BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

APPLICATION OF STATISTICAL PHYSICS IN TIME SERIES ANALYSIS

by

DUAN WANG

B.S., Nanjing University, 2007

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2015

Approved by

First Reader

H. Eugene Stanley, PhD
Professor of Physics, William Fairfield Warren Distinguished Professor

Second Reader

Robert Carey, PhD
Professor of Physics

To the best parents in the world,

Mr. Zhongren Wang and Mrs. Huiming Liu,

and the special one,

my girlfriend Bo Yan.

Acknowledgments

I would like to thank my thesis advisor, Prof. H. Eugene Stanley. He is a gorgeous scientist, for pushing the frontier of interdisciplinary research of complex systems and being a persistent source to provide invaluable advices towards my PhD researches; he is a wonderful advisor, for the financial support and for building such a nice academic family filled with lots of communications and collaborations; more importantly, he is also an amazing friend, for consistently cheering me up, encouraging me, and stimulating me during the seven years.

I also would like to thank my collaborators, Prof. Boris Podobnik, Prof. Davor Harvotic, and Prof. Ivo Grosse. I appreciate the time we were working together, discussing about our researches and physics, as well as brewing new ideas. It's my proud to work with you and I cannot imagine I could achieve anything without any of you.

I want to thank my thesis committee, Prof. H. Eugene Stanley, Prof. Boris Podobnik, Prof. Robert Carey, Prof. William Klein, and Prof. Sholomo Havlin, for your help with the dissertation.

I must acknowledge all the members of the Boston University Physics Department community for all the support, especially Mirtha Cabello and Bob Tomposki, you helped me through my PhD studies here. And Erik Lascaris, without you I can't even log in to our server.

I want to thank all my friends here at Boston University. Fengzhong Wang, Jia Shao, Guanliang Li, Yiping Chen, Di Zhou, Bo Chen, Songbo Jin, Qian Li, Wei Li, Zhiqiang Su, Shuai Shao, Tyler Xuan Gu, Xin Yuan, as well as all the member in Prof. Stanley's group. You are with me and share my pain as well as joy, and I will cherish every moment with you.

In the last, I want to thank my parents, Mr. Zhongren Wang and Mrs. Huiming Liu, for bringing me to this world, so I can experience everything from that moment, and for unconditional love. My girlfriend, Bo Yan, the best thing ever happened in my life. I love

you my family.

APPLICATION OF STATISTICAL PHYSICS IN TIME SERIES ANALYSIS

(Order No.)

DUAN WANG

Boston University, Graduate School of Arts and Sciences, 2015

Major Professor: H. Eugene Stanley, Professor of Physics

ABSTRACT

This dissertation covers the two major parts of my PhD research: i) Modeling instantaneous correlation ii) Quantifying time-lag correlation iii) Modeling time-lag correlation iv) Modeling and application of heteroskedasticity.

For modeling instantaneous correlation, we studied the limitations of random matrix theory (RMT), and proposed autoregressive random matrix theory (ARRMT) which take into account of the impact of autocorrelations in RMT.

For quantifying time-lag correlation, we proposed time-lag random matrix theory (TL-RMT) and found long-range crosscorrelations in financial, physiology and genomic data.

For modeling time-lag correlation, we proposed global factor model (GFM) and built the relationship between the autocorrelation of the global factor and the time-lag cross-correlation among individual time series.

For modeling and application of heteroskedasticity, we proposed a high frequency trading model using two fractionally intergrated autoregressive conditional heteroskedasticity (FIARCH) processes, and explained the fat-tailed distribution of returns and the long memory in volatilities of financial data.

Contents

List of Figures

List of Abbrevations

AR ..... Autoregressive

ARMA ..... Autoregressive Moving Average

PDF ..... Probability Density Function

CDF ..... Cumulative Density Function

ER ..... Erdős Rényi

GARCH Generalized AutoRegressive Conditional Heteroskedasticity

NOI ..... Number Of Iterations

NON ..... Network-Of-Networks

PDF ..... Probability Density Function

RMT ..... Random Matrix Theory

SF ..... Scale-Free

Chapter 1

Introduction

My research focuses on application of statistical physics methods in financial time series modeling. This dissertation will expand in the following way: In the first half, it revisits random matrix theory, a method proposed by Eugene Wigner to model collective behavior in nuclear physics, which was later applied to financial and other time series data to find the significance of correlations in a system. We discussed the limitations of the method and proposed several models based on RMT. In the second half, it covers the application of the heteroskedasticity to trading models.

In part I, In order to study the statistical structure of crosscorrelations in empirical data, we generalize random matrix theory (RMT) and propose a new method of cross-correlation analysis, which we call autoregressive random matrix theory (ARRMT). ARRMT takes into account the influence of auto-correlations in the study of cross-correlations in multiple time series. We first analytically and numerically determine how auto-correlations affect the eigenvalue distribution of the correlation matrix. Then we introduce ARRMT with a detailed procedure of how to implement the method. Finally we illustrate the method using two examples taken from inflation rates for air pressure data for 95 USA cities.

In part II, We study long-range magnitude cross-correlations in collective modes of real-world data from finance, physiology, and genomics using time-lag random matrix theory. We find long-range magnitude cross-correlations (i) in time series of price fluctuations, (ii) in physiological time series, both healthy and pathological, indicating scale-invariant interactions between different physiological time series, and (iii) in ChIP-seq data of the mouse genome, where we uncover a complex interplay of different DNA-binding proteins,

resulting in power-law cross-correlations in x_{ij}, the probability that protein j binds to gene i, ranging up to 10 million base pairs. In finance, we find that the changes in singular vectors and singular values are largest in times of crisis. We find that the largest 500 singular values of the NYSE Composite members follow a Zipf distribution with exponent ≈ 2. In physiology, we find statistically significant differences between alcoholic and control subjects.

In part III, We propose a modified time lag random matrix theory in order to study time lag cross-correlations in multiple time series. We apply the method to 48 world indices, one for each of 48 different countries. We find long-range power-law cross-correlations in the absolute values of returns that quantify risk, and find that they decay much more slowly than cross-correlations between the returns. The magnitude of the cross-correlations constitute "bad news" for international investment managers who may believe that risk is reduced by diversifying across countries. We find that when a market shock is transmitted around the world, the risk decays very slowly. We explain these time lag cross-correlations by introducing a global factor model (GFM) in which all index returns fluctuate in response to a single global factor. For each pair of individual time series of returns, the cross-correlations between returns (or magnitudes) can be modeled with the auto-correlations of the global factor returns (or magnitudes). We estimate the global factor using principal component analysis, which minimizes the variance of the residuals after removing the global trend. Using random matrix theory, a significant fraction of the world index cross-correlations can be explained by the global factor, which supports the utility of the GFM. We demonstrate applications of the GFM in forecasting risks at the world level, and in finding uncorrelated individual indices. We find 10 indices are practically uncorrelated with the global factor and with the remainder of the world indices, which is relevant information for world managers in reducing their portfolio risk. Finally, we argue that this general method can be applied to a wide range of phenomena in which time series are measured, ranging from seismology and physiology to atmospheric geophysics.

In part IV,

Financial markets exhibit a complex hierarchy among different processes, e.g., a trading time marks the initiation of a trade, and a trade triggers the price to change. High-frequency trading data arrive at random times. By combining stochastic and agent-based approaches, we develop a model for trading time, trading volume, and price changes. We generate intertrade time (time between successive trades) Δt_i, and the number of shares traded $q(\Delta t_i)$ as two independent but power-law autocorrelated processes, where Δt_i is subordinated to $q(\Delta t_i)$, and Δt_i is more strongly correlated than $q(\Delta t_i)$. These two power-law autocorrelated processes are responsible for the emergence of strong power-law correlations in (a) the total number of shares traded $N(\Delta T)$ and (b) the share volume $Q_{\Delta T}$ calculated as the sum of the number of shares q_i traded in a fixed time interval ΔT. We find that even though $q(\Delta t_i)$ is weakly power-law correlated, due to strong power-law correlations in Δt_i, the (integrated) share volume $Q(\Delta T) \equiv \sum_{i=1}^{\Delta T} q(\Delta t_i)$ exhibits strong long-range power-law correlations. We propose that intertrade times and bid-ask price changes share the same volatility mechanism, yielding the power-law auto-correlations in absolute values of price change and power-law tails in the distribution of price changes. The model generates the log-linear functional relationship between the average bid-ask spread $\langle S \rangle_{\Delta T}$ and the number of trade occurrences $N_{\Delta T}$, and between $\langle S \rangle_{\Delta T}$ and $Q_{\Delta T}$. We find that both results agree with empirical findings.

In order to avoid redundant and repeating statement in the context, some basic concepts and definition will be introduced here.

1.1 Autocorrelation and Crosscorrelation

Definition, AR, MA, ARFIMA, ...

1.2 Heteroskedasticity

Definition, GARCH, FIARCH,...

1.3 Principal Component Analysis

Equations...

1.4 Random Matrix Theory

Marchenko-Pastur...

Chapter 2

Autoregressive Random Matrix Theory

2.1 Introduction

Cross-correlations have been observed in the outputs of a wide range of phenomena including nanodevices [**?**, 1, **?**], in various fields of wave physics such as ultrasonics [2], underwater acoustics [3], geophysics [4, 5], seismology [6], and finance [7, 8, 9, **?**]. Numerous methods have been introduced to analyze cross-correlations between time series [7, 8, **?**, 10, 11, 12] among which random matrix theory (RMT) is one of the most popular methods in analyzing cross-correlations in multiple time series [7, 8, 13, 14, 15, 16, 17, 18, 19, 20]. The usual approach in RMT is to study the eigenvalue distribution of a Wishart matrix, which is the correlation matrix for finite-length independent and identically distributed (*i.i.d.*) series, and compare it to the eigenvalue distribution of the cross-correlation matrix of an empirical time series. Deviations between these two distributions might then suggest the presence of cross-correlations in the data. In this paper we discuss the limitation of using RMT when empirical data are strongly autocorrelated, and propose a generalization of RMT, autoregressive random matrix theory (ARRMT) to address this problem.

2.2 Impact of Autocorrelations on Random Matrix Theory

When cross-correlations are calculated for empirical data, the degree of cross-correlations between the two time series is usually measured by the cross-correlations coefficient, defined as $\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$, where σ_X σ_Y are the standard deviations of X and Y, respectively, and μ_X and μ_Y are the expected values of X and Y, respectively. The

sample cross-correlation coefficient can be calculated by

$$r = \frac{1}{T-1}\sum_{i=1}^{T}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right). \tag{2.1}$$

For the Wishart matrix for N uncorrelated *i.i.d.* time series, each with length $T \geq N$, the eigenvalues follow a Marchenko-Pastur distribution: $P(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}$ [21], where $Q \equiv \frac{N}{T}$ and

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \tag{2.2}$$

are the maximum and minimum eigenvalues of W. According to RMT, the difference between the eigenvalue distributions of an empirical cross-correlations matrix and a Wishart matrtix indicates the presence of cross-correlations and collective modes in the empirical time series. If cross-correlations are present in the empirical time series, we expect some eigenvalues being larger than λ_+, where the largest eigenvalue λ_L indicates the global behavior of the multiple time series. The eigenvalues smaller than λ_+ and their corresponding eigenvectors are considered as noise.

However, RMT has serious limitation when applied in practise, since it doesn't take into account that the empirical eigenvalue distributions of the cross-correlation matrix can be influenced by auto-correlations in empirical data. The *i.i.d.* time series used to calculate a Wishart matrix generally differ from the empirical time series since commonly in contrast to *i.i.d.* time series, in the empirical time series there are (i) cross-correlations between pairs of time series, and (ii) there are auto-correlations in individual time series. Therefore, the difference between the eigenvalue distributions of the empirical correlation matrix and the Wishart matrix can be caused by either cross-correlations or auto-correlations in empirical data.

Autocorrelations change the eigenvalue distribution of uncorrelated time series by changing the distribution of the sample correlated coefficients between each pair of data. If (X, Y)

has a bivariate *i.i.d* normal distribution, then the Fisher transformation of r, $\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ is approximately normally distributed with mean $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$, and standard error $\frac{1}{\sqrt{N-3}}$ [22]. For a limit when $|r| \ll 1$ and $T \gg 1$, the distribution of the sample correlation coefficients for N *i.i.d.* series is well approximated by a normal distribution with mean zero and standard error $\sqrt{\frac{1}{T}}$. However, the distribution of r will change when both X and Y are autocorrelated time series.

To simplify derivation of the distribution of r between autocorrelated time series, we use a standardized time series $z_t = (X_t - \langle X_t \rangle)/s_X$. The sample cross-correlation coefficient between X_t and X'_t can be expressed as $r = \langle z_t z'_t \rangle = \frac{1}{T}\sum_{t=0}^{T} z_t z'_t$. We assume that X_t and X'_t are not cross-correlated, but both X_t and X'_t are auto-correlated. Thus r is a random variable with expectation zero and variance

$$Var(r) = \frac{1}{T^2}\sum_t \sum_{t'} E(z_t z_{t'}) E(z'_t z'_{t'}) \tag{2.3}$$

$$= \frac{1}{T^2}\sum_t \sum_{t'} d^{|t-t'|}, \tag{2.4}$$

where we use $E(z_t z_{t'}) = A(|t-t'|)$ and $E(z'_t z'_{t'}) = A'(|t-t'|)$, and $A(|t-t'|)$ and $(A'(|t-t'|))$ are the auto-correlations of X_t and X'_t, respectively, where $|t-t'|$ denotes the time lags.

It is straightforward to show that when $|r| \ll 1$ and $T \gg 1$,

$$Var(r) \approx \frac{1}{T}[1 + 2\sum_{\Delta t=1}^{\infty} A(\Delta t) A'(\Delta t)] \tag{2.5}$$

[23]. Compared to an *i.i.d.* time series, the variance of sample correlation coefficients is increased by $\frac{2}{T}\sum_{\Delta t=1}^{\infty} A(\Delta_t) A'(\Delta t)$. We can say that Eq. (2.5) corresponds to *i.i.d.* time series with a different number of observations [24], where the effective number of observations T^* can be obtained from $\frac{1}{T^*} = \frac{1}{T}[1 + 2\sum_{\Delta t=1}^{\infty} A(\Delta t) A'(\Delta t)]$. Therefore by T^* we denote the equivalent length of an autocorrelated time series.

In order to show how the presence of auto-correlations affects the eigenvalue distribution, we assume that empirical time series are generated by the first-order autoregressive

AR(1) process,

$$X_t = \phi X_{t-1} + \epsilon_t, \tag{2.6}$$

where ϕ ($|\phi| < 1$) is a parameter and ϵ is an *i.i.d.* process. The auto-correlation function of an AR(1) process decays with Δt as an exponential function, $A(\Delta t) = \phi^{|\Delta t|}$ [25]. Applying Eq. (2.5), the variance of sample correlation coefficients for two AR(1) process, each defined by coefficients ϕ and ϕ', respectively, is

$$\mathrm{Var}(r) = \frac{1}{T}\frac{1+\phi\phi'}{1-\phi\phi'}. \tag{2.7}$$

Suppose we have N time series X_t, each with the same AR(1) coefficient ϕ, where time series X_t are not cross-correlated. Using Eq. (2.7), where $\phi = \phi'$, with the corresponding expression that holds for *i.i.d.* time series of length T^* which variance is $\frac{1}{T^*}$, we obtain $T^* = T\frac{1-\phi^2}{1+\phi^2}$. Since the eigenvalue distribution of the cross-correlation matrix generated by the *i.i.d.* time series depends only on $Q = T/N$, we can defined a equivalent Q as

$$Q^* = T^*/N = \frac{T}{N}\frac{1-\phi^2}{1+\phi^2}. \tag{2.8}$$

Similarly, the eigenvalue distribution becomes

$$P(\lambda) = \frac{Q^*}{2\pi}\frac{\sqrt{(\lambda_+^* - \lambda)(\lambda - \lambda_-^*)}}{\lambda}, \tag{2.9}$$

where the largest and smallest eigenvalues equal to

$$\lambda'_\pm = 1 + \frac{1}{Q^*} \pm 2\sqrt{\frac{1}{Q^*}}. \tag{2.10}$$

The results above are approximate and hold better for weak autocorrelations. When autocorrelations are large, the distribution of sample correlation coefficients can no longer be approximated by a normal distribution, and therefore the equations will no longer

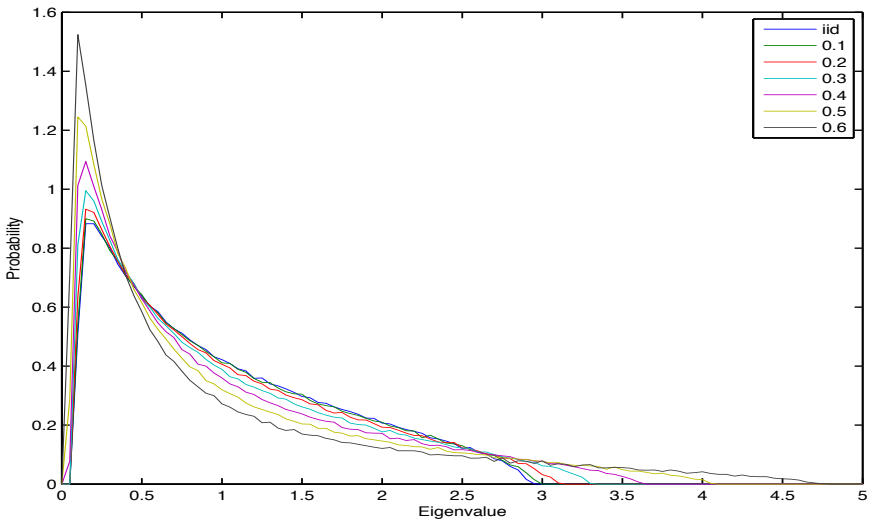


Figure 2.1: Eigenvalue distribution for $N = 2000$ autocorrelate time series each with length $T = 4000$. Time series are simulated using AR(1) processes, with ϕ from 0 to 0.6. As can be seen from the figure, the largest eigenvalue increased from 2.9298 to 4.7257 when ϕ increased from 0 to 0.6.

hold, but the largest eigenvalue still increase with autocorrelation because of the increased variance of the crosscorrelations. We simulate $N = 2000$ time series each with length $T = 4000$. using AR(1) processes, with ϕ from 0 to 0.6, the largest eigenvalue increased from 2.9298 to 4.7257. See Fig. (2.1).

2.3 Method

In order to remove the influence of auto-correlations and study only how cross-correlations affect the data, we introduce the auto-regressive Wishart matrix, that is the correlation matrix of artificial time series $\{Y_t'\}$, with no cross-correlations but with the same auto-correlations as existing in the empirical time series $\{Y_t\}$. By replacing the Wishart matrix in RMT, W, with the autoregressive Wishart matrix, W', we remove the influence of the auto-correlations on the eigenvalue distributions. In this way, the difference between the eigenvalue distributions of the empirical correlation matrix C and W' is purely due to the

cross-correlations between time series $\{Y_t\}$. We call this method autoregressive random matrix theory (ARRMT).

The steps of ARRMT is defined as:

(i) Autocorrelation test: We test whether auto-correlations are significant among N cross-correlated original time series $Y_{i.t}$. One of the most popular autocorrelation test is Ljung-Box test [26].

(ii) Fit autocorrelation model: We fit each time series $Y_{i,t}$ with an appropriate auto-correlation model, the one that is the best fit for $Y_{i,t}$. Based on the fitting we assign to each series i a set of model parameters (e.g., $\phi_i, \theta_i, ...$). The simplest model is AR(1), while higher orders of autoregressivemoving-average (ARMA) models

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2.11}$$

can also be used if AR(1) cannot fit the autocorrelations very well.

(iii) Simulate: Using the fitted model from (i), we simulate N time series $Y'_{i,t}$, each characterized by the same coefficients $(\phi_i, \theta_i, ...)$ as we found in the original time series $Y_{i,t}$. Then, $Y'_{i,t}$ has the same auto-correlation properties as the original time series $Y_{i,t}$.

(iv) Calculate eigenvalues: We calculate the cross-correlation matrix W' of the generated time series $Y'_{i,t}$. Then we calculate the largest eigenvalue λ'_+ of W'.

(vi) Compare: Finally, we compare the largest eigenvalue λ'_+ with the eigenvalues of the correlation matrix C of the empirical time series. Eigenvalues larger than λ'_+ are related to significant factors.

When N and T are small, then the variance of the the simulated λ'_+ will be large. Therefore the last step in the procedure above become:

(v) Repeat: We repeat steps (ii) and (iii) for n times, and calculate the 95th percentile of λ'_+. We call it $\lambda'_{+0.95}$.

(vi) Compare: Finally, we compare $\lambda'_{+0.95}$ with the largest eigenvalue λ_L of the correlation matrix C of the empirical time series. Eigenvalues larger than λ'_+ are related to significant factors.

2.4 Empirical Results

To illustrate the ARRMT method, we apply both RMT and ARRMT to multiple time series characterized by both cross-correlations and auto-correlations, data comprising 649 daily changes in atmospheric pressure $P_{i,t}$ for 95 different cities in the US, defined as

$$R_{i,t} = P_{i,t} - P_{i,t-1}. \tag{2.12}$$

In order to demonstrate the advantage of using ARRMT over RMT, first we apply RMT to air pressure changes, and calculate the 95th percentile of the largest eigenvalues $\lambda_{+0.95} = 1.9174$ of the Wishart matrix using Eq. (2.2). Then we calculate the correlation matrix of empirical time series and the empirical eigenvalues, among which the largest eigenvalue is $\lambda_L = 8.9740$ ($\gg \lambda_+$), indicating the existence of cross-correlations. We find that, among the 20 eigenvalues, there are 13 eigenvalues larger than λ_+, indicating 13 significant factors influencing air pressure of the 95 cities.

Next we apply ARRMT by assuming that AR(1) of Eq. (2.6) is appropriate candidate to model auto-correlations in the data. Thus, by Eq. (2.6) we fit each of the 95 air pressure change time series R_t of Eq. (2.12). For each series $R_{i,t}$ we obtain the AR(1) coefficient (ϕ_i) . Fig.2.2 shows the distribution of AR(1) coefficients, which indicates that the auto-correlations are significant in most of the time series. Then we generate 95 time series $Y'_{i,t}$ using AR(1) model, each with the fitted value of ϕ_i .

Next we calculate the correlation matrix W' of the 95 generated time series $Y'_{i,t}$ and

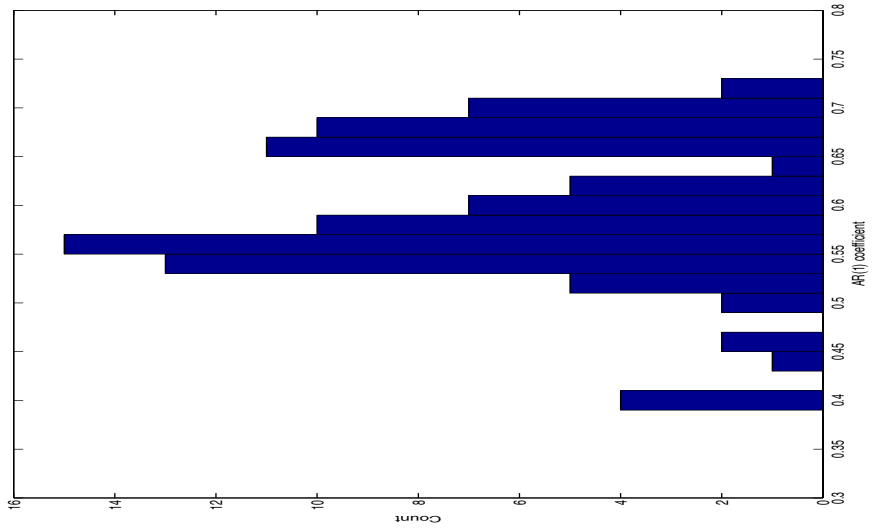


Figure 2.2: Distribution of AR(1) coefficients of the 95 air pressure changes time series. The distribution indicates that most of the air pressure change time series have strong positive autocorrelations.

the eigenvalue distribution. Fig.2.3 shows the largest eigenvalue for the Wishart matrix W and autoregressive Wishart matrix W'. We find, as expected due to the presence of auto-correlations in the data, that the 95th percentile of largest eigenvalues of matrix W', $\lambda'_{+0.95} = 2.5922$, is larger than $\lambda_{+0.95} = 1.9174$ calculated for the Wishart matrix W. Then we compare $\lambda'_{+0.95}$ of Eq. (2.10) with the largest eigenvalues obtained for the empirical correlation matrix of the inflation rates and find that ARRMT reveals that there are only 8 significant eigenvalue larger than λ'_+. Thus, taking into account the presence of auto-correlations in the data, ARRMT finds out that there are only 8 factors that accounts for the air pressure changes in the 95 cities.

In practice, when empirical data exhibit auto-correlations with longer memory, AR(1) should be replaced by the more general AR(n) process $X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i}$, and we fit each time series with a higher-order AR(n) model. In this example, we find that AR(10) fits the data better than AR(1). Applying AR(10) model, we find that the largest eigenvalue is $\lambda'_+ = 2.823$. Although it is larger than 2.592 obtained by AR(1), the number of significant

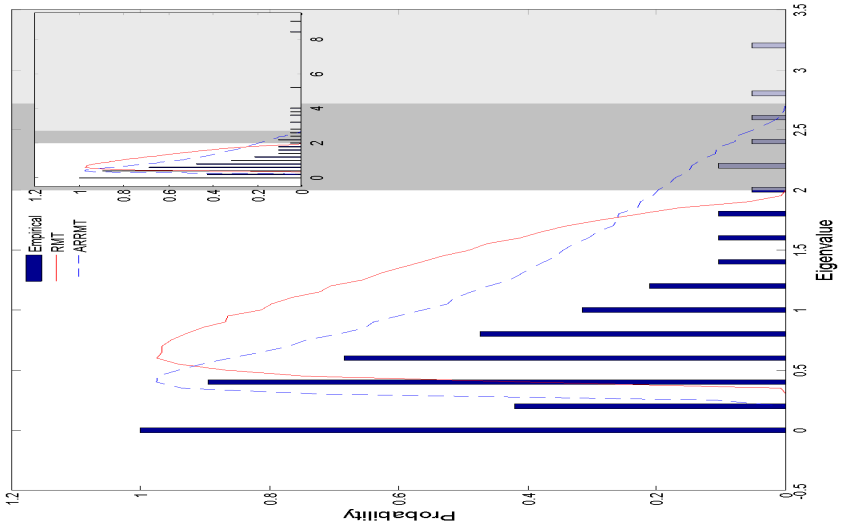


Figure 2.3: Bar: eigenvalue distribution of the correlation matrix for the air pressure changes of 95 US cities. Red solid line: the eigenvalue distribution of 95 simulate random time series repeated 1000 times. Blue dashed line: eigenvalue distribution of 95 simulated uncorrelated time series which has the same autocorrelations as the empirical time series repeated 1000 times. Using ARRMT, we find 8 empirical eigenvalue is larger than λ'_+, indicating that there are only 8 factors that accounts for the air pressures in 95 US cities, as compared to 13 from RMT.

factors is still 8. **(Add a graph here show hline for RMT , ARRMT with AR(1), ARRMT with AR(10), similar to the CPI data.) (What should I say about the result?)**

2.5 Summary

We find that auto-correlations can significantly influence the eigenvalue distribution of the correlation matrix, and that RMT is therefore unreliable in analyzing cross-correlations in multiple time series characterized also by strong auto-correlations. To take into account the presence of auto-correlations in cross-correlated time series, we introduce auto-regressive random matrix theory (ARRMT). In ARRMT we use a modified Wishart matrix which takes into account auto-correlations commonly present in empirical data. The difference between the eigenvalue distributions of empirical correlation matrix and modified Wishart matrix purely reflects the existence of cross-correlations. We illustrate ARRMT using inflation atmospheric pressure data for 95 USA cities.

Chapter 3

Quantifying Long-range Crosscorrelation

3.1 Introduction

labelTLRMT-Introdution

Many complex systems are part of even larger systems where the constituent complex systems mutually interact [27, 28, 29, 30], giving rise to the appearance of "collective modes" [**?**, 31, 8, 7]. Stochastic interactions among related systems are reflected in the presence of cross-correlations, and here we address the question of whether these cross-correlations in the collective modes exhibit power-law scale-invariant properties.

Zero-lag cross-correlations in the collective modes of empirical time series were analyzed by using random matrix theory (RMT) [8, 7, 15, 32]. Recently, RMT became very successful in analysis of cross-correlations between stock price changes, since cross-correlation matrices and associated covariance matrices play important roles in portfolio management [33, 34]. A variety of studies reported the properties of the cross-correlation matrix C of price changes [8, 7, 15, 33, 34, 32, 35, 36]. RMT enables a comparison between the cross-correlation matrix obtained from N empirical time series each of length T and a perfectly random matrix W, called a Wishart matrix, obtained from N mutually uncorrelated time series each of length T [37]. By analyzing the cross-correlations between price changes of the members of the S&P500 index, it has been found that 98% of the eigenvalue spectrum of the correlation matrix C follows the Gaussian orthogonal ensemble of a perfectly random matrix [7, 15].

Recently, time-lag generalizations of RMT were proposed [38, 39, 40, 41, **?**, **?**]. How-

ever, only short-range cross-correlations were found. To quantify long-range collective movements in correlated data sets, we apply time-lag RMT (TLRMT) to the magnitude of three selected examples of real-world data: (i) finance, (ii) physiology, and (iii) genomics.

3.2 Methods

3.2.1 Random Matrix for Time-lag Correlation Matrix

Consider the N-variable time series $X = \{X_{i,t} : i = 1, \ldots, N; t = 1, \ldots, T\}$ of length T, where i indexes the series number, and t denotes the time. The cross-correlation matrices for this time series and for the magnitude time series are

$$C_{ij}(\Delta t) \equiv \frac{\langle X_{i,t} X_{j,t+\Delta t}\rangle - \langle X_{i,t}\rangle\langle X_{j,t+\Delta t}\rangle}{\sigma_i \sigma_j}, \tag{3.1}$$

$$\tilde{C}_{ij}(\Delta t) \equiv \frac{\langle |X_{i,t}||X_{j,t+\Delta t}|\rangle - \langle |X_{i,t}|\rangle\langle |X_{j,t+\Delta t}|\rangle}{\tilde{\sigma}_i \tilde{\sigma}_j}. \tag{3.2}$$

Here σ_i, σ_j, $\tilde{\sigma}_i$, and $\tilde{\sigma}_j$ denote the standard deviations of $X_{i,t}$, $X_{j,t+\Delta t}$, $|X_{i,t}|$, and $|X_{j,t+\Delta t}|$, respectively, and $\langle ... \rangle$ denotes the time average.

In order to quantify cross-correlations for varying lags Δt, we compute the largest singular values $\lambda_L(\Delta t)$ and $\tilde{\lambda}_L(\Delta t)$ of the cross-correlation matrices $C(\Delta t)$ and $\tilde{C}(\Delta t)$ as functions of Δt [**?**]. The squares of the non-zero singular values of C are equal to the non-zero eigenvalues of CC^+ or C^+C, where C^+ denotes the transpose of C. In a singular value decomposition $C = U\Sigma V^+$ the diagonal elements of Σ are equal to singular values of C. The columns of U and V are left and right singular vectors of the corresponding singular values. Consider a matrix C with main diagonal elements 1s and all off-diagonal elements being identical, i.e., $C_{ij}(\Delta t) \equiv C(\Delta t)$. Then we calculate the largest eigenvalue of CC^+ (equal to $\lambda_L^2(\Delta t)$)

$$\lambda_L^2(\Delta t) = 1 + (N-1)^2 C(\Delta t) + 2(N-1)C(\Delta t)^2. \tag{3.3}$$

If C_{ij} follows a power law $C_{ij}(\Delta t) = A(\Delta t)^{-\gamma}$, then for $\Delta t >> 1$, $\lambda_L(\Delta t) = 1 + 0.5A(N-1)^2(\Delta t)^{-\gamma}$, where A is constant.

3.2.2 Adjust for Autocorrelation

Show simulated results to illustrate the difference.

3.2.3 Adjust for Missing Data

Discuss about interpolation, stochastic volatility, and what model?

3.3 Results

We find long-range cross-correlations in the following data series:

(i) 1,340 members $I_{i,t}$ of the New York Stock Exchange (NYSE) Composite. We analyze $1,340$ time series with $2,172$ daily records in the 8.7-year period, 2 Jan 2001 to 24 Aug 2009. We focus on the logarithmic change ("returns") defined as $R_{i,t} \equiv \ln(I_{i,t}/I_{i,t-1})$, where i denotes the index member [**?**], and t stands for time in days. First we apply RMT to calculate $\lambda_L(\Delta t = 0) = 392$ for returns $R_{i,t}$ and $\tilde{\lambda}_L(\Delta t = 0) = 359$ for volatilities $|R_{i,t}|$, and we find that both largest singular values are more than 100 times greater than expected for uncorrelated time series , indicating cross-correlation for $\Delta t = 0$. Next, we apply TLRMT, and we plot $\lambda_L(\Delta t)$ and $\tilde{\lambda}_L(\Delta t)$ as a function of Δt in Fig. **??**(a). We find long-range volatility cross-correlations, implying that $|R_{i,t}|$ affects $|R_{j,t+\Delta t}|$ $(i \neq j)$ for $\Delta t > 0$. Even though RMT shows that $\lambda_L(\Delta t = 0)$ and $\tilde{\lambda}_L(\Delta t = 0)$ are practically the same as found in Ref. [8] for eigenvalues, it is TLRMT that shows that $\tilde{\lambda}_L(\Delta t)$ vs. Δt decays more slowly than $\lambda_L(\Delta t)$, stating that volatility cross-correlations last longer than return cross-correlations, and we find that $\tilde{\lambda}_L(\Delta t)$ can be approximated by a power law $(\Delta t)^{-\gamma}$ with a scaling exponent of $\gamma = 0.64 \pm 0.03$. Note that Ref. [**?**, **?**] reported power-law cross-correlations between pairs of financial time series in magnitudes.

(ii) Physiology Sleep Heart Health Study (SHHS) database[**?**, **?**]. For a single patient, we study 11 time series, $I_{i,t}$, among which are electroencephalography (EEG), heart rate, and electrooculogram. Here, $i = 1, \ldots, 11$ denotes the index of the physiological time series, and t denotes time in seconds. We show $\tilde{\lambda}_L(\Delta t)$ and $\lambda_L(\Delta t)$ in Fig. **??**(b), and we find that $\tilde{\lambda}_L(\Delta t) \sim (\Delta t)^{-\gamma}$ where $\gamma = 0.06 \pm 0.01$. These findings indicate that TLRMT might be useful for distinguishing healthy from pathological behavior for multivariate correlated time series, the same as detrended fluctuation analysis (DFA) [**?**, **?**, **?**, **?**] for a single time series [31].

(iii) ChIP-seq data of mouse chromosome 2[**?**]. The binding affinity of 14 DNA-binding proteins to the DNA of mouse chromosome 2 was obtained by calculating the probability x_{ij} that protein i binds to gene j on mouse chromosome 2 for all $i = 1, ..., 14$ and $j = 1, ..., 1544$. We apply the DFA fluctuation function $F(n)$ [**?**, **?**, **?**, **?**] to these 14 numerical sequences (spatial not temporal) x_{ij} of the same length, and we find that $F(n)$ can be approximated by a power law for all of the 14 numerical sequences, $F(n) \propto n^{\alpha}$. This power-law scaling of $F(n)$ with n indicates the presence of long-range autocorrelations for the 14 individual sequences x_{ij}, and the average DFA scaling exponent $\alpha = 0.69 > 0.5$ indicates that neighboring genes have a higher tendency to be both bound, or to be both unbound, by the same transcription factor than expected by chance. One possible interpretation of this finding is that there is some evolutionary advantage for a species if its genes whose promoters are bound by the same transcription factor are close to each other in the genome. Moreover, the power-law scaling of $F(n)$ with n indicates that this tendency does not decay exponentially with the chromosomal distance between the genes, but this tendency rather decays algebraically.

Next, we focus on nonequal-lag cross-correlations using the TLRMT approach. We show $\tilde{\lambda}_L(\Delta t)$ and $\lambda_L(\Delta t)$ in Fig. **??**(c), and we find approximate power-law cross-correlations with scaling exponents of 0.37 ± 0.01 and of 0.18 ± 0.01, respectively,

implying that the binding or unbinding of protein j to gene i is influenced by the binding or unbinding of other proteins j' to gene i or neighboring genes i'. Interestingly, the neighborhood reaches up to $|i - i'| \approx 100$ genes, corresponding to a chromosomal distance of approximately 10 million base pairs.

In order to investigate if TLRMT might be useful for prediction, we apply it to financial and genomics time series.

(i) 88 companies that contribute to the S&P 500 index in 2009 during the 26-year period 1983-2009. We apply TLRMT for each year, as in Fig. **??**(a), and we show $\tilde{\lambda}_L(\Delta t)$ vs. year in Fig. **??**(a). We find pronounced peaks during the largest market shocks and economic crisis: Black Monday, the Dot-com Bubble, and the 2008 crash. We study different time lags Δt, because the presence of cross-correlations for $\Delta t = 0$ does not imply the presence of cross-correlations for $\Delta t \neq 0$, and indeed we find interesting differences when tuning the time lag $\Delta t \neq 0$. We investigate how $\tilde{\lambda}_L(\Delta t)$ changes over lags (days) for different years, and we show $\tilde{\lambda}_L(\Delta t)$ vs. lag in Fig. **??**(b). We find that $\tilde{\lambda}_L(\Delta t) \sim (\Delta t)^{-\gamma}$ where γ varies from year to year and is greatest in times of crisis.

We calculate the absolute values of the 88 components $|u_i(\tilde{\lambda}_L(\Delta t))|$ of left-singular vectors corresponding to $\tilde{\lambda}_L(\Delta t)$ of the volatility time series in Fig. **??**(a). We show the mean μ and the standard deviation σ of the 88 values of $|u_i(\tilde{\lambda}_L(\Delta t))|$ for each $\Delta t = 0, 4, 16$, in Fig **??**(c), and we find that μ suddenly increases in 2002, whereas σ suddenly decreases. This finding can be partially explained by Fig. **??**(d), where we find that $|u_i(\tilde{\lambda}_L(\Delta t))|$ substantially change after the Dot-com Bubble crash in 2001.

In addition to the largest singular value $\tilde{\lambda}_L(\Delta t)$ shown in Fig. **??**(a), we calculate all singular values $\tilde{\lambda}(\Delta t)$ of the NYSE Composite. We show the rank-ordered distribution of the largest 500 $\tilde{\lambda}(\Delta t)$ for each Δt in Fig. **??**. We find that the distributions for different Δt practically overlap (power-law stability), and they can be approximated by a power law with exponent 2. In comparison, pdfs of returns exhibit power-

law tails with exponent ≈ 4 [**?**]. The first power law is accompanied by power-law volatility cross-correlations (Fig. 1(a)), and the latter by power-law volatility auto-correlations [**?**, **?**, **?**, **?**, **?**].

(ii) EEG time series. Ref. [35] reported power-law auto-correlations in both EEG time series and their magnitudes, with different exponents for healthy subjects and subjects with Alzheimer's disease. Ref. [**?**] reported power-law cross-correlations between pairs of EEG time series in magnitudes. Ref. [**?**] reported that cross-correlations for $\Delta t = 0$ calculated between pairs of EEG time series are inversely related to dissociative symptoms (psychometric measures) in 58 patients with paranoid schizophrenia.

Here we analyze multiple time series of EEG recordings of two groups of subjects: control and alcoholic [**?**]. These data arise from a study to examine EEG correlates of a genetic predisposition to alcoholism. Measurements were obtained from 64 electrodes placed on the scalp, sampled at 256 Hz (3.9-msec epoch) for 1 second. The electrodes were placed at standard sites (Standard Electrode Position Nomenclature, American Electroencephalographic Association 1990). Each of 122 subjects completed 120 trials. Each subject was exposed either to a single stimulus (S1) or to two stimuli (S2) which were different pictures of objects. If two stimuli are equal it is called a matched (S2-M) condition, whereas if two stimuli are different it is called a non-matched (S2-NM) condition.

We randomly choose 15 alcoholic and 15 control subjects. For a given time lag ($\Delta t = 0, 5, 10, 20$) and a given condition (S1, S2-M, S2-NM), we calculate all singular values $\tilde{\lambda}(\Delta t)$ for each alcoholic subject and for each control subject. We calculate the standard deviations σ of $\tilde{\lambda}(\Delta t)$ for each Δt, for each condition, and for alcoholic subjects and control subjects separately, and we show those standard deviations in Fig. **??**(a). We find that σ for control subjects is greater than σ for alcoholic subjects for each Δt and each condition. Our F-test confirms that the differences between alcoholic subjects and control subjects are statistically significant for the S2-M and

the S2-NM condition. We show in Fig. **??**(b) the mean μ of $\tilde{\lambda}_L(\Delta t)$ for each Δt, for each condition, and for alcoholic subjects and control subjects separately. We find that μ for control subjects is substantially greater than for alcoholic subjects for the S2-M and the S2-NM conditions.

Next, for each condition and for alcoholic subjects and control subjects separately, we calculate the magnitudes of components $|u_i(\tilde{\lambda}_L(\Delta t))|$ of all right-singular vectors corresponding to $\tilde{\lambda}_L(\Delta t)$ of the volatility time series. For different conditions in Fig. **??**(c) we show the mean μ of $|u_i(\tilde{\lambda}_L(\Delta t))|$ for varying Δt. For the S2-M and the S2-NM conditions, we find significant difference between alcoholic and control subjects. In case of left-singular vectors for small lags, for the same conditions, we find less substantial difference between alcoholic and control subjects.

3.4 Summary

Cross-correlations are found in a number of studies including nanodevices [**?**, 1, **?**], atmospheric geophysics [4], seismology [6], and finance [8, 7, **?**, 9, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. We study cross-correlations in both temporal and spatial collective modes using time-lag RMT (TLRMT). We find long-range cross-correlations in quite diverse systems, ranging in size from the earth's atmosphere (a volume of approximately $5 \times 10^{18} m^3$) to microscopic systems such DNA sequences (a volume of $5 \times 10^{-18} m^3$), ranging from living to non-living systems, and ranging from physical to non-physical systems such as the financial market.

In genomics data, we find spatial cross-correlations corresponding to a chromosomal distance of ≈ 10 million base pairs. In physiology, TLRMT reveals statistically significant difference in standard deviations and means of singular values between alcoholic and control subjects. In finance, by analyzing cross-correlations in the magnitudes of price fluctuations we find that the largest singular values and their singular vectors substantially change after the Dot-com Bubble crash in 2001. We also find that the largest 500 singular values of the NYSE Composite members nicely follow a Zipf distribution. We find power-law decaying

cross-correlations in the magnitudes of price fluctuations implying that large magnitude fluctuations—commonly taken to quantify risk—created in one stock are transferred to other stocks, and this impact last for many time scales. Such cross-correlations are of potential interest in risk management. TLRMT reveals that cross-correlations are strongest during market crashes and global recessions.

Our findings are consistent with interesting possibility that cross-correlations are ubiquitously present in many systems. Studying these cross-correlations is a necessary prerequisite for understanding them, and a deeper understanding of these cross-correlations enables a deeper understanding of these systems. A deeper understanding of these systems, in turn, enables improved clinical applications and increases our forecasting power. The TLRMT approach developed in statistical physics may contribute to this long-term goal and lead to potential advancements of diverse areas of science.

Chapter 4

Modeling Long-Range Crosscorrelation

4.1 Introduction

When complex systems join to form even more complex systems, the interaction of the constituent subsystems is highly random [27, 28, 29, 30]. The complex stochastic interactions among these subsystems are commonly quantified by calculating the cross-correlations. This method has been applied in systems ranging from nanodevices [**?**, 1, **?**], atmospheric geophysics [4], and seismology [6, **?**, 42], to finance [**?**, **?**, **?**, 7, 8, **?**, 9, **?**, **?**, **?**, 12, 10, **?**]. Here we propose a method of estimating the most significant component in explaining long-range cross-correlations.

Studying cross-correlations in these diverse physical systems provides insight into the dynamics of natural systems and enables us to base our prediction of future outcomes on current information. In finance, we base our risk estimate on cross-correlation matrices derived from asset and investment portfolios [**?**, 7, 8]. In seismology, cross-correlation levels are used to predict earthquake probability and intensity [6]. In nanodevices used in quantum information processing, electronic entanglement necessitates the computation of noise cross-correlations in order to determine whether the sign of the signal will be reversed when compared to standard devices [**?**]. Reference [**?**] reports that cross-correlations for $\Delta t = 0$ calculated between pairs of EEG time series are inversely related to dissociative symptoms (psychometric measures) in 58 patients with paranoid schizophrenia. In genomics data, Ref. [16] reports spatial cross-correlations corresponding to a chromosomal distance of ≈ 10 million base pairs. In physiology, Ref. [16] reports a statistically significant

difference between alcoholic and control subjects.

Many methods have been used to investigate cross-correlations (i) between pairs of simultaneously recorded time series [12, 10] or (ii) among a large number of simultaneously-recorded time series [7, 8, **?**, 11]. Reference [11] uses a power mapping of the elements in the correlation matrix that suppresses noise. Reference [12] proposes detrended cross-correlation analysis (DCCA), which is an extension of detrended fluctuation analysis (DFA) [**?**] and is based on detrended covariance. Reference [10] proposes a method for estimating the cross-correlation function C_{xy} of long-range correlated series x_t and y_t. For fractional Brownian motions with Hurst exponents H_1 and H_2, the asymptotic expression for C_{xy} scales as a power of n with exponents H_1 and H_2.

Univariate (single) financial time series modeling has long been a popular technique in science. To model the auto-correlation of univariate time series, traditional time series models such as autoregressive moving average (ARMA) models have been proposed [**?**]. The ARMA model assumes variances are constant with time. However, empirical studies accomplished on financial time series commonly show that variances change with time. To model time-varying variance, the autoregressive conditional heteroskedasticity (ARCH) model was proposed [**?**]. Since then, many extensions of ARCH has been proposed, including the generalized autoregressive conditional heteroskedasticity (GARCH) model [**?**] and the fractionally-integrated autoregressive conditional heteroskedasticity (FIARCH) model [**?**]. In these models, long-range auto-correlations in magnitudes exist, so a large price change at one observation is expected to be followed by a large price change at the next observation. Long-range auto-correlations in magnitude of signals have been reported in finance [**?**], physiology [**?**, **?**], river flow data [**?**], and weather data [**?**].

Besides univariate time series models, modeling correlations in multiple time series has been an important objective because of its practical importance in finance, especially in portfolio selection and risk management [**?**, **?**]. In order to capture potential cross-correlations among different time series, models for coupled heteroskedastic time series have been introduced [**?**, **?**, **?**]. However, in practice, when those models are employed, the

number of parameters to be estimated can be quite large.

A number of researchers have applied multiple time series analysis to world indices, mainly in order to analyze zero time-lag cross-correlations. Reference [**?**] reported that for international stock return of nine highly-developed economies, the cross-correlations between each pair of stock returns fluctuate strongly with time, and increase in periods of high market volatility. By volatility we mean time-dependent standard deviation of return. The finding that there is a link between zero time lag cross-correlations and market volatility is "bad news" for global money managers who typically reduce their risk by diversifying stocks throughout the world. In order to determine whether financial crises are short-lived or long-lived, Ref. [**?**] recently reported that, for six Latin American markets, the effects of a financial crisis are short-range. Between two and four months after each crisis, each Latin American market returns to a low-volatility regime.

In order to determine whether financial crisis are short-term or long-term at the world level, we study 48 world indices, one for each of 48 different countries. We analyze cross-correlations among returns and magnitudes, for zero and non-zero time lags. We find that cross-correlations between magnitudes last substantially longer than between the returns, similar to the properties of auto-correlations in stock market returns [**?**]. We propose a general method in order to extract the most important factors controlling cross-correlations in time series. Based on random matrix theory [7, 8] and principal component analysis [**?**] we propose how to estimate the global factor and the most significant principal components in explaining the cross-correlations. This new method has a potential to be broadly applied in diverse phenomena where time series are measured, ranging from seismology to atmospheric geophysics.

This paper is organized as follows. In Section II we introduce the data analyzed, and the definition of return and magnitude of return. In Section III we introduce a new modified time lag random matrix theory (TLRMT) to show the time-lag cross-correlations between the returns and magnitudes of world indices. Empirical results show that the cross-correlations between magnitudes decays slower than that between returns. In Section IV we

introduce a single global factor model to explain the short- or long-range correlations among returns or magnitudes. The model relates the time-lag cross-correlations among individual indices with the auto-correlation function of the global factor. In Section V we estimate the global factor by minimizing the variance of residuals using principal component analysis (PCA), and we show that the global factor does in fact account for a large percentage of the total variance using RMT. In Section VI we show the applications of the global factor model, including risk forecasting of world economy, and finding countries who have most the independent economies.

4.2 Data Analyzed

In order to estimate the level of relationship between individual stock markets—either long-range or short-range cross-correlations exist at the world level—we analyze $N = 48$ world-wide financial indices, $S_{i,t}$, where $i = 1, 2, \ldots, 48$ denotes the financial index and t denotes the time. We analyze one index for each of 48 different countries: 25 European indices [**?**], 15 Asian indices (including Australia and New Zealand) [**?**], 2 American indices [**?**], and 4 African indices [**?**]. In studying 48 economies that include both developed and developing markets we significantly extend previous studies in which only developed economies were included—e.g., the seven economies analyzed in Refs. [**?**, **?**], and the 17 countries studied in Ref. [**?**]. We use daily stock-index data taken from *Bloomberg*, as opposed to weekly [**?**] or monthly data [**?**]. The data cover the period 4 Jan 1999 through 10 July 2009, 2745 trading days. For each index $S_{i,t}$, we define the relative index change (return) as

$$R_{i,t} \equiv \log S_{i,t} - \log S_{i,t-1}, \tag{4.1}$$

where t denotes the time, in the unit of one day. By magnitude of return we denote the absolute value of return after removing the mean

$$|r_{i,t}| \equiv |R_{i,t} - \langle R_{i,t} \rangle|. \tag{4.2}$$

4.3 Modified Time-lag Random Matrix Theory

4.3.1 Basic ideas of time-lag random matrix theory

In order to quantify the cross-correlations, random matrix theory (RMT) (see Refs. [**?**] [20] and references therein) was proposed in order to analyze collective phenomena in nuclear physics. Refs. [7, 8] extended RMT to cross-correlation matrices in order to find cross-correlations in collective behavior of financial time series. The largest eigenvalue λ_+ and smallest eigenvalue λ_- of the Wishart matrix W (a correlation matrix of uncorrelated time series with finite length) are

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}, \tag{4.3}$$

where $Q \equiv T/N(> 1)$, and N is the matrix dimension and T the length of each time series. The larger the discrepancy between (a) the correlation matrix C between empirical time series and (b) the Wishart matrix W obtained between uncorrelated time series, the stronger are the cross-correlations in empirical data [7, 8]. Many RMT studies reported equal-time (zero Δt) cross-correlations between different empirical time series [7, 8, **?**, 32, 35, 36].

Recently time-lag generalizations of RMT have been proposed [**?**, 39, **?**]. In one of the generalizations of RMT, based on the eigenvalue spectrum called time-lag RMT (TLRMT), Ref. [16] found long-range cross-correlations in time series of price fluctuations in absolute values of 1340 members of the New York Stock Exchange Composite, in both healthy and pathological physiological time series, and in the mouse genome.

We compute for varying time lags Δt the largest singular values $\lambda_L(\Delta t)$ of the cross-correlation matrix of N-variable time series $X_{i,t}$

$$C_{ij}(\Delta t) \equiv \frac{\langle X_{i,t} X_{j,t+\Delta t}\rangle - \langle X_{i,t}\rangle\langle X_{j,t+\Delta t}\rangle}{\sigma_i \sigma_j}. \tag{4.4}$$

We also compute $\tilde{\lambda}_L(\Delta t)$ of a similar matrix $\tilde{C}(\Delta t)$, where $X_{i,t}$ are replaced by the magnitudes $|X_{i,t}|$. The squares of the non-zero singular values of C are equal to the non-zero

eigenvalues of $\mathsf{C}\mathsf{C}^+$ or $\mathsf{C}^+\mathsf{C}$, where by C^+ we denote the transpose of C. In a singular value decomposition (SVD) [43, ?, 16] $\mathsf{C} = \mathsf{U}\mathsf{D}\mathsf{V}^+$ the diagonal elements of D are equal to singular values of C, where the U and V correspond to the left and right singular vectors of the corresponding singular values. We apply SVD to the correlation matrix for each time lag and calculate the singular values, and the dependence of the largest singular value $\lambda_L(\Delta t)$ on Δt serves to estimate the functional dependence of the collective behavior of C_{ij} on Δt [16].

4.3.2 Modifications of cross-correlation matrices

We make two modifications of correlation matrices in order to better describe correlations for both zero and non-zero time lags.

(i) The first modification is a correction for correlation between indices that are not frequently traded. Since different countries have different holidays, all indices contain a large number of zeros in their returns. These zeros lead us to underestimate the magnitude of the correlations. To correct for this problem, we define a modified cross-correlation between those time series with extraneous zeros,

$$C'_{ij}(\Delta t) \equiv \frac{1}{T'} \frac{\sum_{i=1}^{T} X_{i,t} X_{j,t+\Delta t} - \sum_{i=1}^{T} X_{i,t} \sum_{i=1}^{T} X_{j,t+\Delta t}}{\sigma_i \sigma_j}. \tag{4.5}$$

Here T' is the time period during which both $X_{i,t}$ and $X_{j,t+\Delta t}$ are non-zero. With this definition, the time periods during which $X_{i,t}$ or $X_{j,t+\Delta t}$ exhibit zero values have been removed from the calculation of cross-correlations. The relationship between $C'_{ij}(\Delta t)$ and $C_{ij}(\Delta t)$ is

$$C'_{ij}(\Delta t) = \frac{T}{T'} C_{ij}(\Delta t). \tag{4.6}$$

(ii) The second modification corrects for auto-correlations. The main diagonal elements in the correlation matrix are ones for zero-lag correlation matrices and auto-correlations for non-zero lag correlation matrices. Thus, time-lag correlation matrices allow us to

study both auto-correlations and time-lag cross-correlations. If we study the decay of the largest singular value, we see a long-range decay pattern if there are long-range auto-correlations for some indices but no cross-correlation between indices. To remove the influence of auto-correlations and isolate time-lag cross-correlations, we replace the main diagonals by unity,

$$C''_{ij}(\Delta t) = 1 when$$

i=j$C'_{ij}(\Delta t) when$ i$\neq j$. (4.7)

With this definition the influence of auto-correlations is removed, and the trace is kept the same as the zero time-lag correlation matrix.

4.3.3 Empirical results

In Fig. 1(a) we show the distribution of cross-correlations between zero and non-zero lags. For $\Delta t = 0$ the empirical pdf $P(C_{ij})$ of the cross-correlation coefficients C_{ij} substantially deviates from the corresponding pdf $P(W_{ij})$ of a Wishart matrix, implying the existence of equal-time cross-correlations.

In order to determine whether short-range or long-range cross-correlations accurately characterize world financial markets, we next analyze cross-correlations for ($\Delta t \neq 0$). We find that with increasing Δt the form of $P(C_{ij})$ quickly approaches the pdf $P(W_{ij})$, which is normally distributed with zero mean and standard deviation $1/\sqrt{N}$ [**?**].

In Fig. 1(b) we also show the distribution of cross-correlations between *magnitudes*. In financial data, returns $R_{i,t}$ are generally uncorrelated or short-range auto-correlated, whereas the magnitudes are generally long-range auto-correlated [**?**, **?**]. We thus examine the cross-correlations $\tilde{C}_{ij}(\Delta t)$ between $|r_{i,t}|$ for different Δt. In Fig. 1(b) we find that with increasing Δt, $P(\tilde{C}_{ij})$ approaches the pdf of random matrix $P(W_{ij})$ more slowly than $P(C_{ij})$, implying that cross-correlations between index magnitudes persist longer than

cross-correlations between index returns.

In order to demonstrate the decay of cross-correlations with time lags, we apply modified TLRMT. Fig. 2 shows that with increasing Δt the largest singular value calculated for $\tilde{\mathsf{C}}$ decays more slowly than the largest singular value calculated for C. This result implies that among world indices, the cross-correlations between magnitudes last longer than cross-correlations between returns. In Fig. 2 we find that λ_L vs. Δt decays as a power law function with the scaling exponent equal to 0.25. The faster decay of λ_L vs. Δt for C implies very weak (or zero) cross-correlations among world-index returns for larger Δt, which agrees with the empirical finding that world indices are often uncorrelated in returns. Our findings of long-range cross-correlations in magnitudes among the world indices is, besides a finding in Ref. [**?**], another piece of "bad news" for international investment managers. World market risk decays very slowly. Once the volatility (risk) is transmitted across the world, the risk lasts a long time.

4.4 Global Factor Model

The arbitrage pricing theory states that asset returns follow a linear combination of various factors [**?**]. We find that the factor structure can also model time lag pairwise cross-correlations between the returns and between magnitudes. To simplify the structure, we model the time lag cross-correlations with the assumption that each individual index fluctuates in response to one common process, the "global factor" M_t,

$$R_{i,t} = \mu_i + b_i M_t + \epsilon_{i,t}. \tag{4.8}$$

Here in the global factor model (GFM), μ_i is the average return for index i, M_t is the global factor, and $\epsilon_{i,t}$ is the linear regression residual, which is independent of M_t, with mean zero and standard deviation σ_i. Here b_i indicates the covariance between $R_{i,t}$ and M_t, $\mathrm{Cov}(R_{i,t}, M_t) = b_i \mathrm{Var}(M_t)$. This single factor model is similar to the Sharpe market model [**?**], but instead of using a known financial index as the global factor M_t, we use

factor analysis to find M_t, which we introduce in the next section. We also choose M_t as a zero-mean process, so the expected return $E(R_{i,t}) = \mu_i$, and the global factor M_t is only related with market risk. We define a zero-mean process $r_{i,t}$ as

$$r_{i,t} \equiv R_{i,t} - E(R_{i,t}) = b_i M_t + \epsilon_{i,t}. \tag{4.9}$$

A second assumption is that the global factor can account for most of the correlations. Therefore we can assume that there are no correlations between the residuals of each index, $\mathrm{Cov}(\epsilon_{i,t}, \epsilon_{j,t}) = 0$. Then the covariance between $R_{i,t}$ and $R_{j,t}$ is

$$\mathrm{Cov}(R_{i,t}, R_{j,t}) = \mathrm{Cov}(r_{i,t}, r_{j,t}) = b_i b_j \mathrm{Var}(M_t). \tag{4.10}$$

The covariance between magnitudes of returns depends on the return distribution of M_t and $R_{i,t}$, but the covariance between squared magnitudes $r_{i,t}^2$ indicates the properties of the magnitude cross-correlations. The covariance between $r_{i,t}^2$ and $r_{j,t}^2$ is

$$\mathrm{Cov}(r_{i,t}^2, r_{j,t}^2) = b_i^2 b_j^2 \mathrm{Var}(M_t^2). \tag{4.11}$$

The above results in Eqs. (4.19)-(4.20) show that the variance of the global factor and square of the global factor account for all the zero time lag covariance between returns and squared magnitudes. For time lag covariance between $r_{i,t}$, we find

$$\begin{aligned} \mathrm{Cov}(r_{i,t}, r_{j,t}, \Delta t) &= E(r_{i,t}, r_{j,t-\Delta t}) - E(r_{i,t})E(r_{j,t-\Delta t}) & (4.12) \\ &= b_i b_j A_M(\Delta t). & (4.13) \end{aligned}$$

Here

$$A_M(\Delta t) \equiv E(M_t M_{t-\Delta t}) - E(M_t)E(M_{t-\Delta t}) \tag{4.14}$$

is the autocovariance of M_t. Similarly, we find

$$\mathrm{Cov}(r_{i,t}^2, r_{j,t}^2, \Delta t) = b_i^2 b_j^2 A_{M^2}(\Delta t). \tag{4.15}$$

Here

$$A_{M^2}(\Delta t) = E(M_t^2 M_{t-\Delta t}^2) - E(M_t^2)E(M_{t-\Delta t}^2) \tag{4.16}$$

is the autocovariance of M_t^2.

In GFM, the time lag covariance between each pair of indices is proportional to the autocovariance of the global factor. For example, if there is short-range autocovariance for M_t and long-range autocovariance for M_t^2, then for individual indices the cross-covariance between returns will be short-range and the cross-covariance between magnitudes will be long-range. Therefore, the properties of time-lag cross-correlation in multiple time series can be modeled with a single time series— the global factor M_t.

The relationship between time lag covariance among two index returns and autocovariance of the global factor also holds for the relationship between time lag cross-correlations among two index returns and auto-correlation function of the global factor, because it only need to normalize the original time series to mean zero and standard deviation one.

4.5 Estimation and Analysis of the Global Factor

4.5.1 Estimation of the global factor

The arbitrage pricing theory states that asset returns follow a linear combination of various factors [**?**]. We find that the factor structure can also model time lag pairwise cross-correlations between the returns and between magnitudes. To simplify the structure, we model the time lag cross-correlations with the assumption that each individual index fluctuates in response to one common process, the "global factor" M_t,

$$R_{i,t} = \mu_i + b_i M_t + \epsilon_{i,t}. \tag{4.17}$$

Here in the global factor model (GFM), μ_i is the average return for index i, M_t is the global factor, and $\epsilon_{i,t}$ is the linear regression residual, which is independent of M_t, with mean zero and standard deviation σ_i. Here b_i indicates the covariance between $R_{i,t}$ and M_t, $\mathrm{Cov}(R_{i,t}, M_t) = b_i \mathrm{Var}(M_t)$. This single factor model is similar to the Sharpe market model [**?**], but instead of using a known financial index as the global factor M_t, we use factor analysis to find M_t, which we introduce in the next section. We also choose M_t as a zero-mean process, so the expected return $E(R_{i,t}) = \mu_i$, and the global factor M_t is only related with market risk. We define a zero-mean process $r_{i,t}$ as

$$r_{i,t} \equiv R_{i,t} - E(R_{i,t}) = b_i M_t + \epsilon_{i,t}. \tag{4.18}$$

A second assumption is that the global factor can account for most of the correlations. Therefore we can assume that there are no correlations between the residuals of each index, $\mathrm{Cov}(\epsilon_{i,t}, \epsilon_{j,t}) = 0$. Then the covariance between $R_{i,t}$ and $R_{j,t}$ is

$$\mathrm{Cov}(R_{i,t}, R_{j,t}) = \mathrm{Cov}(r_{i,t}, r_{j,t}) = b_i b_j \mathrm{Var}(M_t). \tag{4.19}$$

The covariance between magnitudes of returns depends on the return distribution of M_t and $R_{i,t}$, but the covariance between squared magnitudes $r_{i,t}^2$ indicates the properties of the magnitude cross-correlations. The covariance between $r_{i,t}^2$ and $r_{j,t}^2$ is

$$\mathrm{Cov}(r_{i,t}^2, r_{j,t}^2) = b_i^2 b_j^2 \mathrm{Var}(M_t^2). \tag{4.20}$$

The above results in Eqs. (4.19)-(4.20) show that the variance of the global factor and square of the global factor account for all the zero time lag covariance between returns and squared magnitudes. For time lag covariance between $r_{i,t}$, we find

$$\begin{aligned} \mathrm{Cov}(r_{i,t}, r_{j,t}, \Delta t) &= E(r_{i,t}, r_{j,t-\Delta t}) - E(r_{i,t}) E(r_{j,t-\Delta t}) && (4.21) \\ &= b_i b_j A_M(\Delta t). && (4.22) \end{aligned}$$

Here

$$A_M(\Delta t) \equiv E(M_t M_{t-\Delta t}) - E(M_t)E(M_{t-\Delta t}) \tag{4.23}$$

is the autocovariance of M_t. Similarly, we find

$$\mathrm{Cov}(r_{i,t}^2, r_{j,t}^2, \Delta t) = b_i^2 b_j^2 A_{M^2}(\Delta t). \tag{4.24}$$

Here

$$A_{M^2}(\Delta t) = E(M_t^2 M_{t-\Delta t}^2) - E(M_t^2)E(M_{t-\Delta t}^2) \tag{4.25}$$

is the autocovariance of M_t^2.

In GFM, the time lag covariance between each pair of indices is proportional to the autocovariance of the global factor. For example, if there is short-range autocovariance for M_t and long-range autocovariance for M_t^2, then for individual indices the cross-covariance between returns will be short-range and the cross-covariance between magnitudes will be long-range. Therefore, the properties of time-lag cross-correlation in multiple time series can be modeled with a single time series— the global factor M_t.

The relationship between time lag covariance among two index returns and autocovariance of the global factor also holds for the relationship between time lag cross-correlations among two index returns and auto-correlation function of the global factor, because it only need to normalize the original time series to mean zero and standard deviation one.

4.5.2 Analysis of the global factor

Next we apply the method of Eq. (**??**) to estimate the global factor of 48 world index returns. We calculate the auto-correlations of M_t and $|M_t|$, which are shown in Figs. 3 and 4. Precisely, for the world indices, Fig. 3(a) shows the time series of the global factor M_t, and Fig. 3(b) shows the auto-correlations in M_t. We find only short-range auto-correlations because, after an interval $\Delta t = 2$, most auto-correlations in M_t fall in the range of $(-1.96\sqrt{1/T}, 1.96\sqrt{1/T})$ [**?**], which is the 95% confidence interval for zero

auto-correlations, Here $T = 2744$.

For the 48 world index returns, Fig. 4(a) shows the time series of magnitudes $|M_t|$, with few clusters related to market shocks during which the market becomes fluctuates more. Fig. 4(b) shows that, in contrast to M_t, the magnitudes $|M_t|$ exhibit long-range auto-correlations since the values $|M_t|$ are significant even after $\Delta t = 100$. The auto-correlation properties of the global factor are the same as the auto-correlation properties of the individual indices, i.e., there are short-range auto-correlations in M_t and long-range power-law auto-correlations in $|M_t|$ [**?**, **?**]. These results are also in agreement with Fig. 1(b) where the largest singular value λ_L vs. Δt calculated for $\tilde{\mathsf{C}}$ decays more slowly than the largest singular value calculated for C. As found in Ref. [16] for $\Delta t >> 1$, $\lambda_L(\Delta t)$ approximately follows the same decay pattern as cross-correlation functions. Although a Ljung-Box test shows that the return auto-correlation is significant for a 95% confidence level [**?**], the return auto-correlation is only 0.132 for $\Delta t = 1$ and becomes insignificant after $\Delta t = 2$. Therefore, for simplicity, we only consider magnitude cross-correlations in modeling the global factor.

We model the long-range market-factor returns $\mathbf{M}$ with a particular version of the GARCH process, the GJR GARCH process [**?**], because this GARCH version explains well the asymmetry in volatilities found in many world indices [**?**, **?**, **?**]. The GJR GARCH model can be written as

$$\epsilon_t = \sigma_t \eta_t, \tag{4.26}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} (\alpha_i + \gamma T_{t-i}) \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2, \tag{4.27}$$

where σ_t is the volatility and η_t is a random process with a Gaussian distribution with standard deviation 1 and mean 0. The coefficients α and β are determined by a maximum likelihood estimation (MLE) and $T_t = 1$ if $\epsilon_{t-1} < 0$, $T_t = 0$ if $\epsilon_{t-1} \geq 0$. We expect the parameter γ to be positive, implying that "bad news" (negative increments) increases volatility more than "good news". For the sake of simplicity, we follow the usual procedure

of setting $p = q = 1$ in all numerical simulations. In this case, the GJR-GARCH(1,1) model for the market factor can be written as

$$M_t = \sigma_t \eta_t, \tag{4.28}$$

$$\sigma_t^2 = \alpha_0 + (\alpha_1 + \gamma T_{t-1})\epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2. \tag{4.29}$$

We estimate the coefficients in the above equations using MLE, where the estimated coefficients are shown in Table. 1.

Next we test the hypothesis that a significant percentage of the world cross-correlations can be explained by the global factor. By using PCA we find that the global factor can account for 30.75% of the total variance. Note that, according to RMT, only the eigenvalues larger than the largest eigenvalue of a Wishart matrix calculated by Eq. (4.3) (and the corresponding $\boldsymbol{\alpha}$s) are significant. To calculate the percentage of variance the significant $\boldsymbol{\alpha}$s account for, we employ the RMT approach proposed in Ref. [7, 8]. The largest eigenvalue for a Wishart matrix is $\lambda_+ = 1.282$ for $N = 48$ and $T = 2744$ we found in the empirical data. From all the 48 eigenvalues, only the first three are significant: $\lambda_1 = 14.762$, $\lambda_2 = 3.453$, and $\lambda_3 = 1.380$. This result implies that among the significant factors, the global factor accounts for $\lambda_1 / \sum_{i=1}^{3} \lambda_i = 75.34\%$ of the variance, confirming our hypothesis that the global factor accounts for most variance of all individual index returns.

PCA is defined to estimate the percentage of variance the global factor can account for zero time lag correlations. Next we study the time lag cross-correlations after removing the global trend, and apply the SVD to the correlation matrix of regression residuals η_i of each index [see Eq. (4.17)]. Our results show that for both returns and magnitudes, the remaining cross-correlations are very small for all time lags compared to cross-correlations obtained for the original time series. This result additionally confirms that a large fraction of the world cross-correlations for both returns and magnitudes can be explained by the global factor.

4.6 Applications of Global Factor Model

4.6.1 Locating and forecasting global risks

The asymptotic (unconditional) variance for the GJR-GARCH model is $\alpha_0/(1-\alpha_1-\beta_1-\gamma/2) = 10.190$ [**?**]. For the market factor the conditional volatility σ_t can be estimated by recursion using the historical conditional volatilities and fitted coefficients in Eq. (4.29). For example, the largest cluster at the end of the graph shows the 2008 financial crisis. In Fig. 5(a) we show the time series of the conditional volatility of Eq. (4.29) of the global factor. The clusters in the conditional volatilities may serve to predict market crashes. In each cluster, the height is a measure of the size of the market crash, and the width indicates its duration. In Fig. 5(b) we show the forecasting of the conditional volatility of the global factor, which asymptotically converges to the unconditional volatility.

4.6.2 Finding uncorrelated individual indices

Next, in Fig. **??** we show the cross-correlations between the global factor and each individual index using Eq. (**??**). There are indices for which cross-correlations with the global factor are very small compared to the other indices; 10 of 48 indices have cross-correlations coefficients with the global factor smaller than 0.1. These indices correspond to Iceland, Malta, Nigeria, Kenya, Israel, Oman, Qatar, Pakistan, Sri Lanka, and Mongolia. The financial market of each of these countries is weakly bond with financial markets of other countries. This is useful information for investment managers because one can reduce the risk by investing in these countries during world market crashes which, seems, do not severely influence these countries.

4.6.3 Multi-Factor Model

4.7 Discussion

We have developed a modified time lag random matrix theory (TLRMT) in order to quantify the time-lag cross-correlations among multiple time series. Applying the modified TLRMT to the daily data for 48 world-wide financial indices, we find short-range cross-correlations between the returns, and long-range cross-correlations between their magnitudes. The magnitude cross-correlations show a power law decay with time lag, and the scaling exponent is 0.25. The result we obtain, that at the world level the cross-correlations between the magnitudes are long-range, is potentially significant because it implies that strong market crashes introduced at one place have an extended duration elsewhere—which is "bad news" for international investment managers who imagine that diversification across countries reduces risk.

We model long-range world-index cross-correlations by introducing a global factor model in which the time lag cross-correlations between returns (magnitudes) can be explained by the auto-correlations of the returns (magnitudes) of the global factor. We estimate the global factor as the first component by using principal component analysis. Using random matrix theory, we find that only three principal components are significant in explaining the cross-correlations. The global factor accounts for 30.75% of the total variance of all index returns, and 75.34% of the variance of the three significant principle components. Therefore, in most cases, a single global factor is sufficient.

We also show the applications of the GFM, including locating and forecasting world risk, and finding individual indices that are weakly correlated to the world economy. Locating and forecasting world risk can be realized by fitting the global factor using a GJR-GARCH(1,1) model, which explains both the volatility correlations and the asymmetry in the volatility response to both "good news" and "bad news." The conditional volatilities calculated after fitting the GJR-GARCH(1,1) model indicates the global risk, and the risk can be forecasted by recursion using the historical conditional volatilities and the fitted

coefficients. To find the indices that are weakly correlated to the world economy, we calculate the correlation between the global factor and each individual index. We find 10 indices which have a correlation smaller than 0.1, while most indices are strongly correlated to the global factor with the correlations larger than 0.3. To reduce risk, investment managers can increase the proportion of investment in these countries during world market crashes, which do not severely influence these countries.

Based on principal component analysis, we propose a general method which helps extract the most significant components in explaining long-range cross-correlations. This makes the method suitable for broad range of phenomena where time series are measured, ranging from seismology and physiology to atmospheric geophysics. We expect that the cross-correlations in EEG signals are dominated by the small number of most significant components controlling the cross-correlations. We speculate that cross-correlations in earthquake data are also controlled by some major components. Thus the method may have significant predictive and diagnostic power that could prove useful in a wide range of scientific fields.

Chapter 5

Modeling Heteroscedasticity

Chapter 6

Application of Heteroscedasticity

6.1 Introduction

The study of price dynamics is the study of price changes [**?**, **?**, **?**, **?**]. Empirical evidence indicates that extremely complex trading activities affect price changes. In one of the first attempts to model this activity, Ref. [**?**] uses a discrete stochastic process t_i to represent times at which trading occurs. Upon this stochastic process, a new stochastic process $X(t_i)$ is defined representing, for example, a stock price at time t_i. The process t_i is said to be subordinated to $X(t_i)$. Clearly, how fast prices respond to trades occurring at t_i determines market liquidity, and liquidity is related to the ease with which securities are bought and sold without substantial price changes. To point out the importance of subordinate stochastic processes, Ref. [**?**] recently proposed a subordinate stochastic process for the model of proportional growth.

There are two main approaches to model price dynamics: the stochastic approach [**?**, **?**, **?**] and an agent-based approach [**?**, **?**, **?**, **?**, **?**]. These two approaches we can understand, e.g., by comparing modeling long-range correlations in price changes, ΔS_t. In the stochastic approach one models these correlations by assuming that ΔS_t depends on its previous values $\Delta S_t \equiv \sum_i a_i \ \Delta S_{t-i}$. The choice for statistical weights a_i determines, first, whether we want long- or short-range dependence in the autocorrelations of ΔS_t, and second, which functional dependence we want to obtain for the autocorrelation function. The agent-based approach models security market microstructure starting from different traders (agents) and defining the trading rules among the agents which, for instance, finally may yield long-

range correlations in price changes. Several papers propose models for artificial markets populated with heterogeneous agents endowed with learning and optimization capabilities [**?**, **?**, **?**, **?**, **?**, **?**].

Here we combine stochastic and agent-based approaches to create a hybrid price dynamics model to simulate empirical evidence reported in bid-ask spread, stock price autocorrelations, and trading volume. We partially follow the subordinated stochastic process proposed in Ref. [**?**]. First we define a process for trading times t_i. When trading occurs, a package of stocks (volume) denoted by q_i changes owner. Thus, in terms of the Clark process [**?**], in our model t_i is subordinated to the number of shares traded $q(t_i)$. However, in contrast to Ref. [**?**], we define both t_i and $q(t_i)$ as long-range correlated processes. When trading occurs it triggers the price to change. In our model a co-movement between intertrade time, defined as $\Delta t \equiv t_i - t_{i-1}$, and volatility, defined as absolute value of a price change, exists because the process controlling Δt also controls the bid-ask spread (the difference between ask and bid). The model generates power-law autocorrelations in absolute returns [**?**, **?**] and power-law tails in distributions of returns [**?**, **?**]. It also yields a log-linear functional relationship between the average bid-ask spread $\langle S \rangle_{\Delta T}$ and the number of trades $N_{\Delta T}$, and between $\langle S \rangle_{\Delta T}$ and the share volume traded $Q_{\Delta T}$.

6.2 Empirical Evidence

When ink particles diffuse in water, the collision of each ink particle with numerous water molecules causes it to move in a random walk pattern [**?**, **?**]. The distance covered by the particle after a time ΔT is $X_{\Delta T} = \sum_{i=1}^{N_{\Delta T}} \Delta x_i$ where $X_{\Delta T}$ is Gaussian distributed and short-range correlated, $N_{\Delta T}$ denotes the number of collisions during the interval ΔT, and Δx_i is the change of position of the ink particle after collision. A more complex variation of the classic diffusion problem exists in finance, with intertrade times—which are the time intervals between two consecutive trades in the market. First, intertrade times are not Gaussian uncorrelated, but are power-law correlated variables [**?**]. Second, financial

markets are characterized by many complex hierarchies among different processes, and the number of trading times is only one variable among others such as the number of shares traded and the share price. The hierarchy is roughly the following: the trading time marks the initiation of the trade, and then a trade triggers the price to change. This implies that in explaining market activities, we must consider not an univariate model, but rather a multivariate model where different time series are subordinated and frequently power-law auto-correlated.

(i) *Empirical evidence in bid-ask spread.* The ability to buy at a low price and sell at a high price is the main compensation to traders for the risk they incur [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. The trader sells at the "ask" (offer) price A and buys at a lower "bid" price B, where the difference is the bid-ask spread. Ref. [**?**] identifies four indicators which determine bid-ask spreads: activity, risk, information, and competition. More specifically,

   (a) greater trading activity (shorter trading times) can lead to lower spreads since the higher the level of trading, the greater the chance that buy and sell orders will tend to balance during a trading period;

   (b) there is a direct relationship between the level of risk and spreads;

   (c) there is a direct relationship between spreads and the amount of information coming to the market—large trades convey more information than small trades; and

   (d) There is an inverse relationship between spreads and the level of competition.

   Competition varies with volume—the number of traders is more active as volume levels increase. In addition, analyzing NYSE stocks, Ref. [**?**] shows that the mean of (ask — bid)/(ask + bid)/2 for each minute of the trading day shows that spreads are relatively high at minute three, decline at a decreasing rate until minute 293 and then increase at an increasing rate until the close of trading. Thus, the plot of

spreads over the trading day exhibits a crude reverse J-shaped pattern. By studying the bid-ask quotations and transactions information during 1988, Ref. [**?**] finds that spreads are negatively associated with the number of exchange listings, share price, and firm size. Different models are proposed to explain bid-ask spread properties [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**].

(ii) *Empirical evidence in stock price correlations.* Analyzing daily recorded SP500 financial index, Ref. [**?**] reports a power-law long memory in auto-correlations of absolute returns. Ref. [**?**] reports power-law cross-correlations of absolute returns between 1,340 members of NYSE. By analyzing the high-frequency S&P500 index and individual U.S. firms, Ref. [**?**] finds a crossover in correlations of absolute returns between two power-law regimes at approximately 1.5 days. Analysis accomplished on the time series of time intervals between consecutive stock trades S&P500 of different US firms revealed the same crossover between power-law regimes, implying a parallel with the crossover in the scaling of absolute price returns [**?**]. Ref. [**?**] reports a Weibull distribution in IBM intertrade times.

(iii) *Empirical evidence in trading volume.* By analyzing a database documenting every transaction for 1000 U.S. stocks for the two-year period 1994–1995, Ref. [44] quantifies the relation between trading activity measured by the number of transactions $N_{\Delta t}$ and the price change $G_{\Delta t}$ for a given stock, over a time interval $[t, \Delta t]$. Denoting by $W^2_{\Delta t}$ the variance of the price changes for all transactions in Δt, it was found that the power-law tails of $P(G_{\Delta t})$ are due to $P(W_{\Delta t})$ and the long-range correlations in $|G_{\Delta t}|$ are due to $N_{\Delta t}$. For the 1000 stocks analyzed, the cumulative distribution of $N_{\Delta t}$ displays a power-law behavior with a mean value 3.40 ± 0.2, close to the exponent of the cubic law found in the tails of $P(G_{\Delta t})$ [**?**]. For the number of shares traded $Q_{\Delta t}$, the distribution $P(Q_{\Delta t})$ displays a power-law decay $P(Q_{\Delta t}) \propto (Q_{\Delta t})^{-1-\alpha}$, where $\alpha = 1.7 \pm 0.1$ [**?**]. Also, the long-range correlations in $Q_{\Delta t}$ are largely due to those of $N_{\Delta t}$. The results are consistent with the interpretation that the large equal-time

correlations between $Q_{\Delta t}$ and the absolute value of price change $|G_{\Delta t}|$ are largely due to $N_{\Delta t}$. However, expressing $Q_{\Delta t}$ as the sum of the number of shares traded for all transactions, $Q_{\Delta t} = \sum q_i$, Ref. [**?**, **?**] reports only weak correlations in q_i. Recently, based on detrending cross-correlations analysis of Ref. [**?**], Ref. [**?**] reports long-range cross-correlations between volatility and the absolute values of volume changes. It also reports the existence of a cubic law in trading volume changes, supporting the intriguing possibility that the cubic law in price changes has its origin in trading activities.

6.3 Model

Our goal is to construct a common framework for modeling trading time, trading volume, and price changes. To test our model, we select Exxon, a stock typical of the U.S. market and, according to the Trades and Quotes database (NYSE, New York, 1993), one of the most traded U.S. companies during the four-year period January 1993 – December 1996. Our model is comprised of three stages: (i) we stochastically generate the duration or intertrade times (the interval between two trading times) Δt_i; (ii) at each Δt_i we stochastically generate the number of shares traded $q(\Delta t_i)$; and (iii) we propose a mechanism that explains how both Δt_i and $q(\Delta t_i)$ affect price change.

(i) We first define trading at times indexed by a set of numbers $t_1, t_2, t_3, ...$ These numbers are a realization of a discrete stochastic process with positive increments (since $t_i \geq 0$) implying that $t_1 < t_2 < t_3....$ In order to reproduce long-range power-law correlations in Δt_i as found for the three-year period January 1993-December 1996 [**?**], we model Δt_i using a fractionally integrated autoregressive conditional duration (FIACD) [**?**, **?**],

$$\Delta t_i = \psi_i(\rho_1)\epsilon_i, \tag{6.1}$$

where ϵ_i is independent and identically distributed (*i.i.d.*) with an exponential probability distribution $(a_1 \exp(-a_1\epsilon))$ (i.e., with one free parameter a_1) that is an approxi-

mation of the Weibull distribution found for U.S. firms [**?**, **?**], $\psi_i(\rho_1)$ is the expectation of duration i [**?**], and Δt_i at each moment t_i depends only on its previous values. The time series $\{\Delta t_i\}$ of Eq. (6.1) in Fig. (6.1)(a) is generated using the fractional parameter $\rho_1 = 0.4$ (from $(1-L)^{\rho_1}$ see [**?**]) which is used to reproduce the power-law scaling in Δt [Fig. 6.1(b)]. To quantify the power-law memory, we use detrended fluctuation analysis (DFA) [**?**]. The fractional parameter $\rho_1 = 0.4$ corresponds to the DFA exponent $\alpha = 0.9$ found for the Exxon company for the three-year period [**?**]. The free parameter a_1 of the (*i.i.d.*) exponential ($a_1 \exp(-a_1\epsilon)$) in Eq. (6.1), can be estimated from the average intertrade times. When a trade occurs at t_i, a number of shares $q(\Delta t_i)$ changes ownership.

(ii) We next model a process for the time series $q(\Delta t_i)$ for the same three-year period. For $q(\Delta t_i)$ of Exxon company trades, we obtain the DFA exponent $\alpha = 0.62$, which implies the presence of weak but long-range power-law correlations. We assume that $q(\Delta t_i)$ depends not on previous Δt values, but on previous q values. Motivated by Clark's subordinated process [**?**], we assume that Δt_i is subordinate not to share price as in Ref. [**?**], but to $q(\Delta t_i)$, and model $q(\Delta t_i)$ using a fractionally integrated moving average process (FIARCH) [**?**],

$$q(\Delta t_i) = \sigma_i(\rho_2)\ \epsilon_i'. \tag{6.2}$$

Here $\sigma_i = \sum_{n=1}^{\infty} a_n(\rho_2) q(\Delta t_{i-n})$, $a_n\Gamma(n-\rho_2)/[\Gamma(-\rho_2)\Gamma(1+n)]$ are statistical weights where Γ denotes the Gamma function, $\rho_2 \in (0, 0.5)$ is a single free parameter [**?**], and ϵ_i' is an *i.i.d.*, for simplicity taken from an exponential distribution $a_2 \exp(-a_2\epsilon')$, the parameter of which (a_2) can be estimated to give the average number of shares traded. Since there is a simple relation between the DFA exponent α and the FIARCH parameter ρ_2—$\alpha = 0.5 + \rho_2$—from $\alpha = 0.62$ calculated for power-law correlations in q_i we obtain $\rho_2 = 0.12$.

Thus we model Δt_i and $q(\Delta t_i)$ as two mutually independent but individually auto-

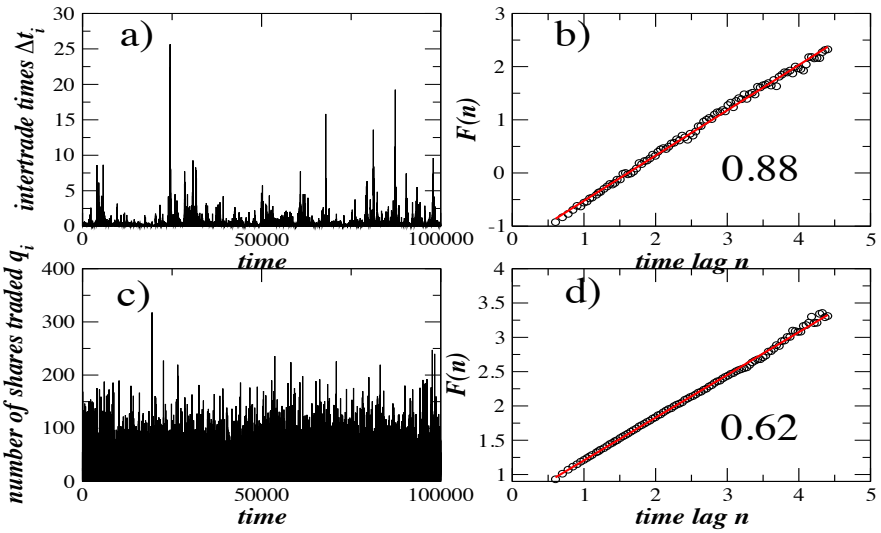


Figure 6.1: Modeling power-law correlations in intertrade time and number of shares traded found in Exon company using the stochastic process of Eq. (6.1) with fractional parameter $\rho_1 = 0.4$ and the stochastic process of Eq. (6.2) with fractional parameter $\rho_1 = 0.12$ (a) Intertrade time Δt of Eq. (6.1). (b) For Δt, detrended fluctuation function $F(n)$ versus time lag n yields strong long-range auto-correlations in Δt. (c) Number of shares traded $q(\Delta t_i)$ of Eq. (6.2). (d) For $q(\Delta t_i)$, we find weak long-range power-law auto-correlations.

correlated power-law processes in which the correlations in Δt_i are much stronger than those in $q(\Delta t_i)$. There are four parameters at this stage: ρ_1 and ρ_2 responsible for power-law scaling in intertrade times Δt_i, the number of shares $q(\Delta t_i)$, and two parameters a_1 and a_2 corresponding to the distributions of *i.i.d.* variables in Eqs. (6.1)-(6.2).

Figures 2(a) and 2(b) show that in Eqs. (6.1) and (6.2) these two power-law scalings are responsible for the strong power-law correlations in the sum of the number of shares q_i traded (the trading volume) in a fixed time interval ΔT (where $\Delta T \gg \langle \Delta t \rangle$),

$$Q(\Delta T) = \sum_{i=1}^{N_{\Delta T}} q_i(\Delta t_i), \tag{6.3}$$

where $N_{\Delta T}$ is the total number of trades within a time interval $\Delta T = \sum_{i=1}^{N_{\Delta T}} \Delta t$. Thus, even though the time series of the individual number of shares traded $q(\Delta t_i)$ is weakly power-law correlated, because of strong power-law correlations in the intertrade time Δt_i the integrated trading volume $Q(\Delta T)$ exhibits strong long-range power-law correlations, which were found empirically by Gopikrishnan et al in Ref. [**?**]. Figures 2(c) and 2(d) show that Eqs. (6.1) and (6.2) also generate long-range power-law auto-correlations in the total number of shares traded in the fixed time interval ΔT, where $\Delta T \gg \langle \Delta t \rangle$.

Trading strategies play a key role in price dynamics, and the literature on this topic is huge. Ref. [**?**] models trading activities by assuming that at the beginning of a trading day traders are greeted with news that is either good or bad, and that long durations are likely to be associated with news that is bad. Ref. [**?**] assumes that informed traders possess non-public information that allows them to better estimate a future security price than uninformed traders. Ref. [**?**] assumes that informed traders trade only when they have information and thus variations in trading rates are associated with the changing number of informed traders. In the model proposed by Bak et

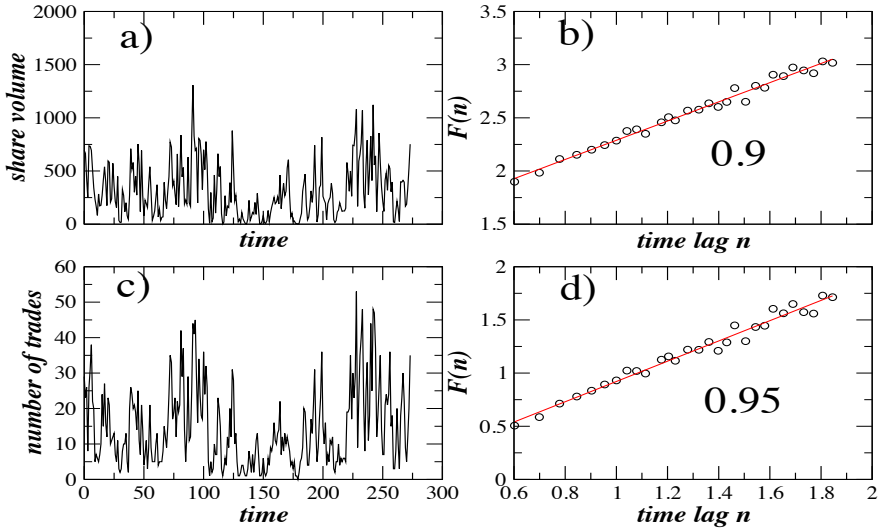


Figure 6.2: Modeling power-law correlations in shares volume $Q(\Delta T)$ and number of trades (transactions) $N_{\Delta T}$ within ΔT using the stochastic process of Eq. (6.1) with fractional parameter $\rho_1 = 0.4$ and the stochastic process of Eq. (6.2) with fractional parameter $\rho_2 = 0.12$ (a) Shares volume $Q(\Delta t)$. (b) For $Q(\Delta t)$, detrended fluctuation function $F(n)$ versus time lag n yields strong long-range power-law auto-correlations in Δt. (c) Number of trades $N_{\Delta T}$. For $N_{\Delta T}$, detrended fluctuation function $F(n)$ versus time lag n yields strong long-range power-law auto-correlations.

al [**?**], buyers and sellers are represented by particles subject to a reaction-diffusion process [**?**]. In Maslov model in Ref. [**?**], traders can either buy or sell stock at the market price or place a limit order to automatically buy or sell a particular amount of stock. In this, traders are allowed to trade only one unit of stock ($q_i = 1$) in each transaction. A mean-field variant of the Ref. [**?**] model proposed by Slanina is found to exhibit a power-law tail with exponent 2 [**?**]. Other models with nontrivial agent strategies have also been proposed [**?**, **?**, **?**, **?**].

We now simplify the trading process, but at a level that can still provide us with the scaling properties found in price and trading dynamics. In this model, bid and ask prices are stochastically generated at each time coordinate. We do this because Gopikrishnan et al in Ref. [**?**] find that the correlations in the absolute values of

price changes are largely due to correlations in trading volume. Engle and Russell Ref. [**?**] reports evidence of co-movements between intertrade time and volatility—the absolute value of price changes. Similarly, Ivanov et al in Ref. [**?**] quantify this co-movement finding an analogy in the power-law scaling between the absolute value of price changes and the time intervals between consecutive stock trades. Finally, for the 116 stocks analyzed, Plerou et al in Ref. [**?**] report that the average bid-ask spread S is characterized by a cumulative distribution that decays as a cubic power law. These results clearly suggest a common origin for price change dynamics and trading time dynamics. In our model, at each trading time a single trader trades stock while other traders put either bid or ask prices. We therefore suggest the following process for generating the trader's (agents') ask and bid price changes, respectively,

$$\Delta S^a = \psi_i(\rho_1)\epsilon_i'', \tag{6.4}$$

$$\Delta S^b = -\psi_i(\rho_1)\epsilon_i'', \tag{6.5}$$

where $\psi_i(d)$—the volatility process shown in Eq. (6.1)—is responsible for the long memory in intertrade times, and ϵ_i'' is from an exponential function. Thus in our model intertrade times and bid-ask price changes share the same volatility mechanism. In our simulations we keep the number of bid and ask traders equal and constant. Clearly this is an approximation, since the number of bid and ask traders changes over time and at certain times, e.g., during market crashes, substantially increases.

(iii) To illustrate how trading influences price changes, consider a simple example with only two ask traders. Suppose trader A puts an ask order with 3000 shares and requires that its price be at least \$100 per share. Trader B puts an ask order with 6000 shares and requires that the price exceed \$110 per share. Trader C decides to buy the cheapest 6000 shares. Clearly, trader C can buy 3000 shares from trader A at \$100 per share and 3000 shares from trader B at \$110 per share. We assume that for the trader who trades shares, the probability of a bid offer is equal to the probability

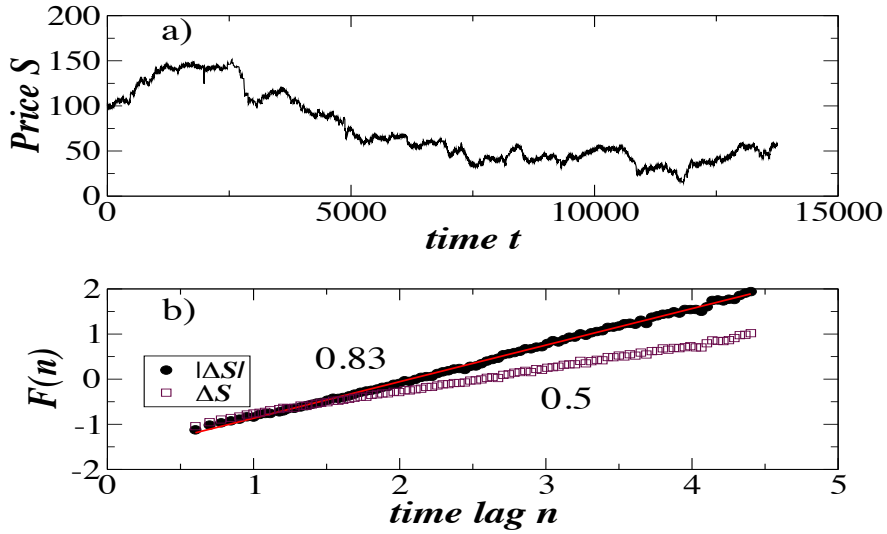


Figure 6.3: Modeling power-law auto-correlations in absolute values of price changes using the stochastic process of Eq. (6.1) with fractional parameter $\rho_1 = 0.4$ and the stochastic process of Eq. (6.2) with fractional parameter $\rho_2 = 0.12$ as in Figs. (1)-(2). (a) Time series of price for 100,000 time steps with average intertrade time $\langle \Delta t \rangle = 0.137$. (b) detrended fluctuation function $F(n)$ versus time lag n yields strong long-range power-law auto-correlations. We also show that there are no correlations in price changes.

of an ask offer, and this assumption assures that there will be no serial correlations [Fig. 6.3(b)]. Based on this trading decision, using the stochastic process of Eq. (6.1) to generate intertrade time Δt_i, the stochastic process of Eq. (6.2) to generate the number of shares traded at Δt_i, $q_i(\Delta t_i)$, and the choice for bid and ask price changes in Eqs. (6.4) and (6.5), we generate a price time series [see Fig. 6.3(a)]. Using the detrended fluctuation function $F(n)$, in Fig. 6.3(b) we show that the absolute values of price changes exhibit strong power-law auto-correlations.

For power-law distributed variables with cumulative distribution $P(s > x) \sim x^{-\zeta'}$, the Zipf plot of size s vs. rank R usually exhibits a power-law scaling regime with a scaling exponent ζ for a large range of R [**?**], $\zeta = 1/\zeta'$. Using the Zipf ranking approach, in Fig. 6.4 we show that the tails of the distribution of absolute values of price change exhibit a power law. The Zipf exponent corresponds to the scaling

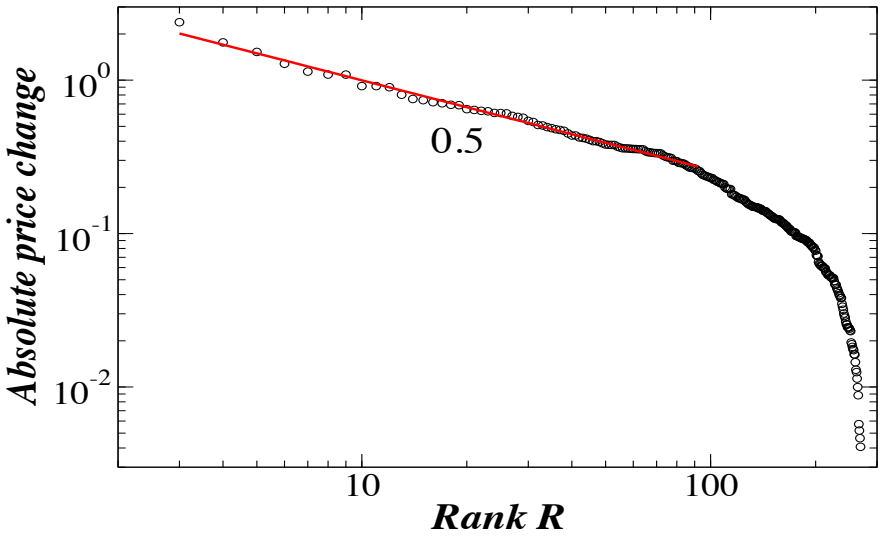


Figure 6.4: Zipf plot with power-law tails in absolute values of price changes using the stochastic process of Eq. (6.1) with fractional parameter $\rho_1 = 0.4$ and the stochastic process of Eq. (6.2) with fractional parameter $\rho_2 = 0.12$ as in Figs. (1)-(2).

exponent $\zeta = 2$. Using different parameters and different i.i.d. distributions in Eqs. (6.1), (6.2), (6.4), and (6.5), it is clear that we can eventually obtain an exponent corresponding to a cubic law [**?**].

Using quote data for the 116 most frequently traded stocks on the New York Stock Exchange over the two-year period 1994–1995, Ref. [**?**] analyzes the relationship between the bid-ask spread and other indicators of liquidity such as the number of trades occurring $N_{\Delta T}$, and the share volume traded $Q_{\Delta T}$. They found $S \propto ln\ N_{\Delta T}$ and $S \propto ln\ Q_{\Delta T}$. They also examined the relationship between the spread expectation conditioned by the time interval between trades. They found that as Δt increases, the bid-ask spread decreases, and the functional relationship is approximately $\langle s \rangle_{\Delta t} \propto -ln\Delta t$. In order to reproduce the last finding and to keep the rest of the findings, we modify the bid-ask process of Eqs. (6.4) and (6.5), which gives the proportional and not the reciprocal dependence between the spread and the intertrade time interval. Then we generate the trader's (agents') ask price

changes,

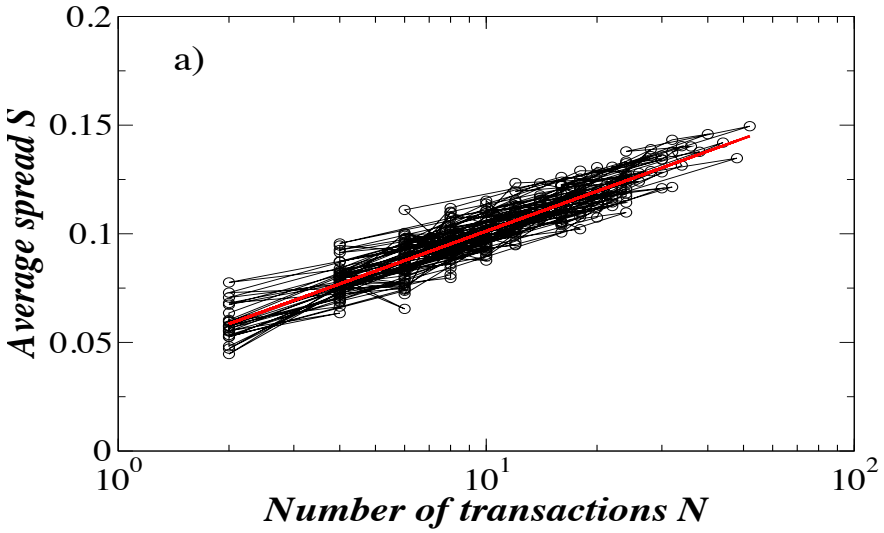
$$\Delta S^a = (\psi_i(\rho_1))^{-\gamma}\epsilon_i'', \tag{6.6}$$

$$\Delta S^b = -(\psi_i(\rho_1))^{-\gamma}\epsilon_i'', \tag{6.7}$$

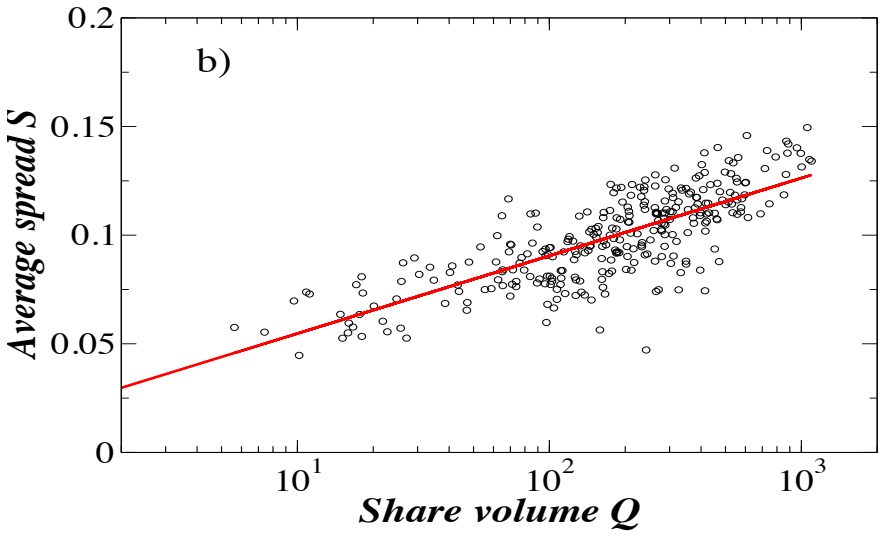
where $\gamma > 0$ and ϵ'' is explained in Eqs.(6.4)-(6.5). In Figs. 6.5(a) and 6.5(b) for $\gamma = 0.25$ we show the log-linear functional relationship between the average of the spread $\langle S\rangle_{\Delta T}$ and the number of trades occurring $N_{\Delta T}$, and between $\langle S\rangle_{\Delta T}$ and the total share volume traded $Q_{\Delta T}$, and both agree with empirical findings. Since the average intertrade time interval Δt can be thought of as a reciprocal of $N_{\Delta T}$, the model accurately gives the the reciprocal dependence between the spread and the intertrade time interval.

6.4 Summary

We have proposed a stochastic process that may offer a guide to modeling the microstructural dynamics of spreads, returns, volume $q(\Delta t_i)$, and volatility. It gives the statistical properties of the intertrade time interval Δt_i, the bid-ask spread, and the volatility in good agreement with empirical findings. We model Δt_i and $q(\Delta t_i)$ as two mutually independent but individually auto-correlated power-law processes in which the correlations in Δt_i are much stronger than those in $q(\Delta t_i)$. There are three exponentially distributed i.i.d. processes in Eqs. (6.1), (6.2), and (6.6)-(6.7), where the parameters a_1 and a_2 defined in Eqs. (6.1), (6.2) can be estimated to fit average intertrade times and the average number of shares traded, respectively. The fractional parameters ρ_1 and ρ_2 in Eqs. (6.1) and (6.2) can be estimated to fit the scaling in the auto-correlations of Δt and $q(\Delta t)$, respectively. The parameter γ in Eqs. (6.6)-(6.7) controls the power-law exponent and the strength of the auto-correlations in absolute values of price changes. The larger the γ, the smaller the exponent for the power-law tails. We believe that subordinated processes with long-range correlations have a broad range of potential applications.



(a)



(b)

Figure 6.5: Model outcomes for the stochastic process of Eq. (6.1) with fractional parameter $\rho_1 = 0.4$ and the stochastic process of Eq. (6.2) with fractional parameter $\rho_2 = 0.12$ as in Figs. (1)-(2) and Eq. (6.6)-(6.7) with $\gamma = 0.25$. (a) Average spread S versus number of transactions (b) Average spread versus share volume for a given ΔT both exhibit log-linear functional dependence in agreement with empirical findings.

Chapter 7

Conclusion

This dissertation is a review of the original research and results I conducted during my PhD studies in Boston University.

In Chapter. 2, we generalize random matrix theory (RMT) and propose a new method of cross-correlation analysis, which we call autoregressive random matrix theory (AR-RMT). ARRMT takes into account the influence of auto-correlations in the study of cross-correlations in multiple time series. We first analytically and numerically determine how auto-correlations affect the eigenvalue distribution of the correlation matrix. Then we introduce ARRMT with a detailed procedure of how to implement the method. Finally we illustrate the method using two examples taken from inflation rates for air pressure data for 95 USA cities.

In Chapter. 3, we study long-range magnitude cross-correlations in collective modes of real-world data from finance, physiology, and genomics using time-lag random matrix theory. We find long-range magnitude cross-correlations (i) in time series of price fluctuations, (ii) in physiological time series, both healthy and pathological, indicating scale-invariant interactions between different physiological time series, and (iii) in ChIP-seq data of the mouse genome, where we uncover a complex interplay of different DNA-binding proteins, resulting in power-law cross-correlations in x_{ij}, the probability that protein j binds to gene i, ranging up to 10 million base pairs. In finance, we find that the changes in singular vectors and singular values are largest in times of crisis. We find that the largest 500 singular values of the NYSE Composite members follow a Zipf distribution with exponent ≈ 2. In physiology, we find statistically significant differences between alcoholic and control

subjects.

In Chapter. 4, we propose a modified time lag random matrix theory in order to study time lag cross-correlations in multiple time series. We apply the method to 48 world indices, one for each of 48 different countries. We find long-range power-law cross-correlations in the absolute values of returns that quantify risk, and find that they decay much more slowly than cross-correlations between the returns. The magnitude of the cross-correlations constitute "bad news" for international investment managers who may believe that risk is reduced by diversifying across countries. We find that when a market shock is transmitted around the world, the risk decays very slowly. We explain these time lag cross-correlations by introducing a global factor model (GFM) in which all index returns fluctuate in response to a single global factor. For each pair of individual time series of returns, the cross-correlations between returns (or magnitudes) can be modeled with the auto-correlations of the global factor returns (or magnitudes). We estimate the global factor using principal component analysis, which minimizes the variance of the residuals after removing the global trend. Using random matrix theory, a significant fraction of the world index cross-correlations can be explained by the global factor, which supports the utility of the GFM. We demonstrate applications of the GFM in forecasting risks at the world level, and in finding uncorrelated individual indices. We find 10 indices are practically uncorrelated with the global factor and with the remainder of the world indices, which is relevant information for world managers in reducing their portfolio risk. Finally, we argue that this general method can be applied to a wide range of phenomena in which time series are measured, ranging from seismology and physiology to atmospheric geophysics.

In Chapter. 5, we propose ...

In Chapter. 6, we propose a stochastic process that may offer a guide to modeling the microstructural dynamics of spreads, returns, volume $q(\Delta t_i)$, and volatility. It gives the statistical properties of the intertrade time interval Δt_i, the bid-ask spread, and the volatility in good agreement with empirical findings. We model Δt_i and $q(\Delta t_i)$ as two mutually independent but individually auto-correlated power-law processes in which the

correlations in Δt_i are much stronger than those in $q(\Delta t_i)$. There are three exponentially distributed i.i.d. processes in Eqs. (6.1), (6.2), and (6.6)-(6.7), where the parameters a_1 and a_2 defined in Eqs. (6.1), (6.2) can be estimated to fit average intertrade times and the average number of shares traded, respectively. The fractional parameters ρ_1 and ρ_2 in Eqs. (6.1) and (6.2) can be estimated to fit the scaling in the auto-correlations of Δt and $q(\Delta t)$, respectively. The parameter γ in Eqs. (6.6)-(6.7) controls the power-law exponent and the strength of the auto-correlations in absolute values of price changes. The larger the γ, the smaller the exponent for the power-law tails. We believe that subordinated processes with long-range correlations have a broad range of potential applications.

Bibliography

[1] A. Cottet, W. Belzig, and C. Bruder. Positive cross correlations in a three-terminal quantum dot with ferromagnetic contacts. *Phys. Rev. Lett.*, 92:206801, May 2004.

[2] Richard Weaver and Oleg Lobkis. Ultrasonics without a source: Thermal fluctuation correlations at mhz frequencies. *Phys. Rev. Lett.*, 87:134301, Sep 2001.

[3] P. Roux, Sabra, K. G., Kuperman, W. A., and A. Roux. Ambient noise cross-correlation in free space: Theoretical approach. *J. Acoust. Soc. Am.*, 117:79–84, 2005.

[4] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate networks around the globe are significantly affected by el niño. *Phys. Rev. Lett.*, 100:228501, Jun 2008.

[5] Kees Wapenaar. Retrieving the elastodynamic green's function of an arbitrary inhomogeneous medium by cross correlation. *Phys. Rev. Lett.*, 93:254301, Dec 2004.

[6] Michel Campillo and Anne Paul. Long-range correlations in the diffuse seismic coda. *Science*, 299(5606):547–549, 2003.

[7] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Lu´Nunes Amaral, and H. Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Phys. Rev. Lett.*, 83:1471–1474, Aug 1999.

[8] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83:1467–1470, Aug 1999.

[9] Mantegna, R. N. Hierarchical structure in financial markets. *Eur. Phys. J. B*, 11(1):193–197, 1999.

[10] Sergio Arianos and Anna Carbone. Cross-correlation of long-range correlated series. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03037, 2009.

[11] Thomas Guhr and Bernd Klber. A new method to estimate the noise in financial correlation matrices. *Journal of Physics A: Mathematical and General*, 36(12):3009, 2003.

[12] Boris Podobnik and H. Stanley. Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Phys. Rev. Lett.*, 100:084102, Feb 2008.

[13] Eugene. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math.*, 62:548–564, Nov 1955.

[14] M. L. Mehta. *Random Matrices*. Academic Press, 2004.

[15] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Lu´Amaral, Thomas Guhr, and H. Stanley. Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, 65:066126, Jun 2002.

[16] B. Podobnik, D. Wang, D. Horvatic, I. Grosse, and H. E. Stanley. Time-lag cross-correlations in collective phenomena. *EPL (Europhysics Letters)*, 90(6):68001, 2010.

[17] Duan Wang, Boris Podobnik, Davor Horvatić, and H. Stanley. Quantifying and modeling long-range cross correlations in multiple time series with applications to world stock indices. *Phys. Rev. E*, 83:046121, Apr 2011.

[18] Safi Bahcall. Random matrix model for superconductors in a magnetic field. *Phys. Rev. Lett.*, 77:5276–5279, Dec 1996.

[19] Valentin Rychkov, Simone Borlenghi, Henri Jaffres, Albert Fert, and Xavier Waintal. Spin torque and waviness in magnetic multilayers: A bridge between valet-fert theory and quantum approaches. *Phys. Rev. Lett.*, 103:066602, Aug 2009.

[20] Thomas Guhr, Axel MllerGroeling, and Hans A. Weidenmller. Random-matrix theories in quantum physics: common concepts. *Physics Reports*, 299(46):189 – 425, 1998.

[21] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb. (N.S.)*, 72:507536, 1967.

[22] R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10:507–521, May 1915.

[23] M. S. Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8:27–41, 1946.

[24] G. V. Bayley and J. M. Hammersley. The "effective" number of independent observations in an autocorrelated time series. *Supplement to the Journal of the Royal Statistical Society*, 8:184–197, 1946.

[25] J. D. Hamilton. *Time Series Analysis*. Princeton, New Jersey, 1994.

[26] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

[27] R. A. Meyers. *Encyclopedia of Complexity and Systems Science*. Springer, 2009.

[28] Christensen K. and Moloney N. R. *Complexity and Criticality*. Imperial College Press, 2010.

[29] Cohen R. and Havlin S. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, 2010.

[30] Heartbeat interval time series, e.g., is one among many time series comprising the functioning human.

[31] Yosef Ashkenazy, Plamen Ch. Ivanov, Shlomo Havlin, Chung-K. Peng, Ary L. Goldberger, and H. Eugene Stanley. Magnitude and sign correlations in heartbeat fluctuations. *Phys. Rev. Lett.*, 86:1900–1903, Feb 2001.

[32] Akihiko Utsugi, Kazusumi Ino, and Masaki Oshikawa. Random matrix theory analysis of cross correlations in financial markets. *Phys. Rev. E*, 70:026110, Aug 2004.

[33] Bernd Rosenow, Parameswaran Gopikrishnan, Vasiliki Plerou, and H.Eugene Stanley. Random magnets and correlations of stock price fluctuations. *Physica A: Statistical Mechanics and its Applications*, 314(1):762–767, 2002.

[34] B. Rosenow, V. Plerou, P. Gopikrishnan, and H. E. Stanley. Portfolio optimization and the random magnet problem. *EPL (Europhysics Letters)*, 59(4):500, 2002.

[35] Raj Kumar Pan and Sitabhra Sinha. Modular networks emerge from multiconstraint optimization. *Phys. Rev. E*, 76:045103, Oct 2007.

[36] J. Shen and B. Zheng. Cross-correlation in financial dynamics. *EPL (Europhysics Letters)*, 86(4):48005, 2009.

[37] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

[38] M. Potters, J. P. Bouchaud, and L. Laloux. Financial applications of random matrix theory: Old laces and new pieces. *Acta Physica Polonica B*, 36(9):2767–2784, 2005.

[39] J. Kwapien, S. Drozdz, A. Z. Gorski, and P. Oswiecimka. Asymmetric matrices in an analysis of financial correlations. *Acta Physica Polonica B*, 37(11):3039–3048, 2006.

[40] K. B. K. Mayya and R. E. Amritkar. Delay correlations in multivariate time series analysis. *AAPPS Bulletin*, 17(2):30–34, 2007.

[41] S. Thurner and C. Biely. The eigenvalue specturm of lagged correlation matrices. *Acta Physica Polonica B*, 38(13):4111–4122, 2007.

[42] E. Lippiello, L. de Arcangelis, and C. Godano. Influence of time and space correlations on earthquake magnitude. *Phys. Rev. Lett.*, 100:038501, Jan 2008.

[43] A. Sengupta and P. Mitra. Distributions of singular values for some random matrices. *Phys. Rev. E*, 60:3389–3392, Sep 1999.

[44] V Plerou, P Gopikrishnan, B Rosenow, L.A.N Amaral, and H.E Stanley. A random matrix theory approach to financial cross-correlations. *Physica A: Statistical Mechanics and its Applications*, 287(34):374 – 382, 2000.

Curriculum Vitae

Contact Duan Wang

Physics Department, Boston University,
590 Commonwealth Avenue, Boston, MA 02215, USA
Telphone: 617-710-2163 E-mail: wangduan@bu.edu

Education

Boston University, PhD candidate, September 2007 – May 2015.
Thesis advisor: H. Eugene Stanley.
Thesis: Application of Statistical Methods in Financial Time Series Analysis

Nanjing University, B.Sc., Physics, September 2003 – June 2007.

Experience

Associate Director, Sun Life Financial, June 2014 – Present.

Quantitative Research Analyst, State Street Corporation, October 2012 – April 2014.

Research Assistant, Boston University, September 2009 – May 2015.

Teaching Assistant, Boston University, September 2007 – August 2009.

Presentations

Time-lag Cross-Correlations in Collective Phenomena, Statphys 24, July 2010, Cairns, Australia.

Publications

1. **D Wang**, B Podobnik, D Horvatić, H E Stanley, *A Generalization of Random Matrix Theory and its Application to Statistical Physics.* on going project
2. **D Wang**, B Podobnik, H E Stanley, *Collective Heteroskedasticity Adjusted Regression Model.* on going project
3. **D Wang**, B Podobnik, D Horvatić, H E Stanley, *Quantifying and Modeling Long-Range Cross-Correlations in Multiple Time Series with Applications to World Stock Indices.* Physical Review E **83** (046121), 066103
4. B Podobnik, **D Wang**, D Horvatić, I Grosse, H E Stanley, *Time-lag Cross-Correlations in Collective Phenomenan.* EPL (Europhysics Letters) **90**, 68001 (2010).
5. B Podobnik, **D Wang**, D Horvatić, , HE Stanley, *High-frequency Trading Model for a Complex Trading Hierarchy.* Quantitative Finance **12**, 559-566 (2012)