A model-free characterization of recurrences in stationary time series. Phys. A Stat. Mech. Appl. **474**, 312–318 (2017)
21. Santhanam, M., Kantz, H.: Return interval distribution of extreme events and long-term memory. Phys. Rev. E **78**(5), 051,113 (2008)
22. Tarnopolski, M.: On the relationship between the hurst exponent, the ratio of the mean square successive difference to the variance, and the number of turning points. Phys. A Stat. Mech. Appl. **461**, 662–673 (2016)
23. Chakraborti, A., Santhanam, M.: Financial and other spatio-temporal time series: long-range correlations and spectral properties. Int. J. Mod. Phys. C **16**(11), 1733–1743 (2005)
24. Torres, L.R.M., Rojas, A.R., Luévano, J.R., Hernández, R.T.P.: Exponente de hurst en series de tiempo electrosísmicas
25. Pharasi, H.K., Sharma, K., Chakraborti, A., Seligman, T.H.: Complex market dynamics in the light of random matrix theory (2018). arXiv:1809.07100
26. Chakraborti, A., Sharma, K., Pharasi, H.K., Das, S., Chatterjee, R., Seligman, T.H.: Characterization of catastrophic instabilities: market crashes as paradigm (2018). arXiv:1801.07213
27. Pharasi, H.K., Sharma, K., Chatterjee, R., Chakraborti, A., Leyvraz, F., Seligman, T.H.: Identifying long-term precursors of financial market crashes using correlation patterns (2018). New J. Phys. **20**, 103041 (2018)

Part III
Miscellaneous

Chapter 18
International Center for Social Complexity, Econophysics and Sociophysics Studies: A Proposal

Bikas K. Chakrabarti

Abstract In the concluding session of the Joint International Conference titled 'Econophys-2017 and Asia Pacific Econophysics Conference (APEC)-2017', held in Jawaharlal Nehru University and Delhi University during November 15–18, 2017, a brief version of this Proposal was presented. There were several enthusiastic comments, received from the participants. This note is based on these discussions.

Introduction

More than twenty years have passed since the formal coining of the term and hence the launch of econophysics as a research topic (since 1995; see the entry by Barkley Rosser on Econophysics in 'The New Palgrave Dictionary of Economics' [1]: *"Econophysics: According to Bikas Chakrabarti, the term 'econophysics' was neologized in 1995 at the second Statphys-Kolkata conference in Kolkata (formerly Calcutta, India) by the physicist H. Eugene Stanley ..."*). Soon, econophysics had been assigned the Physics and Astronomy Classification Scheme (PACS) number 89.65Gh by the American Institute of Physics. According to Google Scholar, typically today more than thousand papers and documents, containing the term 'econophysics', are published each year (many more research papers are, in fact, published today on the topic without ever calling it econophysics) in almost all physics journals covering statistical physics, general science journals and a few economics journals. More than fifteen books on econophysics (with the word econophysics in the title of the book), including some textbooks and monographs written by pioneers and active researchers in the field, have already been published by Cambridge University

B. K. Chakrabarti (✉)
Saha Institute of Nuclear Physics, Kolkata 700064, India
e-mail: bikask.chakrabarti@saha.ac.in

B. K. Chakrabarti
S. N. Bose National Centre for Basic Sciences, Kolkata 700106, India

B. K. Chakrabarti
Economic Research Unit, Indian Statistical Institute, Kolkata 700108, India

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3_18

Press, Oxford University Press, Springer and Wiley. Many more edited books and conference proceedings are published (search of 'econophysics' titles in the 'amazon.com:books' today gives more than 140 entries; with some double counting of course!). Similar has been the story for 'sociophysics'.

Regular interactions and collaborations between the communities of natural scientists and social scientists, however, are rare even today! Though, as mentioned already, interdisciplinary research papers on econophysics and sociophysics are regularly being published at a steady and healthy rate, and a number of universities (including Universities of Bern, Leiden, London, Paris and Tufts University) are offering the interdisciplinary courses on econophysics and sociophysics, not many clearly designated professor positions, or other faculty positions for that matter, are available yet (except for econophysics in Universities of Leiden and London). Neither there are designated institutions on these interdisciplinary fields, nor separate departments or centres of studies for instance. We note however, happily in passing, a recently published highly acclaimed ('landmark' and 'masterful') economics book [2] by Martin Shubik (Seymour Knox Professor Emeritus of Mathematical Institutional Economics, Yale University) and Eric Smith (Santa Fe Institute) discusses extensively on econophysics approaches and in general on the potential of interdisciplinary researches inspired by the developments in natural sciences. Indeed, this massive 580-page book can also serve as an outstanding 'white-paper' document in favor of our intended Proposal.

Though the inter-disciplinary interactions have not grown much, some sure signs of positive impact for the research achievements in econophysics and sociophysics have been documented in the literature. The precise characterizations of stock market fluctuations by Mantegna and Stanley [3] has already made a decisive mark in financial economics and all the related subjects (with more than 4000 citations already for the book [3]; Google scholar). In the section on 'The position of econophysics in the disciplinary space' in the book 'Econophysics and Financial Economics' [4], the authors write (pp. 83, 178): *"To analyze the position of econophysics in the disciplinary space, the most influential authors in econophysics were identified. Then their papers in the literature were tracked by using the Web of Science database of Thomson-Reuters ... The sample is composed of Eugene Stanley, Rosario Mantegna, Joseph McCauley, Jean Philippe Bouchaud, Mauro Gallegati, Benoit Mandelbrot, Didier Sornette, Thomas Lux, Bikas Chakrabarti, and Doyne Farmer."* The book [2] by Shubik and Smith noted (pp. 75–76) that while simple kinetic exchange market model (see e.g., [5]) leads to exponentially decaying distributions, *"it was shown in* [6] *that uniform saving propensity of the agents constrains the entropy maximizing dynamics in such a way that the distribution becomes gamma-like, while (quenched) nonuniform saving propensity of the agents leads to a steady state distribution with a Pareto-like power-law tail* [7]. *A detailed discussions of such steady state distributions for these and related kinetic exchange models is provided in* [8]*"*. Shubik and Smith [2] also noted the important contributions by physicists in the study of multi-agent iterative (and collective) learning game models for efficient resource sharing ([9] for binary choice iterative learning games and [10] for multi-choice

iterative learning games[1]). This book [2] also discusses in details on the impact of the pioneering work by physicist Per Bak and collaborators in the context of self-organizing dynamics of complex markets. The Econophysics course offered by Diego Garlaschelli in the Physics department of the Leiden University, where the first economics Nobel laureate (statistical physicist Jan Tinbergen) came from, follows exclusively the book 'Econophysics: An Introduction' [11] since its inception in 2011 (see e.g., [12] for the 2017–2018 and 2018–2019 e-prospectuses). Discussions on some more impact of econophysics [3, 4, 13, 14] and sociophysics [15–18] researches will be continued later.

Proposal in Brief and Some Earlier Attempts

In view of all these, it seems it is time to try for an international centre for interdisciplinary studies on complexity in social and natural sciences; specifically on econophysics and sociophysics. The model of the Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste (funded by UNESCO and IAEA), could surely be helpful to guide us here. We are contemplating, if an ICTP-type interdisciplinary research institute could be initiated for researches on econophysics and sociophysics (see also [19]).

We note that Dirk Helbing (ETH, Zurich) and colleagues have been trying for an European Union funded 'Complex Techno-Socio-Economic Analysis Center' or 'Economic and Social Observatory' for the last six years (see Ref. [20] containing the White Papers arguing for the proposed centre). We are also aware that Indian Statistical Institute had taken a decision to initiate a similar centre in India (see 'Concluding Remarks' in [21]). Siew Ann Cheong (Nanyang Technological University, Singapore) had tried for a similar Asian Center in Singapore [22]. In view of some recent enthusiasms at the Japan-India Heads of States or Prime Minister level, and signing of various agreements (predominantly for business deals, infrastructure development, technical science and also cultural exchanges) by them, possibility of an Indo-Japan Center for studies on Complex Systems is also being

[1]Important developments have taken place in such many-player, multi-choice iterative learning games for limited resource utilizations, since publication of The Kolkata Paise Restaurant Problem and Resource Utilization, A. S. Chakrabarti, B. K. Chakrabarti, A. Chatterjee and M. Mitra, Physica A, **388**, pp. 2420–2426 (2009). For applications to quantum cryptography physics, computer job scheduling, on-line car hire, etc., see e.g., Strategies in Symmetric Quantum Kolkata Restaurant Problem, P. Sharif and H. Heydari, Quantum Theory: Reconsideration of Foundations 6: AIP Conf. Proc. **1508**, pp. 492–496 (2012); Econophysics of the Kolkata Restaurant Problem and Related Games; B. K. Chakrabarti, A. Chatterjee, A. Ghosh, S. Mukherjee and B. Tamir, Springer (2017); Econophysics and the Kolkata Paise Restaurant Problem: More is Different, B. Tamir, Science and Culture, **84**, pp. 37–47 (2018); The Vehicle for Hire Problem: A Generalized Kolkata Paise Restaurant Problem, L. Martin and P. Karaenke, https://mediatum.ub.tum.de/doc/1437330/1437330.pdf (2018); Kolkata Paise Restaurant Game for Resource Allocation in the Internet of Things, T. Park and W Saad, IEEE Xplore, DOI: https://doi.org/10.1109/ACSSC.2017.8335666, https://ieeexplore.ieee.org/abstract/document/8335666/(2018).

explored, including the possibility of a center in Tokyo with private support [23]. There are several other similar initiatives (e.g., Ref. [24]).

These proposals are, or had been, for regular research centers on such interdisciplinary fields, where regular researchers are expected to investigate such systems. In view of the extreme interdisciplinary nature of econophysics and sociophysics, such efforts may be complemented by another visiting center model.

Unlike the above-mentioned kind of intended centers, this proposed centre may be just a visiting center where natural and social scientists from different universities and institutions of the world can meet for extended periods to discuss and interact on various interdisciplinary issues and collaborate for such researches, following the original ICTP model. Here, as in ICTP, apart from a few (say, about ten to start-with) promising young researchers on econophysics and sociophysics as permanent faculty who will continue active research and active visiting scientist programs (in physics, economics and sociology) etc. can be pursued, The faculty members, in consultation with the advisers from different countries, can choose the invited visitors and workshops or courses, on economics and sociological complexity issues, can be organized on a regular basis (as for basic theoretical sciences in ICTP or in Newton Centre, Cambridge, etc.). In two short communications [25], Martin Shubik (Yale University, New Haven) supported the idea very enthusiastically and encouraged us with some very precise suggestions. He also noted that such a center can play a much more inclusive role for the whole world (as is being done by the ICTP), compared to what the Santa Fe Institute has been successful to do for the US. Gene Stanley (Boston University, Massachusetts) supported enthusiastically such a proposal (*"... you already thought of all the ideas I might have had ... I will continue to think ... congratulations on your ambitious idea ... "* [26]).

Some Responses Received From the Participants

After my brief presentation of this proposal in the Concluding session of our Conferences, there were several appreciative comments made by the participants and a number of precise suggestions mailed to me later by many participants including Frederic Abergel (Centrale Supelec, Chatenay-Malabry Cedex), Bruce Boghosian (Tufts University, Massachusetts), Anirban Chakraborti (Jawaharlal Nehru University, Delhi), Siew Ann Cheong (Nanyang Technological University, Singapore), Acep Purqon (Institute for Technology, Bandung) and Irena Vodenska (Boston University, Massachusetts). I append below parts of a few detailed comments, summarizing the past achievements and some suggestions for possible structural organisation, received from them:

1. Regarding the *"discoveries of important economics and finance phenomena that were unknown to economists and financial economists before, the following few come to my mind:*

a. *The distribution of wealth and income. While Pareto was the first to examine the tail end of the wealth distribution, and found it to be a power law, little was known and understood about the full distribution until you and Victor Yakovenko came along, to (i) examine empirical distributions of wealth and income* [27]*, and (ii) build kinetic theory/agent-based models to show that the full distribution is an exponential distribution crossing over to a power-law tail* [6, 28] *and this arise because for rich people, they can gain from return on investment or through interests generated by savings, whereas the rest of us, repeated random exchange of income/wealth shape the exponential part of the distribution. During Econophys APEC 2017, we heard Bruce talking about his further results showing that if wealth is inadequately redistributed through taxation, oligarchs emerge, leading to the most extreme form of wealth inequality that we can possibly imagine* [29, 30].
b. *Home prices and property bubbles. Following your lead, and more recently the work by Ohnishi et al.* [31]*, my students and I have started looking into the distribution of home prices around various markets. Interestingly, the equilibrium distribution of home prices is similar to the income/wealth distribution, consisting of an exponential body and a power-law tail* [32]*. We see this in Singapore, Hong Kong, Taiwan, UK, and Japan so far, and believe this result is universal. We also found that in bubble years, the home price distribution develop dragon kings, which are strong positive deviations from the equilibrium distribution. We have evidence to suggest that such dragon kings are the results of speculation, but have yet to test regulations that can help defuse them in agent based models that we are currently building. More alarmingly, we have seen from the historical home price data of London and Tokyo that their distributions once contained an exponential body, but after experiencing a couple of property bubbles, have become asymptotic power laws with no exponential body. This is another manifestation of economic inequality, in that for cities like London and Tokyo, homes are priced out of the reach of the middle class. From the historical data for UK, we see this trend repeating itself for cities like Birmingham and Manchester. This calls for action on the part of government, but they cannot act until we understand the processes that drive this trend.*
c. *Louis Bachelier was the first to propose that stock returns perform Brownian motion, and laid the mathematical foundation for finance. However, for a long time, it has not occurred to financial economists to check the validity of Bacheliers assumptions. Benoit Mandelbrot did so in 1967, and found that the tail of the return distribution is a power law* [33]*. Rosario and Gene then demonstrated more convincingly using a large data set of returns for the S and P 500 in their 1995 Nature paper that the return distributions for different time horizons follow a scaling form, and this scaling form can be fitted better to a Levy distribution than to a Gaussian distribution* [34]*. Since then, many different agent based models have been developed to explain the emergence of fat tails in the return distribution. More recently, Hideki and Misako Takayasu examined high-frequency order data, and demonstrated*

convincingly that stock price is an invisible particle performing stochastic motion as a result of it being bombarded on either side by bid and ask orders [35]. *For regular Brownian motion, this noise is uncorrelated in time, and therefore we end up with long autocorrelations in the velocity of the Brownian particle. For stock returns, we know from many previous works that they are nearly uncorrelated in time. The Takayasus explained that this is the consequence of the noise being strongly correlated in time, pointing to what they observe in the order book data. This duality is surprising!*

d. *Economists Ricardo Haussmann and Cesar Hidalgo became world famous for publishing their Atlas of Economic Complexity* [36], *visualising the network of international trade over time. Not convinced that the economists have extracted the most important insights from the data, Luciano Pietronero went in to the data set to plot the economic performances of countries on a two-dimensional plot, with capabilities on the x-axis, and GDP on the y-axis* [37]. *Luciano found that he could classify countries into undeveloped, developing, and developed economies by where they appear on the plot. Undeveloped countries are problematic, and are mostly African, because their GDPs are low, and their capabilities are also low. These countries can potentially be stuck in a poverty trap, because they earn so little that they cannot reinvest into their education system to increase their capabilities. Developing countries like China, India, and Vietnam are countries that have in the past invested heavily into education and are therefore ranked highly in terms of their capabilities. China has already started to benefit from its past investment, to see a steady rise in its GDP. India can be seen to be following suit, and Vietnam will likely take off soon. When Luciano produced such plots using data from different years, he found that the developing countries are in a region where economic trajectories are fairly deterministic, and therefore we can have confidence in the economic futures of India and Vietnam, for example. On the other hand, the undeveloped countries are in a region of the plot where economic trajectories appear to be chaotic and turbulent, where countries can experience periods of enhanced GDP because of exploitation of resources (like Brazil), but can also fall from grace just as quickly because of political turmoil. In creating this list, I am leaving out interesting results obtained by people working on urban complexity, because they rarely attend econophysics conferences. Besides the most important scaling work done by Geoffrey West and Luis Bettencourt, showing that there are urban variables that scale sub linearly with the size of cities, and other urban variables (like GDP, patents, crime, etc.) that scale super linearly with size* [38]. *Hyejin Youn and her collaborators have also found that cities are not equally diverse in terms of job opportunities* [39]. *Small cities tend to have fewer types of jobs, and more people working on the same type of jobs. Large cities tend to have more types of jobs, and fewer people working on each type of job. More importantly, they have discovered that wealth is unequally concentrated in large cities, and that large cities tend to have a better educated populace, and because of this, is more resilient against the ongoing economic restructuring due to automation. Finally, besides telling*

success stories, we also need to frame a few key questions that we hope the international center can address. Here, we should be ambitious, and go for questions that individual investigators, or even individual universities would not have the capability, resource, or correct composition of different experts to address."

2. In this connection, it may be worthy to note that *"the German Physical Society has a working group on Physics of Socio-Economic Systems since 2009 (see e.g.,* [40]*: ... This dedicated scientific community is rapidly growing and involves, besides sociologist and economists, also physicists, mathematicians, computers scientists, biologists, engineers, and the communities working on complex systems and operations research ...). Apart from supporting researches and recognizing regularly active young researchers (with 'Young Scientist Award for Socio- and Econophysics') in such interdisciplinary fields, they organise many conferences within Germany with participants from all over Europe."*
3. Regarding a possible financial structure, *"I note, following Shubik, we want to raise funds for it to be endowed in perpetuity and cost of the regular activities can be met from the (fixed deposit) interests. As we discussed in Delhi after the Conference, this is not easy but I am hopeful. Also, I agree with Shubik, it is worth trying. ... Presumably, to begin with, the founding faculty members would need only a fraction of their salary, and the bulk of the interest money could be used for postdocs, graduate student support, visitor travel, etc. For a different institutional model, have a look at the web page of ICERM at Brown University* (https://icerm.brown.edu/home/index.php). *From our conversations in New Delhi, I understand that you would like to see a more extensive and inclusive model for this purpose, located somewhere in Eurasia, and I am very supportive of this idea. To raise funds for this kind of thing, it will be necessary to create a clear proposal that addresses—at the very minimum—the following items:*

    a. *First, we need a list of names and bios of international faculty who would be willing lend their names to such a center. In fact, it would be better to partition this list into categories: Some more senior faculty with administrative experience could serve on an Advisory Board. Other faculty would be willing to visit the Center from time to time, and perhaps organize conferences there. Some would send their graduate students during the summer, etc.*
    b. *Second, we need a clear business model for the Center, along with a governance model and sample budget. Again, we might learn from the models of ICERM, ICTP and SFI, but we probably want something that is unique to what we have in mind.*
    c. *Third, we need a list of benefits from this proposed Center that would accrue to the hosting institution and the hosting country."*

Concluding Remarks

We think, it is an appropriate time to initiate such a project for the healthy growth of this 'Fusion of Natural and Social Sciences', through active dialogue among the students and experts from different disciplines (e.g., physics, computer science, mathematics, economics and sociology), engaged in researches in their respective disciplines and institutions, from all over the world. We find, both the experts in the related disciplines as well as the researchers already initiated in such interdisciplinary researchers have deep feelings about the need for such a Center, where short and long term visits would be possible and enable them to participate in interdisciplinary schools, workshops, and research collaborations.

Acknowledgements I am grateful to Yuji Aruka, Arnab Chatterjee, Asim Ghosh, Taisei Kaijozi, for several interactive discussions at an earlier stage. Comments on the earlier draft of this Note from Abhik Basu, Soumyajyoti Biswas, Indrani Bose, Anirban Chakraborti and Parongama Sen are also gratefully acknowledged. I am indebted, in particular, to Bruce Boghosian and Siew Ann Cheong for their recent extremely supportive comments and detailed suggestions which, with their kind permission, have been partly included here. I am thankful to J. C. Bose National Fellowship (DST, Govt. India) for support.

References

1. Barkley Rosser Jr, J.: Econophysics. In: Durlauf, S.N., Blume, L.E. (eds.) The New Palgrave Dictionary of Economics, vol. 2, pp. 729–732. Macmillan, London (2008)
2. Shubik, M., Smith, E.: The Guidance of an Enterprise Economy. MIT Press, Cambridge (2016)
3. Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics. Cambridge University Press, Cambridge (2000)
4. Jovanovic, F., Schinckus, C.: Econophysics and Financial Economics. Oxford University Press, Oxford (2017)
5. Yakovenko, V.M., Barkley Rosser, J.: Statistical mechanics of money. Rev. Mod. Phys. **81**, 1703–1725 (2009)
6. Chakraborti, A., Chakrabarti, B.K.: Statistical mechanics of money: how saving propensity affects its distribution. Eur. Phys. J. B **17**, 167–170 (2000)
7. Chatterjee, A., Chakrabarti, B.K., Manna, S.S.: Pareto law in a kinetic model of market with random saving propensity. Physica A **335**, 155–163 (2004)
8. Chakrabarti, B.K., Chakraborti, A., Chakravarty, S.R., Chatterjee, A.: Econophysics of Income and Wealth Distributions. Cambridge University Press, Cambridge (2013)
9. Challet, D., Marsili, M., Zhang, Y.-C.: Minority Games. Oxford University Press, Oxford (2004)
10. Chakraborti, A., Challet, D., Chatterjee, A., Marsili, M., Zhang, Y.-C., Chakrabarti, B.K.: Phys. Rep. **552**, 1–26 (2015)
11. Sinha, S., Chatterjee, A., Chakraborti, A., Chakrabarti, B.K.: Econophysics: An Introduction. Wiley, Berlin (2010)
12. Econophysics: E-Prospectuses, Leiden University (2017–2018 & 2018–2019). https://studiegids.leidenuniv.nl/en/courses/show/69415/econofysica. https://studiegids.leidenuniv.nl/courses/show/81929/econofysica
13. Richmond, P., Mimkes, J., Hutzler, S.: Econophysics and Physical Economics. Oxford University Press, Oxford (2013)

14. Slanina, F.: Essentials of Econophysics Modelling. Oxford University Press, Oxford (2014)
15. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. **81**, 591–646 (2009)
16. Stauffer, D.: Opinion dynamics and sociophysics. In: Meyers, R.A. (ed.) Encyclopedia of Complexity and Systems Science, pp. 6380–6388. Springer, New York (2009)
17. Galam, S.: Sociophysics. Springer, New York (2012)
18. Sen, P., Chakrabarti, B.K.: Sociophysics: An Introduction. Oxford University Press, Oxford (2013)
19. Chatterjee, A., Ghosh, A., Chakrabarti, B.K.: In: Kirman, A., Aruka, Y. (eds.) Economic Foundations for Social Complexity Science: Theory, Sentiments, and Empirical Laws, pp. 51–65. Springer, Tokyo (2017). arxiv.org/pdf/1611.00723.pdf
20. Helbing, D., Balietti, S.: Complex techno-socioeconomics. Eur. Phys. J. Spec. Top. **195**, 1–136 (2011); see also, comments by Chakrabarti, B.K., ibid, pp. 145–146, and the comments by other reviewers for the proposal and the responses of the proposers, ibid, pp. 137–186
21. Ghosh, A.: Econophysics research in India in the last two decades. IIM Kozhikode Soc. Manag. Rev. **2**(2), 135–146 (2013)
22. Cheong, S.A.: Private communications (2013)
23. Aruka, Y., Kaizoji, T.: Private communications (2016–2017)
24. Di Matteo, T.: Private communications (2016–2017). See also: https://econophysicsnetwork.kcl.ac.uk/
25. Shubik, M.: Private communications (2016)
26. Stanley, H.E.: Private communications (2017)
27. Dragulescu, A., Yakovenko, V.M.: Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. Physica A **299**, 213–221 (2001)
28. Dragulescu, A., Yakovenko, V.M.: Statistical mechanics of money. Eur. Phys. J. B **17**, 723–729 (2000)
29. Devitt-Lee, A., Wang, H., Li, J., Boghosian, B.M.: A non-standard description of wealth concentration in large-scale economies. SIAM J. Appl. Math. **78**(2), 996–1008 (2018)
30. Boghosian, B.M., Devitt-Lee, A., Johnson, M., Li, J., Marcq, J.A., Wang, H.: Oligarchy as a phase transition: the effect of wealth-attained advantage in a Fokker-Planck description of asset exchange. Physica A **476**, 15–37 (2017)
31. Ohnishi, T., Mizuno, T., Shimizu, C., Watanabe, T.: Power laws in real estate prices during bubble periods. Int. J. Mod. Phys. Conf. Ser. **16**, 61–81. World scientific, Singapore (2012)
32. Tay, D.J., Chou, C.-I., Li, S.-P., Tee, S.Y., Cheong, S.A.: Bubbles are departures from equilibrium housing markets: evidence from Singapore and Taiwan. PLoS ONE, **11**(11) (3 November 2016)
33. Mandelbrot, B.: The variation of some other speculative prices. J. Bus. **40**, 393–413 (1967)
34. Mantegna, R.N., Stanley, H.E.: Scaling behaviour in the dynamics of an economic index. Nature **376**, 46–49 (1995)
35. Yura, Y., Takayasu, H., Sornette, D., Takayasu, M.: Financial brownian particle in the layered order-book fluid and fluctuation-dissipation relations. Phys. Rev. Lett. **112**, 098703 (2014)
36. The Atlas of Economic Complexity: Mapping Paths to Prosperity. MIT Press, Cambridge (2013). http://atlas.cid.harvard.edu/. https://growthlab.cid.harvard.edu/publications/atlas-economic-complexity-mapping-paths-prosperity
37. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., Pietronero, L.: A new metrics for countries fitness and products complexity. Sci. Rep. **2**, 723 (2012)
38. Bettencourt, L.M., Lobo, J., Helbing, D., Kuhnert, C., West, G.B.: Growth, innovation, scaling, and the pace of life in cities. Proc. Natl. Acad. Sci. **104**, 7301–7306 (2007)
39. Frank, M.R., Sun, L., Cebrian, M., Youn, H., Rahwan, I.: Small cities face greater impact from automation. J. R. Soc. Interface **15**(139), 20170946 (2018)
40. German science foundation physics of socio-economic systems: https://www.dpg-physik.de/dpg/gliederung/fv/soe/index.html. https://www.dpg-physik.de/dpg/gliederung/fv/soe/historie/phbl.html

Epilogue

Kiran Sharma and Anirban Chakraborti

> "The aim of life is inquiry into the Truth, and not the desire for enjoyment in heaven by performing religious rites,
> Those who possess the knowledge of the Truth, call the knowledge of non-duality as the Truth..."
> —Sūta, Bhagavata Purana 1.2.10–11 (800 and 1000 CE); Translated by Daniel Sheridan.

From times immemorial, human beings have been pursuing the Truth. Philosophers and scientists have been discovering the Truth in different forms through various phenomena in life and in nature. While the scientists were more interested in the external nature and tangible things, philosophers have been dwelling on the abstract, intangible things, and internal nature. In actuality, scholars have been taking glimpses of the Truth from different angles. While the spirit of inquiry and the ultimate aim has been similar, the methods of inquiry started differing with the passage of time. In the modern day, few dialogues exist between the philosophers and the scientists.

For that matter, even within the scientific disciplines there is often a lack of cross-dialogues. Even less, when it comes to the exchange of ideas between the natural sciences and social sciences. There have been some progress in this respect, e.g., between the fields of economics and physics, or sociology and physics, leading to the interdisciplinary fields named "Econophysics" and "Sociophysics". The series of conferences that we have been organizing since 2005 have been efforts to bridge the gaps between the parent fields.

Today, data has become easily available and indispensable in many spheres of life. We often hear the buzz words like "Big data" and "Data science". *What is Big data?* Big data is a term used to refer to vast and voluminous data sets that may be

K. Sharma
School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India
e-mail: kiransharma1187@gmail.com

A. Chakraborti
e-mail: anirban@jnu.ac.in

F. Abergel et al. (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, New Economic Windows,
https://doi.org/10.1007/978-3-030-11364-3

structured, semi-structured or unstructured. Big data can typically be characterized by *4Vs*- Volume, Variety, Velocity and Veracity. Big data is getting vast as compared to the traditional sources through which the data used to be captured (Volume); data is captured from various sources (Variety); the speed with which data is generated is phenomenal (Velocity), and given the volume and the variety of data, the quality of data (Veracity) is very important.

One needs to note that merely capturing data is not beneficial, but to understand what insights one can get from that data is of paramount importance in decision making. Big data eliminates intuition so that all imperative decisions can be made through a structured approach and with a data driven insight. Challenges with big data include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy.

"Data science" is an interdisciplinary field that requires data analysis, modeling/statistics and engineering/prototyping to extract knowledge and insights from large sets of data. Diving in at a granular level to mine and understand complex behaviors, trends and inferences. Data science focusses on analyzing the data, making sense of the large amount of data and discovering relevant patterns and designs in them, so that the data can be effectively utilized to realize future goals and objectives. There is a large amount of data available across almost all sectors in the world today and data science is taking on a big and prime role to utilize the data in a proper manner and make better decisions. Data science is all about being inquisitive – asking new questions, making new discoveries, and learning new things.

Data science is also needed to make valid, objective inferences from data, free from human bias. By examining large amounts of data, it is possible to uncover hidden patterns and correlations. So one needs to be very good at that or else the collective data could get wrong and give wrong information and can result in wrong decisions. Data science is a more forward looking approach, which helps answering the open-ended questions as to "what" and "how" events occur.

Our observation has been that with the advent of "Data science" and the era of "Big data" [1, 2], scholars have been coming closer to each other and more and more focus is being laid on interdisciplinary approaches. There has been also a growing need for new statistical and analytical methods, the kinds of which were being developed e.g., in econophysics [3–7]. So, in the coming years, we have many challenges ahead. We sincerely hope that the scientific community can bring forth a change in focus as well as the methods, which will serve mankind in a wholesome and purposeful way.

It was a pleasure co-hosting and co-organizing the joint international workshop. We hope to organize more such events. To reminisce the memories of the past event, we put two photos here (Fig. A.1).

Fig. A.1 Group photographs of the participants in Econophys-2017 and APEC-2017 meeting

Acknowledgements The authors thank Tanushree Mam and Alka Yadav for useful discussions and inputs.

References

1. Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014)
2. Jin, X., Wah, B.W., Cheng, X., Wang, Y.: Significance and challenges of big data research. Big Data Res. **2**(2), 59–64 (2015)
3. Bouchaud, J.P., Potters, M.: Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management. Cambridge University Press, Cambridge (2003)
4. Chakraborti, A., Muni Toke, I., Patriarca, M., Abergel, F.: Econophysics review: I. Empirical facts. Quant. Financ. **11**(7), 991–1012 (2011)
5. Chakraborti, A., Muni Toke, I., Patriarca, M., Abergel, F.: Econophysics review: II. Agent-based models. Quant. Financ. **11**(7), 1013–1041 (2011)
6. Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge (2007)
7. Sinha, S., Chatterjee, A., Chakraborti, A., Chakrabarti, B.K.: Econophysics: An Introduction. Wiley, New Jersey (2010)