Stony Brook University

Essays on the Spillover Effects of Information Arrivals in Security Trading

A Dissertation presented

by

Zhenning Wang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Economics

Stony Brook University

August 2016

Stony Brook University

The Graduate School

Zhenning Wang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Juan Carlos Conesa - Dissertation Advisor
Associate Professor, Department of Economics

Sandro Brusco - Chairperson of Defense
Professor and Chair, Department of Economics

Jake Zhao - Committee Member
Assistant Professor, Department of Economics

Haipeng Xing - Outside Member
Associate Professor, Department of Applied Mathematics and Statistics

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Essays on the Spillover Effects of Information Arrivals in Security Trading

by

Zhenning Wang

Doctor of Philosophy

in

Economics

Stony Brook University

2016

The mixture distribution hypothesis is widely used to explain the behavior of returns and volumes in security trading in single-security settings, and the unobservable information arrival process in models based on this hypothesis is the key factor that determines the conditional distributions of returns and volumes. In order to investigate whether the information arrival processes of different securities may interact with each other, in the first part of this dissertation I extended the framework of mixture distribution hypothesis to a multiple-security setting, in which the unobservable information arrival processes of different securities may potentially interact with each other through vector autoregression.

Then I picked 43 large capitalization stocks publicly traded on US exchanges, grouped them into 11 pairs and 7 triplets with the same industrial sectors, and estimated the multi-security mixture distribution model using these data. My estimation results show that contemporary correlations in the shocks to information arrival processes are more common than cross-security historical dependencies in information arrival processes. However, for 9 out of 18 groups of stocks in industries such as banking, retail, consumer goods and telecommunication services, the cross-security historical dependencies are of both statistical and practical significance. Furthermore, such dependencies are asymmetric in the sense that large capitalization stocks tend to give more impacts to small capitalization stocks than in the other way.

My estimation method is based on transforming the likelihood maximization problem into an equation-solving problem involving a high-dimensional

integral, and then I use Stochastic Approximation and Markov Chain Monte Carlo simulations to search for the equations solution. My simulation study shows that point estimates produced by this method are close to the true parameter values, but the estimated confidence intervals may be not wide enough to cover true parameter values with the corresponding probabilities.

In the second part of this dissertation I applied the same model using ETF data to investigate spillover effects in information arrivals among international stock portfolios and among different US asset classes. My estimation results show that the information arrival process of the US stock portfolio can heavily impact those of other countries stock portfolios. Similarly, the information arrival process of the large-capitalization stock portfolio can heavily impact those of mid- and small-capitalization stock portfolios. However, both of these two kinds of impacts are unidirectional.

Dedication Page

I would like to dedicate my dissertation to my parents for always giving me great support.

Table of Contents

List of Figures

List of Tables

Acknowledgements

During the five years I spent at Stony Brook University, I received great help and support from many individuals. Dr. Haipeng Xing inspired me on the research topic and guided me through every step in carrying out the research and writing the dissertation. Dr. Juan Carlos Conesa is a great educator, mentor, and friend. Without his encouragement, I would not have had enough endurance to finish the study and research process. I would like to also thank Dr. Sandro Brusco and Dr. Jake Zhao for being my committee chairperson and committee member, and for being always responsive when I asked for their help. Dr. Mark Montgomery taught me the programming and typesetting skills that I rely on heavily ever since I learned them.

At last, I would also like to thank my friends Dr. Lunan Jiang, who made my years at Stony Brook University a truly memorable experience.

Part I

The Multiple Security Mixture Distribution Hypothesis with Spillover Effects in Information Arrivals

Chapter 1

Introduction

This research is an investigation into the relationship among multiple financial securities' information flows. To better explain it, I will start with clarifying the meaning of information flow and financial security used here.

Let us assume for one moment that the financial security is the common stock of a publicly traded company. Information flow is a concept that is used by Tauchen and Pitts (1983) among other to explain the behavior of the financial securities' daily returns and trading volumes. The fundamental idea of information flow-centered models is that investors' position choices on the stock are based on their subjective valuations for the stock. The stock's valuations, which in general are different across investors, in turn are calculated using the information investors have received regarding the stock. When new information arrives to the market and is absorbed by the investors (for instance, some news breaks out about the public company whose stocks are traded), it tends to change investors' valuations for the stock idiosyncratically and thus to create trading opportunities among the investors. Investors adjust their positions until the market reaches an equilibrium, and such an position-adjustment and equilibrium process occurs for every arrival of new information. A simplifying assumption is that the effects of each arrival of new information is homogeneous, which is shocking investors' existing valuations by an idiosyncratic random variable. This assumption allows us to focus on information flow, defined as the number of times for which new information regarding a security occurs on the market over a fixed period of time, as information flow in this setting determines the distribution of returns and trading volumes. Since 1970's, information flow has been widely used to explain the distributions of financial securities' daily returns and trading volumes by a well-developed line of literatures on financial econometrics. It provides a way to explain the positive correlation between return volatilities and trading volumes, and also gives an explanation for the autocorrelation in returns' volatility process.[1]

A key feature of models with information flows is that information flows are not directly observable. Nowadays, the trading data for public companies' common stocks, for example, are freely available from Internet financial media.[2] However, the variables included in such data sources are usually limited to open, close, low and high prices, trading volumes and dividend amounts at daily frequencies.[3] Though up-to-date news reports are usually posted on the same page together with other variables, no quantitative measure is taken towards determining the amount of information

[1]See Clark (1973), Tauchen and Pitts (1983), Andersen (1996) and Liesenfeld (2001) for examples of major contributions to this area of research. Reviews on these papers will be presented in Section 1.4.

[2]This research has used Yahoo Finance (`http://finance.yahoo.com`) as the major data source.

[3]Of course, dividends amount is non-zero only on ex-dividend days.

flow, not to mention the availability of historical data. Two reasons may be causing the practical difficulties in implementing the definition of information flow, i.e. the number of times that news breaks out within a given time period regarding a specific security. One is that many different media resources may simultaneously cover the same event, yet producing reports with differentiating information contents. In such cases, it would be difficult to determine whether to count different news reports as a single or multiple information arrivals, due to the overlap in reports. Moreover, the variety of news resources makes it costly to make a complete, fully-covering report count. The other reason is that it is also difficult to count information arrivals over time from a single news source, since a news event itself may develop over time and provide new information to the market. It is thus inappropriate to treat the series reports as a single information arrival, but nor is it easy to separate them as documentations of distinct events. Therefore, an easier way to handle information flow is to assume it is unobservable, and later use special techniques to deal with its unobservability in the estimation stage. After all, no data is available for such variables, but several estimation techniques have been developed for models with unobservable variables.

But the unobservability of information flows does introduce difficulties into the estimation procedure. This research has taken maximum likelihood as the primary estimation technique. With the presence of unobservable information flow variables as well as their autoregressive property, the (log-) likelihood function would contain a very high dimensional integral that cannot be computationally tracked. This research has taken an approach introduced by Gu and Kong (1998), who have applied the Stochastic Approximation algorithm to the likelihood maximization problem. The main idea of their method is to first transform the likelihood maximization problem into an equation solving problem, and then uses Stochastic Approximation (Robbins and Monro 1951) to solve the equation. Since the equation still involves a high dimensional integral, meaning that its value cannot be directly assessed, in each parameter update step, the Stochastic Approximation algorithm uses Markov Chain Monte Carlo simulations (Metropolis et al. 1953, Hastings 1970 and Peskun 1973) to draw an noised observation of the equation's value, and use the noised observation to make parameter updates. Under some regularity conditions, the parameters updated in this manner will converge to the Maximum Likelihood estimate.

The usage of information flow in explaining the behavior of single securities has been very successful.[4] However, the application of the same modeling method to a multiple asset framework have been very limited so far. On one hand, the volatility processes of different securities are often highly correlated, as will be shown in Section 1.3 of this chapter. On the

[4]Please see footnote 1 and Section 1.4.

other, the information flows of different securities could be correlated as well. For example, in the case of common stocks, a companies's financial performance could be related with that of another company through competition, cooperation, shared subjection to macroeconomic shocks and other channels. The existence of such connections between companies implies that the information flows of these securities may also be correlated. This research calls "information spillover" the mechanism through which information flows of different securities are correlated with each other, which will be explained in more detail in Section 1.3 and Chapter 2. The primary goal of this research is testing the existence of information spillover, and using spillovers in information flows to explain the correlations among multiple securities' volatility processes, based on the framework introduced by Tauchen and Pitts (1983).

I have used common stocks as an example for securities in explaining information flow based models, but such models can be applied to analyzing other asset classes as well. For example, Tauchen and Pitts (1983) actually have used this framework to explain the behavior of 90-day Treasury bill contracts. Thus I am planning to use this multi-security model to examine not only the spillover effects of information flows among different stocks, but also among different stock indexes on the US market, among stock markets indexes of different countries, and among indexes of different asset classes (stock, bond and commodity indexes). Since indexes are portfolios of individual securities, in principle they should be able to inherit the interactions among individual securities' information flows. Applying the model to index level data allows us to investigate the joint behavior of different types of securities, without being limited to a few specific choices of each type.

However, index data cannot be directly utilized to estimate information flow based models, since in such models there is another variable that plays an important role, which trading volume. A central idea embedded in these models is that the conditional distributions of both return and trading volume are determined by the unobservable information flow, which in turn indicates that incorporating volume information in the estimation process would produce more efficient estimates. There is no difficulty in using trading volume data of individual stocks, but trading volume is not easily defined for market indexes. Trading volume is the amount of securities that are traded within a given amount of time. For stocks it is the number of shares traded, and for future contracts it is the number of contracts traded. But such a definition cannot be applied to market indexes, since market indexes are not traded in the same way as individual securities. It is very difficult for an investor to trade a market index by directly buying or selling its component securities simultaneously, which would involve a large amount of transactions and also entails a large amount of capital.

Instead, the common alternatives for an investor to trade a market index are through trading an Exchange Traded Fund (ETF), a market index future, or a market index option.[5]

ETF is a type of passively managed funds which target at replicating certain predetermined market indexes. A distinct feature of ETF's is that they are traded in the same way as stocks on public exchanges. In addition, for ETF's there is an arbitrage-free mechanism known as Creation and Redemption, in which an Authorized Participant (a financial institution different from the fund management firm) can creates and redeem shares of the ETF, and essentially limits the price differences between a share in an ETF and a corresponding share in the targeted index into a very small range.[6] Thus ETF's, especially those with large values of asset under management (such that they are more intensively traded), are good proxies for the market indexes they track.

On the other hand, though the theoretical prices of future and option contracts are also uniquely linked with the spot market index value, in practice the linkage could be loose because of transaction costs and uncertainties in parameters. For example, the existence of volatility smiles/skews indicates that option contracts on the same underlying assets and with the same expiration date, may be priced with different volatility values at different strikes. As a matter of fact, the volatility parameter itself is uncertain, the accurate estimation of which is one of the major objectives of this research.

Therefore, ETF's are chosen as proxies for the corresponding market indexes. Three type of ETF's tracking US market indexes are selected, to proxy the performance of stock, bond and commodity markets, respectively. Within the category of stock market ETF's, future granularity is introduced by ETF's tracking stock indexes of large, middle and small capitalization stocks. In addition, ETF's that targeting at replicating the performance of international stock markets are also included. In Chapter **??**, I will present in details the features of such data, and the results of estimating the information flow framework using such data.

[5]Another investment vehicle that allows one to track the performance of a market index is known as index fund, which is a type of mutual fund that is passively managed. The manager of an index fund does not try to beat the performance of the tracked market index, but only to replicate its performance. Index fund cannot be traded (which is also true for the more general mutual funds), since investors can only redeem or purchase shares of an index fund at the end of a trade day from the fund management firm. Nor can an investor hold short positions on an index fund.

[6]For more information on the Creation and Redemption mechanism, please see the following page at `http://www.etf.com/etf-education-center/7540-what-is-the-etf-creationredemption-mechanism.html`.

1.1 Volatility Modeling

Suppose we can draw observations of transaction prices and accumulated transaction volumes over fixed time intervals for a security traded on some speculative market. The length of the time interval used for making such drawings may vary from a few minutes to a month, depending on the data source one can access. Then we can calculate security returns from its price data, either as percentage changes or changes in logarithmic prices between two consecutive observations. The data collected in this way—observations drawn over fixed time intervals, is called *time series* data. And if we see each observation as a realization of a unique random variable, then the collection of all these random variables together with the time index is called a *stochastic process*. A long observed feature of time series return data is that its conditional variances display substantial variations over time.[7]

For example, this research has used the following definition for an asset's return R_t on day t:

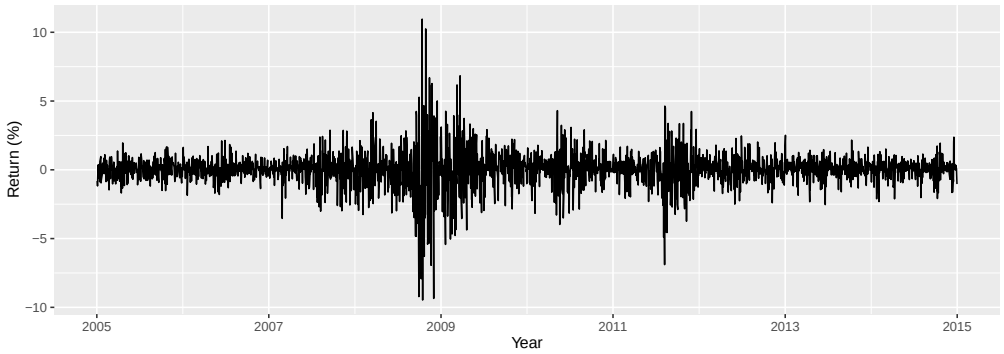
$$R_t = 100 \times \log\left(\frac{S_t}{S_{t-1}}\right),$$

where S_t is the asset's price on day t, and the factor 100 enables the interpretation of R_t as return in percentages. Figure 1.1a then plots the daily returns of the Standard & Poor's 500 Index (S&P 500) between January 3, 2006 to December 31, 2015 calculated in this way. It shows that for most part of the observation period, the returns of S&P 500 fluctuate mildly within the range $(-2.5, 2.5)$, but there are other times during which return observations far out of this range are well documented.[8] The fourth quarter of 2008 is the period that gives the most extreme return observations, while several other periods such as the second half of 2007, the first half of 2008 and 2009, the second quarter of 2010, the second half of 2011 and the third quarter of 2015, also display higher-than-normal return fluctuations. During these heavily fluctuating time periods, extreme return observations do not appear alone but in clusters. An extreme observation is often followed by a few more observations which are also extreme. Mandelbrot (1963) has called such a behavior of security returns "volatility clustering". The existence of volatility clustering indicates that the conditional variances of returns may have increased to and persisted at high levels during certain periods.

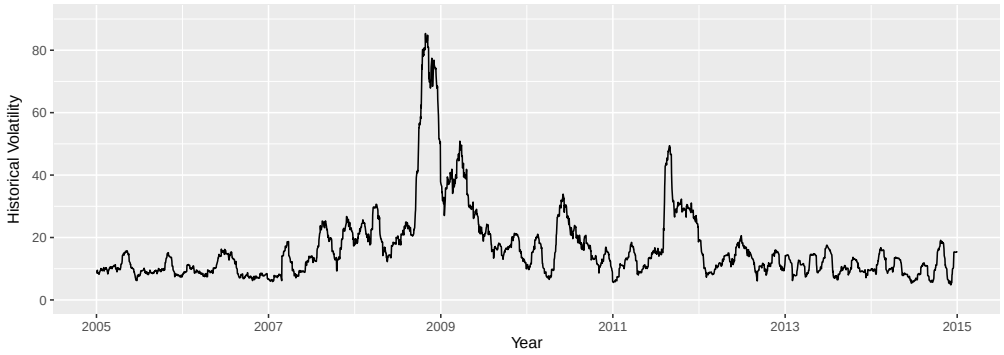
An quantitative way to exhibit the existence of volatility clustering is to examine the levels of partial autocorrelations in the absolute return series. In general, the return series itself is not highly correlated, since if

[7] This feature is also exhibited by some other time series data. See Footnote 10.

[8] Here the S & P 500 index is treated as a single security.



(a) Daily Returns



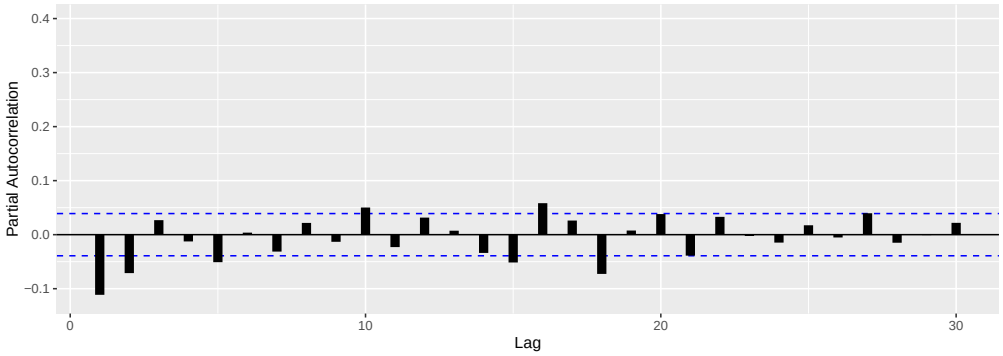
(b) Annualized Historical Volatilities

Figure 1.1: Daily returns and annualized historical volatilities of S&P 500 Index, 2005—2014. Data source: Yahoo Finance (`http://finance.yahoo.com`).

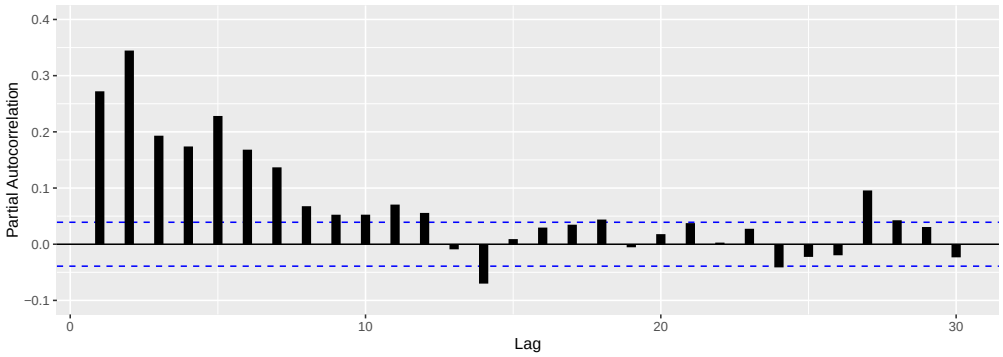
The return and historical volatility are respectively calculated as: $R_t = 100 \cdot \log(S_t/S_{t-1})$ and $\widehat{\sigma}_t = \sqrt{(252} \cdot \left[\sum_{i=0}^{20} \left(R_{t-i} - \sum_{j=0}^{20} R_{t-j}/21\right)^2 /20\right]^{1/2}$, where S_t is S&P500's close value on day t. $\widehat{\sigma}_t$ has a factor of $\sqrt{252}$ since there are about 252 trading days in a year.

it were, traders in the market would be able to make profits from trading on such correlations, and make the correlation vanish. However, the absolute return series can exhibit higher positive autocorrelations, as previously explained in the concept the volatility clustering. Figures 1.2a and 1.2b have plotted the partial autocorrelation coefficients at lags from 1 to 30 for S&P 500's returns and their absolute values, respectively. Compared to autocorrelation, partial autocorrelation cleans the correlation at low lags before examining higher lag correlations. In these two figures, the blue dashed lines mark the 5% rejection boundaries for separately testing the null hypothesis of zero correlation at a given lag. For most lags, the null hypothesis of no correlation cannot be rejected at 5% significance level, as shown in Figure 1.2a. And for the lags with which the null can be rejected, the magnitudes of estimated partial autocorrelation coefficients are not big. Lag 1 shows the highest extent of autocorrelation, which is slightly lower than -0.1. However, Figure 1.2b presents strong evidence against the null

hypothesis of zero autocorrelation for lags from 1 to 7. For these lags the estimated partial autocorrelations fall well above the rejection boundaries, and the magnitude of these estimates are also much greater than those shown in Figure 1.2a. The results shown in two graphs are consistent with the concept of volatility clustering.



(a) The Partial Autocorrelations of Daily Returns



(b) The Partial Autocorrelations of Absolute Daily Returns

Figure 1.2: The partial autocorrelations of S&P 500's returns and their absolute values.

The partial autocorrelation coefficient is defined as follows. Suppose we have observations on variables X, Y and Z. The partial correlation coefficient $\rho_{XY \cdot Z}$ between X and Y adjusted for Z is the correlation coefficient between r_X and r_Y, which are the residuals of regressing X and Y on Z, respectively. The partial autocorrelation coefficient at lag h for time series data R_t is then the partial correlation coefficient between R_t and R_{t-h} adjusted for $R_{t-1}, \cdots, R_{t-h+1}$. The blue dashed lines in the graphs mark the 5% rejection boundaries for partial autocorrelation coefficient estimates at all lags, such that if an estimate falls out, the corresponding null hypothesis should be rejected. Since the partial autocorrelation estimates follow $\mathcal{N}(0, 1/T)$ in large samples, the blue lines are located at $\pm 1.96/\sqrt{T}$, where T is the number of observation periods.

The area of time series research which addresses how conditional variances change over time is called volatility modeling. For return data, the word *volatility* actually refers to an observation's annualized conditional standard deviation[9]. In practice, there are different measures for volatil-

[9]See Ruey S. Tsay (2002, p. 109) for a definition of volatility.

ity, such as realized volatility, implied volatility and historical volatility. Realized volatilities are computed from high frequency data and implied volatilities are calculated using data on option prices. Since these two types of data are not available in this research, historical volatility becomes the suitable tool for measuring volatility empirically. Figure 1.1b has plotted the historical volatilities of S&P 500's returns, showing that the historical volatilities have spiked during the times when returns fluctuate heavily. And then it takes some time for volatilities to return to their long term average level. The idea of volatility modeling is then to describe how volatilities evolve over time using statistical models.

Ever since the groundbreaking work of R. F. Engle (1982) who introduced the *autoregressive conditional heteroscedasticity* (ARCH) model, volatility modeling has become one of the most intensively studied areas in time series research.[10] Depending essentially on the specification of the current period information set, which could be though of as the collection of all currently available information, volatility models can be grouped into two categories: those using the *generalized autoregressive conditional heteroscedasticity* (GARCH) framework and those using the *stochastic volatility* (SV) framework. The GARCH framework is a major development by Bollerslev (1986) to Engle's ARCH model. In the simplest case in which returns are uncorrelated with zero means, the GARCH(p, q) model specifies that

$$R_t = \sigma_t \varepsilon_t, \tag{1.1}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \alpha_j R_{t-j}^2, \tag{1.2}$$

where ε_t is i.i.d. $\mathcal{N}(0, 1)$. Equation (1.2) allows volatilities to be autocorrelated and also to depend on past values of returns. Thus, it can potentially describe the volatility clustering behavior mentioned above. That is, the spikes in volatility are caused by extreme values in returns, and volatilities show persistence due to the autoregression coefficients β_i's.

On the other hand, the origination of the SV framework has been due to Clark (1973) and Taylor (1982)). Clark first proposes that information arrivals during trading days produce a random amount of i.i.d. shocks to compose the daily return distributions. And Taylor introduces autoregression to the information arrival process to model volatility persistence. Taylor's model essentially replaces equation (1.2) with

$$\log \sigma_{t+1}^2 = \omega + \beta(\log \sigma_t^2 - \omega) + \eta_t, \tag{1.3}$$

[10]Thus volatility modeling is not limited to applying to asset return data, but can also be applied to other time series data. For example, R. F. Engle (1982) has used it to estimate the British inflation data.

where η_t's are i.i.d. $\mathcal{N}(0, \sigma_\eta^2)$.[11] Equation (1.3) describes the volatility behavior in a different way: Volatilities still follow an autoregressive process, but are now subject to another source of disturbance η_t rather than R_t, though η_t and R_t are allowed to be correlated.

The critical difference between the two types of models is that, in GARCH the next period conditional variance is fully determined by the current period information set, while in SV the next period conditional variance is driven by another unobservable stochastic process $\{\eta_t\}$.[12]

After being introduced into the academic world, both of these two lines of volatility modeling have experienced significant developments, among which the ones most closely related to my research are their extensions to multivariate models. Early works on volatility modeling have employed a univariate framework, which means they focus on modeling the volatilities of a single stochastic process, though in the case of SV volatilities actually depend on another stochastic process. However, as summarized by Bauwens et al. (2006), there are many cases in which one may be interested in the volatilities of multiple stochastic processes, which entails adopting a multivariate approach. For example, we may be interested in investigating the relations between the return volatilities of multiple assets, of multiple indexes, of different asset classes, or of assets in different countries. We may also would like to study the relations between volatilities of real and financial time series. To attain such objectives, we would have to employ a multivariate framework to model the volatilities of multiple stochastic processes simultaneously. Multivariate volatility modeling also have applications in the areas of portfolio optimization and risk management, since in these areas return series of multiple assets are often involved. In Section 1.4, we will see in more details how multivariate volatility models are developed.

1.2 The Mixture of Distributions Hypothesis

The literature on volatility modeling usually estimates models using only return data. However, there is another widely observed phenomena in financial trading data that squared returns and volatilities are positively correlated with trading volumes. Before presenting a figure showing the

[11] In Taylor's (1982) work, ε_t and η_t are assumed to be independent, but correlations between them are introduced later by other authors to model the leverage effect.

[12] As pointed out by Shephard and Andersen (2008), sometimes the GARCH framework is also described as SV, but here I follow the custom of reserving SV for models with stochastic conditional variances, which is followed by Shepard and Anderson. They have also pointed out other distinctions between the two frameworks. For example, the SV framework is more suitable for continuous time modeling.

correlations between these two variables, an quick introduction on volume detrending is entailed. Trading volumes are not stationary. For example, as the market for a security grows, more traders may participate in trading the security, trading volumes then tend to grow; and vice versa.[13] Such an effect may contaminate the relation between return volatilities and trading volumes this research is trying to uncover, which results from the behavior of traders already participating in trading. Thus detrending the volume data is necessary so as to remove the effects of no interests to us, before using the volume data in any comparison with other variables.

This research has followed a detrending method used by Andersen (1996). It first calculates volume trends as centered moving averages, and then divide trading volumes by their corresponding trend values. Two years of observations (505 trading days) are used in calculating volume trends. Figure 1.3 has plotted the trading volumes of S&P 500 index (black line) as well as its trend values (blue line) calculated using this method. The S&P 500 index has experienced a steadily growing trend in trading volumes until mid-2009, which is then followed by a gradual decrease. Though the mechanism causing such a change of direction in volume trends is not clear, its occurrence seems to coincide with the outbreak of the 2008 financial crisis.

Figure 1.4 then plots the squared returns as well as detrended trading volumes of S&P 500 index. For the periods such as late 2008, early 2009, mid-2010 and late 2011, during which squared returns are higher than adjacent observations, the detrended volumes also tend to be at high levels. To see this more rigorously, the correlation coefficient between detrended trading volumes and squared returns is 0.248, and the correlation coefficient between detrended trading volumes and historical volatilities is 0.285.

Early researchers who have observed such a phenomena include Ying (1966) and Clark (1973). Ying has reported that, "a large increase in volume is usually accompanied by either a large rise in price or a large fall in price", and then conjectured that returns and volumes of stock trading are the results of a single market mechanism. As such, incorporating volume information in model estimation should produce more efficient estimates. Clark further raises a hypothesis that daily returns of security prices are a subordinated process directed by daily numbers of transactions which can be measured by daily trading volumes, and thus explains the positive correlation between squared returns and trading volumes. This hypothesis is later known as the Mixture of Distributions Hypothesis (MDH). An econometric model with microeconomic foundation is then proposed by Tauchen and Pitts (1983), who hypothesize that on a trading day, an equilibrium follows each information arrival to the market, which drives price to fluc-

[13]See Tauchen and Pitts (1983) for an example: As the US Treasury futures market grew, more traders participated in the market, and the trading volumes grew.

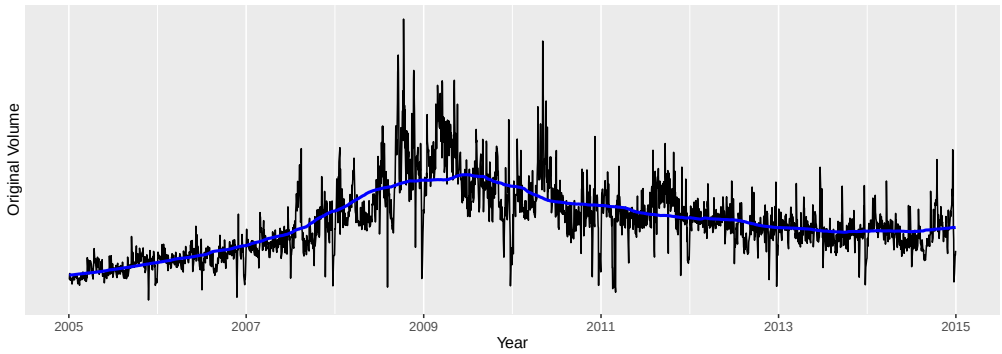


Figure 1.3: Daily Trading Volumes of S&P 500 Index, 2005—2014. Data Source: Yahoo Finance (`http://finance.yahoo.com`).

Trading volumes depend on the numbers of active traders in the market, and are usually non-stationary. This is more evident for individual securities (for example the stocks of International Business Machines investigated by Andersen (1996)) and markets in growth (for example the treasury bill futures market studied by Tauchen and Pitts (1983)). The trading volume trend $\widehat{V}_t^{trend}$ in this figure is calculated as the two-year moving average centered on day t. 505 observations are used in calculating each trend value. The trading volumes of S&P 500 index show a growth trend before mid-2009 and a decline trend after that.

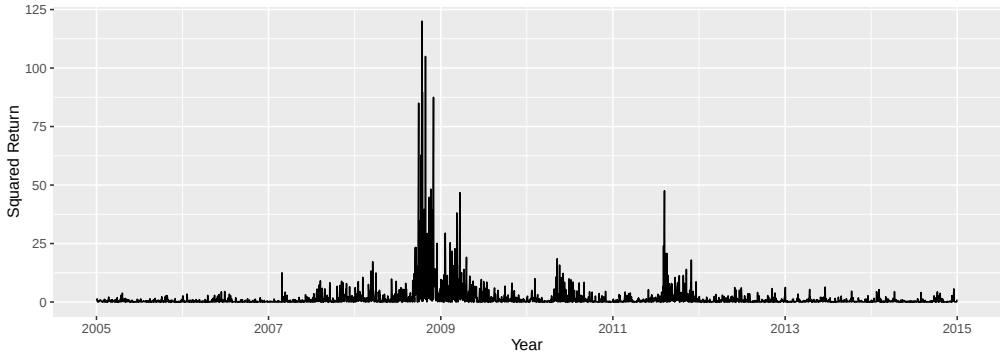
tuate and volume to accumulate. The observed daily returns and trading volumes are then the aggregated effects of these individual information arrivals. The number of information arrivals plays a key role in this model, since it determines the distribution of return and volume.

Later, the work of these authors have been followed by a series of contributions to the MDH framework, which I will explain in more details in Section 1.4. Especially, as the information arrival process is modeled to be autocorrelated, the volatility-volume relation framework would then be able to explain the source of volatility clustering illustrated in Section 1.1.[14] In addition, since it can also take into consideration the information carried by trading volumes, the MDH framework would be able to produce more efficient parameter estimates than the univariate volatility models[15].

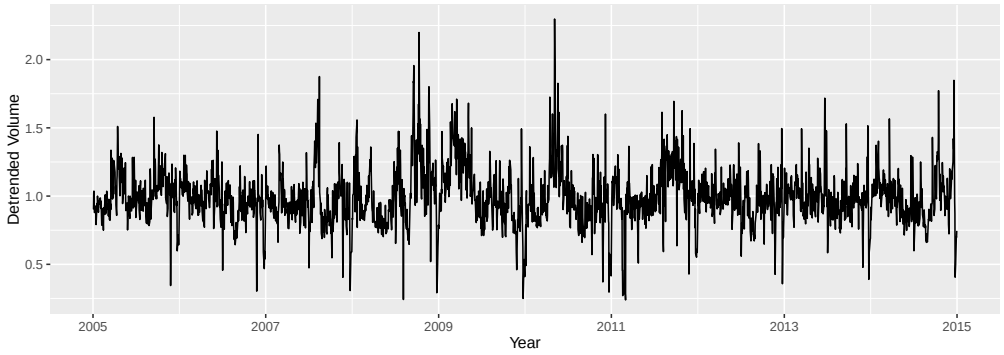
However, most of the researches within the MDH framework consider modeling the trading data of only one security. As mentioned in Section 1.1, there are many cases in which we may be interested in the joint distribution of multiple security's trading data, and especially in the relationship

[14]Shephard and Andersen (2008) actually attribute part of SV's origination to Clark (1973).

[15]Here "univariate" means a model only considers return data in volatility modeling, while the MDH is also known as the Bivariate Mixture Model, since it sees the distributions of volatility and volume both determined by the same information arrival process.



(a) Squared Returns



(b) Detrended Trading Volumes

Figure 1.4: Squared Returns and Detrended Trading Volumes of S&P 500 Index, 2005—2014. Data Source: Yahoo Finance (`http://finance.yahoo.com`).

I have followed one of the detrending methods employed by Andersen (1996), which includes first estimating the trading volume trend $\widehat{V}_t^{trend}$ as the two-year moving average centered on day t, and then computing the detrended trading volume $\widehat{V}_t$ as $V_t/\widehat{V}_t^{trend}$, where V_t is the original trading volume. The correlation between R_t^2 and $\widehat{V}_t$ is 0.248, and the correlation between $\widehat{\sigma}_t$ and $\widehat{V}_t$ is 0.285. The same correlation coefficients computed using the original trading volumes are actually higher, at 0.347 and 0.598 respectively. The original trading volumes having higher correlations with the other two series may be due to some fundamental changes in market participants' behavior, which are out of the scope of this research.

between these securities' return volatilities. Multivariate GARCH and SV models provides options to carry out such studies, but they do not utilize the information contained in trading volumes. My research then focuses on bridging this gap between the MDH and multivariate volatility models, providing a framework to jointly model the volatilities of multiple securities and jointly model the distribution of returns and trading volumes.

1.3 Information, Volatility and Volume for Multiple Securities

As mentioned in the previous section, information arrival plays a key role in determining the joint distribution of returns and trading volumes in the MDH framework. Since each information arrival can drive the fluctuation of price and the accumulation of trading volume, as a greater number of information arrivals occur, both return volatility and trading volume tend to be large. Together with the autocorrelation in the information arrival process, the MDH framework not only explains the correlation between return volatilities and trading volumes, but also reveals the source of volatility clustering to be the autocorrelation of information arrivals.

When justifying the choice of modeling information arrivals as an autoregressive process, Andersen (1996) has elucidated that when some news breaks out about a security, more detailed disclosures tend to follow the initial report over the next few days or even weeks. On the other hand, it also takes time for companies to fully implement changes in their "tactical orientations" such as "take-over battles and proxy fights". Liesenfeld (2001) has used a similar argument to explain why the number of information arrivals on a trading day could depend on its historical values. This research tries to generalize the argument by Andersen and Liesenfeld, such that the number of information arrivals about a security could depend on not only its own past values, but also the past values of other securities' information arrivals.

The reasoning for such a generalization can be explained at multiple levels. For common stocks, it is because companies' businesses are related with each other. They may be supplying similar products or services and thus competing with each other, or be operating in the same industry but taking different positions in the supply chain. A company may also be holding shares in other companies, or be a creditor of other companies. Moreover, they may be subject to the same macroeconomic shocks. These examples are far from complete, but they are sufficient to demonstrate that there are plenty of channels through which companies could be connected with each other. Thus a new piece of information occurred to one company may have substantial impacts on other companies as well. In particular, the type of impacts this research considers is not that one company's number of information arrival directly affects the price or volume of another company's stock, but that it affects the number of information arrivals to other companies that it is related to. There are two forms which such cross-company impacts of information arrivals may take. One is contemporaneous correlations among the random shocks to companies' information arrival processes, such that when one company encounters a positive shock in its number of information arrivals, other related companies also tend to

have positive shocks, and vice versa (in the case of positive correlations).

And the other form is information arrivals' cross-company dependence on their historical values. As emphasized by Andersen and Liesenfeld, an important feature of the role played by information is that it takes time for its impacts to fully develop and disclose, causing autocorrelation in information arrival numbers. There is no reason to limit this time-taking feature of information arrival to only the impacts on the company that new information directly occurred to, because when a news event regarding a company occurs, its impacts on other related companies may also develop over time, creating cross-company historical dependencies in information arrivals. Another reason for cross-company historical dependency is that it takes time for information to spread across companies. For example, compared to information directly occurred to the company she usually analyzes, a business analyst may take more time to pick up a piece of information occurred to another company because of her constrained specialty, though the latter company is related to the one she covers. In addition, she may need communicate with other people to better understand the full impacts of news occurred to other companies, which again is a time-taking process. Therefore, we can generally expect that in the multiple-stock case, the number of information arrivals for one stock would depend not only on its own historical observations, but also on that for other-related companies.

After the forms of cross-security dependence in information arrivals are explained for companies' common stocks, i.e. the contemporaneous correlation and cross-security historical dependence, the case could be easily extended to other financial securities. This research then considers the information spillover effects among stocks, among market indices, among equity and bond markets, and among markets of different countries. Market indices are constructed as portfolios of individual stocks, and thus the connections among individual stocks can be inherited by market indices. The bond market and equity market are both subject to macroeconomic shocks, and thus also have correlated information processes. And finally, markets of different countries are also correlated, because the intense financial and economic interactions among international economies make them all subject to global-wide macroeconomic shocks. Except for the case of stocks, this research will use market indices to represent the corresponding financial markets or hypothetical securities. As in the case of multiple stocks, interactions among the information arrival processes of these securities will take the same two forms, contemporaneous correlations in shocks and cross-security historical dependence.

In order to illustrate the correlation among securities' information arrival processes, I have plotted in Figure 1.5 the squared historical volatilities of all component stocks of Dow Jones Industrial Average Index (DJIA Index), since in Tauchen and Pitts's (1983) model (which is a single-security

MDH model) the number of information arrivals is proportional to a security's squared volatility. All stocks had clearly experienced a volatility spike at the end of 2008, which implies they were all subject to the macroeconomic shock of the 2008 financial crisis. Please see Table 1.2 for correspondence between company names and their stock ticker symbols, as well as their industrial classifications given by Yahoo Finance. In order to more accurately examine the correlations among these squared volatility series, Table 1.1 has presented their correlation coefficients, showing that the correlation coefficients of squared historical volatilities between most pairs of DJIA Index's component stocks are higher than 0.75 (367 out of 435). For some companies in same industries, the correlation coefficients are especially high. For example, BA and UTX (0.93), CVX and XOM (0.99). Some other pairs of stocks also show very high levels of correlations, though they are not in same industries, such as CVX and JNJ (0.96) and DIS and UTX (0.95). Figure 1.7 has made a histogram for all correlation coefficients, showing that the frequencies of correlation coefficients peaked between 0.8 and 0.85, and most correlation coefficients are at relatively high levels. Figure 1.6 is included to facilitate visual examinations of the correlation coefficients in Table 1.2. The evidence shown by these tables and figures strongly support the hypothesis mentioned above, that the securities' information arrival processes are correlated. And the following sections will show in detail the way this research takes to model such correlations.

1.4 Literature Reviews

The Development of Multivariate Volatility Modeling

As mentioned in Section 1.3, while the information arrival process plays an important role in explaining the joint behavior of returns and volumes, it is also closely related to the volatilities of returns. Since the main purpose of this dissertation is to jointly model the information arrival processes of multiple securities, it is appropriate to review the developments of the multivariate volatility literature. For these literature reviews, I will focus on their model specifications.

I will start with multivariate GARCH models. Given a stochastic process $\{\boldsymbol{r}_t\}$ and the information set $\mathcal{F}_{t-1}$ generated by the information up to period $t-1$, a multivariate GARCH specification usually writes

$$\underset{n\times 1}{\boldsymbol{r}_t} = \underset{n\times n}{\boldsymbol{\Sigma}_t}^{1/2} \cdot \underset{n\times 1}{\boldsymbol{\varepsilon}_t},$$

where the i.i.d. process $\{\boldsymbol{\varepsilon}_t\}$ satisfies $\mathrm{E}(\boldsymbol{\varepsilon}_t|\mathcal{F}_{t-1}) = 0$ and $\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'|\mathcal{F}_{t-1}) = \underset{n\times n}{\boldsymbol{I}}$, $\boldsymbol{\Sigma}_t$ is measurable with respect to $\mathcal{F}_{t-1}$, and thus $\mathrm{E}(\boldsymbol{r}_t\boldsymbol{r}_t'|\mathcal{F}_{t-1}) = \boldsymbol{\Sigma}_t$

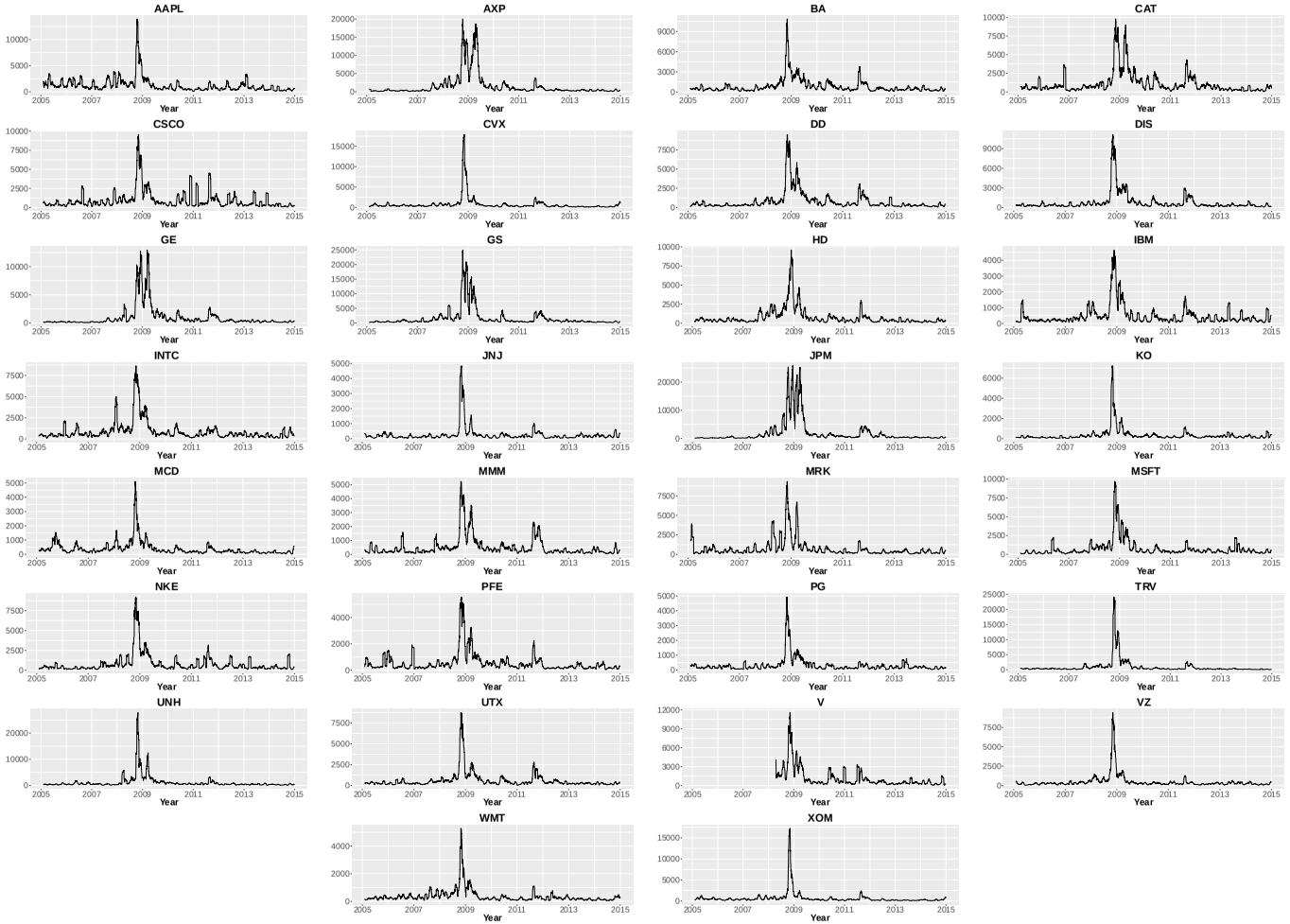


Figure 1.5: The Historical Volatilities of DJIA's Component Stocks, 2005—2014. Data Source: Yahoo Finance (`http://finance.yahoo.com`).

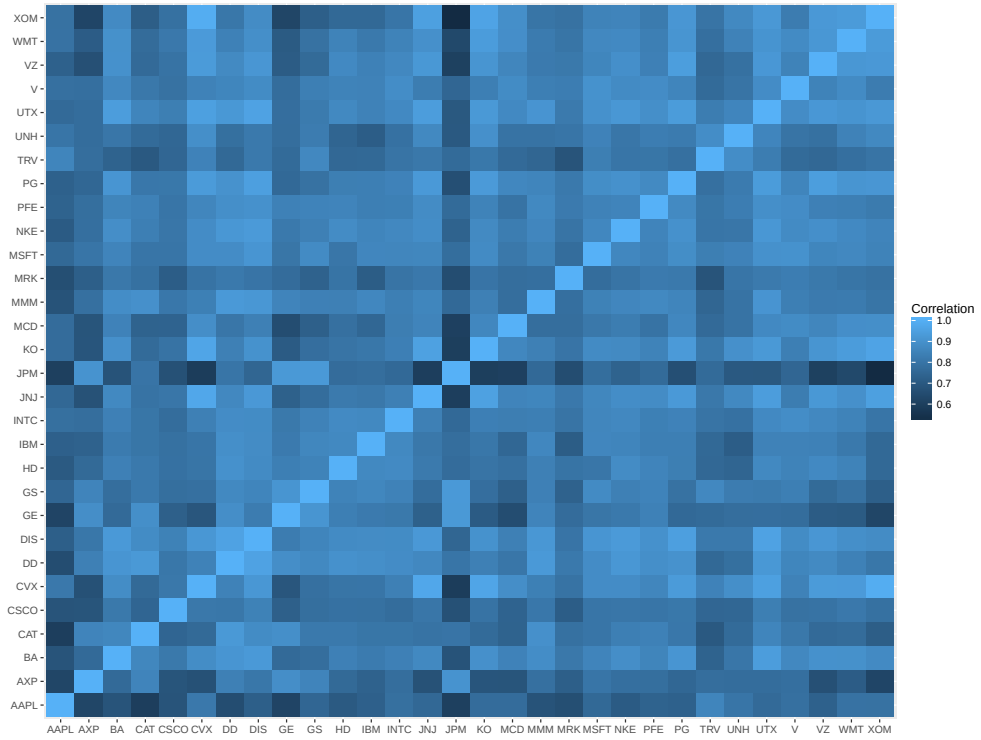


Figure 1.6: The Correlation Among DJIA's Component Stocks. Data Source: Yahoo Finance (`http://finance.yahoo.com`)

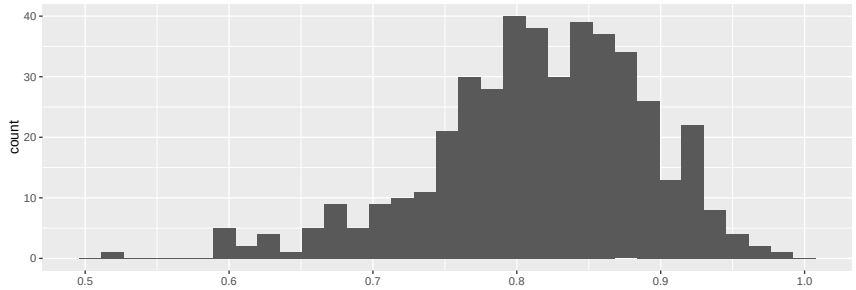


Figure 1.7: The Histogram of Cross-Security Correlations Coefficients for Squared Historical Volatilities. Data Source: Yahoo Finance (`http://finance.yahoo.com`)

is the conditional variance of $\boldsymbol{r}_t$. The focus of the multivariate GARCH literature is then modeling the behavior of $\boldsymbol{\Sigma}_t$.

As stated by Silvennoinen and Teräsvirta (2009), multivariate GARCH models can be caxtegorized into four classes, in terms of the approaches they take to model $\boldsymbol{\Sigma}_t$. But since the semi-parametric and non-parametric approaches are less-related my dissertation, I will review the other three classes of multivariate GARCH models. The first class is known as the VEC-GARCH model by Bollerslev et al. (1988) and the BEKK model by

R. F. Engle and Kroner (1995). Using the vectorization operator Vech$(\cdot)$ which collects the lower triangular entries of a symmetric matrix into a vector, the VEC-GARCH model specifies that

$$\underset{\frac{N(N+1)}{2}\times 1}{\text{Vech}\left(\boldsymbol{\Sigma}_t\right)} = \underset{\frac{N(N+1)}{2}\times 1}{\boldsymbol{c}} + \sum_{j=1}^{q} \underset{\frac{N(N+1)}{2}\times\frac{N(N+1)}{2}}{\boldsymbol{A}_j} \text{Vech}\left(\boldsymbol{r}_{t-j}\boldsymbol{r}_{t-j}'\right) + \sum_{j=1}^{p} \underset{\frac{N(N+1)}{2}\times\frac{N(N+1)}{2}}{\boldsymbol{B}_{t-j}} \text{Vech}\left(\boldsymbol{\Sigma}_{t-j}\right) \tag{1.4}$$

The VEC-GARCH model is flexible as it allows all kinds of cross-entry historical dependency for entries in $\boldsymbol{\Sigma}_t$, but is also limited by the difficulty in guaranteeing the positive definiteness of $\boldsymbol{\Sigma}_t$. In addition, the cross-entry historical dependency introduces a large amount of parameters to be estimated, which is $(p+q)\left(\frac{N(N+1)}{2}\right)^2+\frac{N(N+1)}{2}$. Silvennoinen and Teräsvirta (2009) also mentioned that it is numerically demanding to estimate an VEC-GARCH model, also because of the necessity to invert $\boldsymbol{\Sigma}_t$ for all t in each parameter update step.

In specifying the model for $\boldsymbol{\Sigma}_t$, the BEKK model has taken a different approach by directly embedding the positive definiteness of $\boldsymbol{\Sigma}_t$ in its law of motion:

$$\boldsymbol{\Sigma}_t = \boldsymbol{C}\boldsymbol{C}' + \sum_{j=1}^{q}\sum_{i=1}^{M} \boldsymbol{A}_{ij}'\boldsymbol{r}_{t-j}\boldsymbol{r}_{t-j}'\boldsymbol{A}_{ij} + \sum_{j=1}^{p}\sum_{i=1}^{M} \boldsymbol{B}_{ij}'\boldsymbol{\Sigma}_{t-j}\boldsymbol{B}_{ij}, \tag{1.5}$$

where $\boldsymbol{A}_{ij}, \boldsymbol{B}_{ij}$, and $\boldsymbol{C}$ are all parameter matrices to be estimated and $\boldsymbol{C}$ is lower triangular. In addition, the BEKK model also has a smaller number of parameters to be estimated compared to the VEC-GARCH model (in order of N^2 rather than N^4). But as pointed out by R. F. Engle and Kroner (1995), a given BEKK model has several observationally equivalent parameterizations. The authors have given conditions on eliminate redundant parameterizations.

Another way to circumvent the difficulty of ensuring $\boldsymbol{\Sigma}_t$'s positive definiteness is proposed by Kawakatsu (2006), who specifies that

$$\text{Vech}\left(\ln\boldsymbol{\Sigma}_t - \boldsymbol{C}\right) = \sum_{j=1}^{q} \boldsymbol{A}_i\boldsymbol{\varepsilon}_t + \sum_{j=1}^{q} \boldsymbol{F}_j\left(|\boldsymbol{\varepsilon}_t| - \text{E}|\boldsymbol{\varepsilon}_t|\right) + \sum_{j=1}^{p} \boldsymbol{B}_i\text{Vech}\left(\ln\boldsymbol{\Sigma}_{t-j} - \boldsymbol{C}\right), \tag{1.6}$$

where the logarithm of a symmetric matrix is defined indirectly by

$$\exp(\boldsymbol{A}) = \sum_{i=0}^{\infty} \frac{\boldsymbol{A}^i}{i!}. \tag{1.7}$$

The positive definiteness of $\boldsymbol{\Sigma}_t$ is guaranteed by the positive definiteness of $\exp(\boldsymbol{A})$.

The second class of GARCH models is the factor model introduced by R. Engle et al. (1990), who hope to use a small number of possibly correlated factors $f_{1,t}, \cdots, f_{M,t}$ to model the behavior of $\boldsymbol{\Sigma}_t$. The factors follow the GARCH(1,1) processes:

$$f_{m,t} = \omega_m + \alpha_m(\gamma_m' \boldsymbol{r}_{t-1})^2 + \beta_m f_{m,t-1}. \tag{1.8}$$

And the conditional covariance matrix $\boldsymbol{\Sigma}_t$ follows:

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Omega} + \sum_{m=1}^{M} \boldsymbol{\lambda}_m \boldsymbol{\lambda}_m' f_{m,t}. \tag{1.9}$$

R. Engle et al. (1990) have proposed using two factors, one is the return of a stock index and the other is the return of treasury bills, to model the volatilities of individual assets.

A different approach to factor models of multivariate GARCH is the Generalized Orthogonal GARCH (GO-GARCH) by Weide (2002), which uses uncorrelated factors $\boldsymbol{f}_t$ to write $\boldsymbol{r}_t$ as

$$\underset{N\times N}{\boldsymbol{r}_t} = \underset{N\times N}{\boldsymbol{W}} \boldsymbol{f}_t, \tag{1.10}$$

where $\boldsymbol{W}$ is a non-singular. By restricting the factors to be uncorrelated, the GO-GARCH model can avoid the appearance of several correlated factors showing similar characteristics. Each of the uncorrelated factors follows univariate model, and together they are restricted to be orthogonal. The GARCH models for the factors are written as:

$$\boldsymbol{\Sigma}_t^f = \underset{N\times N}{\boldsymbol{I}} - \underset{N\times N}{\boldsymbol{A}} - \underset{N\times N}{\mathbf{B}} + \boldsymbol{A} \odot \left(\boldsymbol{f}_{t-1}\boldsymbol{f}_{t-1}'\right) + \boldsymbol{B}\boldsymbol{\Sigma}_{t-1}^f, \tag{1.11}$$

where $\boldsymbol{\Sigma}_t^f = \mathrm{E}\left(\boldsymbol{r}_t | \mathcal{F}_{t-1}\right)$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are all orthogonal matrices, and $\odot$ represents the element-wise product of matrices.

The third class is known as the correlation models, the simplest form of which is the Constant Conditional Correlation (CCC-) GARCH model by Bollerslev (1990). The CCC-GARCH model decomposes the conditional covariance matrix of $\boldsymbol{r}_t$ into the product of a constant conditional correlation matrix and a diagonal matrix of standard deviations :

$$\boldsymbol{\Sigma}_t = \boldsymbol{D}_t \boldsymbol{P} \boldsymbol{D}_t,$$

where $\boldsymbol{D}_t = \left[\mathrm{diag}\left(\boldsymbol{\Sigma}_t\right)\right]^{1/2}$, and $\boldsymbol{P}$ is the conditional correlation matrix. The conditional variances are modeled as univariate GARCH for each component of $\boldsymbol{r}_t$. Put in a matrix form, the vector $\boldsymbol{d}_t$ of standard deviations follows:

$$\boldsymbol{d}_t = \boldsymbol{\omega} + \sum_{j=1}^{q} \boldsymbol{A}_j \left(\boldsymbol{r}_{t-j} \odot \boldsymbol{r}_{t-j}\right) + \sum_{j=1}^{p} \boldsymbol{B}_j \boldsymbol{d}_{t-1}, \tag{1.12}$$

where $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ are both diagonal matrices. An extension known as Extended CCC-GARCH model, by Jeantheau (1998) drops the diagonal restriction on $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$. With this extra flexibility, the Extended CCC-GARCH model can describe a rich variety of volatility behaviors.

As empirical evidence indicates the assumption of constant conditional correlation may be too restrictive, some researchers have introduced new models to allow $\boldsymbol{P}$ to vary over time. For example, the Varying-Correlation (VC-) GARCH model by Y. K. Tse and Tsui (2002) assume with the decomposition

$$\boldsymbol{\Sigma}_t = \boldsymbol{D}_t \boldsymbol{P}_t \boldsymbol{D}_t,$$

there is

$$\boldsymbol{P}_t = (1 - a - b)\boldsymbol{S} + a \sum_{s=0}^{t-1} \widehat{\boldsymbol{D}}_s^{-1} \boldsymbol{r}_s \left(\widehat{\boldsymbol{D}}_s^{-1} \boldsymbol{r}_s \right)' + b \boldsymbol{P}_{t-1}, \tag{1.13}$$

where $0 \leq a, b < 1$, $a + b < 1$ and $\boldsymbol{S}$ is a constant correlation matrix. That is, the conditional correlation matrix $\boldsymbol{P}_t$ is a combination of a constant matrix, the observed sample correlation matrix, and its past estimate.

Another interesting extension to the CCC-GARCH model is done by Silvennoinen and Teräsvirta (2005), who gave their model the name Smooth Transition Conditional Correlation (STCC-) GARCH. The essential idea of STCC-GARCH is to introduce another variable s_t which controls a convex combination to produce the conditional correlation matrix $\boldsymbol{P}_t$.

$$\boldsymbol{P}_t = (1 - G(s_t))\boldsymbol{P}_{(1)} + G(s_t)\boldsymbol{P}_{(2)}, \tag{1.14}$$

where $\boldsymbol{P}_{(1)}$ and $\boldsymbol{P}_{(2)}$ are both positive definite correlation matrices, and $G(s_t)$ is a function mapping s_t into the interval $[0, 1]$.

Similar to this convex combination approach, the Regime Switching Dynamic Correlation (RSDC-) GARCH model introduced by Pelletier (2006) specifies that the conditional correlation matrix is determined by a Markov chain $\{X_t\}$ with R states $\{1, 2, \cdots, R\}$, such that

$$\boldsymbol{P}_t = \sum_{i=1}^{R} \mathbf{1}_{\{X_t = i\}} \boldsymbol{P}_i. \tag{1.15}$$

In addition, the Markov chain $\{X_t\}$ has transition probability matrix $\boldsymbol{\Pi}$.

Another comprehensive literature review on multivariate GARCH models is Bauwens et al. (2006). In the following part of this subsection, I will move on to reviewing the developments of multivariate Stochastic Volatility (MSV) models. According to Chib et al. (2009), MSV models can also be categorized into three classes, each of which can be seen as the counterpart class of MGARCH models.

The first class, as investigated by researchers including Harvey et al. (1994) and Danìelsson (1998), is a straightforward extension of univariate SV model into multivariate cases. Given a K-dimensional stochastic process $\{\boldsymbol{r}_t\}$, the multivariate extension for SV specifies that

$$\boldsymbol{r}_t = \operatorname{diag}\left[\exp(h_{t,1}/2), \cdots, \exp(h_{t,K}/2)\right] \boldsymbol{\varepsilon}_t, \tag{1.16}$$

$$\boldsymbol{h}_{t+1} = \boldsymbol{\mu} + \boldsymbol{B}(\boldsymbol{h}_t - \boldsymbol{\mu}) + \boldsymbol{\eta}_t, \tag{1.17}$$

where $\boldsymbol{B}$ is a matrix representing the historical dependence of $\boldsymbol{h}_{t+1}$, $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon\varepsilon})$ with $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ being a correlation matrix, and $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\eta\eta})$. Please note in this model no correlations among components of $\boldsymbol{r}_t$ are assumed. Correlations between $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ could be introduced to model the leverage effect. In order to do this, D. Chan et al. (2006) have built the following model:

$$\boldsymbol{r}_t = \operatorname{diag}\left[\exp(h_{t,1}/2), \cdots, \exp(h_{t,1}/2)\right] \boldsymbol{\varepsilon}_t, \tag{1.18}$$

$$\boldsymbol{h}_{t+1} = \boldsymbol{\mu} + \operatorname{diag}\left(B_{11}, \cdots, B_{KK}\right)(\boldsymbol{h}_t - \boldsymbol{\mu}) + \operatorname{diag}\left(\sigma_{11}, \cdots, \sigma_{KK}\right)\boldsymbol{\eta}_t, \tag{1.19}$$

where

$$\begin{bmatrix}\boldsymbol{\varepsilon}_t \\ \boldsymbol{\eta}_t\end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix}\boldsymbol{\Sigma}_{\varepsilon\varepsilon} & \boldsymbol{\Sigma}_{\varepsilon\eta} \\ \boldsymbol{\Sigma}_{\varepsilon\eta} & \boldsymbol{\Sigma}_{\eta\eta}\end{bmatrix}\right). \tag{1.20}$$

Asai and McAleer (2006) have considered a simpler specification for modeling leverage effects, by assuming the matrix $\boldsymbol{\Sigma}_{\varepsilon\eta}$ to be diagonal.

An important extension to the above strand of MSV model is specifying $\boldsymbol{\varepsilon}$ to follow t-distribution in order to describe its heavy tail behavior, as illustrated in Harvey et al. (1994).

As a matter of fact, Harvey et al. (1994) have also considered another class of MSV model, i.e. the class of factor models. The factor model they consider essentially replaces the specification for the n-dimensional vector $\boldsymbol{h}_t$ with

$$\boldsymbol{h}_t = \boldsymbol{\Phi}\boldsymbol{f}_t + \bar{\boldsymbol{f}}, \tag{1.21}$$

where the L-dimensional vector of factors $\boldsymbol{f}_t$ follows a vector auto-regression process and $L \leq K$.

The above factor model focuses on using a small number of factors to control the transition of conditional variance parameters, while assuming the mean of $\boldsymbol{r}_t$ is always $\mathbf{0}$. In contrast to this, the mean factor model by K. Pitt and Shephard (1999) aims at combining a factor model for $\mathrm{E}(\boldsymbol{r}_t|\mathcal{F}_{t-1})$ with an MSV model.

$$\underset{K\times 1}{\boldsymbol{r}_t} = \underset{K\times L}{\boldsymbol{B}}\underset{L\times 1}{\boldsymbol{g}_t} + \operatorname{diag}\left[\exp\left(\frac{h_{t,1}}{2}\right), \cdots, \exp\left(\frac{h_{t,K}}{2}\right)\right]\boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{I}), \tag{1.22}$$

$$\boldsymbol{g}_t = \text{diag}\left[\exp\left(\frac{h_{t,K+1}}{2}\right), \cdots, \exp\left(\frac{h_{t,K+L}}{2}\right)\right]\boldsymbol{\gamma}_t,\ \boldsymbol{\gamma}_t \sim N(\mathbf{0}, \boldsymbol{I}), \tag{1.23}$$

$$\underset{(K+L)\times 1}{\boldsymbol{h}_{t+1}} = \boldsymbol{\mu} + \text{diag}\left(B_{11}, \cdots, B_{K+L,K+L}\right)(\boldsymbol{h}_t - \boldsymbol{\mu}) + \boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\eta\eta}), \tag{1.24}$$

An extension to the above factor model was introduced by Han (2006), who modeled the mean factors $\boldsymbol{g}_t$ with an vector auto-regression process:

$$\boldsymbol{g}_t = \boldsymbol{c} + \boldsymbol{A}\boldsymbol{g}_{t-1} + \text{diag}\left[\exp\left(\frac{h_{t,K+1}}{2}\right), \cdots, \exp\left(\frac{h_{t,K+L}}{2}\right)\right]\boldsymbol{\gamma}_t,\ \boldsymbol{\gamma}_t \sim N(\mathbf{0}, \boldsymbol{I}). \tag{1.25}$$

The last class of MSV models that I am going to cover in this literature review is known as the Dynamic correlation MSV model. For example, Yu and Meyer (2006) considers a bivariate case of the basic MSV model in which the conditional correlation between the two components of $\boldsymbol{r}_t$ is ρ_t. In addition, they write

$$\rho_t = \frac{\exp(q_t) - 1}{\exp(q_t) + 1}, \tag{1.26}$$

where q_t follows an AR(1) process.

Another approach to modeling the correlation matrix is Cholesky decomposition, which is taken by Ruey S Tsay (2005). In the case of bivariate models, this is to write the conditional covariance matrix of $\boldsymbol{r}_t$ as

$$\text{Cov}\left(\boldsymbol{r}_t|\mathcal{F}_{t-1}\right) = \begin{bmatrix} 1 & 0 \\ q_t & 1 \end{bmatrix}\begin{bmatrix} h_{t,1} & 0 \\ 0 & h_{t,2} \end{bmatrix}\begin{bmatrix} 1 & q_t \\ 0 & 1 \end{bmatrix}. \tag{1.27}$$

And again, q_t follows an AR(1) process. The advantage of the Cholesky decomposition approach is its easy extension to high-dimensional cases.

Similar to MGARCH models, the matrix exponentiation can also be applied to MSV models. That is, define $\text{Cov}\left(\boldsymbol{y}_t|\mathcal{F}_{t-1}\right) = \exp\left(\boldsymbol{A}_t\right)$, and then model the entries of $\boldsymbol{A}_t$ through a VAR process:

$$\text{Vech}\left(\boldsymbol{A}_{t+1}\right) = \boldsymbol{\mu} + \boldsymbol{\Psi}\text{Vech}\left(\boldsymbol{A}_t\right) + \boldsymbol{\eta}_t. \tag{1.28}$$

This approach has been explored by Asai, McAleer, and Yu (2006).

And finally, the Inverse Wishart (IW) distribution has also been employed to build MSV models, by Alexander Philipov (2006). Their model is as such:

$$\boldsymbol{r}_t|\boldsymbol{\Sigma}_t \sim \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{I}), \tag{1.29}$$

$$\boldsymbol{\Sigma}_t|\nu, \boldsymbol{W}_{t-1} \sim \text{IW}\left(\nu, \boldsymbol{W}_{t-1}\right), \tag{1.30}$$

$$\boldsymbol{W}_{t-1} = \frac{1}{\nu}\boldsymbol{A}^{1/2}\left(\boldsymbol{\Sigma}_{t-1}^{-1}\right)^{d}\boldsymbol{A}^{1/2\prime}. \tag{1.31}$$

Besides the researched mentioned in this section, Chib et al. (2009) provide a good review for MSV models.

The Development of the MDH Framework

The MDH framework focuses on investigating the positive correlation between return volatilities and trading volumes. It dates back to Clark (1973), who hypothesizes that due to the variations in the rates of information arrivals to the market, the distribution of daily security returns will consist of a random number of homogeneous shocks. The classical central limit theorems would not apply in this case and the distribution of daily returns would be leptokurtic. Clark has also suggested that trading volumes would be good proxies for information arrivals, and found evidence to support MDH in cotton trading data. To be specific, Clark's results support that daily price changes are observations of a stochastic process subordinated to a normal price change process and directed by a log-normal information arrival process. T. W. Epps and M. L. Epps (1976) have proposed an alternative model based on traders' portfolio optimization behavior, in which they assume that for a single information arrival, trader's disagreement regarding the asset's value is positively correlated with the information's impact on asset value, as well as the amount of traders' position adjustment. And thus return volatilities and trading volumes are positively correlated for each information arrival, but the variations in information arrival rates do not play a role as important as the one in the MDH.

Also observing the positive correlation between price variations (defined as squared price changes) and trading volumes, Tauchen and Pitts (1983) have provided a significant theoretical development for the MDH framework based on information arrival, trader reaction and market equilibrium. They model the price changes and trading volumes of treasury bond future contracts as bivariate normal distributions mixed by the number of information arrivals for the security. Since my model is a multivariate extension to theirs, I will review their model in detail. On a trading day right after the arrival of the i-th piece of information to the market, each of the J traders in the market will form a new reservation price P_{ij}^{*} for the traded security. Then they will adjust their positions on the asset to the level

$$Q_{ij} = c \cdot (P_{ij}^{*} - P_i), \tag{1.32}$$

where c is a constant and P_i is the market price of the asset. The market clearing condition then requires $\sum_{j=1}^{J} Q_{ij} = 0$, which gives $P_i = \sum_{j=1}^{J} P_{ij}^{*}/J$. Tauchen and Pitts decompose the change in trader j's reservation price

$\Delta P^*_{ij} = P^*_{ij} - P^*_{i-1,j}$ resulting from the i-th information arrival as

$$\Delta P^*_{ij} = \phi_i + \psi_{ij}, \tag{1.33}$$

where $\phi_i \sim \mathcal{N}(0, \sigma^2_\phi)$ and $\psi_{ij} \sim \mathcal{N}(0, \sigma^2_\psi)$. The price change ΔP_i and trading volume V_i can thus be written as

$$\Delta P_i = \phi_i + \sum_{j=1}^{J} \psi_{ij}/J, \tag{1.34}$$

$$V_i = \frac{c}{2} \sum_{j=1}^{J} |\Delta P^*_{ij} - \Delta P_i|. \tag{1.35}$$

Then for J being large, they derive the asymptotic distribution of ΔP_i and V_i as independent normals with the following means and variances:

$$\mathrm{E}\left(\Delta P_i\right) = 0, \quad \mathrm{Var}\left(\Delta P_i\right) = \sigma^2_\phi + \sigma^2_\psi/J; \tag{1.36}$$

$$\mathrm{E}\left(V_i\right) = \frac{c}{2}\sigma_\psi \sqrt{\frac{2J(J-1)}{\pi}}, \quad \mathrm{Var}\left(V_i\right) = \left(\frac{c}{2}\right)^2 \sigma^2_\psi (1 - \frac{2}{\pi})J + o(J). \tag{1.37}$$

When there is I_t times of information arrivals to the market on day t, the conditional distributions of price change $\Delta P_t = \sum_{i=1}^{I} \Delta P_i$ and trading volume $V_t = \sum_{i=1}^{I} V_i$ are

$$\Delta P_t | I_t \sim \mathcal{N}\left(0, I_t \sqrt{\mathrm{Var}(\Delta P_i)}\right), \tag{1.38}$$

$$V_t | I_t \sim \mathcal{N}\left(I_t E(V_i), I_t \sqrt{\mathrm{Var}(V_i)}\right). \tag{1.39}$$

Using transaction level data, Harris (1987) has devised a method to test the two alternative models, i.e. MDH and the model by T. W. Epps and M. L. Epps (1976). Based on an assumption that transactions occur at a uniform rate in event time, his method compares the statistical properties of daily data and transaction level data (or transaction-adjusted data), including skewness, kurtosis, correlation and autocorrelation, heteroscedasticity and normality. The differences in such statistical properties between the two groups of data are found to be consistent with the predictions of MDH and thus give strong support for MDH.

In addition, Harris (1986) proposed a method to test the cross-security variations in the information arrival processes. Under the null hypothesis that such processes differ across securities, the following statistics should be all positively correlated: return skewness and kurtosis, volume skewness, return-volume correlation, squared-return volume correlation, return and volume heteroscedasticity. Harris has found evidence supporting the null hypothesis from a sample of 479 stocks traded on the New York or American

Stock Exchanges between January 1, 1976 and December 31, 1977. This study would be part the basis on which my research extends the MDH to a multivariate framework.

Rather than considering prices and volumes rising from Walrasian market equilibria, Andersen (1996) has adopted the market microstructure model proposed by Glosten and Milgrom (1985) in which three types of market participants—informed traders, uninformed traders, and a specialist—interact within a game theoretical framework. The specialist provides liquidity to the market by quoting bid and ask prices, and the informed and uninformed traders decide whether to trade one unit of the asset at the quoted prices based on their respective information sets. Using this framework, Andersen has derived the same conditional distribution for returns, but for trading volumes he has proposed using the following Poisson distribution

$$V_t|I_t \sim \text{Poisson}(\mu_{V0} + \mu_{V1} I_t), \tag{1.40}$$

where μ_{V0} and μ_{V1} correspond to the trading volumes generated by uninformed and informed traders, respectively[16].

In addition, Anderson suggests modeling information arrivals using an autoregressive process, because of two reasons: One is that news arrival itself exhibits autocorrelation, and the other one is that information arrivals are closely related to the volatility process, which has been shown to be autocorrelated by the GARCH literature. Then the author finds out the following information arrival process is best supported by the empirical evidence from the common stock trading data of International Business Machines:

$$I_t^{1/2} = \omega + \beta I_{t-1}^{1/2} + \alpha I_{t-1}^{1/2} \varepsilon_t, \tag{1.41}$$

where ε_t is the absolute value of a Generalized Error distributed random variable $\text{GED}_w(0, 1)$.

Estimated using the generalized method of moments, one of Andersen's key findings is that the persistence level of information process is significantly lowered when volume data is added to estimation. Liesenfeld (1998, 2001) has obtained similar empirical results using maximum simulated likelihood based on importance sampling, and hypothesized that this may be due to trading volumes and return volatilities may have different levels of the persistence. To incorporate this into the bivariate mixture model, Liesenfeld postulates that in addition to the information arrival process

$$\ln I_t = \beta_I \ln I_{t-1} + \varepsilon_t, \tag{1.42}$$

[16]Informed trading depends on the number of information arrivals, but uninformed trading does not.

where $\varepsilon_t \sim$ i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$; the parameters σ_ϕ^2 and σ_ψ^2 which represent traders' sensitivity to information arrivals, are both determined by a second latent process H_t in the following way:

$$\ln \sigma_{\phi,t}^2 = \gamma_\phi + \alpha_\phi H_t, \tag{1.43}$$

$$\ln \sigma_{\psi,t}^2 = \gamma_\psi + \alpha_\psi H_t, \tag{1.44}$$

where

$$H_t = \kappa R_{t-1} + \beta_H H_{t-1} + \nu_t, \tag{1.45}$$

with $\nu_t \sim$ i.i.d. $\mathcal{N}(0, \sigma_\nu^2)$. The autocorrelation of H_t is added to reconcile the different persistence levels between return volatilities and trading volumes, while the dependence of H_t on R_{t-1} is used to model the leverage effect, i.e. returns and future conditional variances are negatively correlated.

Park (2011) asserts that the classical MDH model cannot fully capture of the effects of surprising information, which, according the author's interpretation, is the type of information that changes traders' valuations unanimously and dramatically, and at the same time decreases traders' propensity to trade. Being able to access high frequency foreign exchange rates data and compute realized volatilities, the author has used quantile regression to show that the correlation between return volatilities and trading volumes becomes insignificant at high quantiles. Then the author claims that realized volatilities higher than the threshold quantile indicates arrivals of surprise information, which has distinct signed effects on volatilities and trading volumes.

A Literature Review on Stochastic Approximation

In this section I will review some important Stochastic Approximation algorithms. In addition, I will also review its application to the Expectation Maximization algorithm to solve the estimation problem in models with incomplete data.

The Stochastic Approximation algorithm was introduced by Robbins and Monro in 1951 to address an equation-solving problem in which the equation's value cannot be accurately determined. To illustrate the idea of Stochastic Approximation, consider the following equation of a scalar variable x:

$$M(x) = 0,$$

where $M(x_0) = 0$, and $M'(x) > 0$ for all x. With an initial guess $x^{(0)}$, the Newton-Raphson algorithm searches for x_0 recursively through the updating equation:

$$x^{(k+1)} = x^{(k)} - \frac{M\left[x^{(k)}\right]}{M'\left[x^{(k)}\right]}. \tag{1.46}$$

Based on the assumptions on $M(x)$, $x^{(k)} \to x_0$ as $k \to \infty$. However, there are many cases in which the value of $M(x)$ cannot be directly observed. For example, in this dissertation $M(x)$ actually is a high-dimensional integral which cannot be easily tracked. However, a noised observation $f(x)$ of $M(x)$ is available such that

$$f\left[x^{(k+1)}\right] = M\left[x^{(k)}\right] + \varepsilon_k,$$

where ε_k is random variable with $\mathrm{E}(\varepsilon_k) = 0$. The Newton-Raphson algorithm cannot be directly applied to searching for x_0 in this case. As Lai (2003) has explained, under the Newton-Raphson updating scheme, the presence of ε_k implies that

$$x^{(k+1)} = x^{(k)} - \frac{M\left[x^{(k)}\right]}{M'\left[x^{(k)}\right]} - \frac{\varepsilon_k}{M'\left[x^{(k)}\right]}.$$

Should $x^{(k+1)}$ converge to x_0, it requires that $\varepsilon_k \to 0$, which cannot be guaranteed in many applications.

In order to search for x_0 within such a setting, Robbins and Monro (1951) have proposed modifying Equation (1.46), and using

$$x^{(k+1)} = x^{(k)} - \gamma_k f\left[x^{(k)}\right], \tag{1.47}$$

where $\{\gamma_k\}$ are a series of positive constants such that $\sum_{k=0}^{\infty} \gamma_k = \infty$ and $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$. Assume the distribution function of ε_k has finite tails, then the requirement on $\sum_{k=0}^{\infty} \gamma_k^2$ guarantees the almost sure and L_2 convergence of $\sum_{k=0}^{\infty} \gamma_k \varepsilon_k$, such that the noises ε_k's will eventually "cancel out" with each other. On the other hand, the requirement on $\sum_{k=0}^{\infty} \gamma_k$ makes sure that the $x^{(k)}$ will not converge before reaching x_0. A common choice that satisfies such requirements is $\gamma_k = 1/k^{\alpha}$, with $\alpha \in (1/2, 1]$.

An important development to Robbins and Monro's method was introduced by Kiefer and Wolfowitz (1952), who consider an optimization problem with noises in the objective function. Suppose $M(x)$ has a unique minimum obtained at x_0 (for maximum, multiply -1 to $M(x)$) such that $M'(x_0) = 0$, but still we can only observe $f(x)$. The algorithm for searching for x_0 proposed by Kiefer and Wolfowitz is written as follows:

$$x^{(k+1)} = x^{(k)} - \gamma_k \frac{f\left[x^{(k)} + c_k\right] - f\left[x^{(k)} - c_k\right]}{2c_k}, \tag{1.48}$$

where $\{\gamma_k\}$ and $\{c_k\}$ are both series of positive constants such that $c_k \to 0$, $\sum_{k=1}^{\infty} \gamma_k = \infty$, $\sum_{k=1}^{\infty} \gamma_k c_k < \infty$ and $\sum_{k=1}^{\infty} \left(\gamma_k / c_k\right)^2 < \infty$.

The sequence of $\{\gamma_k\}$, known as the gain sequence, are deterministic in the work of both Robbins and Monro (1951) and Kiefer and Wolfowitz (1952). Kesten (1958) introduced an alternative way to design the gain sequence stochastically, in order to improve the finite-sample performance of

the algorithms. The basic idea of Kesten is, if $\left(x^{(k+1)} - x_0\right) - \left(x^{(k)} - x_0\right) = x^{(k+1)} - x^{(k)}$ changes sign frequently, it indicates $x^{(k)}$ is close to x_0. Otherwise, $x^{(k)}$ is still far from x_0. Thus the gain sequence suggested on Kesten depends on the past number of sign changes in $\left(x^{(i+1)} - x^{(i)}\right)$ for $i \leq k$, such that more sign changes decreases γ_k. Kesten suggests such gain sequences may accelerate the convergence of $x^{(k)}$ to x_0.

Blum (1954) has extended both the Robbins-Monro and Kiefer-Wolfowitz algorithms to multivariate cases. Spall (1987, 1992) has introduced an improvement to the Kiefer-Wolfwitz algorithm using simultaneous perturbation. Specifically, for multivariate cases in which the scaler solution x_0 is replaced by a vector solution $\boldsymbol{x}_0 \in \mathbb{R}^p$, the Kiefer-Wolfwitz algorithm given in Equation (1.48) requires $2p$ evaluations of $f(\cdot)$ in each step of parameter update, to calculate an finite-difference approximation for the gradient of $M(\cdot)$, since each dimension of $\boldsymbol{x}$ has to be evaluated twice. Spall's (1987) simultaneous perturbation method requires only two evaluations in each parameter update step, instead. The idea is to replace the deterministic c_k by a random vector $\boldsymbol{\Delta}_k$, all components of which are mutually independent. And the gradient function $\mathrm{D}M(\cdot)$ is approximated by

$$\widehat{\mathrm{D}M}(\boldsymbol{x}^{(k)}) = \begin{bmatrix} \frac{f\left[\boldsymbol{x}^{(k)}+\boldsymbol{\Delta}_k\right]-f\left[\boldsymbol{x}^{(k)}-\boldsymbol{\Delta}_k\right]}{2\boldsymbol{\Delta}_{k,1}} \\ \vdots \\ \frac{f\left[\boldsymbol{x}^{(k)}+\boldsymbol{\Delta}_k\right]-f\left[\boldsymbol{x}^{(k)}-\boldsymbol{\Delta}_k\right]}{2\boldsymbol{\Delta}_{k,p}} \end{bmatrix}. \tag{1.49}$$

Spall (1992) generalizes the above approximation method to perturbing $x^{(k)}$ with q mutually independent random shocks $\boldsymbol{\Delta}_{k1}, \cdots, \boldsymbol{\Delta}_{kq}$, calculating $\widehat{\mathrm{D}M}(\boldsymbol{x}^{(k)})$ for each perturbation and then using the average over all perturbations to approximate $\mathrm{D}M(\cdot)$.

Spall (2000) has introduced an adaptive stochastic approximation method that is closely related to the one used in this research. In the Newton-Raphson algorithm given by Equation (1.46), $M'(\cdot)$ is used to adjust the magnitude of each update step, which guarantees the fast convergence of the algorithm. But such an adjustment is not possible for the Robbins-Monro algorithm, since $M'(\cdot)$ (or the Jacobian matrix for multivariate cases) is not observable. For optimization problems, the Newton-Raphson algorithm also adjust the parameter update step by the second-order derivative (or the Hessian matrix for multivariate cases) of the objective function, which again is unobservable in the Kiefer-Wolfowitz algorithm. To improve the performance of stochastic approximation algorithms, Spall (2000) has suggested that both the Jacobian and Hessian matrices could be estimated using simultaneous perturbations. However, taking the Hessian matrix for an example, the estimated Hessian matrix $\widehat{\boldsymbol{H}}_k$ at step k should not be directly used in adjusting update steps. Instead, Spall suggests using the

average of all historical estimates:

$$\bar{\boldsymbol{H}}_k = \frac{k-1}{k}\widehat{\boldsymbol{H}}_{k-1} + \frac{1}{k}\widehat{\boldsymbol{H}}_k = \frac{1}{k}\sum_{i=1}^{k}\widehat{\boldsymbol{H}}_k, \tag{1.50}$$

for $k = 1, 2, \cdots$ Equation (1.50) is also used by the method introduced by Gu and Kong (1998), which is employed by this research.

The above part has introduced the forms of some general stochastic approximation algorithms. A large amount of literature has contributed to investigating the conditions under which such algorithms would converge to the solution x_0, as well as the asymptotic distribution of $\sqrt{k}\left[x^{(k)} - x_0\right]$. Please see Lai (2003) for a survey on such results. The remaining part of this section will instead focus on stochastic approximation methods that are more closely related to their applications in maximum likelihood estimations in models with incomplete data. In particular, the maximum likelihood estimate used in this research is based on Expectation Maximization (EM), and thus I would start with a brief introduction on EM.

Introduced by Dempster et al. (1977), the EM algorithm is a widely used method in carrying out maximum likelihood estimation in incomplete data models. Suppose there is a family of distribution densities $f(\boldsymbol{x};\boldsymbol{\theta})$, where the true value of parameter vector $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$. In many applications, the complete data $\boldsymbol{x}$ is not fully observable, but what is observable is $\boldsymbol{y}$ given by a many-to-one mapping $g(\cdot)$ such that $\boldsymbol{y} = g(\boldsymbol{x})$. For example, in this research the distributions of securities' returns and trading volumes are determined by unobservable information arrivals. We can only observe the return and volume data, because information arrival is difficult to quantitatively measure. For more information, please see Section 1. In this case, the vector $\boldsymbol{x}$ corresponds to the collection of all variables including the information variables, while the vector $\boldsymbol{y}$ contains only the observed ones.

The likelihood function with observed data $\boldsymbol{y}$ is written as

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \int_{\boldsymbol{x}\in\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}} f(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}.$$

Directly maximizing $L(\boldsymbol{\theta}|\boldsymbol{y})$ to obtain the maximum likelihood estimate may be infeasible due to the intractability of the involved integral, which is usually very high-dimensional. To address such an issue, the EM algorithm transforms the likelihood maximization problem and approaches it in two iterative steps: expectation (the E-step) and maximization (the M-step).

With an given estimate $\widehat{\boldsymbol{\theta}}_k$, the E-step is to compute the expected log-likelihood of the complete data conditional on the observed data:

$$\mathrm{E}\left(\log f(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right) = \int_{\boldsymbol{x}\in\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}} \log f(\boldsymbol{x};\boldsymbol{\theta})\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}\boldsymbol{x},$$

where conditional distribution

$$\pi\left(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}\right)=\frac{f\left(\boldsymbol{x};\boldsymbol{\theta}\right)}{\int_{\boldsymbol{x}\in\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}}f(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}}$$

is derived using Bayes rule.

The M-step is then to search for

$$\widehat{\boldsymbol{\theta}}_{k+1}=\arg\max_{\boldsymbol{\theta}}\mathrm{E}\left(\log f(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right).$$

Then the EM algorithm is carried out by iterating over the expectation and maximization steps alternately until convergence. The EM algorithm guarantees that the likelihood would be improved during the process, which can be seen from the following inequalities:

$$\begin{aligned}
&\log L\left(\widehat{\boldsymbol{\theta}}_{k+1}|\boldsymbol{y}\right)\\
=&\log\int_{\boldsymbol{x}\in\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}}\frac{f(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{k+1})}{\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)}\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}\boldsymbol{x}\\
\geq&\int_{\boldsymbol{x}\in\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}}\log\left[\frac{f(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{k+1})}{\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)}\right]\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}\boldsymbol{x}\\
\geq&\int_{\boldsymbol{x}\in\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}}\log\left[\frac{f(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_k)}{\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)}\right]\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}\boldsymbol{x}\\
=&\log L\left(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{y}\right).
\end{aligned}$$

The first inequality results from the application of Jensen's inequality, while the second one is due to $\widehat{\boldsymbol{\theta}}_{k+1}$ solves

$$\max_{\boldsymbol{\theta}}\int_{\boldsymbol{x}\in\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}}\log f(\boldsymbol{x};\boldsymbol{\theta})\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}\boldsymbol{x}.$$

Dempster et al. (1977) have proved that the sequence $\{\widehat{\boldsymbol{\theta}}_k\}$ produced by the EM algorithm will converge to the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$.

The original EM algorithm introduced above works well if both the expectation and maximization steps have analytical solutions. However, in practice this may not be the case, and more developed methods have to be employed. To address complexities in the maximization step, possible solutions include replacing the original maximization problem with many smaller and simpler conditional maximization problems (Liu and Rubin 1994; Meng and Rubin 1993), or with a single iteration of approximate Newton's method (Lange 1995). On the other hand, to address complexities

in the expectation step, a few other methods have been proposed, which we will now cover in more details.

Celeux and Diebolt (1985) have suggested using one Monte Carlo draw from $\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)$ to replace the expectation step (known as the Stochastic EM algorithm), while Wei and Martin A. Tanner (1990) (see also Martin A Tanner (1996)) have proposed drawing multiple samples from $\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)$ (known as the Monte Carlo or MCEM algorithm). Basically, the E-step at iteration k is replaced by the following simulations step (Delyon et al. (1999) called this the S-step):

$$\widehat{\mathrm{E}}\left(\log f(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)=\frac{1}{m(k)}\sum_{i=1}^{m(k)}\log f\left(\boldsymbol{x}_i^{(k)};\boldsymbol{\theta}\right),$$

where the $\boldsymbol{x}_i^{(k)}$'s are drawn from $\pi\left(\cdot|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)$. For cases in which $\pi\left(\cdot|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)$ takes a complicated form, MCMC algorithms can be employed to draw samples from the target distribution. And the maximization step remains unchanged.

Delyon et al. (1999) have proposed the application Stochastic Approximation to Expectation Maximization problems in which the E-step is not difficult. In particular, they suggest the following approximation

$$\widehat{\mathrm{E}}\left(\log f(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)=(1-\gamma_k)\widehat{\mathrm{E}}\left(\log f(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_{k-1}\right)+\frac{\gamma_k}{m(k)}\sum_{i=1}^{m(k)}\log f\left(\boldsymbol{x}_i^{(k)};\boldsymbol{\theta}\right),$$

where $\{\gamma_k\}$ is the gain sequence, $x^{(k)}$ is still drawn from $\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)$, and the maximization remains unchanged. This algorithm is called by the authors Stochastic Approximation Expectation Maximization (SAEM). The performance of the SAEM algorithm depends on the choices of γ_k and $m(k)$. As suggested by the authors, one should start with relatively large γ_k and gradually decrease it, and with relatively small $m(k)$ and gradually increase it. Such choices are intuitive, since for the first few steps of iterations, $\widehat{\boldsymbol{\theta}}_k$ may be far from the maximum likelihood estimate $\widehat{\theta}$, and thus $\sum_{i=1}^{m(k)}\log f\left(\boldsymbol{x}_i^{(k)};\boldsymbol{\theta}\right)$ is also potentially far from $\mathrm{E}\left(\log f(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)$. Therefore, larger γ_k makes the past estimates $\sum_{i=1}^{m(k)}\log f\left(\boldsymbol{x}_i^{(k)};\boldsymbol{\theta}\right)$ to fade away faster, and larger $m(k)$ makes the latest estimate more and more precise.

The authors argue that since the SAEM algorithm keeps the drawn samples from previous iteration steps and discount their contributions to $\widehat{\mathrm{E}}\left(\log f(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)$ at rates inversely determined by γ_k, it uses the imputed information more efficiently than the MCEM algorithm, which essentially drops all previously drawn samples. For this reason, the SAEM

converges faster than the MCEM algorithm. In addition, their paper also provides an estimate for the Fisher information matrix that is close to the one used in this research.

Next I am going to review the method that this research has employed, which is due to Gu and Kong (1998). From the M-step of EM algorithm we can see that the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ solves the following equation

$$\int_{\boldsymbol{x}\in\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}} \frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}\boldsymbol{x} = \mathbf{0}. \tag{1.51}$$

This can be seen by assuming the set $\{\boldsymbol{x} : \boldsymbol{y} = g(\boldsymbol{x})\}$ does not depend on the parameter $\boldsymbol{\theta}$, which implies the sequence of integration and differentiation can be interchanged, and then applying the first-order condition to the maximization problem in the M-step. The integral in Equation (1.51) is still non-tractable, yet Gu and Kong (1998) have proposed directly apply the Robbins-Monro algorithm to solving it, as long as one can draw noised observations for the equation's left-hand side expression. They also used an estimated Hessian matrix for $L(\boldsymbol{\theta}|\boldsymbol{y})$ to adjust the parameter updating process.

In particular, Gu and Kong's (1998) algorithm is implemented in the following way:

1. Start with the initial guess $\widehat{\boldsymbol{\theta}}_0$ for parameters and $\widehat{\boldsymbol{\Gamma}}_0$ for the Hessian matrix of $L(\boldsymbol{\theta}|\boldsymbol{y})$.

2. With given $\widehat{\boldsymbol{\theta}}_k$, draw samples of $\left\{\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_m^{(k)}\right\}$ from the distribution $\pi\left(\boldsymbol{x}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}_k\right)$. Again, samples could be drawn using appropriate MCMC algorithms.

3. Update $\widehat{\boldsymbol{\Gamma}}$ using the equation

$$\widehat{\boldsymbol{\Gamma}}_{k+1} = (1-\gamma_k)\,\widehat{\boldsymbol{\Gamma}}_k + \gamma_k \widehat{\boldsymbol{H}}_k\left(\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_m^{(k)}; \widehat{\boldsymbol{\theta}}_k\right),$$

where

$$\begin{aligned}
&\widehat{\boldsymbol{H}}_k\left(\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_m^{(k)}; \widehat{\boldsymbol{\theta}}_k\right) \\
=&\frac{1}{m}\sum_{i=1}^{m}\left[\frac{\partial^2 f(\boldsymbol{x}_i^{(k)};\widehat{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{\partial f(\boldsymbol{x}_i^{(k)};\widehat{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\left(\frac{\partial f(\boldsymbol{x}_i^{(k)};\widehat{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\right)'\right] \\
&-\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial f(\boldsymbol{x}_i^{(k)};\widehat{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\right)\right]\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial f(\boldsymbol{x}_i^{(k)};\widehat{\boldsymbol{\theta}}_k)}{\partial\boldsymbol{\theta}}\right)\right]'.
\end{aligned} \tag{1.52}$$

4. Update the estimate for $\boldsymbol{\theta}_0$ through

$$\widehat{\boldsymbol{\theta}}_{k+1} = (1-\gamma_k)\widehat{\boldsymbol{\theta}}_k + \gamma_k \widehat{\boldsymbol{\Gamma}}_k^{-1} \boldsymbol{G}\left(\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_m^{(k)}; \widehat{\boldsymbol{\theta}}_k\right),$$

where

$$\boldsymbol{G}\left(\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_m^{(k)}; \widehat{\boldsymbol{\theta}}_k\right) = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial \log f\left(\boldsymbol{x}_i^{(k)}; \widehat{\boldsymbol{\theta}}_k\right)}{\partial \boldsymbol{\theta}}.$$

5. Iterate Steps 2 to 4 until convergence.

In order to justify the choice for $\widehat{\boldsymbol{H}}_k\left(\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_m^{(k)}; \widehat{\boldsymbol{\theta}}_k\right)$, note Louis (1982)'s missing information principle gives

$$\frac{\partial^2 \log L(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \mathrm{E}\left[\frac{\partial^2 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} | \boldsymbol{y}; \boldsymbol{\theta}\right] + \mathrm{Var}\left[\frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \boldsymbol{y}; \boldsymbol{\theta}\right].$$

Another paper by Gu and Zhu (2001) has reported that the above scheme does not perform well enough for finite update steps when the $\gamma_k = 1/k$ and the initial guess $\widehat{\boldsymbol{\theta}}_0$ is far away from $\widehat{\boldsymbol{\theta}}$. Thus they have combined two stages of updates, where in stage one the parameters are updated with bigger γ_k and in stage two the parameters are updated with $\gamma_k = 1/k$. ($\{\gamma_k\}$ is the optimal choice of gain sequence theoretically.)

A Literature Review on MCMC

In this part, I will review the development of several classical Markov Chain Monte Carlo (MCMC) algorithms, including the Metropolis algorithm (Metropolis et al. 1953), the Metropolis-Hastings algorithm (Hastings 1970) and the Gibbs Sampling algorithm (S. Geman and D. Geman 1984), and also give a brief introduction on the theoretical background of MCMC.

There are many applications in which one may need compute the expected value of a function $h(\cdot)$ of a random variable X:

$$\mathrm{E}\left[h(X)\right] = \int_x h(x) f(x) \mathrm{d}x,$$

where $f(\cdot)$ is the probability density function of X. Sometimes, analytically computing the expectation is infeasible since no analytical form could be found for the integrand $h(\cdot)f(\cdot)$. And the basic idea of Monte Carlo simulation is to use a random number generator to draw i.i.d. samples $\{X_1, \cdots, X_n\}$ that follow the distribution $f(\cdot)$, and approximate the expectation using the sample mean:

$$\mathrm{E}\left[h(X)\right] \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i). \tag{1.53}$$

Under the Strong Law of Large Numbers there is:

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow[n\to\infty]{a.s.} \mathrm{E}\left[h(X)\right].$$

The method given by Equation (1.53) is not difficult to implement if random number generation algorithms have already been designed for the distribution given by $f(\cdot)$. But more commonly the involved distribution itself is complicated such that no random number generator is readily available for drawing samples from it. In even more complicated cases, the involved distribution $f(\cdot)$ is known only up to a proportion, such that $f(x) = c \cdot g(x)$ with only $g(\cdot)$ being known and c being an unknown constant. This is usually the case for posterior distributions in Bayesian analysis, where $c = \int g(x)\mathrm{d}(x)$ and the definite integral here again has no analytical solution. Then the major difficulty in implementing Monte Carlo simulations in such applications is to draw random samples from a possibly complicated distribution which may be known only up to a proportion. MCMC algorithms are devised to address such issues.

MCMC algorithms are based on the idea of rejection sampling. Consider the above mentioned case in which $f(x) = c \cdot g(x)$ for $x \in \mathbb{R}$ with c be an unknown scaling constant, and we are interested in draw samples from the target distribution $f(\cdot)$. Suppose we can find a known constant $\alpha > 1$ and a known probability density function $\varphi(\cdot)$ known as the proposal distribution with random number generator available, such that $\alpha \cdot \varphi(x) \geq g(x)$ for all $x \in \mathbb{R}$. Then in order to sample from $f(\cdot)$, we can first draw a sample from the proposal distribution $\varphi(\cdot)$, and then accept the sample with probability $g(x)/\left[\alpha \cdot \varphi(x)\right]$ and discard the sample with probability $1 - g(x)/\left[\alpha \cdot \varphi(x)\right]$. The probability density function of a sample drawn in this way is proportional to $\varphi(x) \cdot g(x)/\left[\alpha \cdot \varphi(x)\right] = g(x)/\alpha$, which justifies the method of rejection sampling. But it may be difficult to find an appropriate proposal distribution $\varphi(\cdot)$.

While rejection also occurs in MCMC sampling, the requirement of MCMC sampling on the proposal distribution is much less. Suppose we have a proposal distribution $\varphi(\cdot|x)$ with possibly multiple parameters, and one of the parameters is denoted as x[17]. The requirements of Metropolis algorithm (Metropolis et al. 1953) on $\varphi(\cdot)$ is that it has the same support as $f(\cdot)$, and is symmetric in the sense that for all $x^{(0)}$ and $x^{(1)}$ in its support, there is $\varphi(x^{(1)}|x^{(0)}) = \varphi(x^{(0)}|x^{(1)})$. Then the Metropolis algorithm is implemented in the following way: Start with an initial guess sample $X^{(0)}$. In step $i + 1$ with a previous sample $X^{(i)}$, we first draw a

[17]It is not necessary that the proposal distribution $\varphi(\cdot)$ must have x as one of its parameters. If it does not, then resulting MCMC algorithm is called an independent sampler.

proposal sample $\widetilde{X}^{(i+1)}$ from $\varphi(\cdot|X^{(i)})$, then accept the proposed sample by setting $X^{(i+1)} = \widetilde{X}^{(i+1)}$ with probability $\min\left\{f(\widetilde{X}^{(i+1)})/f(X^{(i)}), 1\right\}$, and reject the proposed sample by setting $X^{(i+1)} = X^{(i)}$ with probability $1 - \min\left\{f(\widetilde{X}^{(i+1)})/f(X^{(i)}), 1\right\}$. The sequence of samples $\left\{X^{(0)}, X^{(1)}, \cdots X^{(N)}\right\}$ drawn in this way though are serial correlated, have the following property under some regularity conditions which will be shown later:

$$\frac{1}{n}\sum_{i=0}^{n} h(X^{(i)}) \xrightarrow[n\to\infty]{a.s.} \mathrm{E}\left[h(X)\right], \quad \text{where } X \sim f(\cdot).$$

The intuition underneath the practice of rejection in Metropolis algorithm is that we want to draw more samples from areas where $f(\cdot)$ is large. While any point in the support of $f(\cdot)$ could be a "good" initial sample regardless of $f(X^{(0)})$ being large of not, when we have already had samples in hand, we want to weigh a proposed sample against our current samples by comparing the probability density at the proposal against that at the most recent sample. By accepting the new sample with probability $\min\left\{f(\widetilde{X}^{(i+1)})/f(X^{(i)}), 1\right\}$, we are essentially biased towards placing more samples in areas where $f(\cdot)$ is large. And the "Markov Chain" in the algorithm family's name comes from fact that only the most recent sample plays a part in determining the acceptance probability of the proposal. In addition, the Markov Chain constructed by MCMC algorithms have the target distribution $f(\cdot)$ as its limit distribution (See Roberts and Smith 1994).

A major generalization to the Metropolis algorithm is known as the Metropolis-Hastings algorithm, introduced by by Hastings (1970). The generalization allows the proposal distribution $\varphi(\cdot)$ to be asymmetric, and in order to do this, changes the acceptance probability to

$$\min\left\{f(\widetilde{X}^{(i+1)})\varphi(X^{(i)}|\widetilde{X}^{(i+1)})/\left[f(X^{(i)})\varphi(\widetilde{X}^{(i+1)}|X^{(i)})\right], 1\right\}.$$

For a symmetric proposal distribution $\varphi(\cdot)$, the acceptance probability will reduce to $\min\left\{f(\widetilde{X}^{(i+1)})/f(X^{(i)}), 1\right\}$. Therefore, the Metropolis algorithm is a special case of the Metropolis-Hastings algorithm.

The above introduction has clearly shown that in the MCMC procedure proposed samples are not always accepted. C. Geyer (2011) has suggested that it is a good practice to keep acceptance rate of the simulated chain at around 20%. One can tune other parameters in the proposal distribution to make acceptance rate close to this level. For multimodal target distributions, if the acceptance rate is too high, then the simulated chain may be only exploring a single mode. If the acceptance rate is too low, then the chain may not have enough distinct samples.

Note that though I have not emphasized in the above introduction to Metropolis and Metropolis-Hastings algorithm, they can be applied to drawing samples from multivariate distributions. However, for a high-dimensional distribution, directly applying an MCMC algorithm to it may have acceptance rate very low. The Gibbs sampling algorithm introduced by S. Geman and D. Geman (1984) was designed to address such an issue. For a high-dimensional target distribution $f(x_1, \cdots, x_d)$, suppose the conditional density of X_k given $\{X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_d\}$ is $f(x_d|$ $x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_d)$. Then the Gibbs sampling algorithm divides every Markov Chain step in the usual Metropolis-Hastings algorithm into d smaller stages to update each dimension of $\boldsymbol{X} = (X_1, \cdots, X_d)$ separately. Specifically, in updating the dimension k, Gibbs sampler uses $f\left(\cdot|x_1^{(i+1)}, \cdots, x_{k-1}^{(i+1)}, x_{k+1}^{(i)}, \cdots, x_d^{(i)}\right)$ as the proposal distribution and acceptance rate is constantly at 1. The acceptance rate in stage k can be derived from the following equation:

$$\begin{aligned}
&\min\left\{\frac{f\left(\widetilde{X}_k^{(i+1)}, \boldsymbol{X}_{-k}\right)\cdot f\left(X_k^{(i)}|\boldsymbol{X}_{-k}\right)}{f\left(X_k^{(i)}, \boldsymbol{X}_{-k}\right)\cdot f\left(\widetilde{X}_k^{(i+1)}|\boldsymbol{X}_{-k}\right)}, 1\right\}\\
=&\min\left\{\frac{f\left(\boldsymbol{X}_{-k}\right)}{f\left(\boldsymbol{X}_{-k}\right)}, 1\right\}\\
=&1,
\end{aligned}$$

where $\boldsymbol{X}_{-k} = \left(X_0^{(i+1)}, \cdots, X_{k-1}^{(i+1)}, X_{k+1}^{(i)}, \cdots, X_d^{(i)}\right)$, and the notation f is abused to represent all of joint, conditional and marginal densities. Therefore, Gibbs sampler is also a special case of Metropolis-Hastings algorithm.

In the remaining part of this section I will present some simple theoretical grounds on why MCMC works, but the presentation is not intended to be fully rigorous. Please see Tierney (1994) for details. The basis of MCMC is the Ergodic Theorem for Markov chains. Consider the stochastic process $\{X_n\}$ following a Markov chain with state space $\mathbb{S}$ and transitional probability matrix $\boldsymbol{P}$ such that $P_{ij} = \Pr\left(X_{n+1} = j|X_n = i\right)$.[18] Then a state S_j is accessible from state S_i if $\left(\boldsymbol{P}^n\right)_{ij} > 0$ for some n. Since $\boldsymbol{P}^n$ gives the n-step transition probability matrix, accessibility means that it is probable to reach S_j from S_i in finite steps. If states S_i and S_j are both accessible from each other, then the two states communicate with each other. If any two states in the state space $\mathbb{S}$ communicate with each other, then the Markov chain is irreducible. Irreducibility essentially requires that it is probable to move from any state to any state in finite steps.

For a state S_i in $\mathbb{S}$ that communicate with itself, denote k as the greatest common divisor for all members of the set $\{n : \left(\boldsymbol{P}^n\right)_{ii} > 0\}$. If $k = 1$, then

[18]Here I use a Markov chain with discrete state space to describe the related definitions.

state S_i is aperiodic. If all states in $\mathbb{S}$ are aperiodic, then the Markov chain is aperiodic. Put in another way, aperiodicity means that it is not allowed for a state to return to itself only at multiples of $k \geq 2$ steps.

A state is transient if there is a nonzero probability that it would never return to itself. A state is recurrent if it is not transient. For a recurrent state, if the expected number of steps to return to itself is finite, then the state is positive recurrent. A Markov chain is positive recurrent if all its states are positive recurrent.

If a Markov chain is irreducible, aperiodic and positive recurrent, then there is the following result:

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow[n\to\infty]{a.s.} \int h(x)\pi(x)\mathrm{d}x, \tag{1.54}$$

where $\{X_1, \cdot, X_n\}$ are the first n observations of the Markov chain, and $\pi(\cdot)$ is the invariant distribution of the Markov chain.

MCMC algorithms are designed to make sure the resulted Markov chains satisfy the above three requirements. They are irreducible if the proposal distribution has the same support as the target distribution, then it is possible to go to any point from any point in one step. They are aperiodic because in each step, there is a possibility of rejection so that the chain will return to its current state. And for conditions on positive recurrence, please see Tierney (1994) and K. S. Chan and C. J. Geyer (1994). The final part is to ensure that the target distribution is the invariant distribution of the Markov chain resulting from MCMC algorithms. This is guaranteed by a sufficient condition called detailed balance condition (also known as reversibility condition), which says if a distribution $f(\cdot)$ satisfies:

$$f(S_i)\boldsymbol{P}_{ij} = f(S_j)\boldsymbol{P}_{ji}$$

for all $S_i, S_j \in \mathbb{S}$, then $f(\cdot)$ is the invariant distribution of the Markov chain with transition probability matrix $\boldsymbol{P}$. Metropolis-Hastings algorithm has its transitional kernel and the target distribution meet the detailed balance condition (see Andrieu et al. 2003).

Table 1.1: The Correlation Table of DJIA Component Stocks' Squared Volatilities. (Data Source: Yahoo Finance (`http://finance.yahoo.com`), Continued on Next Page)

	AAPL	AXP	BA	CAT	CSCO	CVX	DD	DIS	GE	GS	HD	IBM	INTC	JNJ	JPM
AAPL	1.00	0.63	0.68	0.60	0.68	0.81	0.66	0.72	0.62	0.75	0.70	0.73	0.78	0.75	0.60
AXP	0.63	1.00	0.76	0.85	0.68	0.67	0.84	0.81	0.88	0.85	0.76	0.73	0.77	0.67	0.90
BA	0.68	0.76	1.00	0.86	0.81	0.88	0.91	0.92	0.76	0.77	0.84	0.82	0.84	0.87	0.68
CAT	0.60	0.85	0.86	1.00	0.74	0.76	0.92	0.88	0.89	0.81	0.81	0.80	0.80	0.79	0.80
CSCO	0.68	0.68	0.81	0.74	1.00	0.81	0.81	0.84	0.72	0.78	0.79	0.78	0.77	0.80	0.67
CVX	0.81	0.67	0.88	0.76	0.81	1.00	0.84	0.91	0.69	0.78	0.80	0.80	0.84	0.97	0.59
DD	0.66	0.84	0.91	0.92	0.81	0.84	1.00	0.95	0.88	0.87	0.89	0.89	0.88	0.86	0.80
DIS	0.72	0.81	0.92	0.88	0.84	0.91	0.95	1.00	0.82	0.85	0.87	0.88	0.88	0.92	0.74
GE	0.62	0.88	0.76	0.89	0.72	0.69	0.88	0.82	1.00	0.91	0.83	0.82	0.81	0.73	0.92
GS	0.75	0.85	0.77	0.81	0.78	0.78	0.87	0.85	0.91	1.00	0.85	0.86	0.84	0.77	0.92
HD	0.70	0.76	0.84	0.81	0.79	0.80	0.89	0.87	0.83	0.85	1.00	0.87	0.87	0.82	0.77
IBM	0.73	0.73	0.82	0.80	0.78	0.80	0.89	0.88	0.82	0.86	0.87	1.00	0.87	0.81	0.78
INTC	0.78	0.77	0.84	0.80	0.77	0.84	0.88	0.88	0.81	0.84	0.87	0.87	1.00	0.84	0.76
JNJ	0.75	0.67	0.87	0.79	0.80	0.97	0.86	0.92	0.73	0.77	0.82	0.81	0.84	1.00	0.60
JPM	0.60	0.90	0.68	0.80	0.67	0.59	0.80	0.74	0.92	0.92	0.77	0.78	0.76	0.60	1.00
KO	0.77	0.68	0.89	0.76	0.79	0.96	0.84	0.90	0.71	0.78	0.80	0.81	0.84	0.95	0.60
MCD	0.77	0.68	0.85	0.74	0.74	0.88	0.80	0.84	0.66	0.72	0.78	0.75	0.83	0.85	0.61
MMM	0.68	0.78	0.88	0.89	0.81	0.84	0.92	0.92	0.85	0.84	0.84	0.86	0.83	0.86	0.76
MRK	0.66	0.72	0.81	0.79	0.71	0.79	0.81	0.80	0.77	0.73	0.79	0.71	0.80	0.81	0.66
MSFT	0.76	0.80	0.85	0.80	0.80	0.88	0.88	0.91	0.80	0.87	0.80	0.86	0.86	0.86	0.77
NKE	0.71	0.78	0.88	0.83	0.80	0.88	0.92	0.92	0.81	0.84	0.88	0.85	0.86	0.88	0.73
PFE	0.73	0.78	0.85	0.84	0.80	0.86	0.89	0.89	0.84	0.85	0.85	0.83	0.84	0.88	0.76
PG	0.73	0.75	0.91	0.81	0.81	0.93	0.90	0.94	0.76	0.79	0.83	0.83	0.85	0.92	0.66
TRV	0.85	0.77	0.73	0.70	0.75	0.85	0.76	0.81	0.76	0.86	0.75	0.76	0.80	0.81	0.76
UNH	0.80	0.77	0.80	0.76	0.75	0.89	0.78	0.81	0.77	0.82	0.74	0.71	0.79	0.87	0.70
UTX	0.76	0.77	0.93	0.86	0.83	0.94	0.92	0.95	0.78	0.82	0.87	0.84	0.87	0.93	0.70
V	0.78	0.78	0.87	0.81	0.79	0.84	0.86	0.88	0.77	0.83	0.85	0.84	0.88	0.83	0.75
VZ	0.73	0.67	0.89	0.76	0.79	0.93	0.87	0.91	0.71	0.76	0.87	0.84	0.86	0.92	0.61
WMT	0.79	0.71	0.89	0.77	0.81	0.93	0.84	0.89	0.71	0.79	0.85	0.81	0.85	0.89	0.64
XOM	0.79	0.63	0.87	0.72	0.79	0.99	0.81	0.88	0.63	0.72	0.76	0.76	0.80	0.94	0.52

Table 1.1 (Continued): The Correlation Table of DJIA Component Stocks' Squared Volatilities.

	KO	MCD	MMM	MRK	MSFT	NKE	PFE	PG	TRV	UNH	UTX	V	VZ	WMT	XOM
AAPL	0.77	0.77	0.68	0.66	0.76	0.71	0.73	0.73	0.85	0.80	0.76	0.78	0.73	0.79	0.79
AXP	0.68	0.68	0.78	0.72	0.80	0.78	0.78	0.75	0.77	0.77	0.77	0.78	0.67	0.71	0.63
BA	0.89	0.85	0.88	0.81	0.85	0.88	0.85	0.91	0.73	0.80	0.93	0.87	0.89	0.89	0.87
CAT	0.76	0.74	0.89	0.79	0.80	0.83	0.84	0.81	0.70	0.76	0.86	0.81	0.76	0.77	0.72
CSCO	0.79	0.74	0.81	0.71	0.80	0.80	0.80	0.81	0.75	0.75	0.83	0.79	0.79	0.81	0.79
CVX	0.96	0.88	0.84	0.79	0.88	0.88	0.86	0.93	0.85	0.89	0.94	0.84	0.93	0.93	0.99
DD	0.84	0.80	0.92	0.81	0.88	0.92	0.89	0.90	0.76	0.78	0.92	0.86	0.87	0.84	0.81
DIS	0.90	0.84	0.92	0.80	0.91	0.92	0.89	0.94	0.81	0.81	0.95	0.88	0.91	0.89	0.88
GE	0.71	0.66	0.85	0.77	0.80	0.81	0.84	0.76	0.76	0.77	0.78	0.77	0.71	0.71	0.63
GS	0.78	0.72	0.84	0.73	0.87	0.84	0.85	0.79	0.86	0.82	0.82	0.83	0.76	0.79	0.72
HD	0.80	0.78	0.84	0.79	0.80	0.88	0.85	0.83	0.75	0.74	0.87	0.85	0.87	0.85	0.76
IBM	0.81	0.75	0.86	0.71	0.86	0.85	0.83	0.83	0.76	0.71	0.84	0.84	0.84	0.81	0.76
INTC	0.84	0.83	0.83	0.80	0.86	0.86	0.84	0.85	0.80	0.79	0.87	0.88	0.86	0.85	0.80
JNJ	0.95	0.85	0.86	0.81	0.86	0.88	0.88	0.92	0.81	0.87	0.93	0.83	0.92	0.89	0.94
JPM	0.60	0.61	0.76	0.66	0.77	0.73	0.76	0.66	0.76	0.70	0.70	0.75	0.61	0.64	0.52
KO	1.00	0.86	0.84	0.80	0.87	0.87	0.85	0.92	0.81	0.89	0.92	0.83	0.91	0.93	0.96
MCD	0.86	1.00	0.77	0.77	0.81	0.83	0.79	0.86	0.76	0.79	0.87	0.88	0.86	0.89	0.89
MMM	0.84	0.77	1.00	0.78	0.84	0.86	0.87	0.85	0.75	0.79	0.91	0.83	0.82	0.82	0.80
MRK	0.80	0.77	0.78	1.00	0.77	0.79	0.82	0.81	0.68	0.80	0.82	0.83	0.81	0.80	0.79
MSFT	0.87	0.81	0.84	0.77	1.00	0.86	0.84	0.88	0.83	0.85	0.89	0.90	0.86	0.86	0.85
NKE	0.87	0.83	0.86	0.79	0.86	1.00	0.85	0.89	0.80	0.80	0.92	0.87	0.89	0.87	0.85
PFE	0.85	0.79	0.87	0.82	0.84	0.85	1.00	0.87	0.80	0.83	0.89	0.88	0.84	0.83	0.82
PG	0.92	0.86	0.85	0.81	0.88	0.89	0.87	1.00	0.78	0.82	0.93	0.86	0.94	0.91	0.91
TRV	0.81	0.76	0.75	0.68	0.83	0.80	0.80	0.78	1.00	0.88	0.83	0.76	0.75	0.78	0.80
UNH	0.89	0.79	0.79	0.80	0.85	0.80	0.83	0.82	0.88	1.00	0.86	0.80	0.79	0.84	0.87
UTX	0.92	0.87	0.91	0.82	0.89	0.92	0.89	0.93	0.83	0.86	1.00	0.88	0.92	0.90	0.92
V	0.83	0.88	0.83	0.83	0.90	0.87	0.88	0.86	0.76	0.80	0.88	1.00	0.85	0.87	0.82
VZ	0.91	0.86	0.82	0.81	0.86	0.89	0.84	0.94	0.75	0.79	0.92	0.85	1.00	0.91	0.92
WMT	0.93	0.89	0.82	0.80	0.86	0.87	0.83	0.91	0.78	0.84	0.90	0.87	0.91	1.00	0.93
XOM	0.96	0.89	0.80	0.79	0.85	0.85	0.82	0.91	0.80	0.87	0.92	0.82	0.92	0.93	1.00

Table 1.2: The List of All DJIA's Component Stocks. Data Source: Yahoo Finance (`http://finance.yahoo.com`).

Sector	Company	
Basic Materials	Chevron Corporation (CVX)	E.I. du Pont de Nemours and Company (DD)
	Exxon Mobil Corporation (XOM)	
Consumer Goods	Apple Inc. (AAPL)	The Coca-Cola Company (KO)
	NIKE, Inc. (NKE)	The Procter & Gamble Company (PG)
Financial	American Express Company (AXP)	The Goldman Sachs Group, Inc. (GS)
	The Travelers Companies, Inc. (TRV)	JPMorgan Chase & Co. (JPM)
	Visa Inc. (V)	
Health Care	Johnson & Johnson (JNJ)	Merck & Co. Inc. (MRK)
	Pfizer Inc. (PFE)	United Health Group Incorporated (UNH)
Industrial Goods	The Boeing Company (BA)	Caterpillar Inc. (CAT)
	General Electric Company (GE)	United Technologies Corporation (UTX)
	3M Company (MMM)	
Services	The Walt Disney Company (DIS)	The Home Depot, Inc. (HD)
	McDonald's Corp. (MCD)	Wal-Mart Stores Inc. (WMT)
Technology	Cisco Systems, Inc. (CSCO)	Intel Corporation (INTC)
	Microsoft Corporation (MSFT)	International Business Machines Corporation (IBM)
	Verizon Communications Inc. (VZ)	

Chapter 2

Model specification

In this Chapter I will extend Tauchen and Pitts's (1983) framework to a multi-security case, in order to study the spillover effects of information flows across different securities. Tauchen and Pitts's model has been introduced in Section 1.4, thus here I will only briefly review the essence of their framework, to facilitate the following multi-security extension.

In the market there is a group of traders who can take long or short positions on a single security. The position they would like to take depends on how they interpret the information they received. All information is public, and arrives to all traders in the market simultaneously. But traders may still have different valuations for the security based on the same information, because of the differences in their preferences. After receiving a new piece of information, a trader examines her new valuation, or reservation price for the security, and takes a position in the security that is proportional to the difference between her reservation price and the security's current market price. The market reaches an equilibrium for the traded security when the sum of all traders' desired positions equals the total supply of the security, which is usually assumed fixed. The model abstracts away from the mechanism through which the market reaches a new equilibrium, but focuses on the effects of the new information arrival on security's price and trading volume. The effect of a new information arrival on an individual trader's reservations prices is decomposed into two parts, one remains the same across all traders and the other is idiosyncratic. Based on this decomposition, the effects of the new information arrival on the security under law of large numbers is to shock its return r_t and trading volume by two independent random variables.[1] When a series of information arrives to the market on a trading day, the daily return and trading volume of the security would then be the aggregate effects of all information arrivals. To sum up, their model describes the joint behavior of a security's daily return and trading volume using the following distributions:

$$r_t|I_t \sim \mathcal{N}(I_t\mu_r, I_t\sigma_r^2), \tag{2.1}$$

$$V_t|I_t \sim \mathcal{N}(I_t\mu_V, I_t\sigma_V^2), \tag{2.2}$$

$$I_t \sim \text{i.i.d. } lognormal(\mu_I, \sigma_I^2), \tag{2.3}$$

where I_t, which is unobservable, is the number of information arrivals on day t.

My plan to extend the above single-security framework is to consider the case in which there are K securities traded on K markets separately. The traders in each market considers trading the security on the corresponding market only. This implies that the traders do not consider doing

[1] In Tauchen and Pitts's (1983) work, they have modeled price changes rather than returns. Liesenfeld (2001) considers logarithmic prices and new information arrivals would then affect returns.

portfolio optimizations across securities. Their positions in a given security is completely determined by the difference between the security's market price and their reservation prices for the security. These assumptions are not completely realistic, but they allow the model to be more tractable. In addition, the following facts can help justifying such assumptions. One is that it is costly to formulate accurate valuations regarding a security. When gathering and analyzing information cost too much, a trader's optimal choice may be specializing in trading only one security, if she tries to buy low and sell high for this security. If the trader intends to hold some security portfolio for a long time instead and is satisfied with receiving the expected return, then she can choose to invest in one of the ETF's introduced in Chapter 1. The ETF market is rapidly growing,[2] and there is an increasingly large amount of ETF products tracking various kinds of market indexes. The existence of such investment vehicles would decrease the necessity for traders to do portfolio optimizations on their own. Unless a trader intends to outperform a market index, she can choose to invest in an ETF tracking the corresponding index.[3]

Based on the assumption that the markets for these K securities are separate, then the following model is built. For each security I assume there is a positive stochastic process $\{I_{tk}\}$, where I_{tk} measures the amount of information arrivals to the market on day t that are related to Security k. Information arrivals then determine the conditional distributions of the securities' daily returns and trading volumes. In particular, by writing $\boldsymbol{Y}_{tk} \coloneqq (R_{tk},\ V_{tk})'$ where R_{tk} and V_{tk} are Security k's return and trading volume on the trading day t respectively, I assume $\boldsymbol{Y}_{tk}$'s distribution conditional on I_{tk} is:

$$\boldsymbol{Y}_{tk} \,\Big|\, I_{tk} \sim \mathcal{N}\left(I_{tk} \cdot \underset{2\times 1}{\boldsymbol{\mu}_k},\ I_{tk} \cdot \underset{2\times 2}{\boldsymbol{\Sigma}_k} \right);\ k = 1, \ldots, K,\ t = 1, \cdots, T \qquad (2.4)$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are stock-specific parameters determining the means and covariance matrices of the distributions. Equation (2.4) is an example of the mixture models explained in Section 1.4, with $\{I_t\}$ being the continuous mixing variable.

The next step is to specify the distributions of information arrivals. As pointed out by Andersen (1996) and Liesenfeld (2001), information arrivals are auto-regressive, because when a news event breaks out and brings new

[2]For example, see the following report by Wall Street Journal on : *Another Milestone-Leaping Year for ETFs*.

[3]Of course, if a trader intends to beat the performance of a market index, she still needs active portfolio optimizations. But there are plenty of reports showing that the majority of actively managed funds actually underperform their benchmark indexes. For example, see the following report by Financial Times on : *Nine out of 10 active funds underperform benchmark*. Actually, the unsatisfactory performance of active asset management is part of the reason why index funds and ETF's were introduced.

information to the market, it tends to develop over a following time period and continues to shock the market with new information within that time period.

Another property that information arrival variables should possess is that they are correlated. As explained in Chapter 1, the cross-security correlation in information arrivals takes two forms. One is that the random shocks to different securities' information arrival processes are contemporarily correlated. If the valuations of some securities are correlated, then when a news event breaks out about one security, shocking its information arrival process with a positive random variable, the information arrival processes of other related securities also tend to be subject to positive shocks. Thus the shocks to information arrival processes are positively correlated.

The second form of correlation is that the information arrival of one security may depend on historical information arrivals of other securities, during to the following reasons. First, as a news event develops, it continues to impact other related securities' information arrivals; and second, it also takes traders time to fully realize the cross-security impacts of the occurred new event.

Based on the above explanations on the forms of cross-security correlations in information arrivals, I have the following specification for the information arrival processes. With the definition $\Lambda_{tk} = \log I_{tk}$ and the notation $\boldsymbol{\Lambda}_t = (\Lambda_{t1}, \ldots, \Lambda_{tK})'$, the dynamics of information flow follows the vector auto-regression process:

$$\underset{K\times 1}{\boldsymbol{\Lambda}_t} = \underset{K\times 1}{\boldsymbol{c}} + \underset{K\times K}{\boldsymbol{B}} \cdot \underset{K\times 1}{\boldsymbol{\Lambda}_{t-1}} + \boldsymbol{\varepsilon}_t, \tag{2.5}$$

where $\boldsymbol{c}$ is a constant term, and the shock term $\boldsymbol{\varepsilon}_t \sim$ i.i.d. $\mathcal{N}(\mathbf{0}, \underset{K\times K}{\boldsymbol{V}})$. The diagonal elements in the coefficient matrix $\boldsymbol{B}$ represent the persistence in the information processes, which already exist in Andersen's (1996) and Liesenfeld's (2001) work. The off-diagonal entries of $\boldsymbol{B}$ represent the cross-security historical dependencies in information arrivals, and those of $\boldsymbol{V}$ corresponds to the contemporary correlations in the shocks to information arrival processes. Based on the previous explanation, we should expect that all entries of the matrices $\boldsymbol{B}$ and $\boldsymbol{V}$ are positive.

Similar to other Mixture Distribution models, the information arrival process $\{\boldsymbol{\Lambda}_t\}$ is not directly observable. This in turn causes the constant term $\boldsymbol{c}$ in Equation (2.5) not identifiable. This can be seen by defining $\widetilde{\boldsymbol{\Lambda}}_t = \boldsymbol{\Lambda}_t + \boldsymbol{d}$, with $\boldsymbol{d}$ being a constant vector. Then $\widetilde{\boldsymbol{\Lambda}}_t$ satisfies the following equation:

$$\widetilde{\boldsymbol{\Lambda}}_t = \widetilde{\boldsymbol{c}} + \boldsymbol{B} \cdot \widetilde{\boldsymbol{\Lambda}}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{2.6}$$

where $\widetilde{\boldsymbol{c}} = \boldsymbol{c} + (\boldsymbol{I} - \boldsymbol{B})\,\boldsymbol{d}$. Similarly, with $\widetilde{I}_{tk} = \exp\left(\widetilde{\lambda}_{tk}\right) = \exp(d_k)\, I_t$, the conditional distributions of $\boldsymbol{Y}_{tk}$ on $\widetilde{I}_t$ can be written as

$$\boldsymbol{Y}_{tk} \,\Big|\, \widetilde{I}_{tk} \sim \mathcal{N}\left(\widetilde{I}_{tk} \cdot \widetilde{\boldsymbol{\mu}_k},\ \widetilde{I}_{tk} \cdot \widetilde{\boldsymbol{\Sigma}_k}\right).$$

By writing $\widetilde{\boldsymbol{\mu}_k} = \exp(-d_k) \cdot \boldsymbol{\mu}_k$ and $\widetilde{\boldsymbol{\Sigma}_k} = \exp(-d_k) \cdot \boldsymbol{\Sigma}_k$, we have the the same conditional distribution for $\boldsymbol{Y}_{tk}$ as the one in (2.4).

Therefore, for the model to be identifiable, the term $\boldsymbol{c}$ in Equation (2.5) has to be dropped. The model to be used in the following estimation is

$$\boldsymbol{Y}_{tk} \,\Big|\, I_{tk} \sim \mathcal{N}\left(I_{tk} \cdot \boldsymbol{\mu}_k,\ I_{tk} \cdot \boldsymbol{\Sigma}_k\right), \tag{2.7}$$

$$\boldsymbol{\Lambda}_t = \boldsymbol{B} \cdot \boldsymbol{\Lambda}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{2.8}$$

where $k = 1, \ldots, K$, and $t = 1, \cdots, T$. Suppose $\widetilde{\boldsymbol{\Lambda}}_t$ in Equation (2.6) is the true value of logarithmic information arrivals, then $\boldsymbol{\Lambda}_t$ in Equation (2.8) is equal to $\widetilde{\boldsymbol{\Lambda}}_t + (\boldsymbol{B} - \boldsymbol{I})^{-1}\,\boldsymbol{c}$.

Chapter 3

Inference method

There are mainly two types of methods used in estimating MDH models. One is maximum likelihood and its simulation-based counterparts, such as the estimation techniques adopted by Tauchen and Pitts (1983), and Liesenfeld (1998, 2001). The other is Generalized Method of Moments, which is adopted by Andersen (1996). Maximum likelihood based methods are more efficient since they assume the parametric form of the distribution family from which data are generated, rather than only assuming some moment conditions. Therefore, in this dissertation, I will adopt an estimation method based on maximum likelihood.

3.1 The Maximum Likelihood Problem

In order to carry out the maximum likelihood estimation, I have to first write the likelihood function in terms of the parameters to be estimated.

The matrices $\{\boldsymbol{\Sigma}_k\}_{k=1..K}$ and $\boldsymbol{V}$ are constrained to be symmetric and positive definite. This allows us to work with their inverses, which turns out to be more convenient. Define

$$\boldsymbol{\Omega}_k \coloneqq \boldsymbol{\Sigma}_k^{-1} = \begin{bmatrix} \Omega_{k11} & \Omega_{k12} \\ \Omega_{k12} & \Omega_{k22} \end{bmatrix}$$

and

$$\boldsymbol{W} \coloneqq \boldsymbol{V}^{-1} = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1K} \\ W_{12} & W_{22} & \cdots & W_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1K} & W_{2K} & \cdots & W_{KK} \end{bmatrix}.$$

Since we only need the upper triangular entries of $\boldsymbol{\Omega}_k$ and $\boldsymbol{W}$, I define $\boldsymbol{\theta} \coloneqq \left(\left\{ \boldsymbol{\mu}_k^T, \mathrm{vech}_U(\boldsymbol{\Omega}_k)^T \right\}_{k=1,\cdots,K}, \mathrm{vech}_F(\boldsymbol{B})^T, \mathrm{vech}_U(\boldsymbol{W})^T \right)^T$ as the vector consisting of all individual parameters, where the operator $\mathrm{vech}_U(\cdot)$ takes a symmetric matrix as its argument and collects the upper triangular entries of the matrix by line and stack them into a column vector. The operator $\mathrm{vech}_F(\cdot)$ does a similar work as $\mathrm{vech}(vech)_U(\cdot)$, except that its stacks all entries of its argument. In addition, assume the true value of $\boldsymbol{\theta}$ is equal to $\boldsymbol{\theta}_0$.

Also, let $\boldsymbol{Y}_t \coloneqq (\boldsymbol{Y}_{t1}, \cdots, \boldsymbol{Y}_{tK})$ be the collection of all securities's return and volume observations on Day t, $\boldsymbol{Y} \coloneqq (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_T)$ be the collection of observations on all days, and $\boldsymbol{\Lambda} \coloneqq (\boldsymbol{\Lambda}_1, \cdots, \boldsymbol{\Lambda}_T)$ be the collection of all securities' latent information variables on all days.

Then the likelihood function is written as:

$$l(\boldsymbol{\theta}|\boldsymbol{Y}) = \int f_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{\lambda}) g_{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \mathrm{d}\boldsymbol{\lambda}, \tag{3.1}$$

where $f_{\boldsymbol{\theta}}(\cdot|\boldsymbol{\lambda})$ is the p.d.f. of $\boldsymbol{Y}$ conditional on $\boldsymbol{\Lambda} = \boldsymbol{\lambda}$ and $g_{\boldsymbol{\theta}}(\cdot)$ is the p.d.f. of $\boldsymbol{\Lambda}$. According to Equations (2.4) and (2.5), the detailed expressions for the two p.d.f.'s are:

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_{t=1}^{T}\prod_{k=1}^{K} \frac{\exp\left\{-\frac{e^{-\lambda_{tk}}}{2}(\boldsymbol{y}_{tk} - e^{\lambda_{tk}}\boldsymbol{\mu}_k)'\boldsymbol{\Omega}_k(\boldsymbol{y}_{tk} - e^{\lambda_{tk}}\boldsymbol{\mu}_k)\right\}}{2\pi e^{\lambda_{tk}}|\boldsymbol{\Omega}_k|^{-1/2}}, \tag{3.2}$$

$$g_{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \prod_{t=1}^{T} \frac{\exp\left\{-\frac{1}{2}\left(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1}\right)'\boldsymbol{W}\left(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1}\right)\right\}}{(2\pi)^{K/2}|\boldsymbol{W}|^{-1/2}}, \tag{3.3}$$

$\boldsymbol{\Lambda}_0$ is assumed to be $\mathbf{0}$ for convenience. For large T, choosing a specific value for $\boldsymbol{\Lambda}_0$ or assuming it is drawn from the invariant distribution of $\boldsymbol{\Lambda}_t$ should not have much difference.

Then, to carry out the maximum likelihood estimation for $\boldsymbol{\theta}_0$, one can solve

$$\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{Y}). \tag{3.4}$$

3.2 Missing Information

Assume $\widehat{\boldsymbol{\theta}}$ solves the maximization problem in (3.4). Finding $\widehat{\boldsymbol{\theta}}$ by directly solving (3.4) is impractical as the high-dimensional integral in Equation (3.1) has no analytical form. However, if we assume that $\widehat{\boldsymbol{\theta}}$ is an interior solution, and the support of $\boldsymbol{\Lambda}$ does not depend on $\boldsymbol{\theta}$, then we can exchange the order of expectation and differentiation, and $\widehat{\boldsymbol{\theta}}$ must satisfy the following first-order condition:

$$\int \frac{\partial h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} \mathrm{d}\boldsymbol{\lambda} = 0, \tag{3.5}$$

where for convenience, I redefine $h_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{\lambda}) = f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\lambda})g_{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ as the joint density function of $\boldsymbol{Y}$ and $\boldsymbol{\Lambda}$.

By dividing and then multiplying $h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})$ the integrand in Equation (3.5), and then dividing both sides of the equation by $\int h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{\lambda}$, we have

$$\int \frac{\frac{\partial h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}}}{h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})} \frac{h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})}{\int h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{\lambda}} \mathrm{d}\boldsymbol{\lambda} = 0, \tag{3.6}$$

where

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\lambda}|\boldsymbol{Y}) = \frac{h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})}{\int h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{\lambda}} \tag{3.7}$$

is actually the p.d.f. of the conditional distribution of $\boldsymbol{\Lambda}$ on $\boldsymbol{Y}$ with given $\boldsymbol{\theta}$. Solving Equation (3.6), rather than solving the maximization problem in (3.4), is then the missing information principle introduced by Louis (1982). In the following part, I will use $\Delta(\boldsymbol{\theta})$ to denote the left hand side of (3.6) such that

$$\Delta(\boldsymbol{\theta}) = \int \frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(\boldsymbol{\lambda}|\boldsymbol{Y}) \mathrm{d}\boldsymbol{\lambda}. \tag{3.8}$$

Note that the expression in (3.8) still involves a high-dimensional integration and we cannot easily evaluate $\Delta(\boldsymbol{\theta})$. Yet Gu and Kong (1998) have proposed to apply the method of stochastic approximation introduced by Robbins and Monro (1951) to solve this equation. The idea is that though we are not able to observe the true value of $\Delta(\boldsymbol{\theta})$, we can obtain an unbiased estimate for it by drawing Monte Carlo simulations from $\pi_{\boldsymbol{\theta}}(\boldsymbol{\lambda}|\boldsymbol{Y})$. Then we can use such estimates for $\Delta(\boldsymbol{\theta})$ to update our estimates for $\boldsymbol{\theta}$.

In particular, given a current estimate $\widehat{\boldsymbol{\theta}}_n$, if we draw m simulations of $\boldsymbol{\Lambda}$ from $\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\lambda}|\boldsymbol{Y})$, denoted as $\boldsymbol{\Lambda}_n \coloneqq (\boldsymbol{\Lambda}_{n,1}, \cdots, \boldsymbol{\Lambda}_{n,s})$, then $\Delta(\widehat{\boldsymbol{\theta}}_n)$ can be estimated as

$$\widehat{\Delta}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\Lambda}_n) = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\Lambda}_{n,i})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \right).$$

In addition, since

$$\frac{\partial \Delta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \mathrm{E}_{\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{Y})} \left(\frac{\partial^2 \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\Lambda})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right) + \mathrm{Var}_{\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{Y})} \left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\Lambda})}{\partial \boldsymbol{\theta}} \right), \tag{3.9}$$

Gu and Kong (1998) have suggested approximating the derivative matrix $\frac{\partial \Delta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$ using:

$$\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\Lambda}_n) = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{\partial^2 \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\Lambda}_{n,i})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\Lambda}_{n,i})}{\partial \boldsymbol{\theta}} \right)^{\otimes 2} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \right] \tag{3.10}$$

$$- \left[\frac{1}{m} \sum_{i=1}^{m} \left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\Lambda}_{n,i})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \right) \right]^{\otimes 2}. \tag{3.11}$$

With the above equations, the estimation procedure can be carried out by the following algorithm.

1. Start with an initial guess $\widehat{\boldsymbol{\theta}}_0$, an arbitrary matrix $\boldsymbol{\Gamma}_0$ and a sequence of step sizes $\{\gamma_n\}$. A typical choice for step size is $\gamma_n = \frac{1}{(n+1)^\alpha}$ with $\alpha \in (0.5, 1]$.

2. Given an estimate $\widehat{\boldsymbol{\theta}}_n$, draw Monte Carlo simulations $\boldsymbol{\Lambda}_n$ from the distribution $\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\lambda}|\boldsymbol{Y})$, the expression of which is given in Equation (3.7). Since we are unable to evaluate the normalizing constant $\int h_{\boldsymbol{\theta}}(\boldsymbol{Y},\boldsymbol{\lambda})\mathrm{d}\boldsymbol{\lambda}$ in Equation (3.7), we employ MCMC algorithms to do sampling from $\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\lambda}|\boldsymbol{Y})$. A classical algorithm to do such samplings is the Metropolis-Hastings algorithm introduced in Section 1.4, which is restated as follows:

   (a) Start with an arbitrary point $\boldsymbol{\Lambda}^{(0)}$, and a proposal distribution $p\left(\cdot|\boldsymbol{\Lambda};\boldsymbol{\beta}\right)$. As indicated by its notation, the proposal distribution may have parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\beta}$. The proposal distribution used in this research is $\mathcal{N}\left(\cdot|\boldsymbol{\Lambda}^{(i)},\sigma^2\boldsymbol{E}\right)$, where $\boldsymbol{E}$ is the identity matrix of order KT, and σ^2 is a parameter to be tuned.

   (b) Given $\boldsymbol{\Lambda}^{(i)}$, draw a new sample $\widetilde{\boldsymbol{\Lambda}}^{(i)}$ from the proposal distribution $p\left(\cdot|\boldsymbol{\Lambda}^{(i)};\boldsymbol{\beta}\right)$.

   (c) With probability

$$\frac{\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\Lambda}^{(i+1)}|\boldsymbol{Y})p\left(\boldsymbol{\Lambda}^{(i)}|\boldsymbol{\Lambda}^{(i+1)};\boldsymbol{\beta}\right)}{\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\Lambda}^{(i)}|\boldsymbol{Y})p\left(\boldsymbol{\Lambda}^{(i+1)}|\boldsymbol{\Lambda}^{(i)};\boldsymbol{\beta}\right)},$$

   accept the proposed new sample by setting $\boldsymbol{\Lambda}^{i+1}=\widetilde{\boldsymbol{\Lambda}}^{i+1}$. And with probability

$$1-\frac{\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\Lambda}^{(i+1)}|\boldsymbol{Y})p\left(\boldsymbol{\Lambda}^{(i)}|\boldsymbol{\Lambda}^{(i+1)};\boldsymbol{\beta}\right)}{\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\Lambda}^{(i)}|\boldsymbol{Y})p\left(\boldsymbol{\Lambda}^{(i+1)}|\boldsymbol{\Lambda}^{(i)};\boldsymbol{\beta}\right)},$$

   reject the proposed new sample by setting $\boldsymbol{\Lambda}^{i+1}=\boldsymbol{\Lambda}^{i}$.

   When the symmetric proposal distribution $\mathcal{N}\left(\cdot|\boldsymbol{\Lambda}^{(i)},\sigma^2\cdot\boldsymbol{E}\right)$ is employed, the probability of accepting the proposed new sample is equal to

$$\frac{\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\Lambda}^{(i+1)}|\boldsymbol{Y})}{\pi_{\boldsymbol{\theta}_n}(\boldsymbol{\Lambda}^{(i)}|\boldsymbol{Y})}.$$

   (d) Repeat steps (b) to (c) until getting m samples $\boldsymbol{\Lambda}^{(0)},\cdots,\boldsymbol{\Lambda}^{(m-1)}$.

   (e) Compute the percentage p_a for proposed samples being accepted among the newly sampled chain $\boldsymbol{\Lambda}^{(0)},\cdots,\boldsymbol{\Lambda}^{(m-1)}$. If p_a is not in the interval $[0.21,0.25]$, then tune the parameter σ^2 and repeat the steps (a) to (d). When the proposal distribution $\mathcal{N}\left(\cdot|\boldsymbol{\Lambda}^{(i)},\sigma^2\cdot\boldsymbol{E}\right)$ is used, increasing σ^2 will decrease p_a and vice versa. The necessity to keep p_a in a range close to 20% has been explained in Section 1.4.

3. Compute $\widehat{\Delta}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\Lambda}_n)$ and $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\Lambda}_n)$, and update $\boldsymbol{\Gamma}_n$ through the equation

$$\boldsymbol{\Gamma}_{n+1} = \gamma_n \boldsymbol{H}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\Lambda}_n) + (1 - \gamma_n)\boldsymbol{\Gamma}_n. \tag{3.12}$$

4. Compute $\widehat{\boldsymbol{\theta}}_{n+1}$ through the equation

$$\widehat{\boldsymbol{\theta}}_{n+1} = \widehat{\boldsymbol{\theta}}_n - \gamma_n \boldsymbol{\Gamma}_{n+1}^{-1} \widehat{\Delta}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\Lambda}_n).$$

5. Iterate through Steps 2 to 4 until convergence of $\widehat{\boldsymbol{\theta}}_n$.

According to Theorem 1 in Gu and Kong (1998), we have $\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \widehat{\boldsymbol{\theta}}$, and $\boldsymbol{\Gamma}_n \xrightarrow{a.s.} \frac{\partial \Delta(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'}$, as n goes to infinity. In addition, we can use $-\boldsymbol{\Gamma}_n^{-1}$ as an estimate of $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$. Recall that I have included the entries of $\boldsymbol{\Sigma}_k = \boldsymbol{\Omega}_k^{-1}$ and $\boldsymbol{W} = \boldsymbol{V}^{-1}$ in $\boldsymbol{\theta}$. I will use the delta method to obtain the covariance matrix of the original parameter vector.

The analytical expressions for the first order and second order partial derivatives $\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\Lambda})}{\partial \boldsymbol{\theta}}$ and $\frac{\partial^2 \log h_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{\Lambda})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ are shown in Section 3.3, and the details of the delta method to derive the covariance matrix of the original parameter vector are shown in Section 3.4.

3.3 Detailed analytical expressions

Since the stochastic approximation procedure for parameter estimation involves many partial derivatives, in this part I will give the detailed expressions for all such derivatives. In notation, I will follow the tradition that $\left[\frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}}\right]_{ij} = \frac{\partial f(\boldsymbol{A})}{\partial A_{ji}}$, where $\boldsymbol{A}$ is a matrix and $f(\cdot)$ is a scalar-valued function.

The log-likelihood function of the complete data is written as:

$$\begin{aligned}\log h_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{\lambda}) = &-\frac{1}{2}\sum_{t=1}^{T}\Big\{(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1})'\boldsymbol{W}(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1}) - \log|\boldsymbol{W}| + K\log 2\pi \\ &+ \sum_{k=1}^{K}\left[\mathrm{e}^{-\lambda_{tk}}(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k)'\boldsymbol{\Omega}_k(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k) + 2\lambda_{tk} - \log|\boldsymbol{\Omega}_k| + 2\log 2\pi\right]\Big\}.\end{aligned} \tag{3.13}$$

For the first-order partial derivatives of $\log h_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{\lambda})$, we have: For $k \in \{1, \cdots, K\}$,

$$\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k} = \sum_{t=1}^{T}(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k)'\boldsymbol{\Omega}_k; \tag{3.14}$$

$$\begin{bmatrix} \frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} & \frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} \\ 0 & \frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{1}{2}\sum_{t=1}^{T}\left[e^{-\lambda_{tk}}(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k)^{\otimes 2} - \boldsymbol{\Omega}_k^{-1}\right]_{11} & -\sum_{t=1}^{T}\left[e^{-\lambda_{tk}}(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k)^{\otimes 2} - \boldsymbol{\Omega}_k^{-1}\right]_{12} \\ 0 & -\frac{1}{2}\sum_{t=1}^{T}\left[e^{-\lambda_{tk}}(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k)^{\otimes 2} - \boldsymbol{\Omega}_k^{-1}\right]_{22} \end{bmatrix}; \tag{3.15}$$

$$\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} = \sum_{t=1}^{T} \boldsymbol{\lambda}_{t-1}(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1})'\boldsymbol{W}; \tag{3.16}$$

$$\begin{bmatrix} \frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{11}} & \cdots & \frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{ij}} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{KK}} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{1}{2}\sum_{t=1}^{T}\left[(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1})^{\otimes 2} - \boldsymbol{W}^{-1}\right]_{11} & \cdots & -\sum_{t=1}^{T}\left[(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1})^{\otimes 2} - \boldsymbol{W}^{-1}\right]_{ij} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{1}{2}\sum_{t=1}^{T}\left[(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1})^{\otimes 2} - \boldsymbol{W}^{-1}\right]_{KK} \end{bmatrix}, \tag{3.17}$$

where $i,\ j \in \{1, \cdots, K\}$ and $i < j$.

And for the second-order partial derivatives of $\log h_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{\lambda})$, we have: For $k \in \{1, \cdots, K\}$,

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' = -\sum_{t=1}^{T} e^{\lambda_{tk}}\boldsymbol{\Omega}_k; \tag{3.18}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_j}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' = \underset{2\times 2}{\mathbf{0}}, \quad \text{for } j \neq k; \tag{3.19}$$

$$\begin{bmatrix} \frac{\partial}{\partial \Omega_{k11}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' & \frac{\partial}{\partial \Omega_{k12}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' \\ \underset{2\times 1}{\mathbf{0}} & \frac{\partial}{\partial \Omega_{k22}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{t=1}^{T}\left(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k\right)_1 & \sum_{t=1}^{T}\left(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k\right)_2 \\ 0 & \sum_{t=1}^{T}\left(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k\right)_1 \\ 0 & 0 \\ 0 & \sum_{t=1}^{T}\left(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k\right)_2 \end{bmatrix}; \tag{3.20}$$

$$\begin{bmatrix} \frac{\partial}{\partial \Omega_{j11}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' & \frac{\partial}{\partial \Omega_{j12}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' \\ \underset{2\times 1}{\mathbf{0}} & \frac{\partial}{\partial \Omega_{j22}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' \end{bmatrix} = \underset{4\times 2}{\mathbf{0}}, \quad \text{for } j \neq k; \tag{3.21}$$

$$\left[\frac{\partial}{\partial B_{ji}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)'\right]_{2K\times K} = \underset{2K\times K}{\mathbf{0}}; \tag{3.22}$$

$$\begin{bmatrix} \frac{\partial}{\partial W_{11}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' & \cdots & \frac{\partial}{\partial W_{ij}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial}{\partial W_{KK}}\left(\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_k}\right)' \end{bmatrix}_{2K\times K} = \underset{2K\times K}{\mathbf{0}}; \tag{3.23}$$

$$\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\mu}_k}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} & \frac{\partial}{\partial \boldsymbol{\mu}_k}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k21}} \\ \underset{1\times 2}{\mathbf{0}} & \frac{\partial}{\partial \boldsymbol{\mu}_k}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \end{bmatrix} \tag{3.24}$$

$$= \begin{bmatrix} \sum_{t=1}^{T}\left(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k\right)_1 & 0 & \sum_{t=1}^{T}\left(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k\right)_2 & \sum_{t=1}^{T}\left(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k\right)_1 \\ 0 & 0 & 0 & \sum_{t=1}^{T}\left(\boldsymbol{y}_{tk} - \mathrm{e}^{\lambda_{tk}}\boldsymbol{\mu}_k\right)_2 \end{bmatrix}; \tag{3.25}$$

$$\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\mu}_j}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} & \frac{\partial}{\partial \boldsymbol{\mu}_j}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k21}} \\ \underset{1\times 2}{\mathbf{0}} & \frac{\partial}{\partial \boldsymbol{\mu}_j}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \end{bmatrix} = \underset{2\times 4}{\mathbf{0}}, \quad \text{for } j \neq k; \tag{3.26}$$

$$\begin{bmatrix} \frac{\partial}{\partial \Omega_{k11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} & \frac{\partial}{\partial \Omega_{k12}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} \\ 0 & \frac{\partial}{\partial \Omega_{k22}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} \end{bmatrix} = \begin{bmatrix} -\frac{T}{2}\frac{\Omega_{k22}^2}{|\boldsymbol{\Omega}_k|^2} & T\frac{\Omega_{k12}\Omega_{k22}}{|\boldsymbol{\Omega}_k|^2} \\ 0 & -\frac{T}{2}\frac{\Omega_{k12}^2}{|\boldsymbol{\Omega}_k|^2} \end{bmatrix}; \tag{3.27}$$

$$\begin{bmatrix} \frac{\partial}{\partial \Omega_{k11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} & \frac{\partial}{\partial \Omega_{k12}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \\ 0 & \frac{\partial}{\partial \Omega_{k22}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \end{bmatrix} = \begin{bmatrix} -\frac{T}{2}\frac{\Omega_{k12}^2}{|\boldsymbol{\Omega}_k|^2} & T\frac{\Omega_{k11}\Omega_{k12}}{|\boldsymbol{\Omega}_k|^2} \\ 0 & -\frac{T}{2}\frac{\Omega_{k11}^2}{|\boldsymbol{\Omega}_k|^2} \end{bmatrix}; \tag{3.28}$$

$$\begin{bmatrix} \frac{\partial}{\partial \Omega_{k11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} & \frac{\partial}{\partial \Omega_{k12}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} \\ 0 & \frac{\partial}{\partial \Omega_{k22}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} \end{bmatrix} = \begin{bmatrix} T\frac{\Omega_{k12}\Omega_{k22}}{|\boldsymbol{\Omega}_k|^2} & -T\frac{\Omega_{k11}\Omega_{k22}+\Omega_{k12}^2}{|\boldsymbol{\Omega}_k|^2} \\ 0 & T\frac{\Omega_{k11}\Omega_{k12}}{|\boldsymbol{\Omega}_k|^2} \end{bmatrix}; \tag{3.29}$$

$$\begin{bmatrix} \frac{\partial}{\partial \Omega_{j11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} & \frac{\partial}{\partial \Omega_{j12}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} \\ 0 & \frac{\partial}{\partial \Omega_{j22}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \Omega_{j11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} & \frac{\partial}{\partial \Omega_{j12}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \\ 0 & \frac{\partial}{\partial \Omega_{j22}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \end{bmatrix} \tag{3.30}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \Omega_{j11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} & \frac{\partial}{\partial \Omega_{j12}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} \\ 0 & \frac{\partial}{\partial \Omega_{j22}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} \end{bmatrix} = \underset{2\times 2}{\mathbf{0}}, \quad \text{for } j \neq k; \tag{3.31}$$

$$\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{B}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} & \frac{\partial}{\partial \boldsymbol{B}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} \\ \underset{K\times K}{\mathbf{0}} & \frac{\partial}{\partial \boldsymbol{B}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \end{bmatrix} = \underset{2K\times 2K}{\mathbf{0}}; \tag{3.32}$$

$$\begin{bmatrix} \frac{\partial}{\partial W_{ij}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k11}} & \frac{\partial}{\partial W_{ij}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k12}} \\ 0 & \frac{\partial}{\partial W_{ij}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \Omega_{k22}} \end{bmatrix} = \underset{2\times 2}{\mathbf{0}}, \quad \text{where } i \leq j; \tag{3.33}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} = \underset{K\times 2K}{\mathbf{0}}; \tag{3.34}$$

$$\begin{bmatrix} \frac{\partial}{\partial \Omega_{k11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} & \frac{\partial}{\partial \Omega_{k12}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} \\ \underset{K\times K}{\mathbf{0}} & \frac{\partial}{\partial \Omega_{k22}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} \end{bmatrix} = \underset{2K\times 2K}{\mathbf{0}}, \tag{3.35}$$

$$\begin{bmatrix} \frac{\partial}{\partial B_{11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} & \cdots & \frac{\partial}{\partial B_{K1}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial B_{1K}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} & \cdots & \frac{\partial}{\partial B_{KK}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1\boldsymbol{w}_1' & \cdots & \boldsymbol{u}_1\boldsymbol{w}_K' \\ \vdots & \ddots & \vdots \\ \boldsymbol{u}_K\boldsymbol{w}_1' & \cdots & \boldsymbol{u}_K\boldsymbol{w}_K' \end{bmatrix}, \tag{3.36}$$

where $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_K$ and $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K$ are the column and row vectors decomposing the following two matrices such that $\sum_{t=1}^{T} -\boldsymbol{\lambda}_{t-1}\boldsymbol{\lambda}_{t-1}' = \begin{bmatrix}\boldsymbol{u}_1 & \cdots & \boldsymbol{u}_K\end{bmatrix}$, and $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K\end{bmatrix}'$;

$$\begin{bmatrix} \frac{\partial}{\partial W_{11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} & \cdots & \frac{\partial}{\partial W_{ij}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} \\ \vdots & \ddots & \vdots \\ \underset{K\times K}{\mathbf{0}} & \cdots & \frac{\partial}{\partial W_{KK}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial \boldsymbol{B}} \end{bmatrix} \tag{3.37}$$

$$= \begin{bmatrix} \sum_{t=1}^{T} \boldsymbol{\lambda}_{t-1}(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1})'\tilde{\boldsymbol{E}}_{11} & \cdots & \sum_{t=1}^{T} \boldsymbol{\lambda}_{t-1}(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1})'\tilde{\boldsymbol{E}}_{ij} \\ \vdots & \ddots & \vdots \\ \underset{K\times K}{\mathbf{0}} & \cdots & \sum_{t=1}^{T} \boldsymbol{\lambda}_{t-1}(\boldsymbol{\lambda}_t - \boldsymbol{B}\boldsymbol{\lambda}_{t-1})'\tilde{\boldsymbol{E}}_{KK} \end{bmatrix}, \tag{3.38}$$

where $\tilde{\boldsymbol{E}}_{ij}$ is the K-order identity matrix whose i-th and j-th columns are interchanged and all other columns are set to zero;[1]

$$\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\mu}_k}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{11}} & \cdots & \frac{\partial}{\partial \boldsymbol{\mu}_k}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{ij}} \\ \vdots & \ddots & \vdots \\ \underset{1\times 2}{\mathbf{0}} & \cdots & \frac{\partial}{\partial \boldsymbol{\mu}_k}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{KK}} \end{bmatrix} = \underset{K\times 2K}{\mathbf{0}}; \tag{3.39}$$

$$\begin{bmatrix} \frac{\partial}{\partial \Omega}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{11}} & \cdots & \frac{\partial}{\partial \Omega}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{ij}} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial}{\partial \Omega}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{KK}} \end{bmatrix} = \underset{K\times K}{\mathbf{0}}, \quad \text{where } \Omega \in \{\Omega_{k11}, \Omega_{k12}, \Omega_{k22}\}; \tag{3.40}$$

$$\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{B}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{11}} & \cdots & \frac{\partial}{\partial \boldsymbol{B}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{ij}} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial}{\partial \boldsymbol{B}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\lambda})}{\partial W_{KK}} \end{bmatrix} \tag{3.41}$$

$$= \begin{bmatrix} \sum_{t=1}^{T} \boldsymbol{e}_1'(\boldsymbol{\lambda}_t - \boldsymbol{B}\lambda_{t-1})\boldsymbol{\lambda}_{t-1}\boldsymbol{e}_1' & \cdots & \sum_{t=1}^{T} \left[\boldsymbol{e}_i'(\boldsymbol{\lambda}_t - \boldsymbol{B}\lambda_{t-1})\boldsymbol{\lambda}_{t-1}\boldsymbol{e}_j' + \boldsymbol{e}_j'(\boldsymbol{\lambda}_t - \boldsymbol{B}\lambda_{t-1})\boldsymbol{\lambda}_{t-1}\boldsymbol{e}_i'\right] \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{t=1}^{T} \boldsymbol{e}_K'(\boldsymbol{\lambda}_t - \boldsymbol{B}\lambda_{t-1})\boldsymbol{\lambda}_{t-1}\boldsymbol{e}_K' \end{bmatrix}, \tag{3.42}$$

$$\tag{3.43}$$

[1] Thus when $i = j$ only the i-th column is kept unchanged.

where $\boldsymbol{e}_k$ is the k-th column of the K-order identity matrix.

$$\begin{bmatrix} \frac{\partial}{\partial W_{11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\theta})}{\partial W_{ii}} & \cdots & \frac{\partial}{\partial W_{mn}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\theta})}{\partial W_{ii}} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial}{\partial W_{KK}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\theta})}{\partial W_{ii}} \end{bmatrix} = \begin{bmatrix} -\frac{T}{2}V_{1i}^2 & \cdots & -TV_{mi}V_{ni} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T}{2}V_{Ki}^2 \end{bmatrix}, \tag{3.44}$$

where $i, m, n \in \{1, 2, \cdots, K\}$, $m < n$, and V_{mi} is the corresponding entity of $\boldsymbol{V} = \boldsymbol{W}^{-1}$;

$$\begin{bmatrix} \frac{\partial}{\partial W_{11}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\theta})}{\partial W_{ij}} & \cdots & \frac{\partial}{\partial W_{mn}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\theta})}{\partial W_{ij}} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial}{\partial W_{KK}}\frac{\partial \log h_{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{\theta})}{\partial W_{ij}} \end{bmatrix} = \begin{bmatrix} -TV_{1i}V_{1j} & \cdots & -T(V_{mi}V_{nj} + V_{ni}V_{mj}) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -TV_{Ki}V_{Kj} \end{bmatrix}, \tag{3.45}$$

where $i, j, m, n \in \{1, 2, \cdots, K\}$, $i < j$ and $m < n$.

3.4 Covariance Matrix for Parameter Estimates

In addition, since the estimation procedure has used entries of the inverses of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{V}$, I need run the delta method to derive the covariance matrix for estimate of the original parameter vector $\boldsymbol{\xi} \coloneqq \Big(\left\{ \boldsymbol{\mu}_k^T, \text{vech}_U(\boldsymbol{\Sigma}_k)^T \right\}_{k=1,\cdots,K}, \text{vech}_F(\boldsymbol{B})^T, \text{vech}_U(\boldsymbol{V})^T \Big)^T$. The delta method is a procedure used to derive the asymptotic distributions of functions of estimated parameters.

In order to apply the delta method, I first have to define a function that relates $\boldsymbol{\xi}$ and the estimated parameter $\boldsymbol{\theta}$. The function is defined as follows: $\boldsymbol{F} : \mathbb{R}^{(2+3)K} \times \mathbb{R}^{K^2} \times \mathbb{R}^{(K^2+K)/2} \to \mathbb{R}^{(2+3)K} \times \mathbb{R}^{K^2} \times \mathbb{R}^{(K^2+K)/2}$ such that for vectors $\boldsymbol{a}_k$, $\boldsymbol{b}_k$, $\boldsymbol{c}$ and $\boldsymbol{d}$ of length 2, 3, K^2 and $(K^2 + K)/2$, respectively, there is:

$$\begin{aligned} &\boldsymbol{F}\left[(\boldsymbol{a}_1^T, \boldsymbol{b}_1^T; \ldots; \boldsymbol{a}_K^T, \boldsymbol{b}_K^T; \boldsymbol{c}^T; \boldsymbol{d}^T)^T\right] \\ &= \begin{bmatrix} \boldsymbol{a}_1 \\ \text{vech}_U\left\{[\text{vech}_U^{-1}(\boldsymbol{b}_1)]^{-1}\right\} \\ \vdots \\ \boldsymbol{a}_K \\ \text{vech}_U\left\{[\text{vech}_U^{-1}(\boldsymbol{b}_K)]^{-1}\right\} \\ \boldsymbol{c} \\ \text{vech}_U\left\{[\text{vech}_U^{-1}(\boldsymbol{d})]^{-1}\right\} \end{bmatrix}. \end{aligned} \tag{3.46}$$

The operator $\text{vech}_U^{-1}(\cdot)$ is the inverse operator of $\text{vech}_U(\cdot)$ so that it takes a vector argument and constructs a symmetric matrix by filling by line its upper triangular entries with the provided vector argument.[2]

Suppose the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ is $\mathcal{N}\left[\boldsymbol{\theta}_0, \text{Var}(\widehat{\boldsymbol{\theta}})\right]$. Then the first-order Taylor extension of $\boldsymbol{F}(\widehat{\boldsymbol{\theta}})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ gives the following approximation:

$$\boldsymbol{F}(\widehat{\boldsymbol{\theta}}) \approx \boldsymbol{F}(\boldsymbol{\theta}_0) + D\boldsymbol{F}(\boldsymbol{\theta}_0) \cdot (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{3.47}$$

where $D\boldsymbol{F}$ is the Jacobian matrix of $\boldsymbol{F}$. The delta method essential says: From Equation (3.47) one can derive the asymptotic distribution of $\widehat{\boldsymbol{\xi}} = \boldsymbol{F}(\widehat{\boldsymbol{\theta}})$ is $\mathcal{N}\big[\boldsymbol{\xi}_0, D\boldsymbol{F}(\boldsymbol{\theta}_0)\text{Var}(\widehat{\boldsymbol{\theta}})\left(D\boldsymbol{F}(\boldsymbol{\theta}_0)\right)^T\big]$, with $\boldsymbol{\xi}_0 = \boldsymbol{F}(\boldsymbol{\theta}_0)$ being the true value of $\boldsymbol{\xi}$. In order to determine the covariance matrix of this normal distribution, we only need to derive the expressions for entries of the Jacobian matrix $D\boldsymbol{F}$, which in the case of Equation (3.46) is block diagonal. In addition, many blocks on the diagonal of $D\boldsymbol{F}(\cdot)$ are identity matrices, except for the ones corresponding to the operator vech_U^{-1}.

The main functional transformation in operator vech_U^{-1} is matrix inversion. For a symmetric invertible matrix $\boldsymbol{A} = [A_{ij}]$, we can write

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}. \tag{3.48}$$

Taking derivatives of both sides of Equation (3.48) with respect to A_{ij}, we have

$$\frac{\partial \boldsymbol{A}}{\partial A_{ij}}\boldsymbol{A}^{-1} + \boldsymbol{A}\frac{\partial \boldsymbol{A}^{-1}}{A_{ij}} = \boldsymbol{0},$$

which gives

$$\frac{\partial \boldsymbol{A}^{-1}}{\partial A_{ij}} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial A_{ij}}\boldsymbol{A}^{-1}. \tag{3.49}$$

Then we can apply Equation (3.49) to deriving the blocks of $D\boldsymbol{F}$ and have

$$\begin{bmatrix} \frac{\partial \Sigma_{k11}}{\partial \Omega_{k11}} & \frac{\partial \Sigma_{k11}}{\partial \Omega_{k12}} & \frac{\partial \Sigma_{k11}}{\partial \Omega_{k22}} \\ \frac{\partial \Sigma_{k12}}{\partial \Omega_{k11}} & \frac{\partial \Sigma_{k12}}{\partial \Omega_{k12}} & \frac{\partial \Sigma_{k12}}{\partial \Omega_{k22}} \\ \frac{\partial \Sigma_{k22}}{\partial \Omega_{k11}} & \frac{\partial \Sigma_{k22}}{\partial \Omega_{k12}} & \frac{\partial \Sigma_{k22}}{\partial \Omega_{k22}} \end{bmatrix}$$

[2]Given a vector of length $(K^2+K)/2$, the output of $\text{vech}_U^{-1}(\cdot)$ is uniquely determined, since the following equation in K has only one positive solution for $n \in \mathbb{N}$:

$$\frac{K^2 + K}{2} = n.$$

The unique positive solution is $K = \left(\sqrt{1+8n} - 1\right)/2$, meaning if the vector argument is of length n, $\text{vech}_U^{-1}(\cdot)$ will produce a symmetric matrix of order $\left(\sqrt{1+8n} - 1\right)/2$.

$$= |\Omega_k|^{-2} \begin{bmatrix} -\Omega_{k22}^2 & 2\Omega_{k12}\Omega_{k22} & -\Omega_{k12}^2 \\ \Omega_{k12}\Omega_{k22} & -\Omega_{k11}\Omega_{k22} - \Omega_{k12}^2 & \Omega_{k11}\Omega_{k12} \\ -\Omega_{k12}^2 & 2\Omega_{k11}\Omega_{k12} & -\Omega_{k11}^2 \end{bmatrix}, \tag{3.50}$$

where $k \in \{1, 2, \cdots K\}$. For entries in $\frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial \text{vec}_U(\boldsymbol{W})}$, there is

$$\frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial W_{ij}} = \begin{cases} -\text{vech}_U\left(\boldsymbol{v}_i \boldsymbol{v}_i'\right), & \text{if } i = j, \\ -\text{vech}_U\left(\boldsymbol{v}_i \boldsymbol{v}_j' + \boldsymbol{v}_j \boldsymbol{v}_i'\right), & \text{otherwise}, \end{cases} \tag{3.51}$$

where $\boldsymbol{v}_i$ is the i-th column vector of $\boldsymbol{V}$, $i, j \in \{1, 2, \cdots K\}$, and $i < j$. In construct the matrix $\frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial \text{vec}_U(\boldsymbol{W})}$, expressions in Equation (3.51) are placed in the following way:

$$\begin{aligned} &\frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial \text{vec}_U(\boldsymbol{W})} \\ &= \left[\frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial W_{11}}, \cdots, \frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial W_{1K}}; \frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial W_{22}}, \cdots, \frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial W_{2K}}; \cdots; \frac{\partial \text{vech}_U(\boldsymbol{V})}{\partial W_{KK}}\right]. \end{aligned} \tag{3.52}$$

So far, all the blocks on the diagonal of $D\boldsymbol{F}$ have been defined, while all other blocks are zero matrices.

Chapter 4

Simulation Studies

Since the method described in Chapter 3, i.e. applying stochastic approximation to a maximum likelihood estimation problem is relatively new, its practical performance have not been extensively studied so far, especially in complicated cases like the multivariate volatility-volume model introduced in Chapter 2. Therefore, in this chapter, I will use Monte Carlo simulations to investigate the method's performance.

The first question of interest is whether the algorithm can run smoothly without moving into un-allowed areas in the parameter space. There are mainly two restrictions on the parameters. One is that the covariance matrices $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ and $\boldsymbol{V}$ have to be positive definite. And the other is that the auto-regression matrix $\boldsymbol{B}$ must have all its eigenvalues in the interval $(-1, 1)$, in order to guarantee that the information arrival process described by Equation (2.5) is covariance stationary. It turned out that there are some parameters in the algorithm that could be tuned to guarantee the smooth updates of the algorithm, which will be explained in detail in Section 4.1.

The convergence of the algorithm is theoretically guaranteed, as explained in Section 1.4 and Chapter 3. Thus as long as it can run without moving into prohibited areas in the parameter space, we do not need worry about its convergence. But we are still interested in whether the parameter estimates produced by the SA algorithm can converge within a reasonable number of steps, and whether confidence intervals produced by the SA algorithm can effectively cover the true values. These two questions will be addressed in Sections 4.2.1.

4.1 Primary Simulation Results

Since the smooth updating of parameter estimates depends on the true parameter value, as well as the parameters in the estimation algorithm, I have used some primary simulations to study how to tune the parameters in the estimation algorithm to keep the algorithm running smoothly. It turned out the smooth running of the algorithm depends primarily on the true values of the parameters $\boldsymbol{B}$, $\boldsymbol{V}$ and the length of MCMC chains m. In particular, for a given m, the algorithm tends to fail when the values of $\boldsymbol{B}$ and $\boldsymbol{V}$ become too extreme, which in turn can be alleviated by increasing m.

4.1.1 Sensitivity of the Algorithm w.r.t. B

The simulations in this part focus on testing the sensitivity of the algorithm regarding $\boldsymbol{B}$, so the other parameters remain fixed as follows:

$$\boldsymbol{\mu}_k = [1, 1]', \ \boldsymbol{\Sigma}_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ for } k = 1, \ 2;$$

$$\boldsymbol{V} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

For all the five settings listed below, I simulated $m = 120000$ samples of information variables in each parameter update step, with each sample being 753-period observations of bivariate normals, i.e. $T = 753$.

1. **The benchmark case:** $\boldsymbol{B} = \begin{bmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{bmatrix}$, with eigenvalues 0.75 and 0.25.

   The algorithm updates estimates smoothly in this case. This case can be interpreted as: The persistence in information flows is at a level comparable to that in other researchers' findings. For example, Andersen (1996) estimated that the persistence coefficient in a model with a single security is between 0.7 to 0.8. In a multi-security model, the information, persistence of the information arrival variables are determined by the eigenvalues of $\boldsymbol{B}$. As the off-diagonal entries of $\boldsymbol{B}$ increase with diagonal entries fixed, the eigenvalues of $\boldsymbol{B}$ will also increase. As mentioned in Chapter 2, the off-diagonal entries of $\boldsymbol{B}$ also correspond to one of the cross-security spillover effects that this research tries to identify.

2. **Cases with increased off-diagonal entries.**

   (a) $\boldsymbol{B} = \begin{bmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{bmatrix}$, with eigenvalues equal to 0.8 and 0.2; and $\boldsymbol{B} = \begin{bmatrix} 0.5 & 0.35 \\ 0.35 & 0.5 \end{bmatrix}$, with eigenvalues equal to 0.85 and 0.15.
   The algorithm can still update parameter estimates smoothly.

   (b) $\boldsymbol{B} = \begin{bmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}$, with eigenvalues equal to 0.9 and 0.1.
   The algorithm fails with simulation number being 120000, because the updated parameters violate the restrictions on the parameter vector. The algorithm can be restored to working as I increase the number of simulations m from 120000 to 200000. However, even with increased m, the estimation error is bigger than the cases with smaller eigenvalues. Especially, the estimate $\widehat{\boldsymbol{\Omega}}_k = \widehat{\boldsymbol{\Sigma}}_k^{-1}$ becomes less accurate for $k = 1, 2$. Estimates for other parameters seem not affected much.
   For example, with the true value $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, after three steps of parameter update, there is $\widehat{\boldsymbol{\Omega}}_1 = \begin{bmatrix} 0.923 & -0.196 \\ -0.196 & 0.923 \end{bmatrix}$ and

$\widehat{\boldsymbol{\Omega}}_2 = \begin{bmatrix} 0.876 & -0.227 \\ -0.227 & 0.911 \end{bmatrix}$. The off-diagonal entries of $\widehat{\boldsymbol{\Omega}}_k$ are more erratic than diagonal entries. And when the simulation number was not big enough (e.g. $m = 120000$), it was also the off-diagonal values that became largely negative, such that their absolute values are bigger than the diagonal elements, making these two matrices no longer positive definite.

For the four settings of $\boldsymbol{B}$ in Cases 1, 2a and2b, the cross-security historical dependency in information arrivals becomes more and more important, and simultaneously the bigger eigenvalue increased from 0.75 in the benchmark case to 0.80, 0.85 and 0.90, meaning the information process becomes more and more persistent. This indicates that to study cases of heavier cross-security historical dependency, one has to use larger number of simulations m.

3. **Cases with more dispersed diagonal entries.**

   (a) $\boldsymbol{B} = \begin{bmatrix} 0.6 & 0.25 \\ 0.25 & 0.4 \end{bmatrix}$, with eigenvalues equal to 0.77 and 0.23, and $\boldsymbol{B} = \begin{bmatrix} 0.65 & 0.25 \\ 0.25 & 0.35 \end{bmatrix}$, with eigenvalues equal to 0.79 and 0.21.
   The algorithm works in these two settings. These cases can be interpreted as: The coefficient of auto-regression differs across securities, causing one security to be more and more persistent while the other one in the opposite direction.

   (b) $\boldsymbol{B} = \begin{bmatrix} 0.8 & 0.25 \\ 0.25 & 0.2 \end{bmatrix}$. The algorithm fails. The eigenvalues are 0.89 and 0.11. In addition, for Security 2, cross-security dependency is even more important and auto-dependency. Again, with increased m, the algorithm can be restored to working.

4. **Cases in which $\boldsymbol{B}$ is asymmetric.**

   $\boldsymbol{B} = \begin{bmatrix} 0.5 & 0.3 \\ 0.2 & 0.5 \end{bmatrix}$ with eigenvalues being 0.74 and 0.26, $\boldsymbol{B} = \begin{bmatrix} 0.5 & 0.35 \\ 0.15 & 0.5 \end{bmatrix}$ with eigenvalues being 0.73 and 0.27, and $\boldsymbol{B} = \begin{bmatrix} 0.5 & 0.45 \\ 0.05 & 0.5 \end{bmatrix}$ with eigenvalues being 0.65 and 0.35.

   With the three settings for $\boldsymbol{B}$ in this case, the cross-security historical dependencies in information variables become increasingly asymmetric. However, the larger eigenvalue of $\boldsymbol{B}$ becomes further away from 1. Not surprisingly, the algorithm also works well in these three settings.

5. **A general case for $\boldsymbol{B}$.**

$\boldsymbol{B} = \begin{bmatrix} 0.6 & 0.2 \\ 0.3 & 0.4 \end{bmatrix}$, with eigenvalues 0.76 and 0.24.

This is the most general setting for $\boldsymbol{B}$ tested so far. The two hypothetical securities have different levels of auto-regression (0.6 versus 0.4), and the cross-security historical dependencies for their information arrivals are also different (0.2 versus 0.3). Since the eigenvalues of $\boldsymbol{B}$ in this case are not too extreme, the estimation algorithm still works well.

From the above simulation results, we can see that the bigger eigenvalue of $\boldsymbol{B}$, which determines the closeness of the information process to non-stationarity, affects the reliability of the estimation algorithm. As the information process approaches non-stationarity, it requires a bigger number of simulations m in each parameter update step, in order to prevent updated parameters from going into un-allowed areas.

4.1.2 Sensitivity of the Algorithm w.r.t. V

Another parameter that affects the reliability of the estimation algorithm is the matrix $\boldsymbol{V}$, which is the covariance matrix of the shocks to information arrival processes. In principle, smaller $\boldsymbol{V}$ should impose no problem to the success of the algorithm, because the information arrival process becomes less noisy with smaller $\boldsymbol{V}$, which in turn makes the unobservability of information arrivals a less severe problem. Therefore, in this part I have focused on testing the reliability of the estimation algorithm in cases with $\boldsymbol{V}$ having bigger entries. For convenience I have copied here the benchmark parameter settings from Section 4.1.1:

$$\boldsymbol{\mu}_k = [1, 1]', \ \boldsymbol{\Sigma}_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ for } k = 1, \ 2;$$
$$\boldsymbol{V} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{bmatrix}.$$

To test the reliability of the estimation algorithm in cases with "bigger" V, I have multiplied the matrix $\boldsymbol{V} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$ by 1.2, 1.4, 1.6, 1.8 and 2 respectively, and keep other parameters unchanged. The algorithm works with $m = 120000$ except in the last case. And again, as I increase m, the algorithm can be restored to working.

Another interesting thing to note is that as the entries of $\boldsymbol{V}$ increase, the accuracy of $\boldsymbol{\Omega}_k$'s are affected the most, while other parameters are still close to their complete model MLE estimates (i.e. when information variables are observable). For example, when $\boldsymbol{V} = \begin{bmatrix} 7.2 & 3.6 \\ 3.6 & 7.2 \end{bmatrix}$, I obtained the estimation

results shown in Table 4.1 after 10 parameter updates. Please note that I have used entries of $\boldsymbol{W} = \boldsymbol{V}^{-1}$ in the table.

Table 4.1: Estimation Results for Increased $\boldsymbol{V}$

	SA	MLE		SA	MLE
$\boldsymbol{\mu}_{1,1}$	1.01	1.00	$\boldsymbol{\Sigma}_{2,22}$	0.64	0.99
$\boldsymbol{\mu}_{1,2}$	1.01	1.00	$\boldsymbol{B}_{11}$	0.46	0.48
$\boldsymbol{\mu}_{2,1}$	1.01	1.00	$\boldsymbol{B}_{12}$	0.26	0.23
$\boldsymbol{\mu}_{2,2}$	1.00	0.99	$\boldsymbol{B}_{21}$	0.25	0.26
$\boldsymbol{\Sigma}_{1,11}$	0.55	1.02	$\boldsymbol{B}_{22}$	0.53	0.52
$\boldsymbol{\Sigma}_{1,12}$	-0.34	0.00	$\boldsymbol{W}_{11}$	0.33	0.35
$\boldsymbol{\Sigma}_{1,22}$	0.59	0.94	$\boldsymbol{W}_{12}$	-0.18	-0.19
$\boldsymbol{\Sigma}_{2,11}$	0.59	1.01	$\boldsymbol{W}_{22}$	0.35	0.38
$\boldsymbol{\Sigma}_{2,12}$	-0.30	-0.01			

Numbers in the columns labeled "MLE" are parameter estimates when information variables are observable.

As we can see from the above simulation results as I vary $\boldsymbol{B}$ and $\boldsymbol{W}$, the estimation algorithm tends to fail if the information process becomes too close to random walk or its shock term has too "big" variances. A common feature shared by these two types of changes is that they make the information process more noisy. Since the MCMC simulations are used to approximate a high-dimensional integral, and its approximation error is determined by the noisiness (or variance) of the unobservable information variables, it makes sense that we need run more simulations to obtain a certain level of approximation accuracy as the unobservable information process becomes more noisy.

4.2 Large Scale Simulations

The previous section has explored different settings of model parameters and algorithm parameters to examine whether the estimation algorithm can update parameter estimates smoothly in various settings. When the estimation algorithm has no problem in updating parameter estimates, the convergence of updated parameter estimates to the true maximum likelihood estimate is theoretically guaranteed. However, it is still of great interest to investigate the practical performance of the algorithm in two aspects. One is whether the algorithm can converge within a practical time period. And the other one is whether the estimated information matrix can provide reliable estimates for the covariance matrix of parameter estimates. This section is used to address these two questions.

4.2.1 Simulation Parameter Settings

To carry out such studies, I have chosen sets of values for each parameter, which are listed below. The primary consideration for choosing such values is that the exhibit enough variations to cover distinct cases, so that the performance of the SA algorithm can be examined in a thorough way. In particular, some parameter values are chosen to be close to estimates given in other researchers' work.

In the general case, the model in Chapter 2 has K securities. The study in this chapter has set $K = 2$.

1. One choice for $\boldsymbol{\mu}$:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 & 1 \end{bmatrix}', \ \boldsymbol{\mu}_2 = \begin{bmatrix} 0.1 & 1 \end{bmatrix}'.$$

   The two components of $\boldsymbol{\mu}_k$ correspond to the expectations of Security K's return and trading volume, respectively, conditional on that the information arrival I is equal to 1. A great amount of researches, including Tauchen and Pitts (1983), Andersen (1996) and Liesenfeld (2001), have shown that for daily data, the conditional expected return is not significantly different from 0, thus μ_{11} is set equal to 0. μ_{21} is set to be slightly positive, in order to introduce some variations. The detrending procedure used in estimation, which will be introduced in Chapter **??** will bring the unconditional mean of trading volume close to 1. Since the expectation of information arrival I is approximately 1, it implies the conditional mean of trading volume is also close to 1. Therefore, both μ_{12} and μ_{22} are set to 1 in parameter choice.

2. Two choices for $\boldsymbol{\Sigma}$:

   (a) $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 0.05 \end{bmatrix}, \ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 0.2 \end{bmatrix}.$

   (b) $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}, \ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.1 \end{bmatrix}.$

   (c) $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.6 & 0 \\ 0 & 0.4 \end{bmatrix}, \ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.3 \end{bmatrix}.$

   (d) $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.2 & 0 \\ 0 & 0.6 \end{bmatrix}, \ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}.$

   It has also been observed by Andersen (1996), Liesenfeld (2001) and Park (2010) among others that the conditional variance of return (Σ_{k11}) is bigger than that of trading volume (Σ_{k22}). Thus all of the above four pairs of $\boldsymbol{\Sigma}_k$ have set Σ_{k11} bigger than Σ_{k22}. The ratio between these two entries takes value from the set $\{1, 2, 3, 4, 5, 10, 20, 40\}$,

which should provide a good coverage for all practical cases. In Tauchen and Pitts (1983)'s original model, the authors have derived that the correlation between returns and trading volumes conditional on information arrival is zero. Other researches on the MDH framework have adopted the same assumption. Thus the above four pairs of $\boldsymbol{\Sigma}_k$ have all set the off-diagonal values equal to 0.

3. Seven choices for $\boldsymbol{B}$:

   (a) $\boldsymbol{B} = \begin{bmatrix} 0.55 & 0.15 \\ 0.15 & 0.55 \end{bmatrix}$, with eigenvalues 0.70 and 0.40.

   (b) $\boldsymbol{B} = \begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.60 \end{bmatrix}$, with eigenvalues 0.80 and 0.50.

   (c) $\boldsymbol{B} = \begin{bmatrix} 0.40 & 0.10 \\ 0.20 & 0.30 \end{bmatrix}$, with eigenvalues 0.50 and 0.20.

   (d) $\boldsymbol{B} = \begin{bmatrix} 0.20 & 0.50 \\ 0.60 & 0.10 \end{bmatrix}$, with eigenvalues 0.70 and -0.40.

   (e) $\boldsymbol{B} = -\begin{bmatrix} 0.20 & 0.50 \\ 0.60 & 0.10 \end{bmatrix}$, with eigenvalues 0.40 and -0.70.

   (f) $\boldsymbol{B} = -\begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.60 \end{bmatrix}$, with eigenvalues -0.50 and -0.80.

   (g) $\boldsymbol{B} = -\begin{bmatrix} 0.40 & 0.10 \\ 0.20 & 0.30 \end{bmatrix}$, with eigenvalues -0.20 and -0.50.

   In their study of uni-asset models, Andersen (1996), Liesenfeld (1998) and Liesenfeld (2001) have found out that in the MDH framework, the persistence of the information arrival process is lower than that of the volatility process in univariate models in which only return data is considered. In particular, their estimated autoregression parameter for the information arrival process is about 0.8 or lower. In a multi-asset MDH framework, the eigenvalues of the $\boldsymbol{B}$ matrix determines the persistence level of the information arrival process. This can be seen by replacing $\boldsymbol{B}$ with its singular value decomposition such that $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$, where $\boldsymbol{D}$ is a diagonal matrix having $\boldsymbol{B}$'s eigenvalues on its diagonal, and the columns of the invertible matrix $\boldsymbol{P}$ are the corresponding eigenvectors of $\boldsymbol{B}$.[1] By multiplying $\boldsymbol{P}^{-1}$ to both sides

[1] Here I assume $\boldsymbol{B}$ is diagonalizable. If it is not, $\boldsymbol{B}$ can be written in Jordan canonical form such that $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1}$, where the columns of $\boldsymbol{P}$ are eigenvectors or generalized eigenvectors of $\boldsymbol{B}$, and the matrix $\boldsymbol{J}$ is no longer diagonal but may have 1's for entries on the diagonal immediately above the main diagonal. Please see Simon and Blume (1994, Chapter 23, p. 605–608) for details on Jordan canonical form. However, in this case, the validity of the analysis that follows are not affected, except that the notation may become more tedious.

of Equation (2.5), we have

$$\boldsymbol{P}^{-1}\boldsymbol{\Lambda}_t = \boldsymbol{D}\boldsymbol{P}^{-1}\boldsymbol{\Lambda}_{t-1} + \boldsymbol{P}^{-1}\boldsymbol{\varepsilon}_t.$$

By defining the $\boldsymbol{X}_t \coloneqq \boldsymbol{P}^{-1}\boldsymbol{\Lambda}_t$ and $\boldsymbol{\eta}_t \coloneqq \boldsymbol{P}^{-1}\boldsymbol{\varepsilon}_t$, we have a new vector autoregression process

$$\boldsymbol{X}_t = \boldsymbol{D}\boldsymbol{X}_{t-1} + \boldsymbol{\eta}_t. \tag{4.1}$$

However, since the matrix $\boldsymbol{D}$ is diagonal, the components of $\boldsymbol{X}_t$ can be solved individually, which can be used to recover $\boldsymbol{\Lambda}_t = \boldsymbol{P}\boldsymbol{X}_t$. In this procedure, it is evident that the diagonal elements of $\boldsymbol{D}$, which are the eigenvalues of $\boldsymbol{B}$ determines the autocorrelation of $\boldsymbol{X}_t$ and thus also the autocorrelation of $\boldsymbol{\Lambda}_t$.

Therefore, I have made choices for $\boldsymbol{B}$ based on its eigenvalues, to make sure they cover persistence levels of the information arrival processes reported in other researches.

The other consideration in choosing $\boldsymbol{B}$ is to introduce distinct combinations of the signs of its two eigenvalues. The above list have covered all three possible cases, i.e. both positive, only one positive and both negative. The inclusion of such combinations is still used to examine the SA algorithm's reliability in practice.

4. Four choices for $\boldsymbol{V}$:

   (a) $\boldsymbol{V} = \begin{bmatrix} 0.2 & 0.01 \\ 0.01 & 0.2 \end{bmatrix}$, with correlation being 0.05.

   (b) $\boldsymbol{V} = \begin{bmatrix} 0.4 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$, with correlation being 0.25.

   (c) $\boldsymbol{V} = \begin{bmatrix} 0.16 & 0.08 \\ 0.08 & 0.25 \end{bmatrix}$, with correlation being 0.4.

   (d) $\boldsymbol{V} = \begin{bmatrix} 0.32 & 0.24 \\ 0.24 & 0.50 \end{bmatrix}$, with correlation being 0.6.

   There are mainly two considerations in setting values for entries in $\boldsymbol{V}$. One is regarding the diagonal values. Liesenfeld (1998) and Liesenfeld (2001) have estimated models that are close to a uni-asset version of my model, and the diagonal values of $\boldsymbol{V}$ are chosen to be close to the parameter estimates in these two papers. However, since other researchers so far have not studied multi-asset specifications close to mine, there are not many established benchmarks to consider in choosing the off-diagonal values for $\boldsymbol{V}$. The values given in the above list are chosen so as to cover a relatively large range of correlations between the random shocks to the information processes, from 0.05 in the least correlated case, to 0.6 to the most correlated case.

In addition, I have made the following choices for other parameters:

1. The length T of the observation period is set to 755, which is approximately equal to the number of trading days in three years.

2. The parameter m, which is the number of MCMC simulations used in each parameter update step, is set to 1.6×10^5. This is a very large number when compared to the suggested range of 10–100 by Gu and Kong (1998) (GK hereafter). But my application of the SA algorithm to the multi-asset MDH model in this paper is very different from the example given in GK's work. First, GK's example essentially has only 200 periods of observations, while my setting is about as 3.5 times long as that. Second, in GK's example, there is only 1 unobservable variable, while mine has 2. Third, unobservable variables in GK's example are i.i.d, but my application has all these variables correlated. And fourth, GK's example has only 1 parameter to be estimated, while my application has 17 for the case with $K = 2$.[2]

   These differences create large computational difficulties for my application. In particular, the SA algorithm when applied to the multi-asset MDH model in this dissertation has to maintain that the updated parameter vector still has positive definite $\boldsymbol{\Sigma}_k$ and $\boldsymbol{V}$. When m is small, this requirement is seldom met, implying that the Monte Carlo estimates for $\boldsymbol{\Delta}(\widehat{\boldsymbol{\theta}})$ and $\partial\boldsymbol{\Delta}(\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ are too non-accurate. In order to avoid this problem, I increased the m to its current level, since the variance of the Monte Carlo estimates for $\boldsymbol{\Delta}(\widehat{\boldsymbol{\theta}})$ and $\partial\boldsymbol{\Delta}(\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ are of order $1/m$.

   Another consideration in choosing m is the associated computational costs. When m is large, it requires more computing time and storage space. The simulation results in Section 4.2.2 show that the SA algorithm always converges with my current choice of m.

3. The MCMC algorithm uses a normal random walk proposal distribution and has maintained acceptance rates of chains between 21% and 25%.

   To be specific, in the n-th parameter update step, given a current sample $\boldsymbol{\Lambda}_{n,i}$ in the MCMC chain, the next sample $\boldsymbol{\Lambda}_{n,i+1}$ is drawn from the distribution $\mathcal{N}\left(\boldsymbol{\Lambda}_{n,i}, \sigma^2 \cdot \boldsymbol{E}_{2T}\right)$, where σ_p is a scalar parameter to be tuned, and $\boldsymbol{E}_{2T}$ is the identity matrix of order $2T$, since there are $2T$ unobservable information variables in my model.

   The procedure for tuning the parameter σ_p in the proposal distribution is designed as follows. σ_p is set equal to a specific value in the first

[2]Given K, the number of parameters to be estimated in my model is $\left(3K^2 + 11K\right)/2$.

parameter update step.[3] Then the MCMC procedure is run with the chosen σ_p to produce m samples of $\mathbf{\Lambda}$, as well as an acceptance rate. If the produced acceptance rate is just between 21% and 25%, then the initial value for σ_p has worked well and the simulated chain is then used for computing the first-step parameter update. However, if the produced acceptance rate is below the range 21%, then the MCMC procedure is re-run with a decreased σ_p, and vice-versa.[4] A binary search algorithm for σ_p is used when two consecutive trials produce two acceptance rates that completely cover the range between 21% to 25%. When an acceptance rate within this range is found, the simulated MCMC chain is then used in computing parameter updates, and the σ_p just used is kept to the next parameter update step as the initial value for MCMC trials.

Choosing an appropriate range for acceptance rates is a non-trivial issue in MCMC. For example, see *Reference problematic here.* I have followed the suggestion by these authors to choose the current range.

In addition, though authors like Brooks et al. (2011) has suggested that it is not necessary to burn-in the simulated chain of samples, I have burned the first 20% of all simulated chains to be more cautionary. The length of unburned MCMC chain is 2.0×10^5, so as to make the after burn-in length to be 1.6×10^5.

4. The number of parameter update steps is set to 30. There is a trade-off between two considerations in choosing this number. One is to run the algorithm long enough to insure its convergence. Since the effectiveness of the SA algorithm essentially depends on a law of large numbers, the longer the algorithm is run, the closer would the parameter estimates be to their limit values, which are the maximum likelihood estimates that would be obtained by solving Problem (3.4). And the other consideration is to maintain the computation costs at a manageable level. Since I have already choose $m = 1.6 \times 10^5$, in each parameter update step, it takes a large amount of computation costs both to generate long a chain of MCMC simulations, and to make calculations based on the simulated chain. The results shown in Section 4.2.2 suggest that the improvements in parameter estimates is quite small in general after 25 steps of parameter updates. Thus 30 seems a conservative choice guaranteeing the convergence of the algorithm,

[3]The choice of this initial value for σ_p depends on the values of $\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \boldsymbol{B}$ and $\boldsymbol{V}$. For example, when the latter four parameters all take their first choice listed above, the initial value for σ_p is set at 0.0125.

[4]It has turned out that the acceptance rate depends negatively on σ_p, which is not difficult to understand, since the normal proposal distribution with a small σ_p tends to draw samples close to the current sample, which in turn keeps the acceptance rate high.

which is also computationally viable.

5. The starting point of parameter update is equal to be the parameter true value $\boldsymbol{\theta}_0$ plus a normal random shock, with mean zero and standard deviation equal to 0.05 multiplying the absolute value of $\boldsymbol{\theta}_0$.

The simulations are then generated in this way: For each choice of parameters, simulate the information arrival process, which will be used to determine the conditional distributions of returns and trading volumes. Then returns and trading volumes are simulated, which are then fed into the SA algorithm to produce parameter estimates. Based on the above numbers of choices for each parameter, there are 112 such simulations in total.

4.2.2 Convergence of the SA Algorithm

After fully describing the settings used for running simulations, I can now present the simulation results. The most desirable result is that the parameter estimates $\widehat{\boldsymbol{\theta}}$ produced by the SA algorithm can quickly converge to the true maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_{MLE}$ obtained from solving Problem (3.4). The convergence is theoretically guaranteed, but cannot be easily verified in practice, since the direct solution to Problem (3.4) is not practically feasible in the problem studied in this research. Therefore, the following part instead focuses on presenting the changes in $\widehat{\boldsymbol{\theta}}$, the difference between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, and the difference between $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^c_{MLE}$, where $\widehat{\boldsymbol{\theta}}^c_{MLE}$ is the maximum likelihood estimate of the complete model in which the information variables are observable. Please note that since in the incomplete model with unobservable information variables, $\widehat{\boldsymbol{\theta}}$ will be different from $\widehat{\boldsymbol{\theta}}^c_{MLE}$ in general due to the lost information, there is no guarantee that $\widehat{\boldsymbol{\theta}}$ will converge to $\widehat{\boldsymbol{\theta}}^c_{MLE}$.

The first choice of update steps γ_n's is defined by

$$\gamma_n = 1/n. \tag{4.2}$$

And the second choice of update steps is defined by

$$\gamma_n = \frac{b/n + a}{b/n + a + n - 1}.$$

where $a = 1$ and $b = 2000$.

4.2.3 Comparison of Update Steps

As noted in the literature, one of the main factor that affects the performance of the stochastic approximation algorithm is the sequence of update

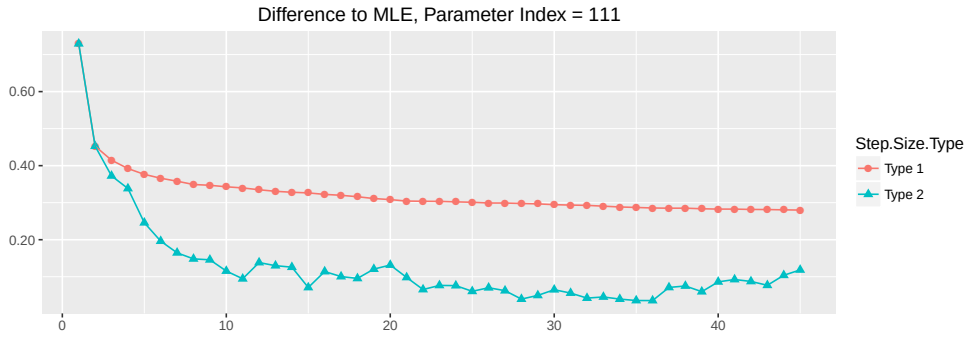


Figure 4.1: Convergence of the Stochastic Approximation Algorithm

steps. Unfortunately, there is no computationally tractable way to construct the optimal series of update steps systematically, and the choice of update steps are mainly ad hoc. The convergence results shown in Figure 4.1 suggest that the update steps defined by $\gamma_n = 1/n$ probably might have decreased to zero too fast when $\widehat{\boldsymbol{\theta}}_n$ was still far from $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\boldsymbol{\theta}}^c_{MLE}$. Thus I tried an alternative way to define update steps which is suggested by Powell (2007):

$$\gamma_n = \frac{b/n + a}{b/n + a + n - 1}, \tag{4.3}$$

where a and b are tunable positive parameters. $\{\gamma_n\}$ given Equation (4.3) satisfies the restrictions on update steps, such that

$$\sum_{n=1}^{\infty} \gamma_n = \infty, \text{ and } \sum_{n=1}^{\infty} \gamma_n^2 < \infty.$$

The parameters I actually used are $a = 1$ and $b = 2000$ which are also included in Powell's work. In order to show the difference in the behaviors of step sizes defined in Equations (4.2) and (4.3), I have plotted them in Figure 4.2. Clearly, the updating steps given by Equation (4.3) decreases to zero much slower than those given in Equation (4.2).

The performances of the stochastic approximation algorithm with these two schemes of step sizes are also compared, as shown in Figures 4.3 and 4.4. In these figures, a label having the form of "xyz" indicates that $\boldsymbol{\Sigma}$, $\boldsymbol{B}$ and $\boldsymbol{V}$ taking their x-th, y-th and z-th settings, respectively. I have chosen two benchmarks for comparison: One is the true parameter values, and the other is the complete model maximum likelihood estimates. There is no guarantee that parameter estimates should converge to either of these two values, but generally they should not be too far from them.

From Figure 4.3 we can see that for the parameter settings labeled as "111", step sizes given in Equation (4.3) performs better than those given

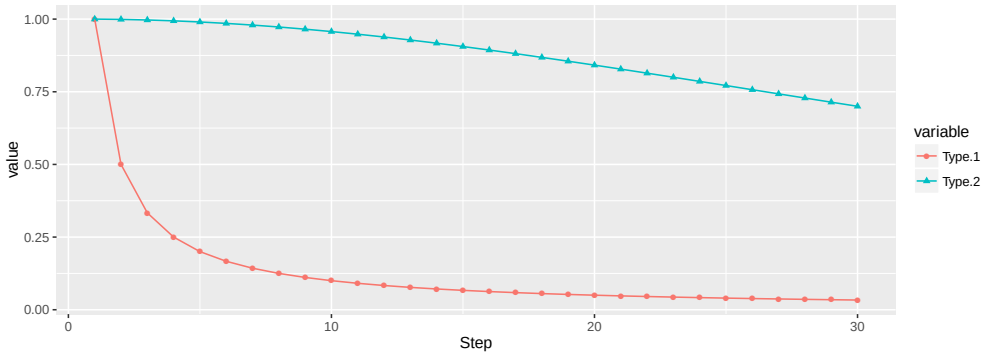


Figure 4.2: Comparison of Step Sizes.

by Equation (4.2). Because the initial parameter estimate starts relatively far from the true parameters, and steps sizes in (4.2) decreases to 0 too fast, the parameter estimates cannot be updated quickly enough by the simulated gradients. On the contrary, the step sizes given in (4.3) stay close to 1 for an extended period, and thus allows the influence of a bad initial guess to fade away more quickly.

However, for the remaining three settings labeled as "242", "313" and "444", the step sizes given in (4.3) fail to provide better performances. They cannot make parameter estimates closer to the true parameter values. Their inferior performance appear to be more apparent for parameter sets "242" and "313", for which the distance between parameter estimates with update steps given by (4.2) is still decreasing. However, we should note that the sub-figures in Figure 4.3 are plotted in different scales, such that the initial distance for parameter set "111" is much bigger.

Figure 4.4 provides very similar results, that if the initial parameter estimates is far from the true value, update steps given in (4.3) seem capable of quickly moving parameter estimates into an area that is close to the true parameter values. But if the initial guess is already close the to true parameter value, then update steps given in (4.2) are more stable in keeping parameter estimates within a small area around the true parameter values. The unstable performance of update steps given in Equation (4.3) may be due to that fact that the matrix $\mathbf{\Gamma}_n$ in Equation (3.12) is an approximation to Hessian matrix in the Newton-Raphson algorithm, and it is also updated using the same update steps. When (4.3) discounts historical values at a faster rate, it also drops too much information in the historical approximations of $\mathbf{\Gamma}_n$, making the direction of parameter update not as accurate as those using (4.2).

In principle, one can use different update steps for $\widehat{\boldsymbol{\theta}}_n$ and $\mathbf{\Gamma}_n$. Or alternatively, one can use a piecewise function to force quick parameter updates, and then switch to Equation (4.2). In the following part of large scale simulation, as well as in real data estimation presented in Chapter

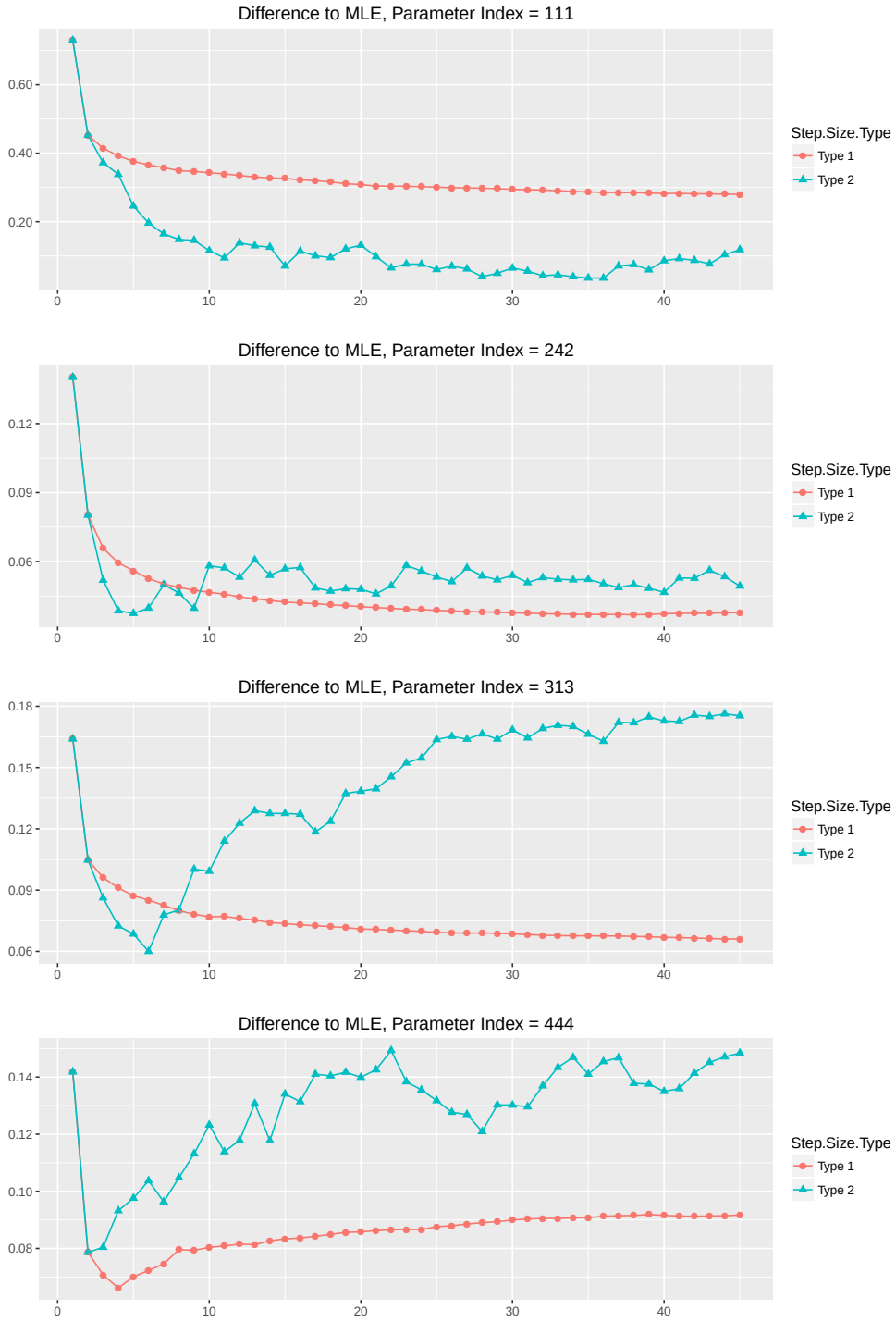


Figure 4.3: Comparison of Step Sizes: Distance to $\widehat{\boldsymbol{\theta}}^{c}_{MLE}$.

??, I have used the following update steps

$$\gamma_n = \begin{cases} 1, & \text{for } n \leq 3; \\ \frac{1}{n-3}, & \text{for } n \geq 4. \end{cases} \tag{4.4}$$

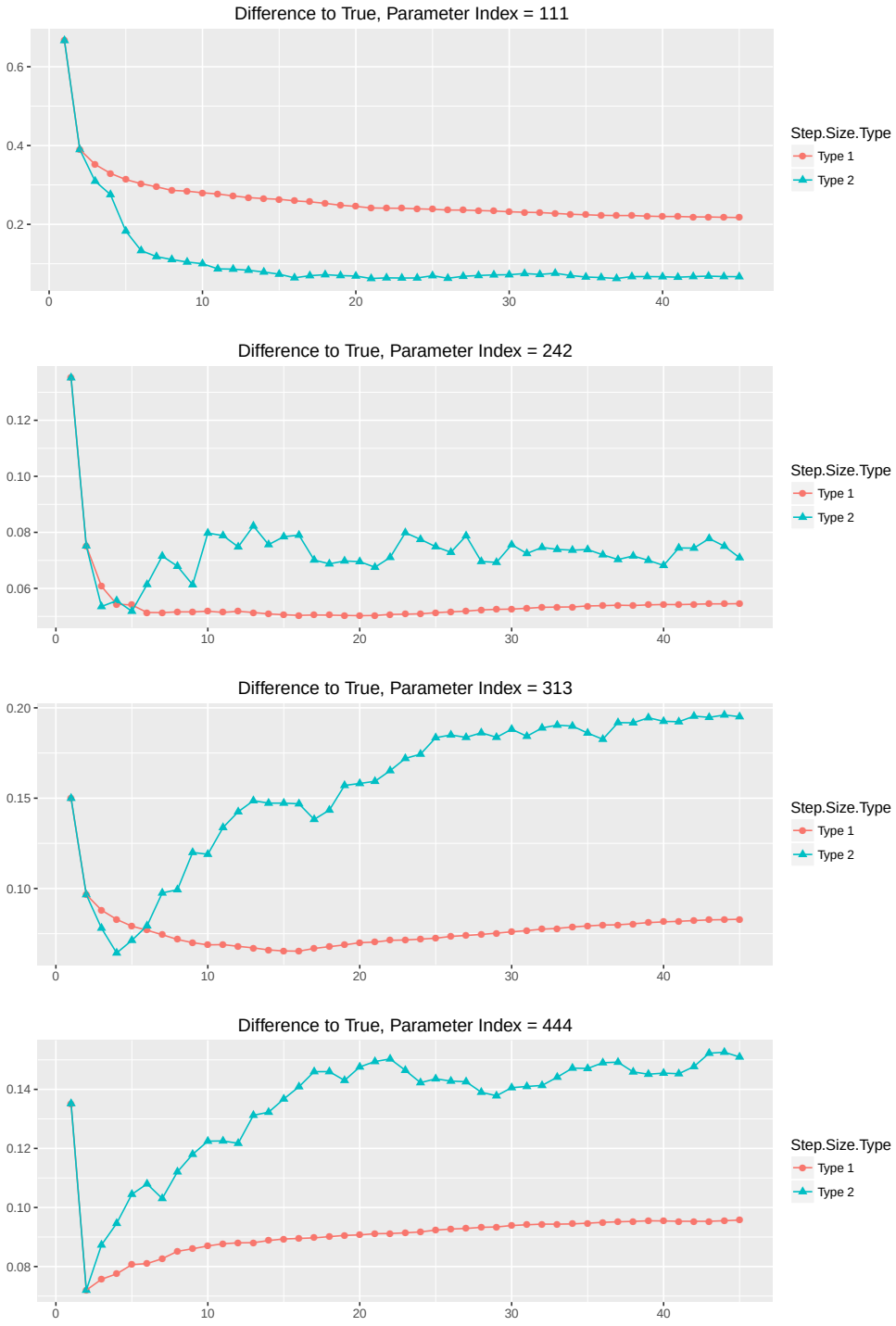


Figure 4.4: Comparison of Step Sizes: Distance to $\widehat{\boldsymbol{\theta}}_0$.

4.2.4 Effectiveness of the Information Matrix Estimate

A desirable feature of the EM algorithm is that it can easily accommodate the computation of the information matrix, which can then be used for computing confidence intervals and implementing hypothesis tests. Since the application of SA algorithm in this research is closely related to EM, it also produces estimates for the information matrix in the process of parameter updates. One question of interest is then whether the information matrix estimates are accurate enough, such that the confidence intervals produced using such estimates can cover the true parameter values $\boldsymbol{\theta}_0$ with the corresponding confidence levels. In this section, I will show the results of some large scale simulations intended to address this question.

I have chosen four sets of parameters to carry out large-scale simulations. For convenience, I have copied the parameters in the following list.

1. Setting "111".
   $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 0.05 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 0.2 \end{bmatrix}$, $\boldsymbol{B} = \begin{bmatrix} 0.55 & 0.15 \\ 0.15 & 0.55 \end{bmatrix}$, $\boldsymbol{V} = \begin{bmatrix} 0.2 & 0.01 \\ 0.01 & 0.2 \end{bmatrix}$.
2. Setting "242".
   $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.1 \end{bmatrix}$, $\boldsymbol{B} = \begin{bmatrix} 0.20 & 0.50 \\ 0.60 & 0.10 \end{bmatrix}$, $\boldsymbol{V} = \begin{bmatrix} 0.4 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$.
3. Setting "313".
   $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.6 & 0 \\ 0 & 0.4 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.3 \end{bmatrix}$, $\boldsymbol{B} = \begin{bmatrix} 0.55 & 0.15 \\ 0.15 & 0.55 \end{bmatrix}$, $\boldsymbol{V} = \begin{bmatrix} 0.16 & 0.08 \\ 0.08 & 0.25 \end{bmatrix}$.
4. Setting "444".
   $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.2 & 0 \\ 0 & 0.6 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$, $\boldsymbol{B} = \begin{bmatrix} 0.20 & 0.50 \\ 0.60 & 0.10 \end{bmatrix}$, $\boldsymbol{V} = \begin{bmatrix} 0.32 & 0.24 \\ 0.24 & 0.50 \end{bmatrix}$.

For each of the above four settings, at least 300 simulations have been run. And each simulation is run in this way: Generate information variables using the true parameter values for $T = 755$ and $K = 2$, then generate return and volume variables also using the true parameter values and their conditional distributions on information variables, and finally estimate the model under the restriction that the information variables are unobservable. The update steps given in Equation (4.4) are used, and each parameter estimation uses 30 steps of updates.

In practice, I did not directly use confidence intervals, but p-values for testing the hypothesis $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_0$ to carry out the accuracy checks. Again, such p-values are calculated separately for each dimension of $\boldsymbol{\theta}$. This method is justified by observing that the confidence interval for a parameter θ is calculated from the following equation:

$$\Pr\left\{|\widehat{\theta} - \theta_0| < C_{1-\alpha/2}\right\} = 1 - \alpha,$$

where $\widehat{\theta}$ follows its asymptotic normal distribution which is symmetric, and $C_{1-\alpha/2}$ is the $(1-\alpha/2)$-th percentile of this distribution. For a specific realization $\widehat{\theta}^*$ of $\widehat{\theta}$, the inequality

$$|\widehat{\theta}^* - \theta_0| < C_{1-\alpha/2}$$

is equivalent to

$$\Pr\left\{|\widehat{\theta} - \theta_0| < |\widehat{\theta}^* - \theta_0|\right\} < 1 - \alpha,$$

or

$$1 - \Pr\left\{|\widehat{\theta} - \theta_0| < |\widehat{\theta}^* - \theta_0|\right\} > \alpha.$$

The detailed procedure for determining the accuracy of information matrix estimates is as such: Fix a set of parameters, run a large number of simulations, and then determine in each simulation the p-value of each parameter. Suppose the significance level is α, then if the portion of simulations with p-value greater than α is approximately equal to $1 - \alpha$, it would imply the information matrix estimation is accurate.

From Table 4.2 we can see that the confidence interval estimates for μ_{k1} are most accurate, because for these parameters, the table entries are most close to $1 - \alpha$ at all three significance levels. In contrast, the confidence interval estimates for μ_{k2} are least accurate, especially for the first three settings. This is somewhat counter-intuitive, since these two parameter play very similar roles in the model, except for that fact that their numerical values are different. An examination into the actual parameter estimates shows that for μ_{k2}, the estimated standard errors are very small, and thus the confidence intervals can easily miss the true parameter values. The parameters that we are most interested in are B_{12} and B_{21}, the estimation accuracy for the confidence interval of which are at intermediate levels. Though they also miss the true parameter values with probabilities higher than significance levels, the differences are not very big. In the real data estimation, we should only reject hypotheses with extra high level of confidence.

A possible explanation for the discrepancy between the estimated and Monte Carlo simulation confidence intervals is that the distribution of the estimated parameters are not normal, which is only the limiting distribution for $T = \infty$. To investigate this point, I have plotted histograms of parameter estimates from the simulations, which are shown in Figures 4.5, 4.6, 4.7 and 4.8.

There are 17 parameters in a two-security model, among which there are entries in $\boldsymbol{\Sigma}_k$ and $\boldsymbol{V}$. The figures have also plotted histograms for entries in the inverse matrices of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{V}$. Estimated kernel densities are plotted together with histograms as black curves. Such densities are estimated using a Gaussian kernel and a bandwidth given by Silverman's

Table 4.2: The Effectiveness of Estimated Confidence Intervals

	Setting 1 (305 Sim.)			Setting 2 (307 Sim.)			Setting 3 (307 Sim.)			Setting 4 (325 Sim.)		
	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
μ_{11}	0.88	0.94	0.98	0.90	0.96	0.99	0.87	0.94	0.99	0.90	0.94	0.98
μ_{12}	0.23	0.26	0.31	0.27	0.31	0.39	0.57	0.64	0.77	0.62	0.70	0.83
μ_{21}	0.92	0.96	0.99	0.89	0.95	1.00	0.93	0.97	1.00	0.91	0.97	0.99
μ_{22}	0.42	0.49	0.60	0.34	0.39	0.49	0.47	0.54	0.65	0.79	0.85	0.94
$\Sigma_{1,11}$	0.78	0.84	0.93	0.73	0.83	0.92	0.79	0.88	0.96	0.81	0.87	0.94
$\Sigma_{1,12}$	0.65	0.75	0.89	0.69	0.81	0.92	0.85	0.91	0.98	0.81	0.89	0.96
$\Sigma_{1,22}$	0.64	0.74	0.81	0.61	0.68	0.79	0.71	0.79	0.89	0.71	0.79	0.91
$\Sigma_{2,11}$	0.78	0.85	0.94	0.79	0.88	0.94	0.75	0.84	0.95	0.79	0.85	0.94
$\Sigma_{2,12}$	0.79	0.87	0.93	0.77	0.82	0.92	0.80	0.87	0.94	0.84	0.90	0.97
$\Sigma_{2,22}$	0.68	0.76	0.89	0.62	0.72	0.87	0.71	0.77	0.87	0.73	0.80	0.93
B_{11}	0.81	0.87	0.94	0.82	0.87	0.97	0.72	0.79	0.91	0.73	0.82	0.90
B_{12}	0.77	0.86	0.94	0.79	0.87	0.95	0.70	0.79	0.90	0.66	0.74	0.85
B_{21}	0.75	0.84	0.95	0.69	0.77	0.86	0.68	0.78	0.88	0.69	0.76	0.86
B_{22}	0.73	0.82	0.93	0.72	0.82	0.92	0.69	0.79	0.88	0.70	0.77	0.90
V_{11}	0.74	0.81	0.90	0.76	0.83	0.92	0.62	0.70	0.82	0.59	0.68	0.80
V_{12}	0.73	0.80	0.92	0.71	0.81	0.92	0.65	0.74	0.88	0.56	0.66	0.81
V_{22}	0.67	0.74	0.90	0.64	0.73	0.86	0.64	0.73	0.91	0.58	0.67	0.79
$\Omega_{1,11}$	0.77	0.84	0.94	0.74	0.83	0.91	0.81	0.87	0.95	0.80	0.87	0.93
$\Omega_{1,12}$	0.65	0.74	0.89	0.69	0.81	0.92	0.85	0.91	0.98	0.81	0.89	0.96
$\Omega_{1,22}$	0.64	0.72	0.82	0.61	0.68	0.79	0.71	0.79	0.92	0.71	0.78	0.90
$\Omega_{2,11}$	0.78	0.85	0.95	0.79	0.87	0.96	0.73	0.84	0.95	0.78	0.86	0.95
$\Omega_{2,12}$	0.79	0.87	0.93	0.77	0.82	0.92	0.80	0.87	0.94	0.84	0.90	0.97
$\Omega_{2,22}$	0.68	0.76	0.88	0.62	0.72	0.87	0.72	0.76	0.88	0.71	0.80	0.91
W_{11}	0.75	0.80	0.91	0.72	0.81	0.93	0.65	0.71	0.84	0.64	0.72	0.86
W_{12}	0.71	0.79	0.91	0.71	0.83	0.92	0.74	0.81	0.93	0.74	0.80	0.92
W_{22}	0.68	0.76	0.90	0.64	0.73	0.88	0.68	0.79	0.91	0.66	0.75	0.88

For each parameter setting, 300 simulations have been run. The numbers shown in the table are the ratio between the number of simulations with true parameter's p-values greater the corresponding significance levels. A number close to $1-\alpha$ implies that the confidence interval estimate is accurate.

(1986) "rule of thumb". The rule chooses as the bandwidth 0.9 multiplying the smaller value between sample standard deviation and the interquartile range divided by 1.34 to the power of the -0.2. The red curves in these figures are the density function of normal distributions with the same means and standard deviations.

From these graphs we can see that the difference between kernel density estimates and normal densities are actually not very big, which implies that the accuracy of confidence intervals may be negatively impacted by some

factors other than the non-normal distributions of parameter estimates. For example, 30 steps of parameter updates may be too less for the algorithm to approximate the Hessian matrix accurately using $\mathbf{\Gamma}_n$. Since $\mathbf{\Gamma}_n$ is used in calculating the confidence intervals, the noise in $\mathbf{\Gamma}_n$ may cause confidence intervals to be inaccurate. To alleviate this, a straightforward method is to increase the number of parameter update steps. The real data estimation presented in Chapter **??** are all estimated using 100 parameter update steps.

4.2.5 Simulated Information Variables

Another way to inspect the performance of the estimation algorithm is examining if the sampled Markov chains in MCMC have been run long enough. If this is not the case, then these MCMC samples may not be closely approximating the unobservable information variables, which in turn entails more parameter updating steps. In order to do such an examination, I will plot the percentiles of the simulated information variables, and see if the true information variables can be effectively covered by some percentile bands of the simulated information variables.

Figure 4.9 has plotted the information variables as well as the percentile bands of simulated information variables for the model with Parameter Index "111" (see Section 4.2.4 for the parameter values). The simulated information variables are those drawn after 30 steps of parameter updates. The red lines marked the maximum and minimum simulated information variables in each period, and the blue lines are the true information variables. Though not easy to spot, two percentile bands are also plotted in the graph, with the darker one marks the $(25\%, 75\%)$ percentiles of simulated information variables, and the lighter one marks the $(2.5\%, 97.5\%)$ percentile.

Compared to the volatilities of the information variables, the conditional distribution of information variable on parameter values and observed return and volume data are very concentrated. Thus in Figure 4.9 it is difficult to distinguish the percentile bands of simulated information variables from the true information variable values. This in turn suggests that the information variables have been well-approximated. Because otherwise we will see the blue lines deviate from the bands delimited by the red dashed lines, or the gray bands frequently. The same figure has been plotted for the other three model settings too. Since they are all qualitatively very similar, these figures are not included in this dissertation.

The narrowness of percentile bands compared to the volatility of information variables suggests breaking the graphs in Figure 4.9 into smaller time periods, which is done in Figures 4.10 to 4.13. From these graphs, we can see more clearly that the simulated percentile bands have covered

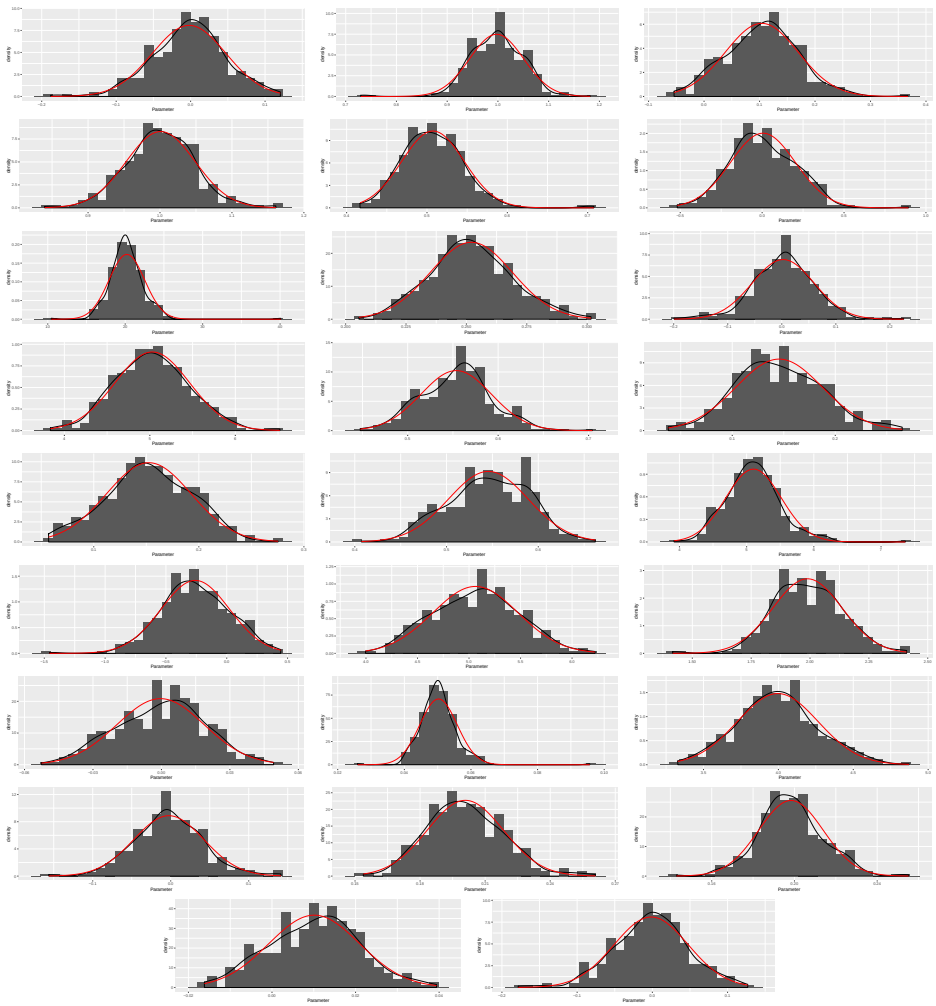


Figure 4.5: The Histograms of Parameter Estimates for Setting One. Parameters Index = 111.

There are 17 parameters in the two-asset model, where 9 of them are entries of covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{V}$. Histograms for entries in the inverses of these matrices are also plotted. For a full list of parameters, please see the first column in Table 4.2. The black curves are estimated Gaussian kernel densities using bandwidth given by Silverman's (1986) "rule of thumb" for bandwidth, while the red curves are normal densities with the same mean and variance. In general, the above histograms show that the distributions of parameter estimates are close to normal distributions.

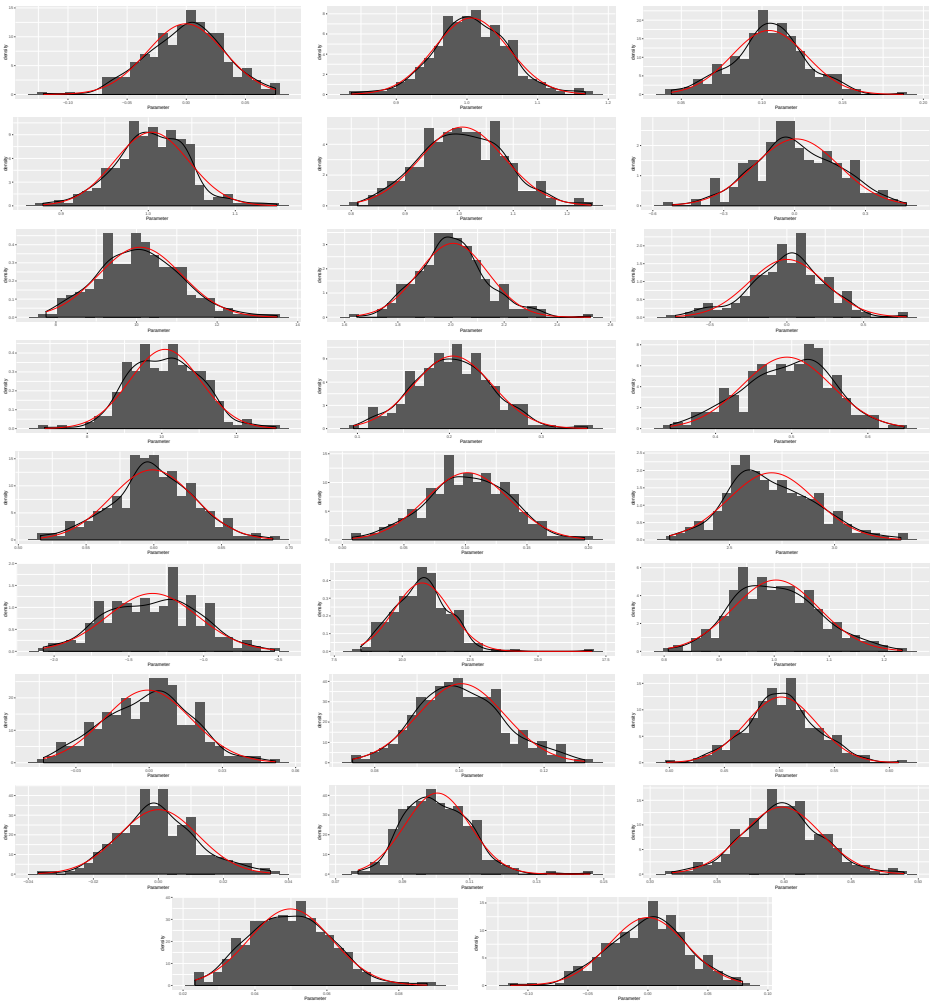


Figure 4.6: The Histograms of Parameter Estimates for Setting Two. Parameters Index = 242.

There are 17 parameters in the two-asset model, where 9 of them are entries of covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{V}$. Histograms for entries in the inverses of these matrices are also plotted. For a full list of parameters, please see the first column in Table 4.2. The black curves are estimated Gaussian kernel densities using bandwidth given by Silverman's (1986) "rule of thumb" for bandwidth, while the red curves are normal densities with the same mean and variance. In general, the above histograms show that the distributions of parameter estimates are close to normal distributions.

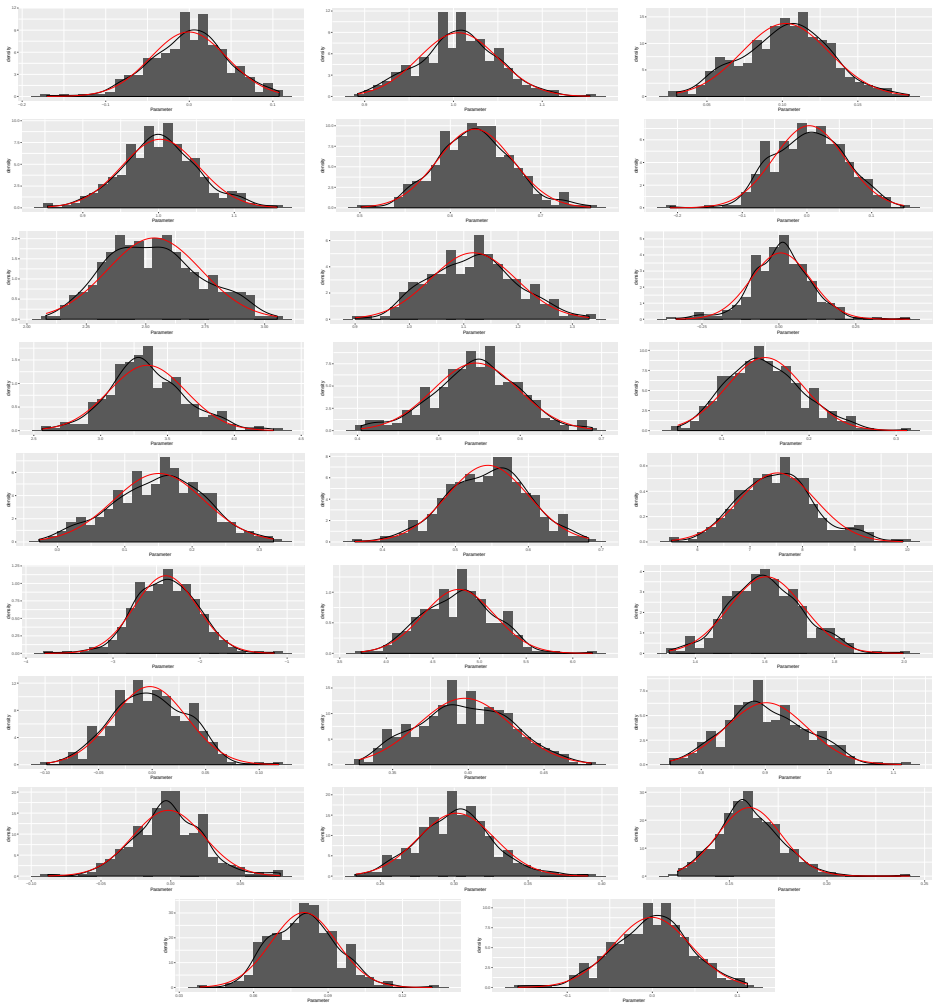


Figure 4.7: The Histograms of Parameter Estimates for Setting Three. Parameters Index = 313.

There are 17 parameters in the two-asset model, where 9 of them are entries of covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{V}$. Histograms for entries in the inverses of these matrices are also plotted. For a full list of parameters, please see the first column in Table 4.2. The black curves are estimated Gaussian kernel densities using bandwidth given by Silverman's (1986) "rule of thumb" for bandwidth, while the red curves are normal densities with the same mean and variance. In general, the above histograms show that the distributions of parameter estimates are close to normal distributions.

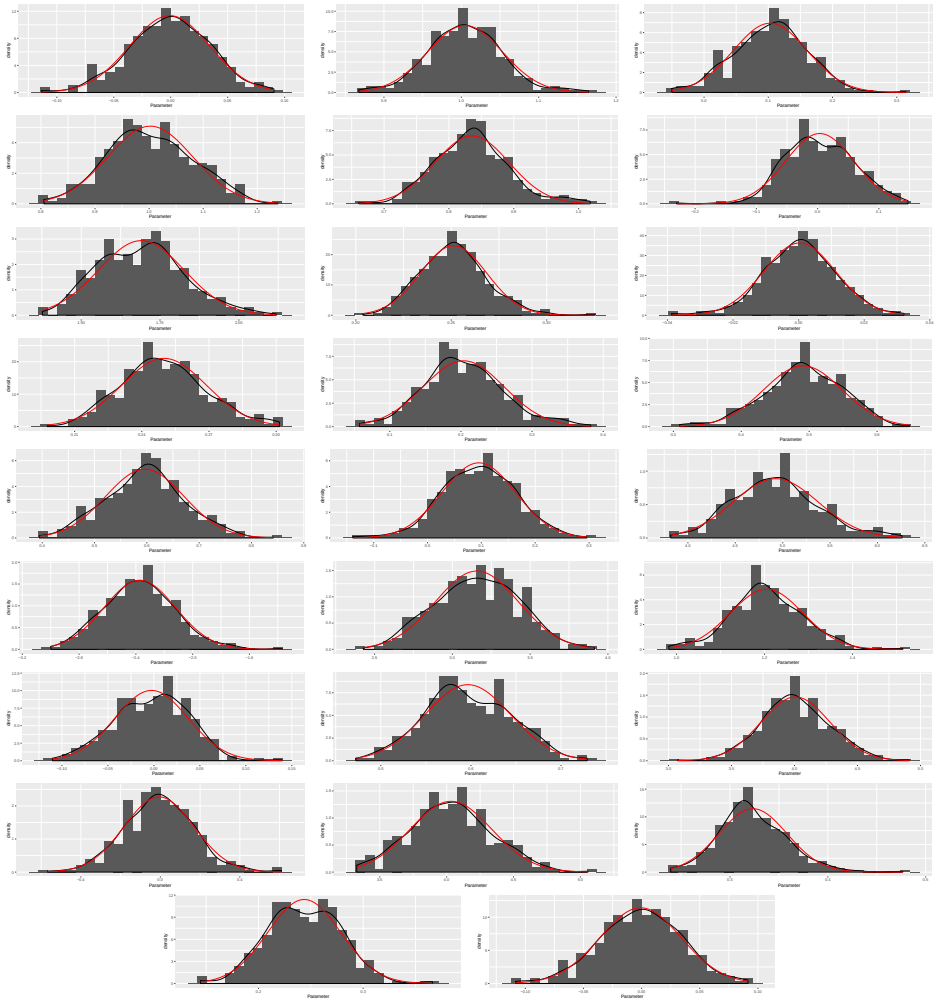


Figure 4.8: The Histograms of Parameter Estimates for Setting Four. Parameters Index = 444.

There are 17 parameters in the two-asset model, where 9 of them are entries of covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{V}$. Histograms for entries in the inverses of these matrices are also plotted. For a full list of parameters, please see the first column in Table 4.2. The black curves are estimated Gaussian kernel densities using bandwidth given by Silverman's (1986) "rule of thumb" for bandwidth, while the red curves are normal densities with the same mean and variance. In general, the above histograms show that the distributions of parameter estimates are close to normal distributions.

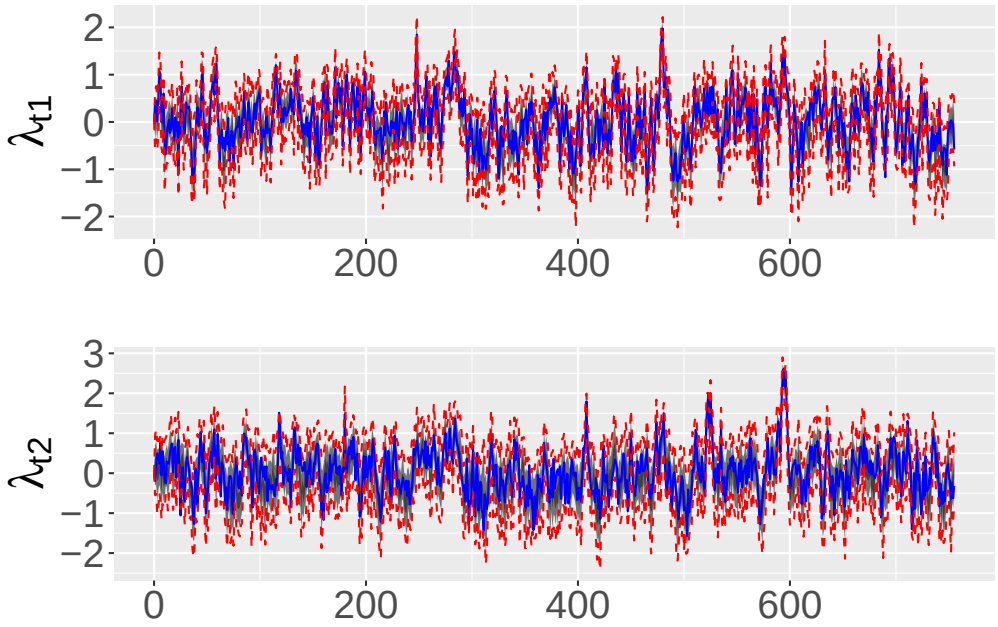


Figure 4.9: The True and Simulated Information Variables, Parameter Index = 111. Asset 2.

For the two assets in the model with Parameter Index 111, these two graphs have plotted the true values (solid blue line), maximum and minimum simulated values (dashed red lines), $(2.5\%, 97.5\%)$ percentile band (light gray) and $(25\%, 75\%)$ percentile band (dark gray) of information variables. During $T = 755$ periods of observations, the posterior distributions of information variables in each period are very concentrated around the true values. Thus the percentile bands are very small such that it is not easy to visually distinguish them. But on the other hand, this also suggests that the MCMC algorithm has worked well, drawn samples successfully from the target distribution and approximated the true information variables closely.

the true information variables effectively. As the true information variables fluctuate, the simulated percentile bands can make quick adjustments accordingly. I will give more specific explanations for the graphs in each graph's footnote.

To examine the performance of the MCMC part in the estimation algorithm more quantitatively, I have calculated the percentages of information variables being covered by respective percentile bands over the 755 observation periods, the results of which are listed in Table 4.3. It is confirmed that the MCMC algorithm has approximated the unobservable information variables well, because the percentage numbers are close to the width of percentile bands.

Another way to examine the performance of the MCMC algorithm is comparing the distribution of the true and simulated information variables over time, which is done in Figure 4.14. In this graph, I have plotted the kernel density estimates for both the true and simulated information vari-

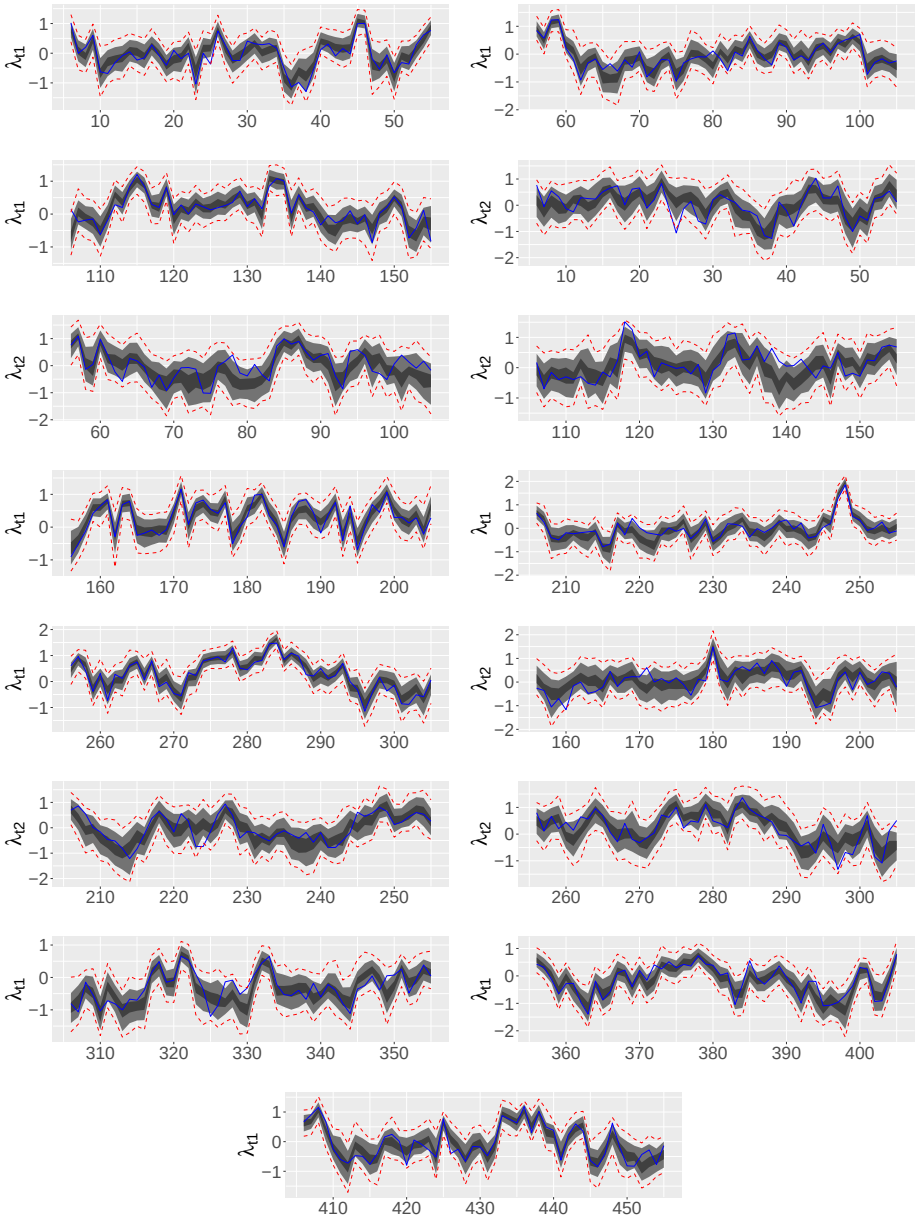


Figure 4.10a: The evolution of two assets' true and simulated information variables. Parameters Index = 111.

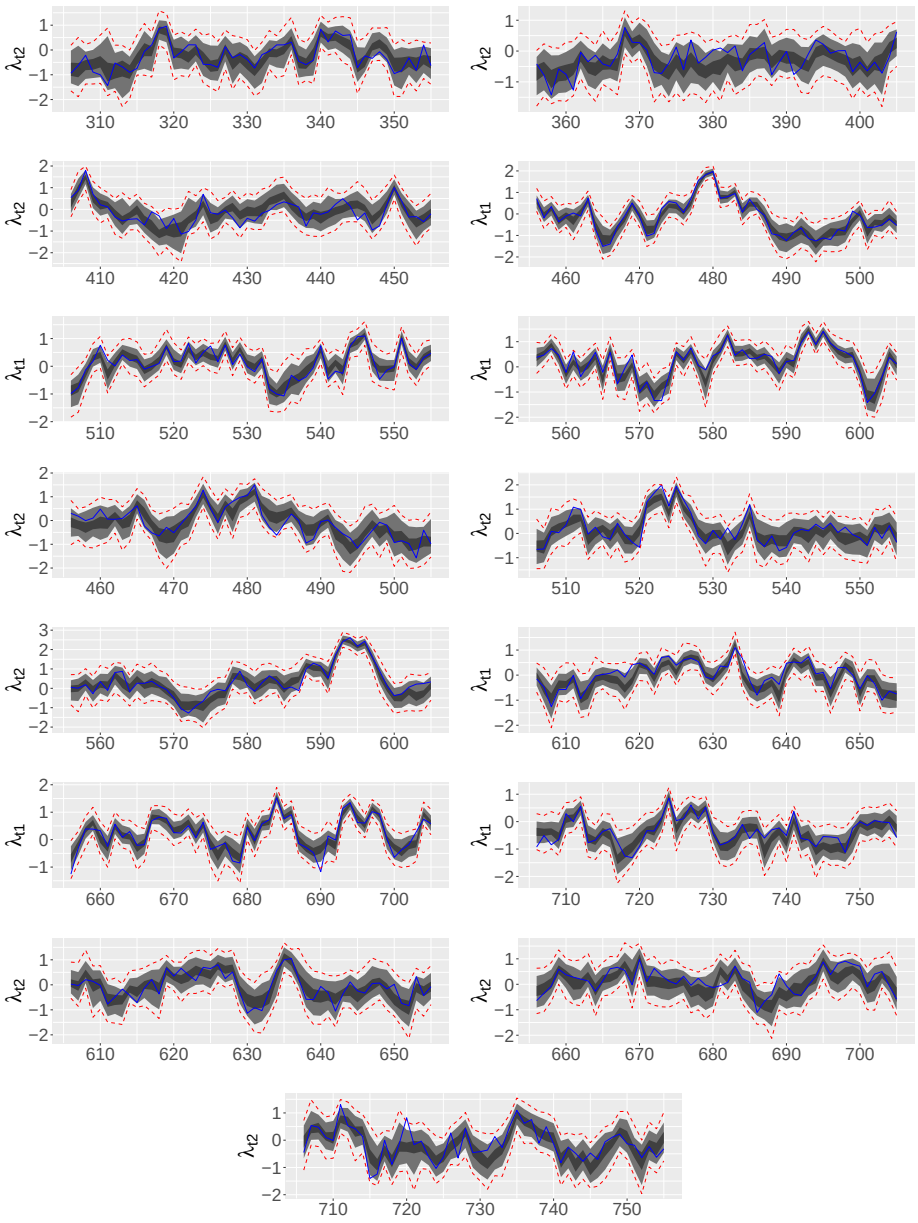


Figure 4.10b: The evolution of two assets' true and simulated information variables. Parameters Index = 111.

These graphs compare the true and simulated information variables. Each of the small graphs in this panel plots the true values (solid blue line), maximum and minimum simulated values (dashed red lines), (2.5%, 97.5%) percentile band (light gray) and (25%, 75%) percentile band (dark gray) of information variables. The graphs in Lines 1, 3, 5, 7 and 9 are for Asset 1, and the remaining lines are for Asset 2. The percentile bands for Asset 1's information variables are narrower than those for Asset 2's, though the two information variables are completely symmetric, as determined by $\boldsymbol{B}$ and $\boldsymbol{V}$. This is because entries in $\boldsymbol{\Sigma}_2$ are bigger than those in $\boldsymbol{\Sigma}_1$, meaning the unobservability of Asset 2's information variables are more severe. Consequently, the conditional variance of Asset 2's information variable is bigger.

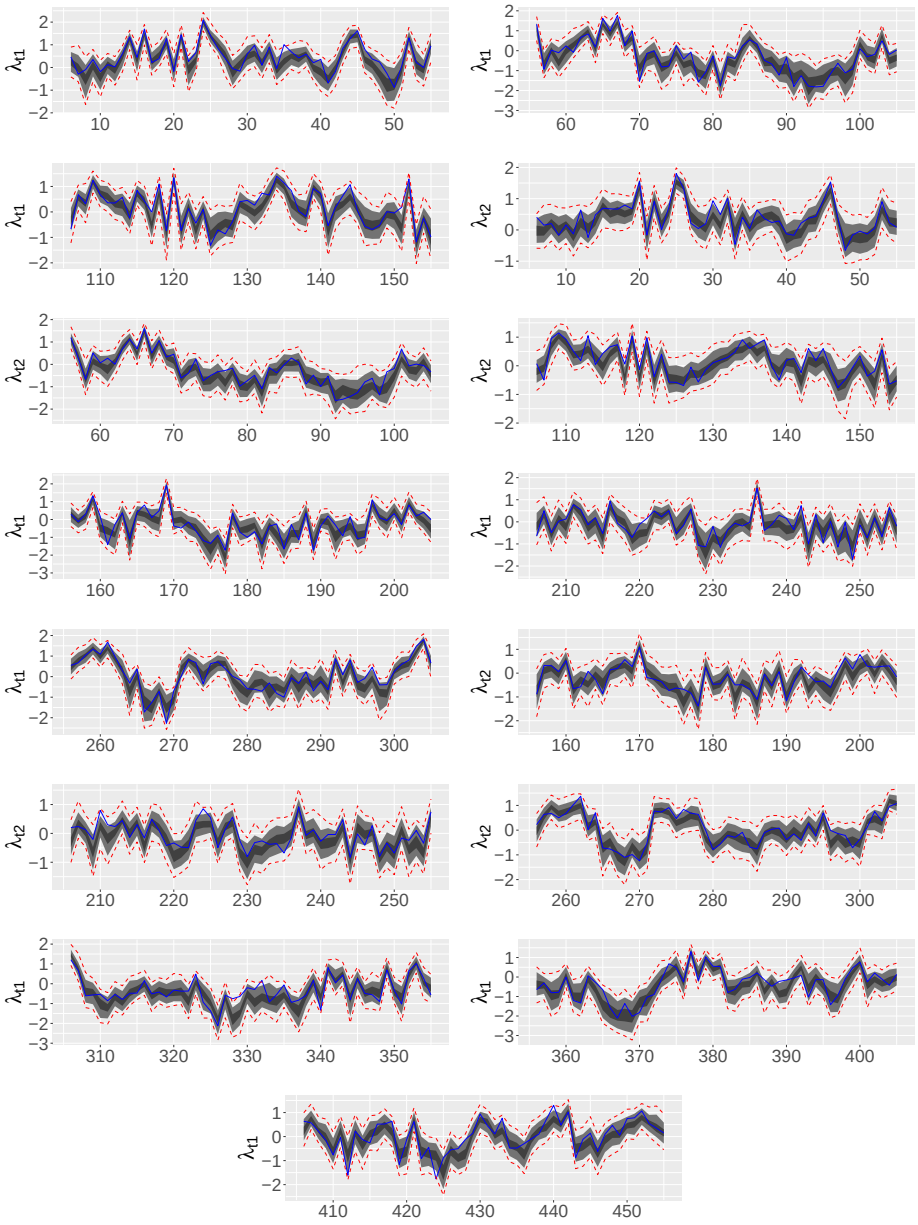


Figure 4.11a: The evolution of two assets' true and simulated information variables. Parameters Index = 242.

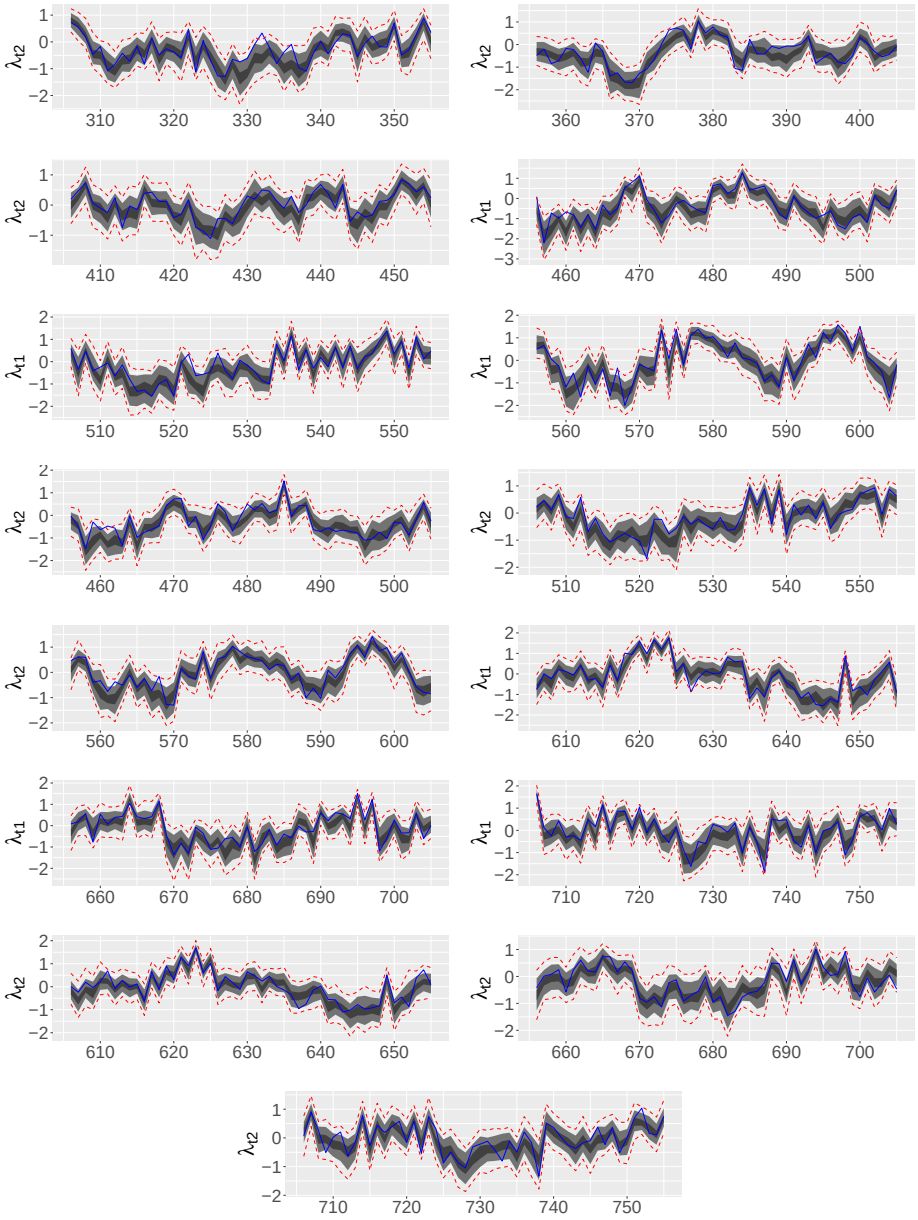


Figure 4.11b: The evolution of two assets' true and simulated information variables. Parameters Index = 242.

These graphs compare the true and simulated information variables. Each of the small graphs in this panel plots the true values (solid blue line), maximum and minimum simulated values (dashed red lines), (2.5%, 97.5%) percentile band (light gray) and (25%, 75%) percentile band (dark gray) of information variables. The graphs in Lines 1, 3, 5, 7 and 9 are for Asset 1, and the remaining lines are for Asset 2. The width of percentile bands for the two asset's simulated information variables are close, though V_{11} is four time as big as V_{22}.

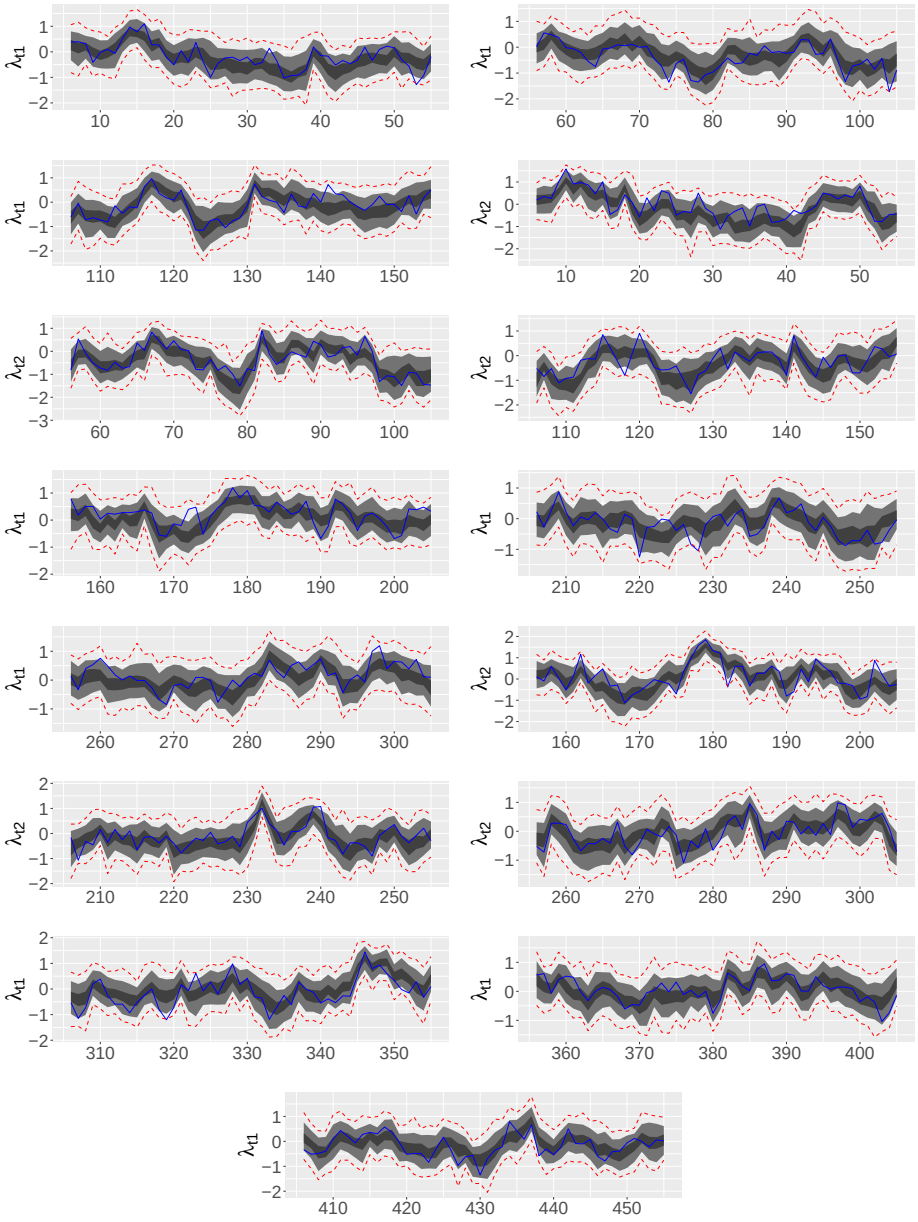


Figure 4.12a: The evolution of two assets' true and simulated information variables. Parameters Index = 313.

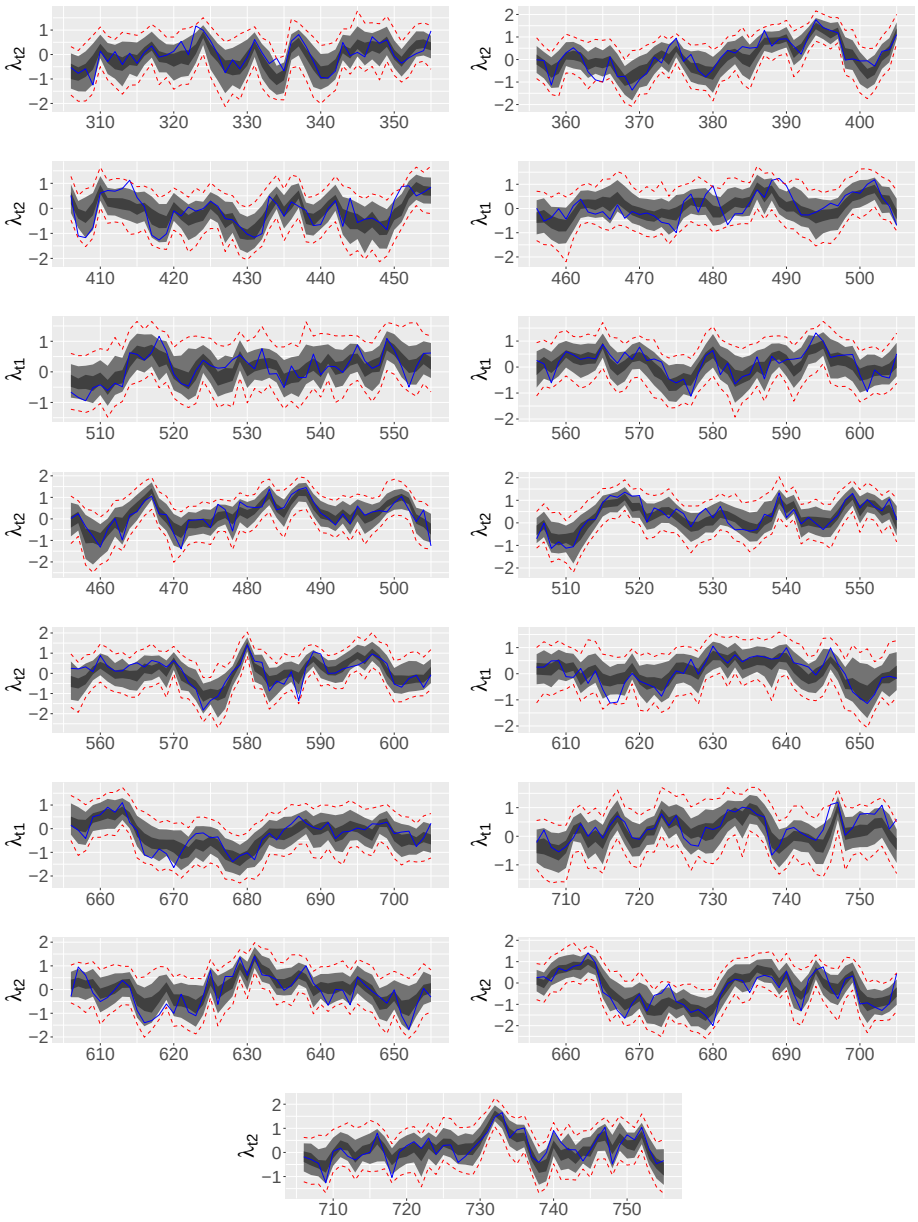


Figure 4.12b: The evolution of two assets' true and simulated information variables. Parameters Index = 313.

These graphs compare the true and simulated information variables. Each of the small graphs in this panel plots the true values (solid blue line), maximum and minimum simulated values (dashed red lines), $(2.5\%, 97.5\%)$ percentile band (light gray) and $(25\%, 75\%)$ percentile band (dark gray) of information variables. The graphs in Lines 1, 3, 5, 7 and 9 are for Asset 1, and the remaining lines are for Asset 2. The width of percentile bands for the two asset's simulated information variables are also close. But compared to Figures 4.10 and 4.11, the percentile bands in Figure 4.12 are generally wider. This may suggest that the minimum diagonal value of $\mathbf{\Sigma}_k$ is important in determining the width of percentile bands.

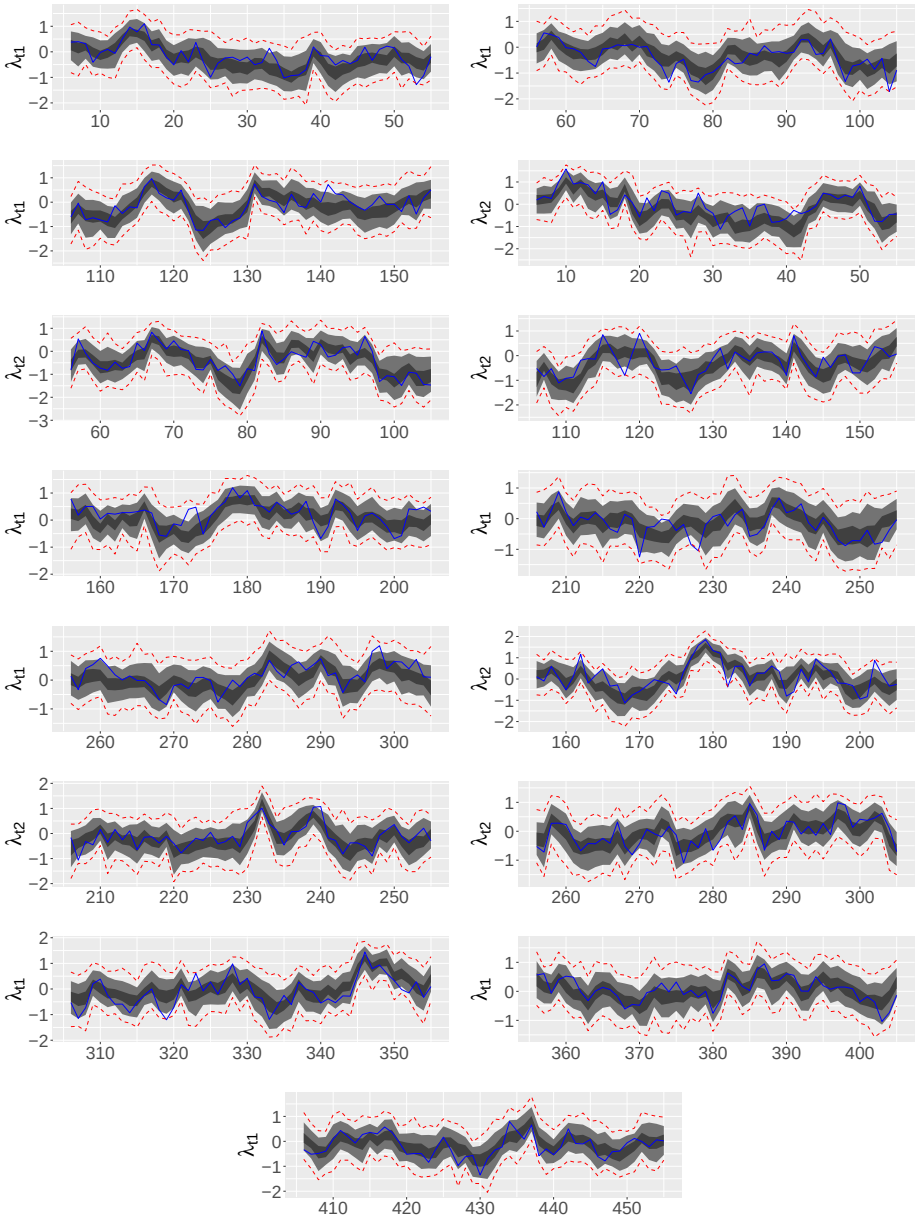


Figure 4.13a: The evolution of two assets' true and simulated information variables. Parameters Index = 444.

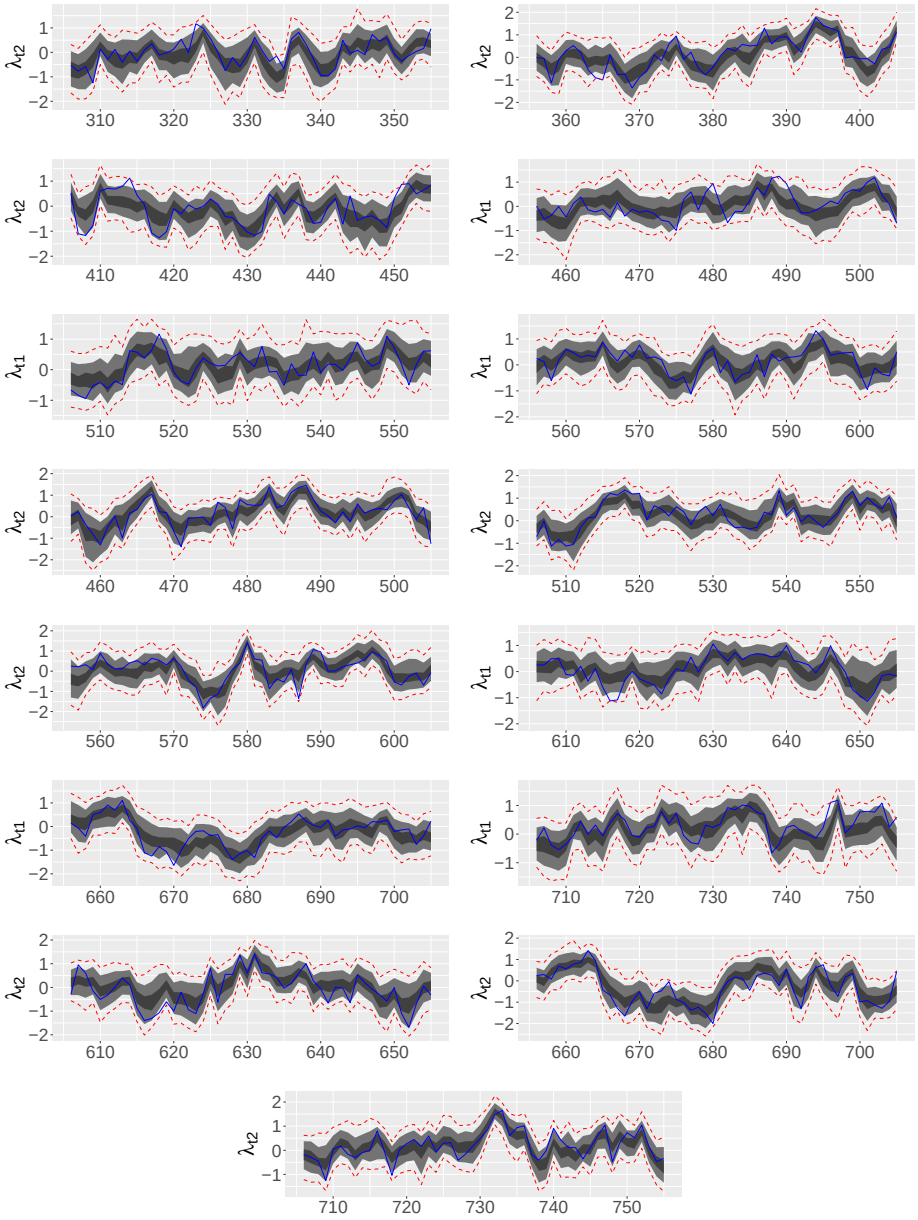


Figure 4.13b: The evolution of two assets' true and simulated information variables. Parameters Index = 444.

These graphs compare the true and simulated information variables. Each of the small graphs in this panel plots the true values (solid blue line), maximum and minimum simulated values (dashed red lines), (2.5%, 97.5%) percentile band (light gray) and (25%, 75%) percentile band (dark gray) of information variables. The graphs in Lines 1, 3, 5, 7 and 9 are for Asset 1, and the remaining lines are for Asset 2. Again, the width of percentile bands are close for the two assets.

Table 4.3: The Effectiveness of Percentile Bands

	25% – 50%	2.5% – 97.5%	0–100%
Model 111, Asset 1	0.48	0.93	1.00
Model 111, Asset 2	0.48	0.92	0.99
Model 242, Asset 1	0.47	0.91	0.99
Model 242, Asset 2	0.46	0.93	0.99
Model 313, Asset 1	0.50	0.94	0.99
Model 313, Asset 2	0.48	0.92	0.99
Model 444, Asset 1	0.50	0.94	0.99
Model 444, Asset 2	0.48	0.92	0.99

ables (taking the last step in MCMC chains). Again, the graphs conform with my previous finding that MCMC part of the algorithm has worked well, since the two estimated kernel densities are close to each other.

To sum up the simulation results, the success of estimation algorithm depends on three things. One is the length of MCMC chains, which determines if the true information variables can be well-approximated. Longer chains are always better, at the cost of longer computation time. The second one is the step size scheme. If step sizes decreases too fast, the parameter updating may become too slow while the parameters are still far from their limit value. On the other hand, if the step sizes decreases too slow, past simulated information may be discounted at a high rate, causing the algorithm to be unstable. The third one is the number update steps. More parameter update steps is always better. The examination on the large-scale simulation results and the simulated information variables suggest that 30 parameter update steps may be too few. Based on this result, in the real data estimation part, I will increase the number of update steps to 100.

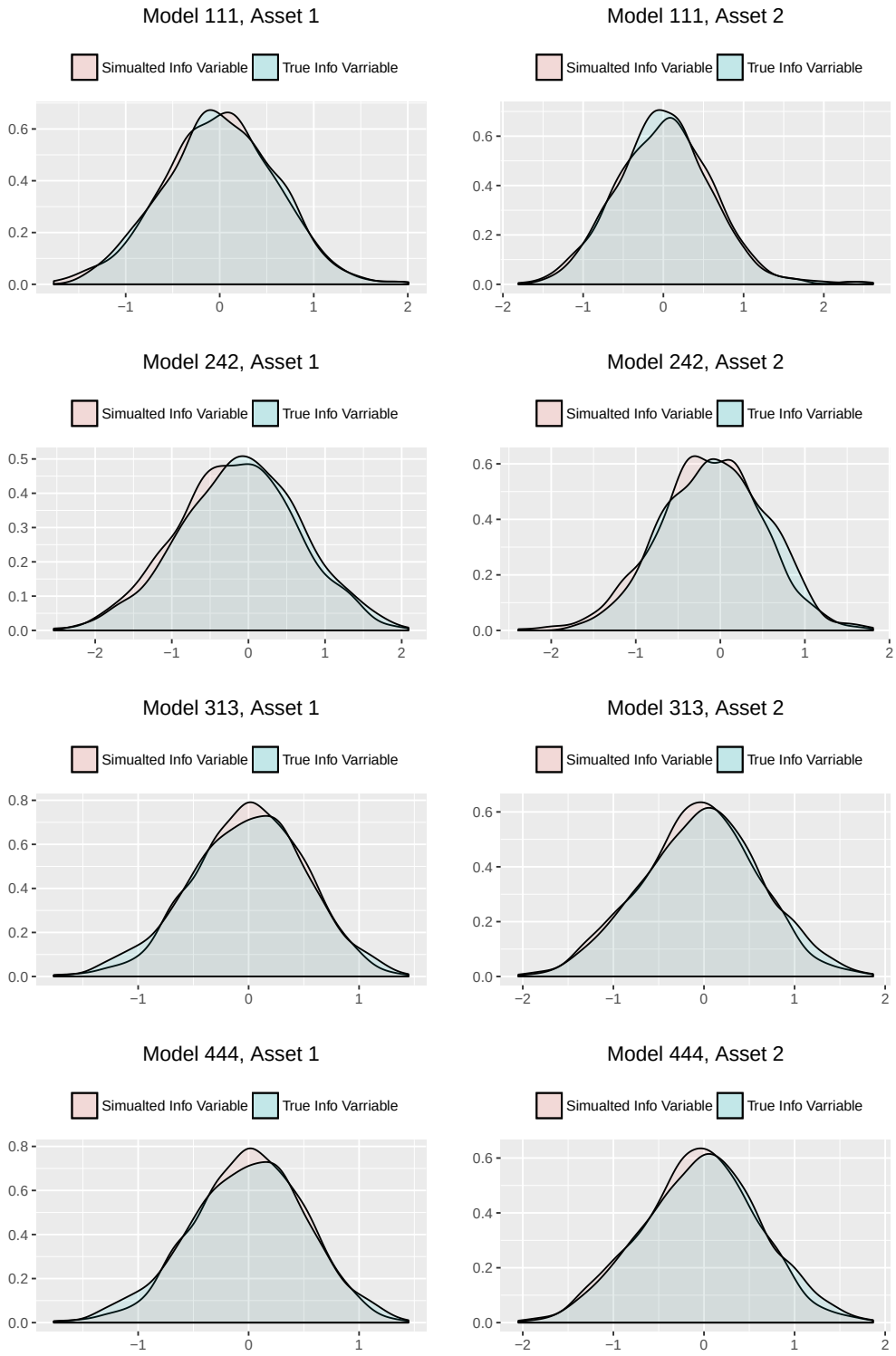


Figure 4.14: Kernel Density Estimates for the True and Simulated Information Variables.

These graphs compare the kernel density estimates between the true and simulated information variables (last steps in MCMC chains are used) in the four models. The kernel densities are close to each other in general, again suggesting that the MCMC part of the algorithm has worked well.

Chapter 5

Real Data Estimation

In this chapter, I will present the results obtained from estimating the model in Chapter 2 using real security trading data. As mentioned in Chapter 1, the real data are organized into three categories. One is individual stock data. The second one is data of ETFs tracking U.S. equity and bond market indexes. And the third one is data of ETFs tracking U.S. and international stock market indexes. In the remaining part of this chapter, I will present in sequence the results on these three categories of real data estimates.

5.1 Models Estimated Using Stock Data

I have built real data estimations mainly using the component stocks of the DJIA index. Table 5.1 shows a record of all real data models estimated. The models are labeled using stock ticker symbols. For a correspondence between these ticker symbols and the companies they represent, please see Table 5.2 on the following page. Also listed in Table 5.2 are the industrial sectors that each company belongs to. All the data used in this chapter is taken from Yahoo Finance (`http://finance.yahoo.com`).

The real data estimations are designed in such a way that each of the seven industrial sectors will be covered by at least one two-security model and one three-security model. The exceptions are the health care industry which has two three-security models, and the consumer good industry which has two two-security models. To meet such a purpose, I have included some other stocks with large market capitalizations that are not DJIA components, but are in the same industries as some DJIA component stocks. For example, the basic materials sector has three component stocks, CVX, DD and XOM. However, the three companies they represent are not still not very similar, since Chevron and Exxon Mobil are primarily oil companies while DuPont focuses on the production of chemical materials. It is better to estimate them in separate models, in order to find information spillover effects. Thus another chemical material company's stock DOW is added to be estimated together with DD, and RDS-B is also added to be estimated together with CVX and XOM. Similar actions have been taken for other industries as well. Another exception is the model composed of AAPL and PG. The companies these two stocks represent also differ very much in their major products, but I kept the model, with an expectation that the information spillover effects are not significant between these two companies.

Table 5.1: Read Data Pairs and Triplets Estimated

Sector Name	Estimated Models	
Basic Materials	DD & DOW	CVX &RDS-B &XOM
Consumer Goods	AAPL & PG	
	KO & PEP	
Financial	AXP &V	BAC & C & WFC
	GS & JPM	
Health Care		AET & ANTM & UNH
		JNJ & MRK & PFE
Industrial Goods	BA & GE	CAT & MMM & UTX
Services	DIS & FOXA	MCD & SBUX & YUM
	COST & WMT	
	HD & LOW	
Technology	T & VZ	CSCO & INTC & IBM
	GOOGL & MSFT	

5.2 Stock Data Description

The returns of Stock k on Day t is defined as

$$r_{t,k} = 100 \ln \left(\frac{S_{t,k}}{S_{t,k-1}} \right),$$

where $S_{t,k}$ is Stock k's closing price on day t.[1] For each stock, daily observations of return $r_{t,k}$ and trading volume $V_{t,k}$ are collected for the period between Jan 1, 2012 and Dec 31, 2014, giving 754 observations per stock. Table 5.3 has listed some descriptive statistics for the stocks.

A feature that is shared by all stocks' return data is that, compared to the magnitudes of average returns, standard deviations of returns are usually larger by an order of magnitude. Also interesting is that stocks' standard deviations are concentrated in the relatively small interval of $(0.78, 1.83)$. Yet stocks differ in their levels of skewness, with 11 positively skewed and the rest negatively skewed. GOOGL, FOXA and COST are the three most right-skewed stocks, and IBM, ANTM and YUM are the most left-skewed ones. As far as kurtosis is concerned, all stocks are leptokurtic, but of varying degrees. Kurtosis values between 4.0 and 6.0 are most common, and stocks with the highest kurtosis values are CSCO, GOOGL and IBM.

Figures 5.1a to 5.1d have plotted the autocorrelations of returns, absolute returns and squared returns for lags between 1 and 30. For all stocks, the autocorrelations of returns at most lags are not significantly different from zero. A limited amount of exceptions exist. For example, the autocorrelation of AAPL at lags 8 and 11 are slightly out of the range marked by

[1] The data source has adjusted the raw closing prices for dividends and splits.

5% critical values. But it is more common for absolute and squared returns to have significant autocorrelations. For example, BAC's autocorrelations for squared and absolute returns are still high at lag 29. The p-values lines in Table 5.3 report the p-Values of Ljung-Box tests carried out for each stocks returns with lag 30. 6 stocks out of 43 have p-Values smaller than 0.1.

Figures 5.2a to 5.2d compare the sample distributions of returns and normal distributions with the same means and variances. All stocks' kernel density estimates have more probability mass concentrated around the means than the normal distributions do. The difference in tail behaviors between sample and normal distributions are difficult to visually examine. But the sample Kurtosis given in Table 5.3 indicate that the sample distributions all have fatter tails.

Qualitatively, the data used in this research show the same features as those investigated by Andersen (1996) and Liesenfeld (2001).

Now I will move on to describing the volume data. The first step in processing trading volume data is de-trending. I have followed the method used by Andersen (1996) and taken the centered two-year moving average as the trend values of trading volumes. De-trended volumed are then calculated as the ratios between original volumes and the corresponding trend values.

Table 5.4 has listed some descriptive statistics for the de-trended volume series. The means of all stocks' de-trended volumes are close to one, which is due to the de-trending process calculating the ratio between raw volumes and trend values. Remarkably, the sample variances of de-trended volumes are smaller than those of returns. As the means de-trended volumes are larger in the meantime than those of returns, this indicates that return series are more volatile than trading volumes. All stocks' de-trended volumes are right-skewed, which is of no surprise as they are positive random variables. And kurtosis are all greater than 3, indicating the high probabilities of extreme observations.

The p-Value lines give the p-Values of Ljung-Box tests for autocorrelation, the null hypothesis of which (no autocorrelation) is rejected for all stocks. And the Cor. lines give the correlation between de-trended volumes and squared returns. Most stocks show positive correlations between these two series, with correlation coefficient in the range between 0.4 and 0.7. Four exceptions are provided by PFE, RDSB, UTX and VZ, the correlations of which are below 0.30.

Figures 5.3a to 5.3d plot the autocorrelation functions for the stocks' de-trended volumes. As consistent with the results of Ljung-Box tests, all stocks show high levels of autocorrelations, at least for low lags (lags smaller than 10). Interestingly, the auto-correlations in de-trended volumes are higher than that of absolute or squared returns, which may imply

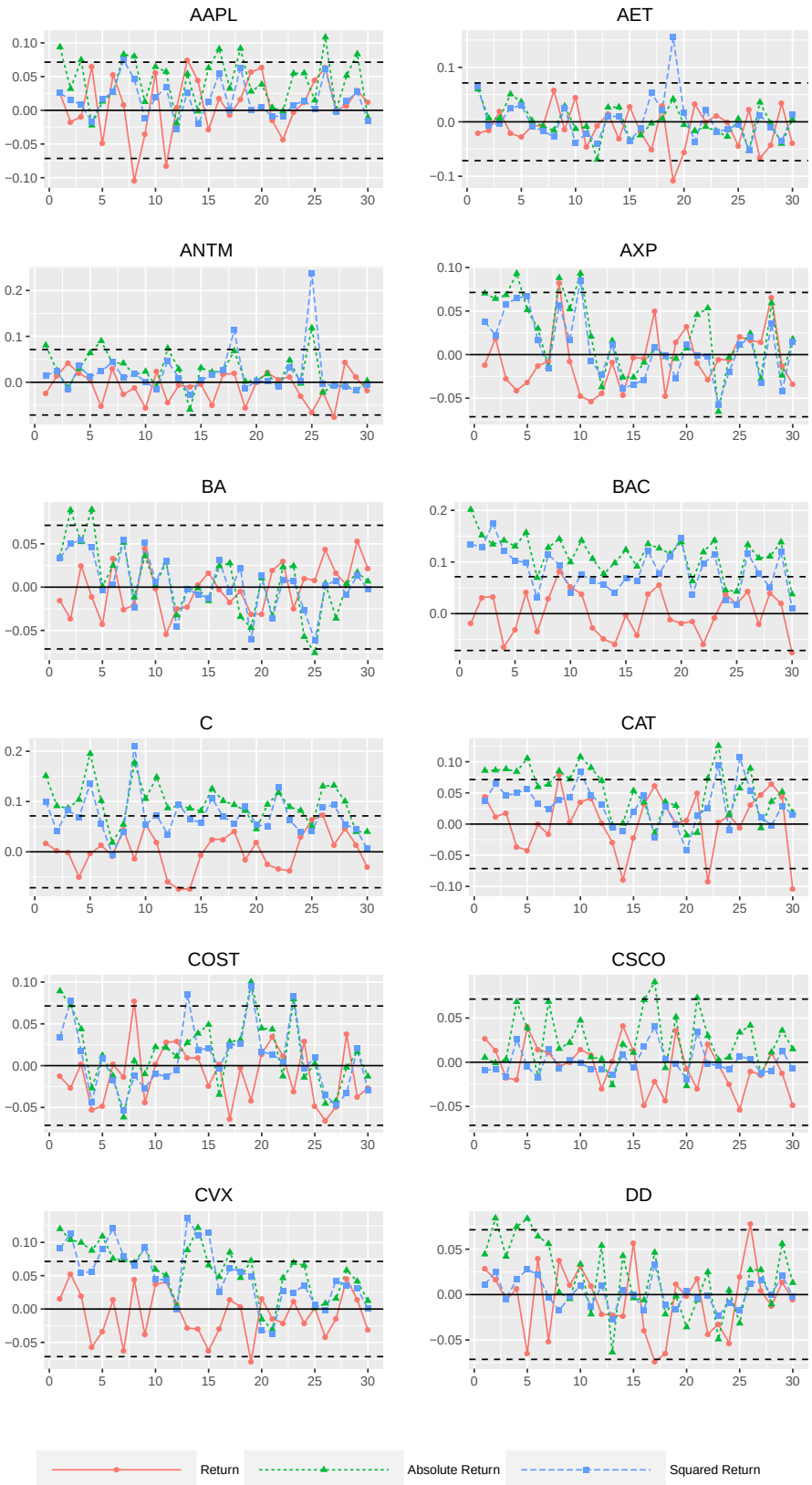


Figure 5.1a: Autocorrelation Functions of Returns, Absolute Returns and Squared Returns.

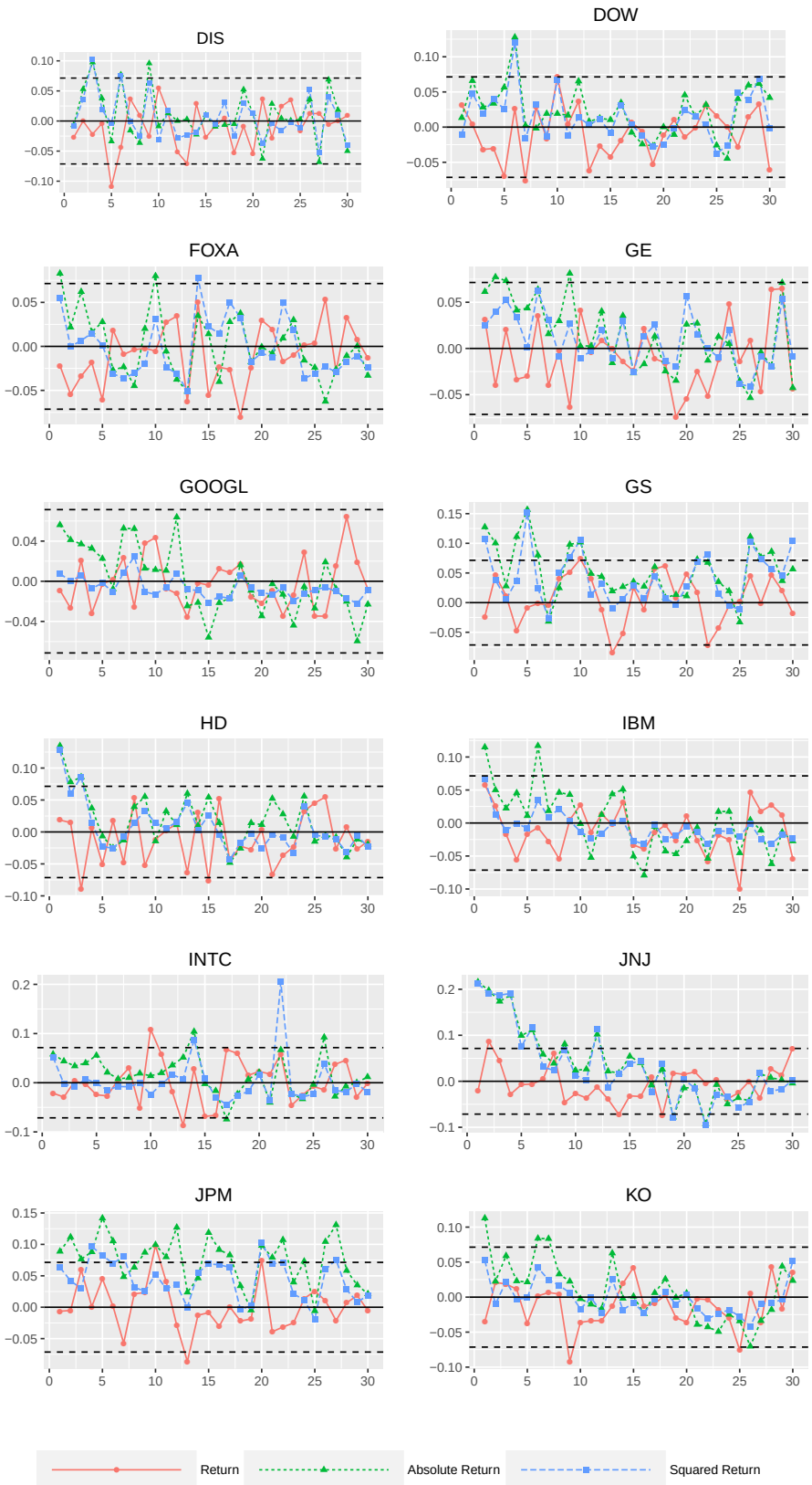


Figure 5.1b: Autocorrelation Functions of Returns, Absolute Returns and Squared Returns.

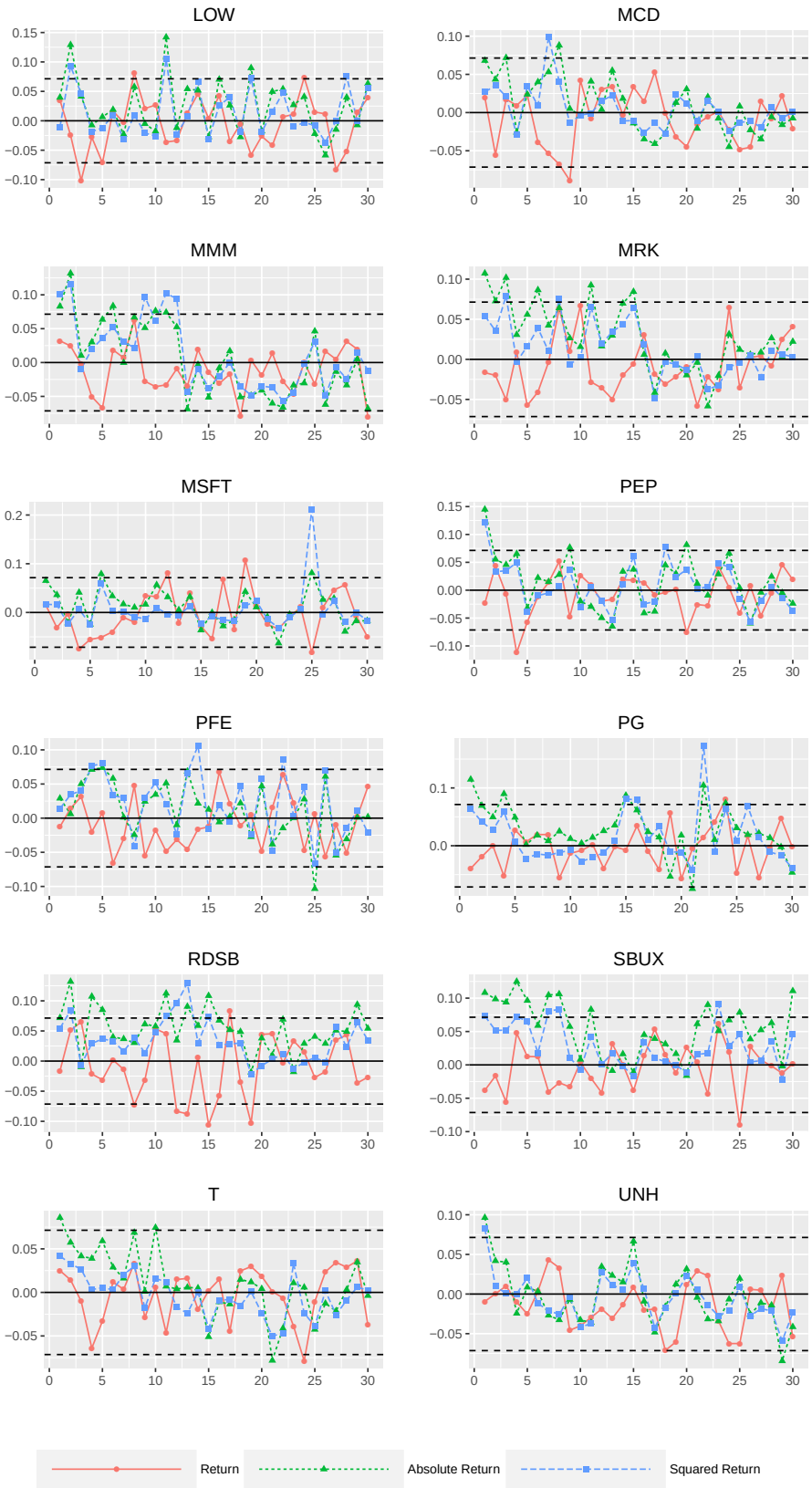


Figure 5.1c: Autocorrelation Functions of Returns, Absolute Returns and Squared Returns.

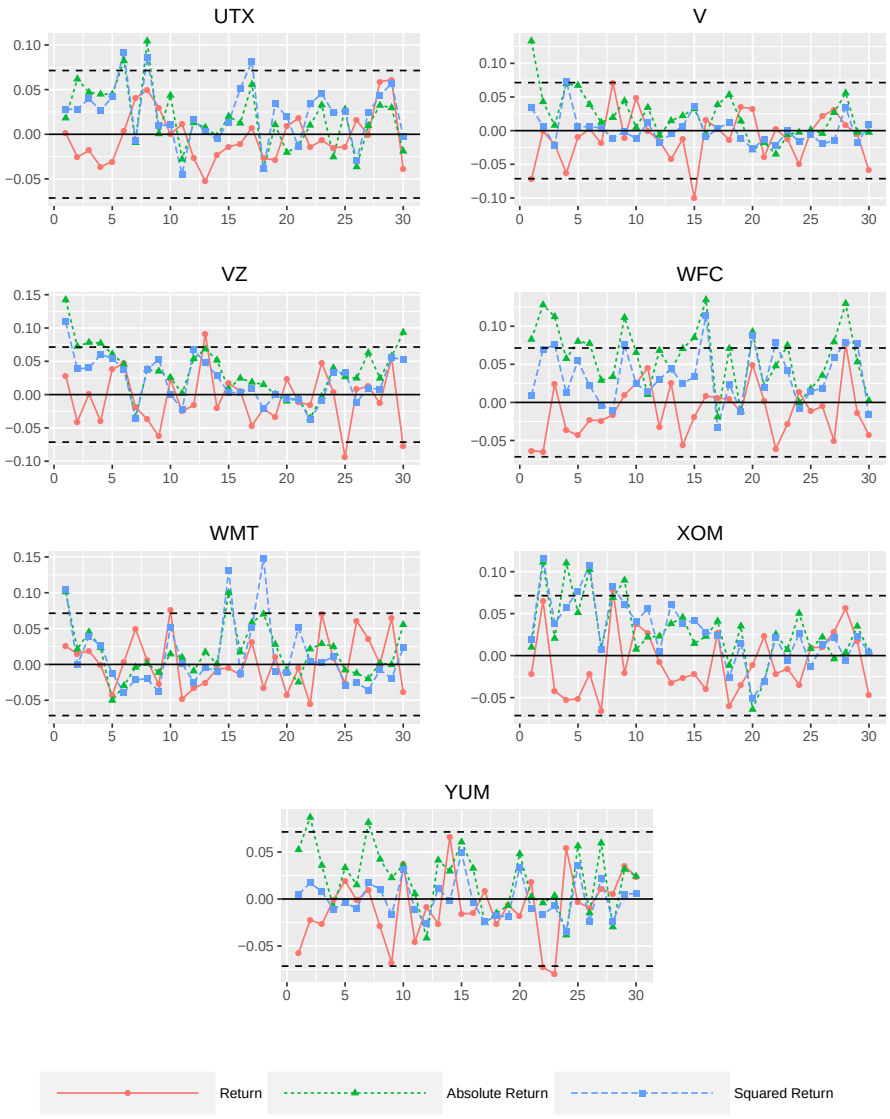


Figure 5.1d: Autocorrelation Functions of Returns, Absolute Returns and Squared Returns.

The autocorrelation functions in Figures 5.1a to 5.1d show that autocorrelations are more significantly different from zero in absolute and squared returns than returns. The black dashed lines mark the 5% critical values for the null hypothesis of zero autocorrelation.

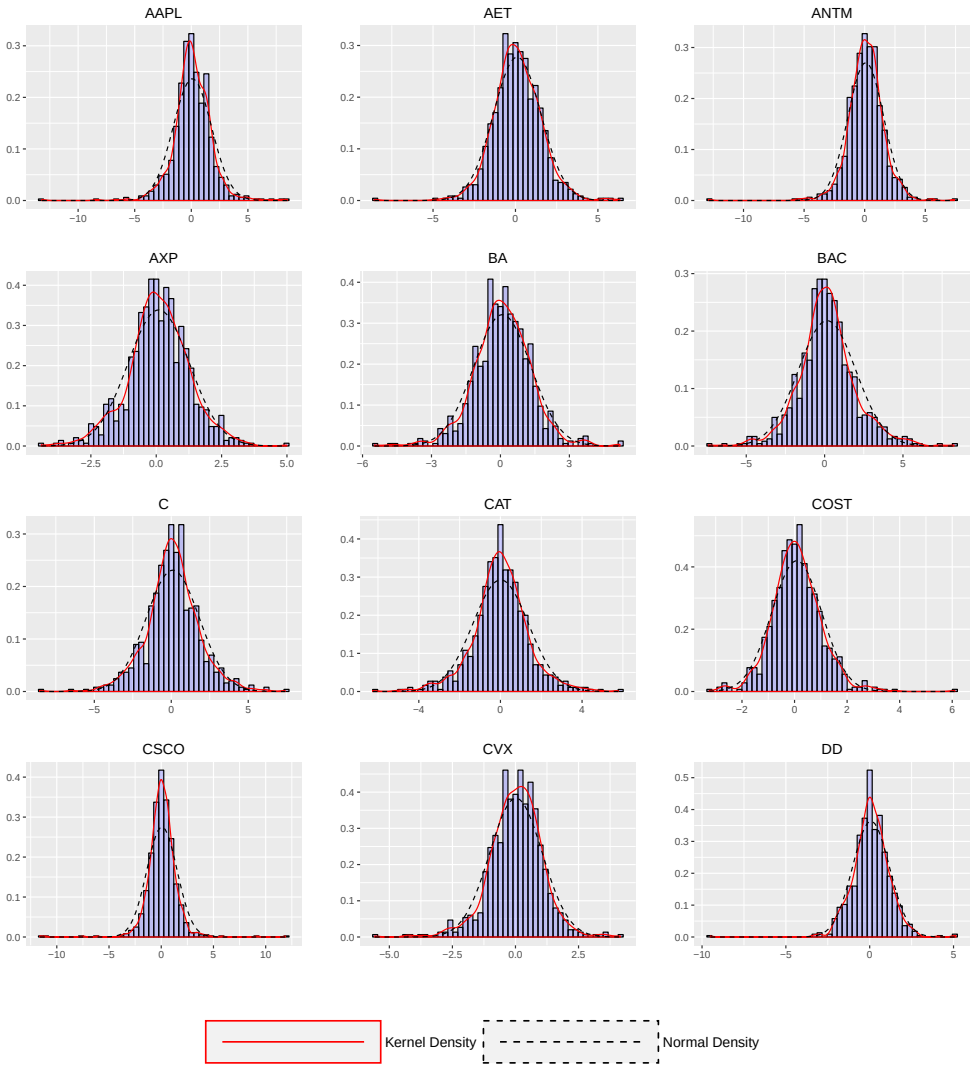


Figure 5.2a: Comparisons between the Sample Return Distributions and Normal Distributions

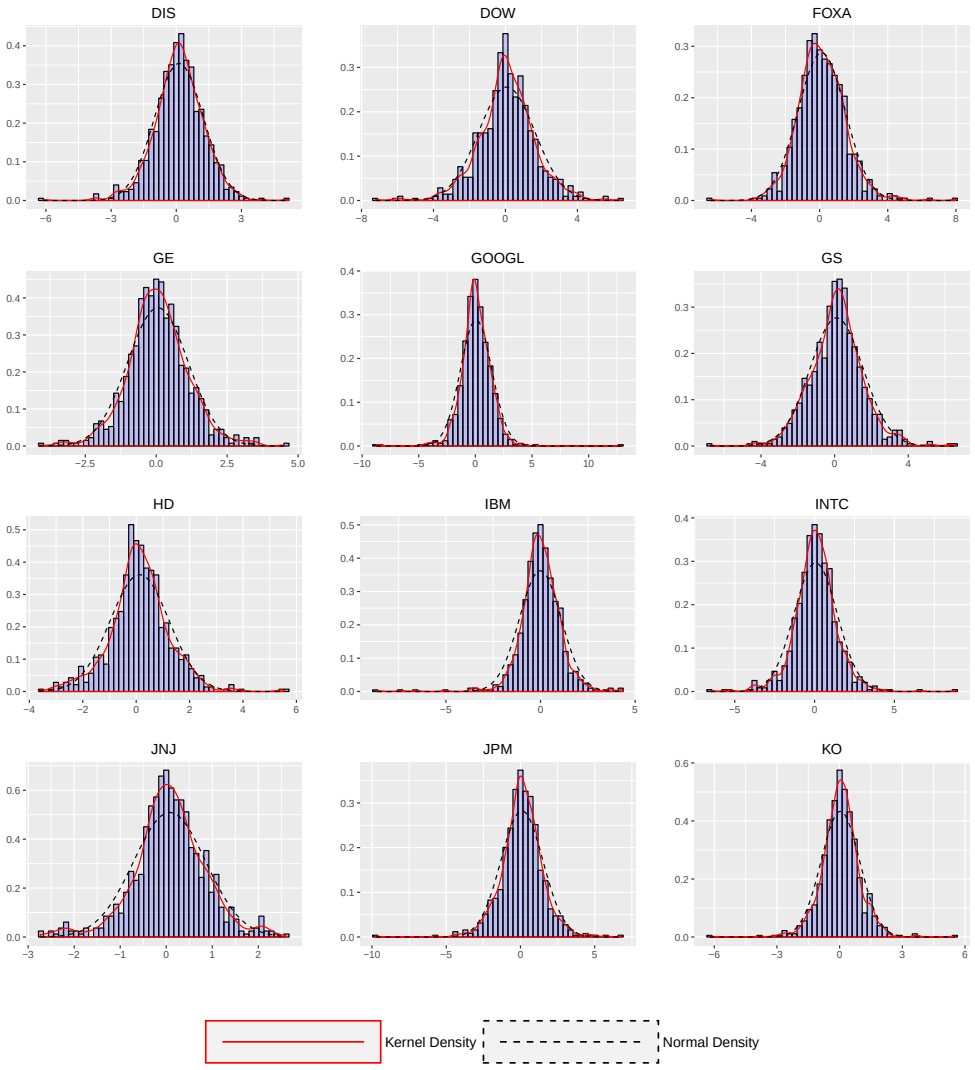


Figure 5.2b: Comparisons between the Sample Return Distributions and Normal Distributions

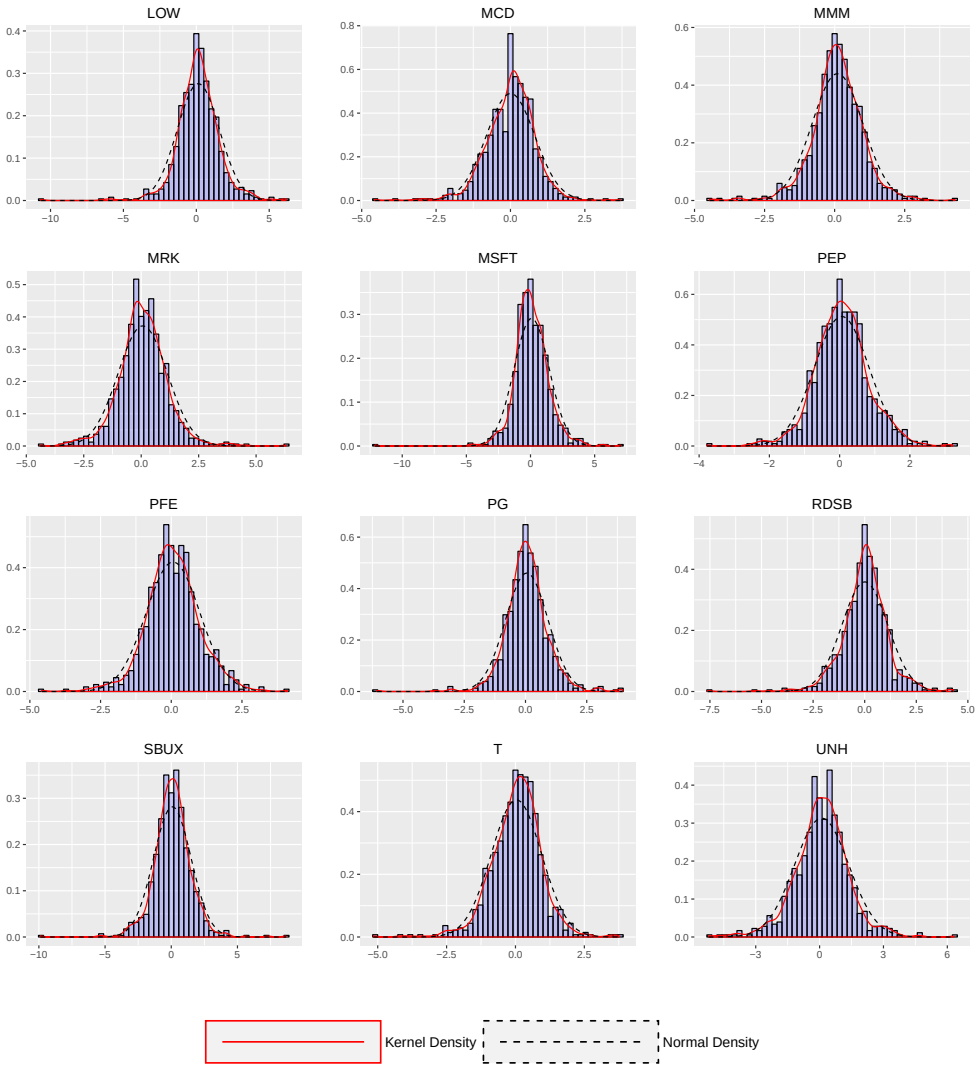


Figure 5.2c: Comparisons between the Sample Return Distributions and Normal Distributions

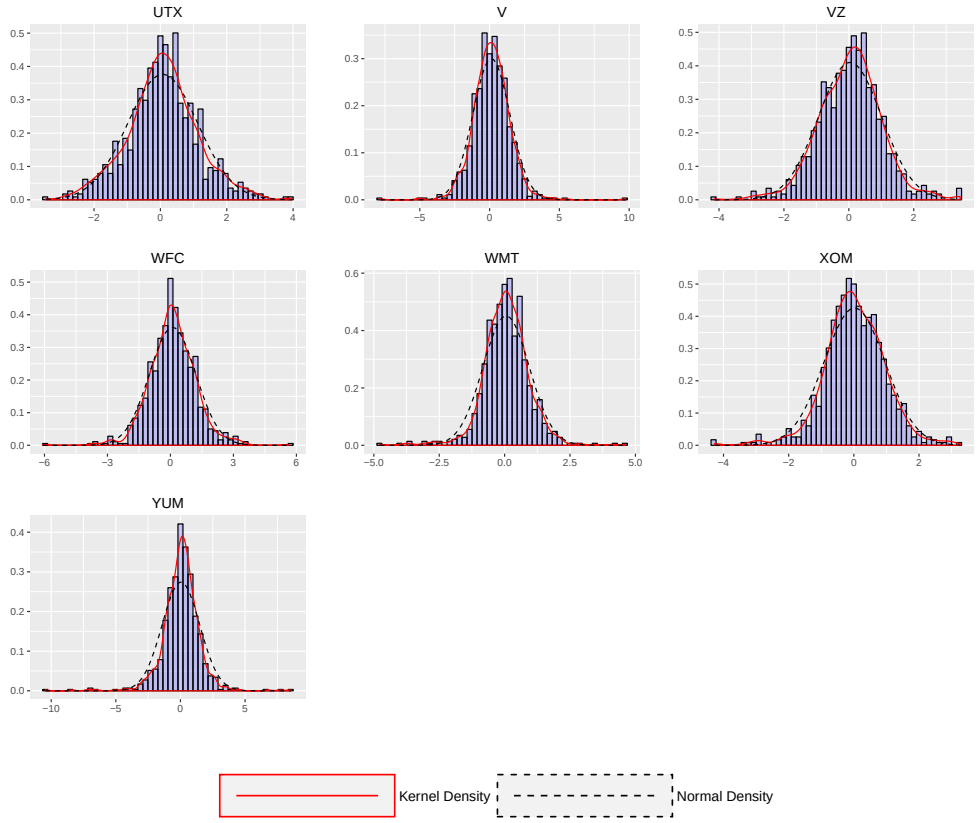


Figure 5.2d: Comparisons between the Sample Return Distributions and Normal Distributions

The histograms shown in Figures 5.2a to 5.2d exhibits similar qualitative features possessed by all stocks. That is, all stocks' kernel density estimates have higher probability mass concentrated around means than normal distributions.

that return volatilities and volumes are driven by different processes, as suggested by Tauchen and Pitts (1983) and Liesenfeld (2001).

Figures 5.4a to 5.4d plot the histograms for all stocks' logarithmic de-trended volumes. Here I took the logarithms to counter-balance the right-skewness in the distributions of de-trend volumes, as shown in Table 5.4. The procedure seems worked well, since the resulting histograms and kernel density estimates resembles normal distribution densities with the same means and variances. This also suggests that the distributions of de-trended volumes may be well-described by log-normal distributions. Again, this feature also contrasts with that of returns, which cannot be closely described by normal distributions.

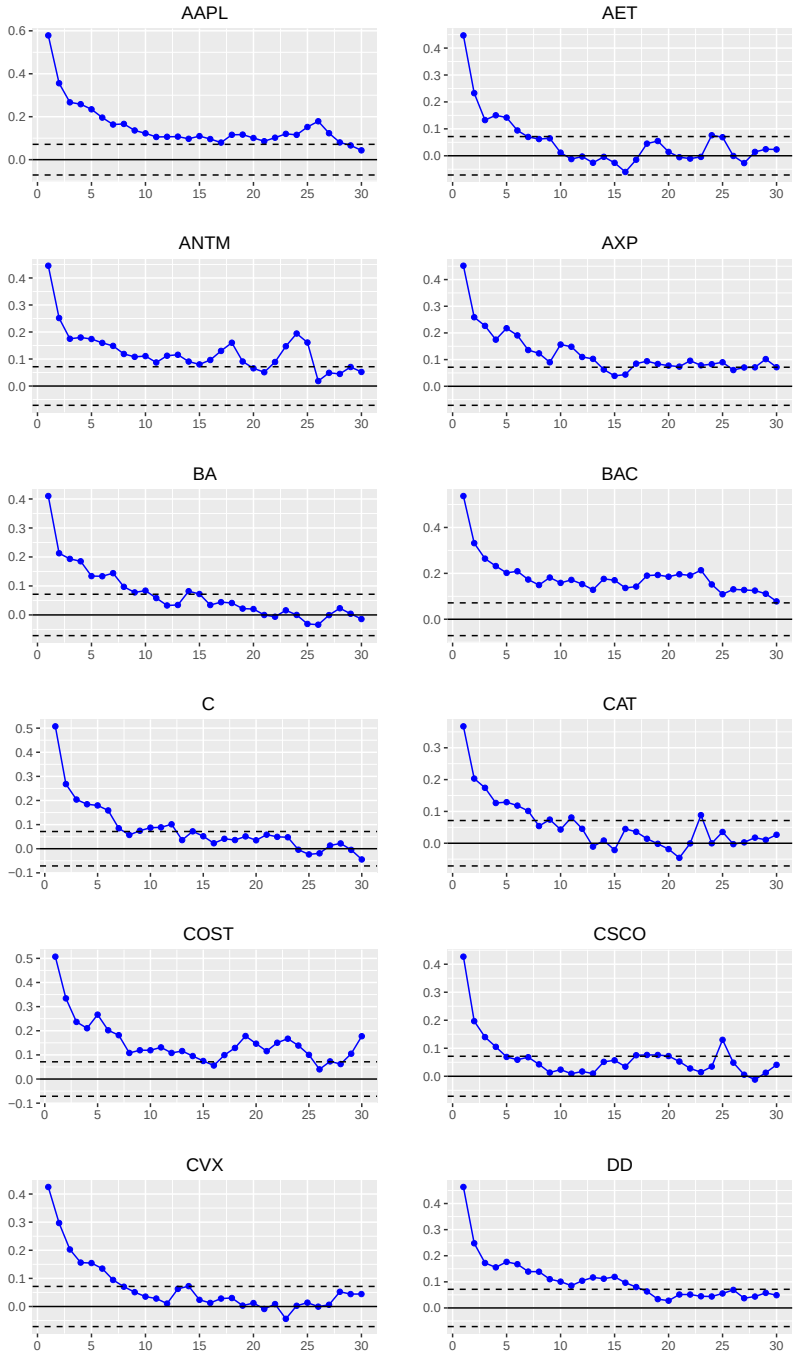


Figure 5.3a: Autocorrelation Functions of Detrended Volumes
These graphs plot the auto correlation coefficients of de-trended volumes for lags between 1 and 30. The black dotted lines mark the 5% confidence intervals.

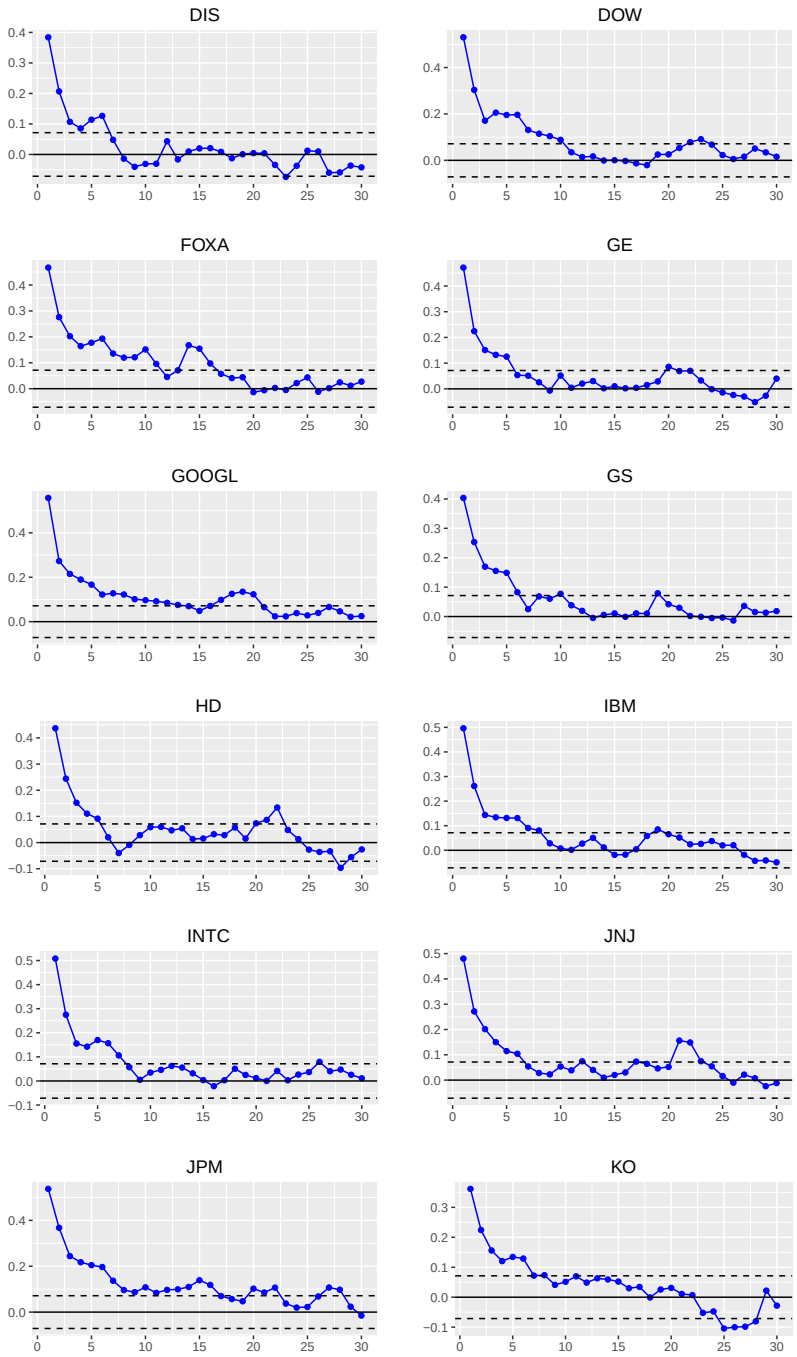


Figure 5.3b: Autocorrelation Functions of Detrended Volumes
These graphs plot the auto correlation coefficients of de-trended volumes for lags between 1 and 30. The black dotted lines mark the 5% confidence intervals. The original data is downloaded from Yahoo Finance.

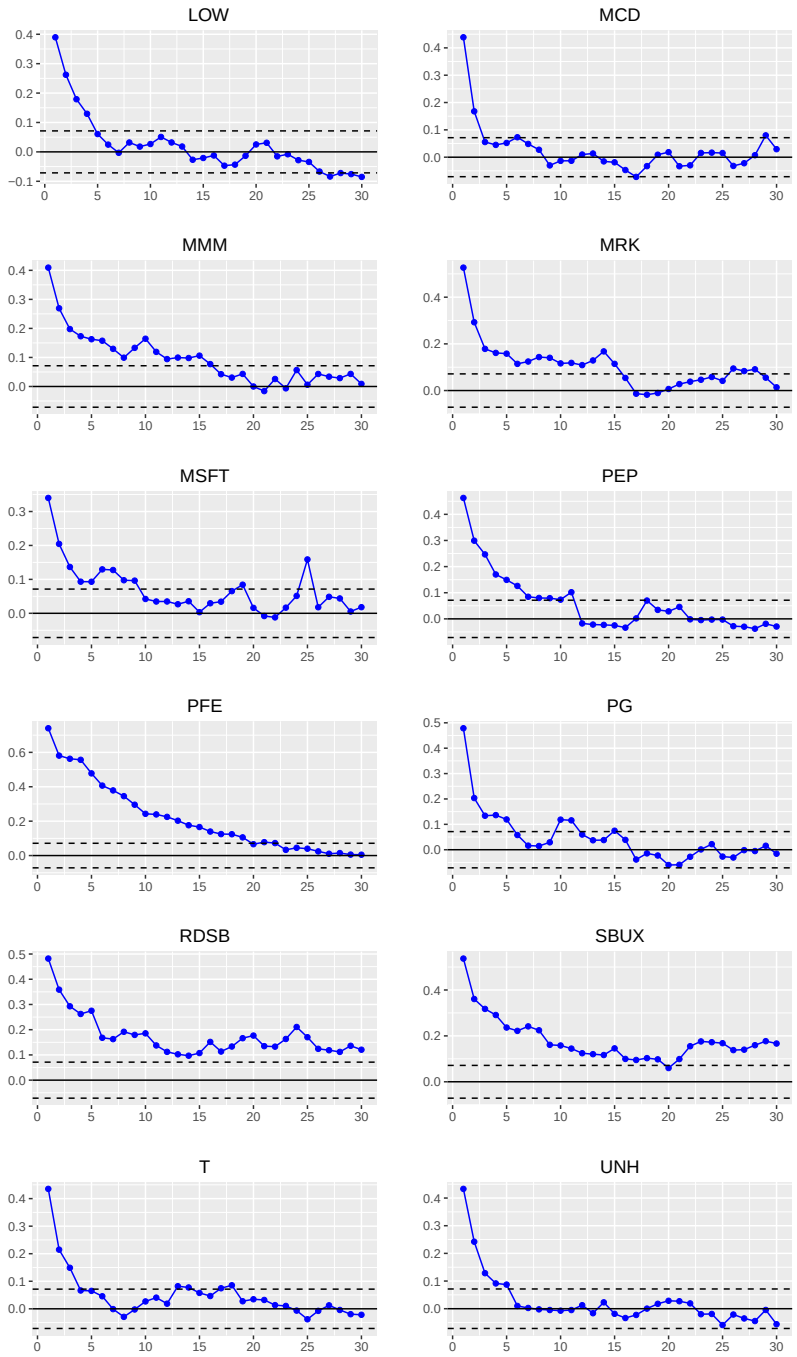


Figure 5.3c: Autocorrelation Functions of Detrended Volumes
These graphs plot the auto correlation coefficients of de-trended volumes for lags between 1 and 30. The black dotted lines mark the 5% confidence intervals. The original data is downloaded from Yahoo Finance.

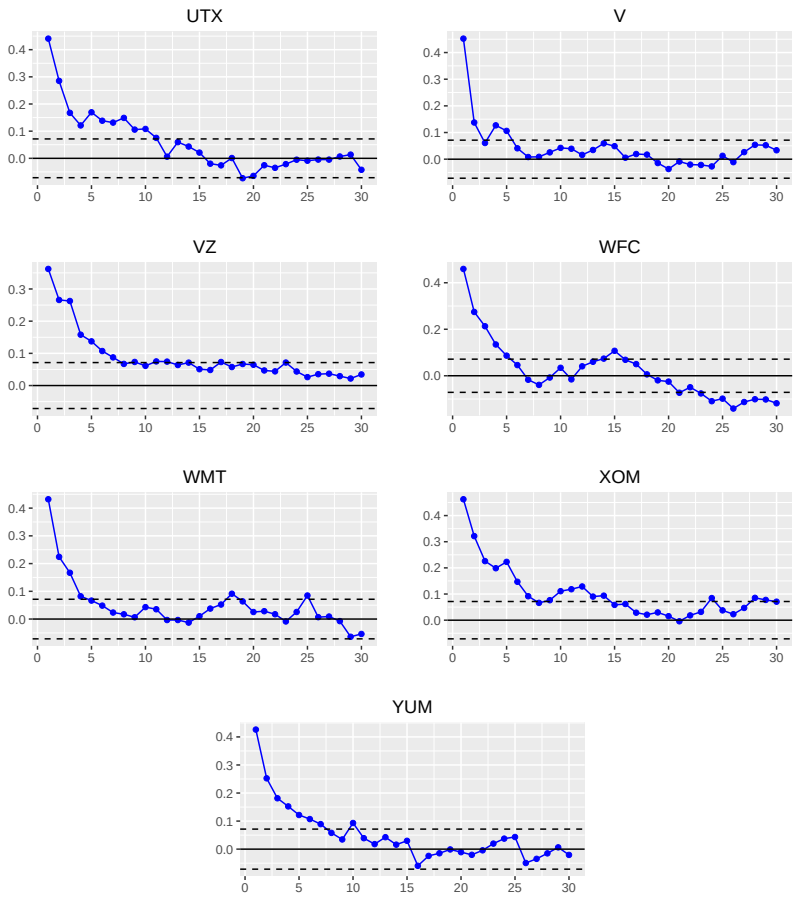


Figure 5.3d: Autocorrelation Functions of Detrended Volumes
These graphs plot the auto correlation coefficients of de-trended volumes for lags between 1 and 30. The black dotted lines mark the 5% confidence intervals. The original data is downloaded from Yahoo Finance.

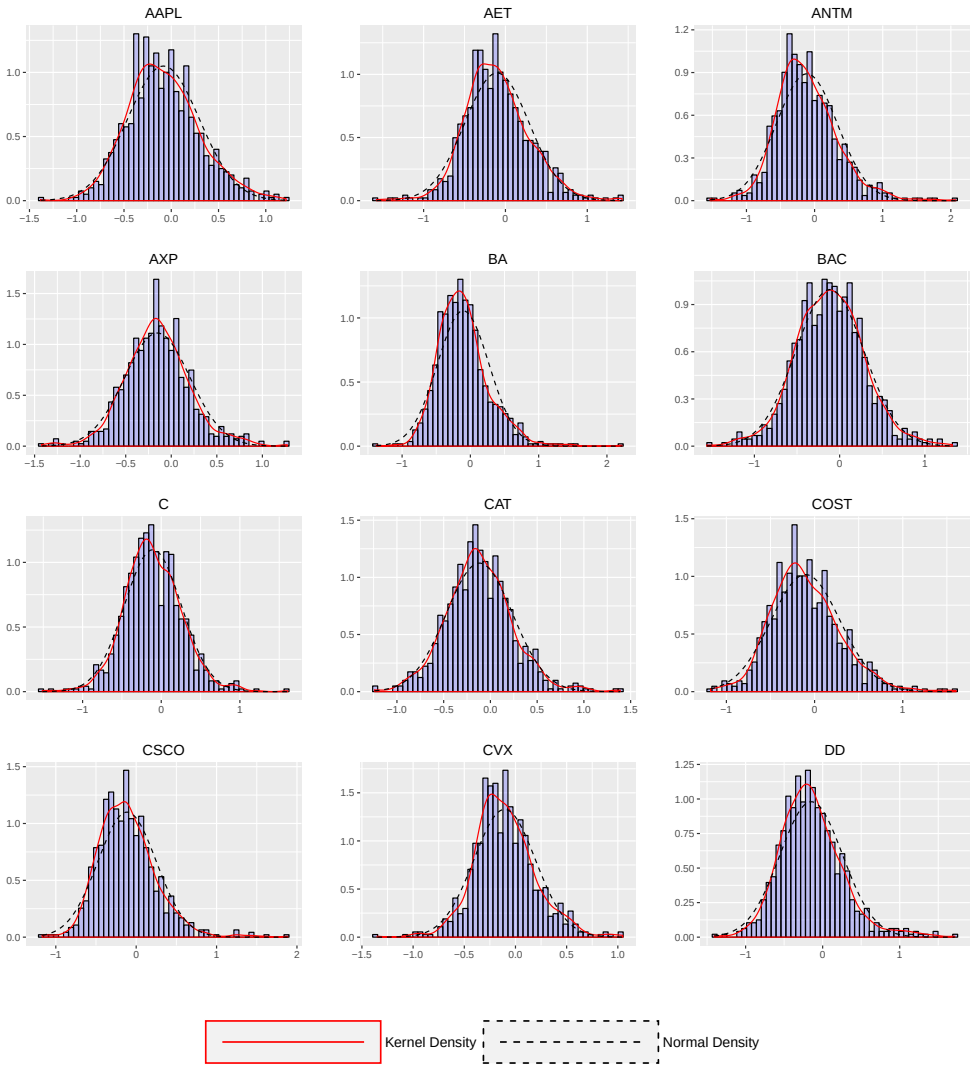


Figure 5.4a: Comparisons between the Sample Distributions of Logarithmic De-trended Volumes and Normal Distributions

These graphs plot the histograms of logarithmic de-trended volumes. Two lines were added to the plots. The red solid lines are the kernel density estimates for logarithmic de-trended volumes, while the black dotted lines are the densities of normal distributions of the same means and variances. The original data is downloaded from Yahoo Finance.

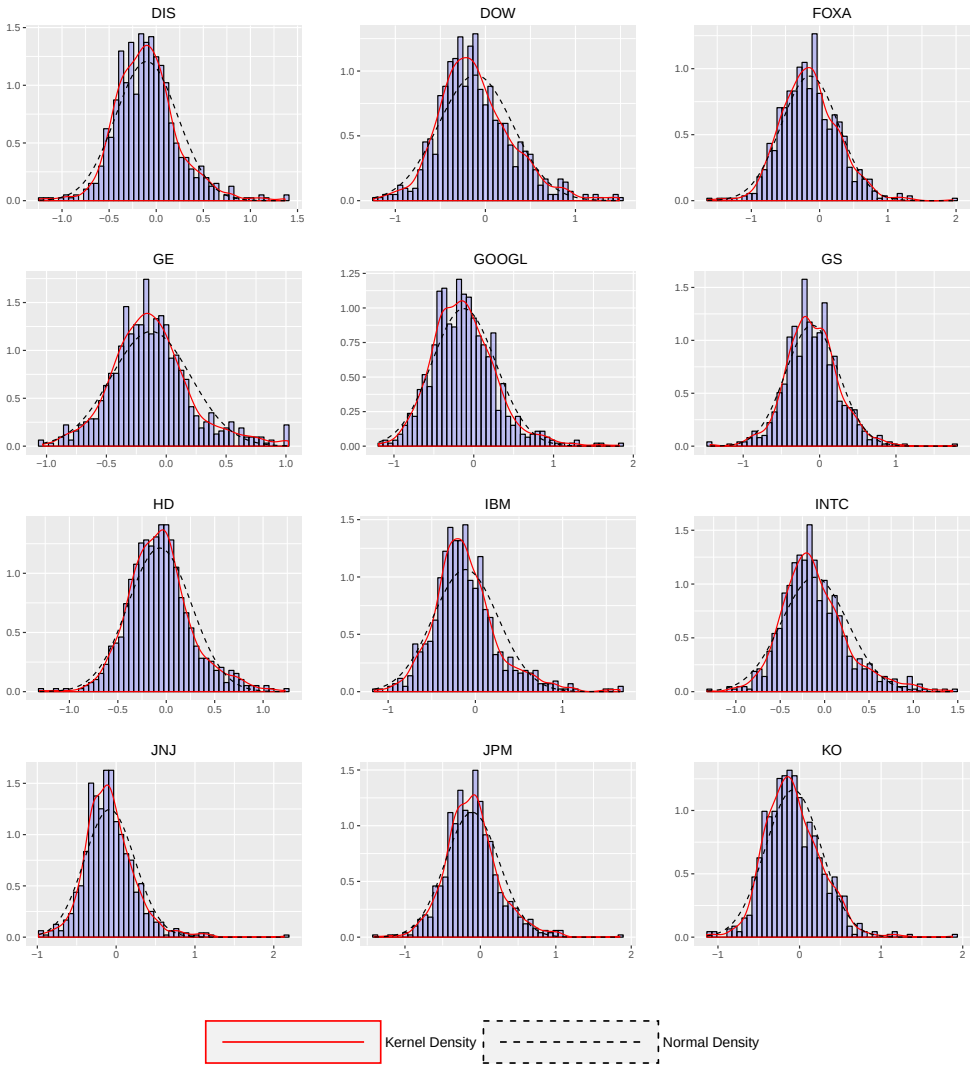


Figure 5.4b: Comparisons between the Sample Distributions of Logarithmic De-trended Volumes and Normal Distributions

These graphs plot the histograms of logarithmic de-trended volumes. Two lines were added to the plots. The red solid lines are the kernel density estimates for logarithmic de-trended volumes, while the black dotted lines are the densities of normal distributions of the same means and variances. The original data is downloaded from Yahoo Finance.

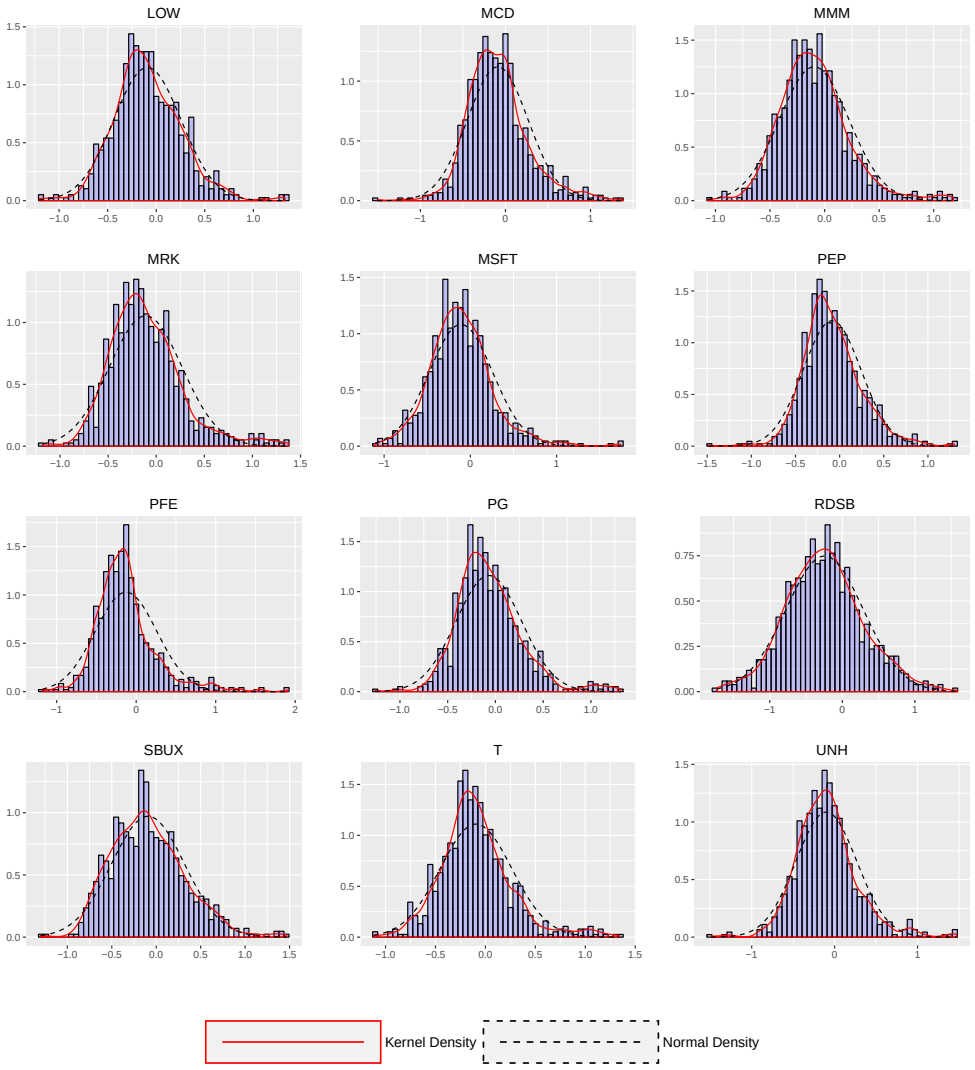


Figure 5.4c: Comparisons between the Sample Distributions of Logarithmic De-trended Volumes and Normal Distributions

These graphs plot the histograms of logarithmic de-trended volumes. Two lines were added to the plots. The red solid lines are the kernel density estimates for logarithmic de-trended volumes, while the black dotted lines are the densities of normal distributions of the same means and variances. The original data is downloaded from Yahoo Finance.

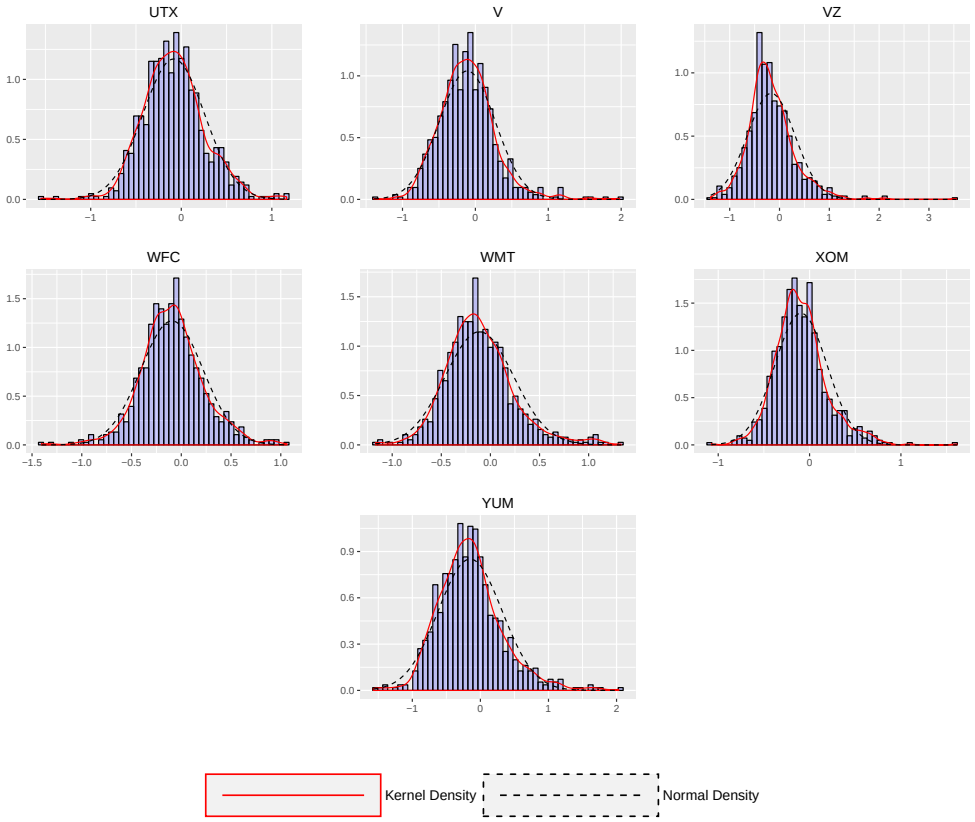


Figure 5.4d: Comparisons between the Sample Distributions of Logarithmic De-trended Volumes and Normal Distributions

These graphs plot the histograms of logarithmic de-trended volumes. Two lines were added to the plots. The red solid lines are the kernel density estimates for logarithmic de-trended volumes, while the black dotted lines are the densities of normal distributions of the same means and variances. The original data is downloaded from Yahoo Finance.

Table 5.2: The List of All Stocks Used in Real Data Estimation. Data Source: Yahoo Finance (`http://finance.yahoo.com`).

Sector	Company	
Basic Materials	Chevron Corporation (CVX)	E.I. du Pont de Nemours and Company (DD)
	Exxon Mobil Corporation (XOM)	Royal Dutch Shell plc (RDS-B)
	The Dow Chemical Company (DOW)	
Consumer Goods	Apple Inc. (AAPL)	The Coca-Cola Company (KO)
	The Procter & Gamble Company (PG)	
	Pepsico, Inc. (PEP)	
Financial	American Express Company (AXP)	The Goldman Sachs Group, Inc. (GS)
	Bank of America Corporation (BAC)	JPMorgan Chase & Co. (JPM)
	Visa Inc. (V)	
	Citigroup Inc. (C)	Wells Fargo & Company (WFC)
Health Care	Johnson & Johnson (JNJ)	Merck & Co. Inc. (MRK)
	Pfizer Inc. (PFE)	United Health Group Incorporated (UNH)
	Aetna Inc. (AET)	Anthem, Inc. (ANTM)
Industrial Goods	The Boeing Company (BA)	Caterpillar Inc. (CAT)
	General Electric Company (GE)	United Technologies Corporation (UTX)
	3M Company (MMM)	
Services	The Walt Disney Company (DIS)	The Home Depot, Inc. (HD)
	McDonald's Corp. (MCD)	Wal-Mart Stores Inc. (WMT)
	Twenty-First Century Fox, Inc. (FOXA)	Starbucks Corporation (SBUX)
	Yum! Brands, Inc. (YUM)	Costco Wholesale Corporation (COST)
	Lowe's Companies, Inc. (LOW)	
Technology	Alphabet Inc. (GOOGL)	AT&T Inc. (T)
	Cisco Systems, Inc. (CSCO)	Intel Corporation (INTC)
	Microsoft Corporation (MSFT)	International Business Machines Corporation (IBM)
	Verizon Communications Inc. (VZ)	

Table 5.3: Descriptive Statistics of Stock Return Data

	AAPL	AET	ANTM	AXP	BA	BAC	C	CAT	COST	CSCO	CVX
Mean	0.09	0.10	0.09	0.10	0.08	0.16	0.10	0.01	0.08	0.07	0.02
Std.Dev	1.69	1.44	1.47	1.18	1.24	1.83	1.73	1.36	0.95	1.45	1.04
Skewness	-0.51	-0.03	-0.84	-0.09	-0.10	0.23	-0.03	-0.09	0.48	-0.35	-0.41
Kurtosis	7.64	2.53	9.04	0.92	1.78	1.72	1.89	1.97	2.86	18.58	2.45
Maximum	8.50	6.33	7.40	5.00	5.20	8.26	7.40	5.77	6.10	11.90	4.16
Minimum	-13.19	-8.56	-12.86	-4.39	-5.48	-7.41	-8.55	-6.26	-3.24	-11.61	-5.57
p-Value	0.04	0.29	0.67	0.67	0.97	0.12	0.30	0.02	0.49	0.98	0.50
	DD	DIS	DOW	FOXA	GE	GOOGL	GS	HD	IBM	INTC	JNJ
Mean	0.08	0.13	0.07	0.12	0.06	0.07	0.11	0.13	-0.01	0.07	0.07
Std.Dev	1.10	1.13	1.56	1.39	1.07	1.38	1.44	1.10	1.10	1.34	0.78
Skewness	-0.66	-0.25	-0.16	0.43	0.10	0.41	0.10	0.18	-1.23	0.16	-0.21
Kurtosis	8.33	1.93	1.84	2.41	1.28	13.54	1.78	1.90	10.13	4.51	1.31
Maximum	5.14	5.16	6.43	7.99	4.51	12.92	6.57	5.54	4.34	8.87	2.60
Minimum	-9.50	-6.14	-7.22	-6.43	-4.14	-8.75	-6.77	-3.65	-8.64	-6.51	-2.73
p-Value	0.38	0.43	0.47	0.60	0.34	0.99	0.15	0.13	0.46	0.04	0.29
	JPM	KO	LOW	MCD	MMM	MRK	MSFT	PEP	PFE	PG	RDSB
Mean	0.09	0.04	0.14	0.00	0.10	0.07	0.09	0.06	0.06	0.05	0.01
Std.Dev	1.41	0.92	1.45	0.81	0.91	1.07	1.37	0.78	0.95	0.87	1.11
Skewness	-0.31	-0.23	-0.51	-0.53	-0.31	0.29	-0.51	-0.03	-0.07	-0.14	-0.58
Kurtosis	4.54	5.21	5.98	3.15	2.90	3.00	9.66	1.58	1.69	5.45	4.64
Maximum	6.79	5.53	6.18	3.69	4.30	6.29	7.03	3.22	4.11	3.96	4.30
Minimum	-9.74	-6.22	-10.66	-4.56	-4.46	-4.39	-12.10	-3.77	-4.57	-6.06	-7.60
p-Value	0.35	0.73	0.05	0.48	0.49	0.51	0.02	0.37	0.38	0.53	0.00
	SBUX	T	UNH	UTX	V	VZ	WFC	WMT	XOM	YUM	
Mean	0.08	0.03	0.10	0.07	0.13	0.04	0.10	0.06	0.02	0.04	
Std.Dev	1.42	0.91	1.27	1.06	1.33	0.97	1.10	0.89	0.94	1.45	
Skewness	-0.06	-0.49	-0.08	0.08	0.24	-0.07	-0.05	-0.24	-0.30	-0.67	
Kurtosis	6.15	3.13	2.06	0.57	5.69	1.21	2.48	4.15	1.90	9.37	
Maximum	8.67	3.76	6.32	3.96	9.75	3.43	5.62	4.61	3.27	8.54	
Minimum	-9.88	-5.16	-5.21	-3.44	-7.84	-4.13	-6.08	-4.77	-4.26	-10.45	
p-Value	0.63	0.85	0.66	0.94	0.36	0.16	0.45	0.43	0.29	0.51	

The daily observations on returns and volumes of the 43 stocks cover the period between Jan 1, 2012 and Dec 31, 2014, with 754 observations in total for each stock. Returns are calculated as the natural logarithms of two consecutive close prices, which have been adjusted for stock splits and dividends by the data source, Yahoo Finance. The p-Value lines report the p-Values of Ljung-Box tests with 30 lags. 6 of 43 stocks have p-Values smaller than 0.1.

Table 5.4: Descriptive Statistics for De-trended Volumes

	AAPL	AET	ANTM	AXP	BA	BAC	C	CAT	COST	CSCO	CVX
Mean	0.99	0.98	1.00	0.92	0.98	0.97	0.96	0.95	0.98	0.96	0.94
Std.Dev	0.42	0.45	0.61	0.37	0.54	0.42	0.40	0.39	0.47	0.48	0.31
Skewness	1.73	2.39	4.38	2.15	6.43	1.83	2.72	2.41	2.91	4.73	1.81
Kurtosis	4.46	10.16	33.52	8.82	78.96	6.20	16.96	11.47	15.17	38.68	5.91
Maximum	3.37	4.15	7.94	3.58	9.20	3.75	5.00	4.01	4.98	6.59	2.85
Minimum	0.25	0.21	0.22	0.24	0.25	0.21	0.22	0.29	0.31	0.31	0.26
p-Value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cor.	0.49	0.56	0.48	0.45	0.54	0.56	0.50	0.57	0.52	0.69	0.42
	DD	DIS	DOW	FOXA	GE	GOOGL	GS	HD	IBM	INTC	JNJ
Mean	0.95	0.96	0.99	0.96	0.94	0.97	0.97	0.99	0.97	0.97	0.98
Std.Dev	0.50	0.38	0.50	0.50	0.36	0.51	0.39	0.37	0.49	0.45	0.45
Skewness	3.41	2.82	2.75	3.89	2.11	3.87	3.70	2.00	3.83	2.60	7.32
Kurtosis	18.51	14.09	11.73	33.43	6.35	26.21	35.31	6.45	23.60	10.11	103.24
Maximum	5.55	3.92	4.42	7.06	2.78	6.20	5.95	3.41	5.24	4.23	8.43
Minimum	0.24	0.29	0.29	0.19	0.36	0.30	0.24	0.27	0.31	0.27	0.37
p-Value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cor.	0.50	0.47	0.52	0.53	0.54	0.62	0.41	0.53	0.67	0.57	0.40
	JPM	KO	LOW	MCD	MMM	MRK	MSFT	PEP	PFE	PG	RDSB
Mean	0.96	0.99	0.97	0.98	0.96	0.97	0.97	0.97	0.97	1.00	0.91
Std.Dev	0.42	0.42	0.40	0.42	0.37	0.48	0.48	0.38	0.58	0.43	0.56
Skewness	4.01	4.32	2.64	2.49	2.65	2.87	4.43	2.49	4.88	2.88	2.27
Kurtosis	36.93	42.00	12.79	9.38	10.68	11.06	33.11	10.93	34.86	11.52	7.86
Maximum	6.35	6.52	3.76	3.98	3.26	3.95	5.71	3.66	6.75	3.72	4.90
Minimum	0.25	0.32	0.30	0.22	0.34	0.31	0.33	0.23	0.31	0.28	0.18
p-Value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cor.	0.67	0.45	0.54	0.52	0.48	0.46	0.60	0.52	0.28	0.48	0.26
	SBUX	T	UNH	UTX	V	VZ	WFC	WMT	XOM	YUM	
Mean	1.00	0.98	0.98	0.98	0.97	1.00	0.96	0.96	0.95	0.98	
Std.Dev	0.49	0.44	0.45	0.37	0.53	1.39	0.32	0.40	0.33	0.64	
Skewness	2.50	2.74	3.23	1.85	5.18	19.45	1.66	2.57	3.73	4.20	
Kurtosis	10.24	10.41	16.53	6.23	44.54	458.43	5.01	9.86	32.49	28.29	
Maximum	4.32	3.82	4.36	3.20	7.23	34.72	2.89	3.78	5.02	7.64	
Minimum	0.27	0.33	0.23	0.21	0.25	0.25	0.25	0.31	0.34	0.21	
p-Value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Cor.	0.57	0.41	0.58	0.27	0.63	0.20	0.40	0.64	0.40	0.64	

The first six lines of the table has self-explanatory line names. Lines with name p-Value give the p-Values of Ljung-Box tests for no autocorrelation. The lines with name Cor. give the correlation coefficients between squared returns and de-trended volumes. The raw volume data is downloaded from Yahoo Finance, which is then fed to a de-trending procedure. Specifically, the de-trending procedure first finds two-year centered moving averages as trend values, then de-trend the raw volumes by taking the ratios between raw volumes and corresponding trend values.

5.3 Estimation Results on Individual Stocks

Using return and volume data of the 43 individual stocks described above, I have estimated 11 bi-stock models and 7 tri-stock models. Tables 5.5 (bi-stock models) and 5.6 (tri-stock models) have collected the results of these estimates. In the following part, I will move on to explaining the results of these estimations. The large sample properties of Maximum Likelihood Estimation indicate these parameter estimates are asymptotically normally distributed.

Starting from the estimation for the conditional mean return parameters, μ_{k1} for $k = 1, 2, 3$, a result to observe is that, the signs of $\widehat{\mu}_{k1}$ are consistent with the mean returns shown in Table 5.3. Among all the stocks used in estimation, only *IBM* has negative negative mean returns in Table 5.3, and it also has a negative $\widehat{\mu}_{k1}$ in Table 5.5, though not significantly different from 0. As a matter of fact, $\widehat{\mu}_{k1}$ are close to the mean returns given in Section 5.2 for all stocks. And the stocks having significant estimates for μ_{k1} overlap by a large extent with the stocks having significant estimates for conditional mean returns in Table 5.3. For example, the most significant estimates include μ_{11} in columns 3 (*COST*), 5 (*DIS*), 9 (*HD*), and μ_{21} in columns 2 (*V*), 5 (*FOXA*), 9 (*LOW*) in Table 5.5, and μ_{11} in columns 2 (*BAC*), 6 (*JNJ*), μ_{21} in column 3 (*MMM*), and μ_{31} in column 2 (*WFC*) in Table 5.6. These stocks also have the most significant estimates for expected returns in Table 5.3.

Tauchen and Pitts (1983) have assumed the parameter μ_{k1} to be zero when they built a single security model to explain the trading data of future contracts on treasury notes. Results in Tables 5.5 and 5.6 suggest that such an assumption may need reconsideration. Andersen (1996), Liesenfeld (1998) and Liesenfeld (2001) have obtained similar results that the conditional mean parameter estimate $\widehat{\mu}_{k1}$ may be significantly different from zero.

The estimated conditional mean parameter of trading volumes, $\widehat{\mu}_{k2}$ for $k = 1, 2, 3$ are close to one for all stocks, which are also close to the descriptive statistics in Table 5.4. The standard errors for $\widehat{\mu}_{k2}$ are small, indicating the significant difference of these estimates from zero.

It is interesting to see that the parameter estimates for both the conditional mean return and the conditional mean volume are close to the corresponding descriptive statistics. This is actually caused by the specification of Equation (2.8), which I have copied here:

$$\boldsymbol{\Lambda}_t = \boldsymbol{B} \cdot \boldsymbol{\Lambda}_{t-1} + \boldsymbol{\varepsilon}_t.$$

The unconditional mean of $\boldsymbol{\Lambda_t}$ is $\mathrm{E}(\boldsymbol{\Lambda}_t) = \mathbf{0}$. And since the information variable $I_{tk} = \exp(\lambda_{tk})$, the unconditional mean $\mathrm{E}(\boldsymbol{I}_{tk})$ is approximately

equal to **1**. According to the distribution specification for the vector of return and volume $\boldsymbol{Y}_{tk} = (R_{tk},\ V_{tk})$ in Equation (2.4):

$$\boldsymbol{Y}_{tk} \mid I_{tk} \sim \mathcal{N}\left(I_{tk} \cdot \underset{2\times 1}{\boldsymbol{\mu}_k},\ I_{tk} \cdot \underset{2\times 2}{\boldsymbol{\Sigma}_k}\right),$$

the unconditional mean of $\boldsymbol{Y}_{tk}$ is:

$$\mathrm{E}\left(\boldsymbol{Y}_{tk}\right) = \mathrm{E}\left[\mathrm{E}\left(\boldsymbol{Y}_{tk}|I_{tk}\right)\right] = \mathrm{E}\left(I_{tk}\right) \cdot \boldsymbol{\mu_k},$$

which is close to $\boldsymbol{\mu}_k$.

Next, let us move on to examining the estimates for entries of $\boldsymbol{\Sigma}_k$. For the variance of $\boldsymbol{Y}_{tk}$, there is

$$\begin{aligned}
&\mathrm{Var}(\boldsymbol{Y}_{tk}) \\
=&\mathrm{E}\left[\mathrm{Var}\left(\boldsymbol{Y}_{tk}|I_{tk}\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\boldsymbol{Y}_{tk}|I_{tk}\right)\right] \\
=&\mathrm{E}\left[I_{tk}\right]\boldsymbol{\Sigma}_k + \mathrm{Var}\left(I_{tk}\right)\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T
\end{aligned}$$

For I_{tk} is distributed as log-normal, the exact expressions for its mean and variance are:

$$\begin{aligned}
\mathrm{E}\left(I_{tk}\right) &= \exp\left[\frac{\mathrm{Var}\left(\lambda_{tk}\right)}{2}\right], \\
\mathrm{Var}\left(I_{tk}\right) &= \left\{\exp\left[\mathrm{Var}\left(\lambda_{tk}\right)\right] - 1\right\}\exp\left[\mathrm{Var}\left(\lambda_{tk}\right)\right],
\end{aligned}$$

where I have used the condition that $\mathrm{E}\left(\lambda_{tk}\right) = 0$. To find the value of $\mathrm{Var}\left(I_{tk}\right)$, I have used the following method: With the singular value decomposition of $\boldsymbol{B}$, such that $\boldsymbol{B} = \boldsymbol{P}^{-1}\boldsymbol{D}\boldsymbol{P}$, rewrite Equation (2.5) as

$$\boldsymbol{\eta}_t = \boldsymbol{D}\boldsymbol{\eta}_{t-1} + \boldsymbol{\delta}_t,$$

where $\boldsymbol{\eta}_t = \boldsymbol{P}\boldsymbol{\lambda}_t$, and $\boldsymbol{\delta}_t = \boldsymbol{P}\boldsymbol{\varepsilon}_t$. Then $\mathrm{Var}(\boldsymbol{\eta}_t)$ satisfies the following equation:

$$\mathrm{Var}(\boldsymbol{\eta}_t) = \boldsymbol{D}\mathrm{Var}(\boldsymbol{\eta}_t)\boldsymbol{D}^T + \mathrm{Var}(\boldsymbol{\delta}_t), \tag{5.1}$$

where $\mathrm{Var}(\boldsymbol{\delta}_t) = \boldsymbol{P}\mathrm{Var}(\boldsymbol{\varepsilon}_t)\boldsymbol{P}^T = \boldsymbol{P}\boldsymbol{V}\boldsymbol{P}^T$. Since $\boldsymbol{D}$ is diagonal, the entries of $\mathrm{Var}(\boldsymbol{\eta}_t)$ can be solved for individually, which can then be used to compute

$$\mathrm{Var}(\boldsymbol{\lambda}_t) = \boldsymbol{P}^{-1}\mathrm{Var}(\boldsymbol{\eta}_t)\left(\boldsymbol{P}^{-1}\right)^T$$

Thus to compare the estimation results with the descriptive statistics, I have used the point estimates shown in Tables 5.5 and 5.6 to compute $\mathrm{E}(\boldsymbol{Y}_{tk})$ and $\mathrm{Var}\left(\boldsymbol{Y}_{tk}\right)$, the results of which are shown in Table 5.9. Columns 2 to 5 of Table 5.9 are the sample moments (means and standard deviations, actually) of returns and trading volumes copied from Tables 5.3 and 5.4,

while Columns 6 to 9 show the same moments calculated from estimation results. The sample moments and estimated moments of returns are close to each other, while those of trading volumes are more discrepant. Please note that there are 40 stock tickers listed in Table 5.9, with *CSCO*, *INTC* and *IBM* missing. The reason is that $\widehat{\boldsymbol{B}}$ in the estimated model for these three stocks has complex eigenvalues, which eventually causes the estimated moments of these three stocks to be complex numbers. Thus I dropped them from the comparison.

Due to the large number of stock tickers, it is difficult to see straightforwardly on how much the differences are between the sample and estimated moments. Thus using the data in Table 5.9, I have made four scatter plots, as shown in Figure 5.6 to examine the estimation results. In each of the scatter plots, estimated moments (either mean of standard deviation of returns or trading volumes) are plotted on the horizontal axis, and the vertical axis plots the corresponding sample moments. For example, the top-left panel plots sample means of returns (vertical axis) against estimated means of returns (horizontal axis). For each scatter plot, I have run an Ordinary Least Squares (OLS) regression of sample moments on the corresponding estimated moments, the intercept ($\widehat{\alpha}$), slope ($\widehat{\beta}$) and goodness of fit (R^2) of which have been put in the subtitles. Ideally, the estimated intercept should be close to 0, while the estimated slope should be close to 1. From the plots and parameter estimates, we can clearly see that the estimated moments of returns conform better with the sample means, while the estimated moments of volumes is relatively far from their sample counterparts. This is partly due to the de-trending procedure of trading volume, which rescaled sample mean volumes to a very small range, and thus causing the R^2 in the top-right panel to be small. For the same reason, the standard deviations in the top-right panels are much bigger than those in other figures. Another reason that might have caused the discrepancy in estimated moments vs. sample moments comparison between returns and volumes may be due the hypothesis raised by Liesenfeld (2001), that the volumes and returns are driven by different latent processes, and thus my current model is not versatile enough to capture that.

In addition to the above comparison between sample moments and estimated moments, there are two features shown in Tables 5.5 and 5.6 regarding the conditional variance matrix $\boldsymbol{\Sigma}_k$ that should be mentioned. One is that the conditional return variance parameter $\Sigma_{k,11}$ is much bigger than the conditional volume variance parameter $\Sigma_{k,22}$. Together with the fact that average de-trended volumes are close to one while average returns are close to zero, this indicates that the return series are more volatile than the volume series. Tauchen and Pitts (1983)'s model has provided an explanation for this: Arrivals of common information will move all traders' reservation prices, and thus they will affect the equilibrium price but will

not increase the trading volume.

The other feature shown by the entries of $\widehat{\boldsymbol{\Sigma}}_k$ is that the conditional covariance between return and volume is small for all cases and insignificant for most cases (30 out of 43 stocks have the Z-stat for $\boldsymbol{\Sigma}_k = 0$ smaller than 2). Previous researches, such as Tauchen and Pitts (1983), Andersen (1996) and Liesenfeld (1998, 2001) all assumed conditional independence between returns and volumes, which implies $\Sigma_{k,12} = 0$ for all k. Thus the results in Table 5.5 and 5.6 conform well with this assumption. To better examine this hypothesis, I have calculated the conditional correlations between returns and volumes, the result of which are shown in Table 5.7. The conditional correlations are small for all stocks, with the biggest absolute value being 0.117. But they are all significantly different from zero, according to their standard errors shown in parentheses. The standard errors are calculated using the delta method. Specifically, with the estimate $\widehat{\Sigma}_k$ and its standard error matrix $\widehat{\text{Var}}\left(\widehat{\Sigma}_k\right)$, the standard error of the conditional correlation between r_k and V_k

$$\widehat{\rho} = \frac{\widehat{\Sigma}_{k,12}}{\sqrt{\widehat{\Sigma}_{k,11}\widehat{\Sigma}_{k,22}}}$$

is

$$\widehat{\text{Var}}\left(\widehat{\rho}\right) = \begin{bmatrix} \frac{1}{\sqrt{\widehat{\Sigma}_{k,11}\widehat{\Sigma}_{k,22}}} \\ -\frac{\widehat{\Sigma}_{k,12}}{2\sqrt{\widehat{\Sigma}_{k,11}^3\widehat{\Sigma}_{k,22}}} \\ -\frac{\widehat{\Sigma}_{k,12}}{2\sqrt{\widehat{\Sigma}_{k,11}\widehat{\Sigma}_{k,22}^3}} \end{bmatrix}^T \widehat{\text{Var}}\left(\widehat{\Sigma}_k\right) \begin{bmatrix} \frac{1}{\sqrt{\widehat{\Sigma}_{k,11}\widehat{\Sigma}_{k,22}}} \\ -\frac{\widehat{\Sigma}_{k,12}}{2\sqrt{\widehat{\Sigma}_{k,11}^3\widehat{\Sigma}_{k,22}}} \\ -\frac{\widehat{\Sigma}_{k,12}}{2\sqrt{\widehat{\Sigma}_{k,11}\widehat{\Sigma}_{k,22}^3}} \end{bmatrix}$$

Next, let us move on to examining the entries of $\widehat{\boldsymbol{B}}$. As I have explained in Chapter 2, we should expect the diagonal entries $\widehat{B}_{kk}$ of $\widehat{\boldsymbol{B}}$ to be positive due to the autocorrelation of information arrival processes. This expectation is well confirmed by the estimation results. The 43 stocks in 18 models have their corresponding diagonal entries ($\widehat{B}_{kk}$) ranging between 0.435 and 0.865, and the corresponding standard deviations ($\widehat{\text{s.e.}}(\widehat{B}_{kk})$) ranging between 0.026 and 0.044. The ratio between $\widehat{B}_{kk}$ and $\widehat{\text{s.e.}}(\widehat{B}_{kk})$ ranges from 10.273 to 28.238. Since these ratios are asymptotically distributed as standard normal under the null hypothesis $B_{kk} = 0$, it implies that the p-values for testing $B_{kk} \leq 0$ individually are essentially zero and the evidence supporting the alternative hypothesis $B_{kk} > 0$ is strong.

In addition, we also expect that the off-diagonal values of $\widehat{\boldsymbol{B}}$ to be non-negative. If there are spillover effects in the form of cross-security historical dependency in information arrivals, then the off-diagonal values of $\boldsymbol{B}$ should be zero. Otherwise, they should be positive. In the 11 bi-stock

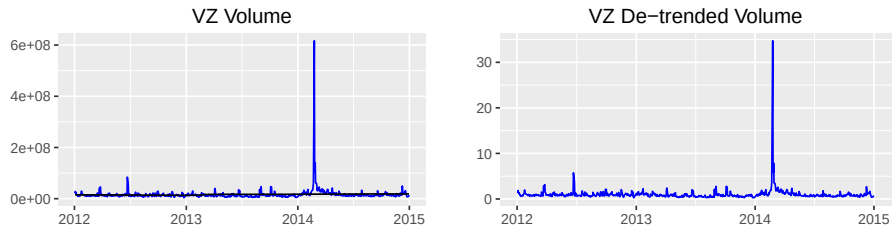


Figure 5.5: The Abnormal Observation in *VZ*'s Trading Volumes
Note: The extraordinarily high trading volume was on 02/14/2014.

and 7 tri-stock models estimated, there are 64 off-diagonal entries of $\boldsymbol{B}$ in total, among which 14 $\widehat{B}_{ij}$'s are negative. Except for $\widehat{B}_{21} = -0.16$ with s.e. $\left(\widehat{B}_{21}\right) = 0.047$ in the model of T and VZ, other $\widehat{B}_{ij}$'s range between -0.045 and -0.002, with standard errors ranging between 0.027 and 0.041. The ratio $\widehat{B}_{ij}$/s.e. $\left(\widehat{B}_{ij}\right)$ ranges between -1.430 and -0.079 so that the null hypothesis $B_{ij} = 0$ could not be rejected for these models, when tested individually.

As a matter of fact, the negative $\widehat{\boldsymbol{B}}_{21}$ in the model of T and VZ seems to be caused by VZ's abnormally high volume observation on 02/14/2014, which is plotted in Figure 5.5.

Thus the non-negativity of $\boldsymbol{B}$'s off-diagonal values is largely confirmed. Furthermore, to find the stocks between which evidence for the spillover effects is strong, I tested the hypothesis $\boldsymbol{B}_{kl} \leq 0$ with $k \neq l$, $k, l \in \{1, 2, 3\}$ for the positive $\widehat{B}_{kl}$'s. Among all the estimated stock models, 50 point estimates of $\widehat{B}_{kl}$ are positive, ranging between 0.001 and 0.229. Their standard errors range between 0.018 and 0.068, and the ratio $\widehat{B_{kl}} \Big/ \widehat{\text{s.e.}}(\widehat{B_{kl}})$ ranges from 0.024 to 5.764. Under the asymptotic normality of MLE estimator, 29 out of 50 hypothesis tests have p-values smaller than 0.1, 24 have p-values smaller than 0.05, 18 have p-values smaller than 0.01, and 13 have p-values smaller than 0.001, which indicates that at least for some of the stock models, the evidence for the existence of spillover effects is strong. The parameter estimates showing the strongest evidence for the spillover effects (p-value smaller than 0.001) include $\widehat{B}_{12}$ in columns 3 (*BA* and *GE*), 8 (*GS* and *JPM*), 10 (*KO* and *PEP*), 11 (*T* and *VZ*), and $\widehat{B}_{21}$ in columns 3 (*BA* and *GE*), 7 (*GOOGL* and *MSFT*), 9 (*HD* and *LOW*), 10 (*KO* and *PEP*) of Table 5.5, and $\widehat{B}_{13}$ in column 5 (*CVX*, *RDSB* and *XOM*), $\widehat{B}_{21}$ in column 3 (*CAT*, *MMM* and *UTX*), $\widehat{B}_{23}$ in columns 2 (*BAC*, *C* and *WFC*), 6 (*JNJ*, *MRK* and *PFE*), and $\widehat{B}_{32}$ in columns 2 (*BAC*, *C* and *WFC*), 4 (*CSCO*, *INTC* and *IBM*), 5 (*CVX*, *RDSB* and *XOM*) in Table 5.6.

To sum up the results of individual test on off-diagonal values of $\boldsymbol{B}$, we have the stocks in the two models, *BA* and *GE*, and *KO* and *PEP*, both having the most significant spillover effects in the form of cross-security his-

torical dependencies in information arrival processes. For tri-stock models, among *BAC*, *C* and *WFC*, and among *CVX*, *RDSB* and *XOM*, there are two stocks having the most significant spillover effects.

We also observe from these test results that the spillover effects are not symmetric between stocks. While the information arrivals of one stock may depend on the historical value of another stock's information arrivals, the second stock's information arrivals are not necessarily affected as much by the first one's. Such asymmetric patterns of spillover effects are most apparent for *GOOGL* vs. *MSFT*, *GS* vs. *JPM*, *HD* vs. *LOW*, *CVX* vs. *XOM*, *INTC* vs. *IBM*, and *MRK* vs. *PFE*. For example, in the model of *GS* vs. *JPM*, $\widehat{B}_{12} = 0.154$ which implies a considerably large spillover effect of JPM to GS, but the impact of GS on JPM is only 0.01. An examination into the market capitalization of these stocks shows that the information arrival process of the stock with larger capitalization tends to affect that of the stock with smaller capitalization, but not vice versa.

The final part of the estimation results are entries of $\widehat{\boldsymbol{V}}$. The diagonal entries of $\widehat{\boldsymbol{V}}$ are concentrated in a fairly small interval, between 0.054 and 0.142 (with only 4 bigger than 0.1). Since their standard errors of $\widehat{V}_{kk}$ are very small compared to $\widehat{V}_{kk}$, these estimates are all significantly greater than zero, which is a desired result.

The off-diagonal values of $\widehat{\boldsymbol{V}}$ are associated with the second form of spillover effects in information arrival processes. That is, the shocks to stocks' information arrival processes are correlated contemporarily. Again, we expect this form of spillover effects to be measured by positive numbers, which implies that the off-diagonal values of $\boldsymbol{V}$ are non-negative. The second form of information arrival's spillover effects is well documented by the off-diagonal values of $\widehat{\boldsymbol{V}}$ as shown by Tables 5.5 and 5.6. The values of V_{ij} range between 0.012 and 0.062, while their standard errors range from 0.002 to 0.004. Correspondingly, the ratio $\widehat{V}_{ij}/\text{s.e.}(\widehat{V}_{ij})$ range from to . As a result, the p-values for testing individually the null hypotheses $V_{kl} < 0$ with $k, l \in \{1, 2, 3\}$ and $k < l$, are essentially zero, giving strong evidence favoring the alternative hypothesis that $V_{kl} \geq 0$.

It is also of interest to investigate the estimate $\widehat{\text{Corr}}(\varepsilon_{tk}, \varepsilon_{tl}) = \widehat{V}_{kl}/\sqrt{\widehat{V}_{kk}\widehat{V}_{ll}}$ for the contemporary correlation between shocks to the information arrival processes of stocks k and l. Thus in Table 5.8 I have calculated the point estimates of such correlations as well as their standard deviations for each pair of stocks shown in Tables 5.5 and 5.6. The correlations are generally of considerable magnitude, ranging from 0.170 to 0.643. The z-tests for testing the null hypothesis $\text{Corr}(\varepsilon_{tk}) = 0$ individually are all rejected essentially with p-value zero.

To sum up, the two forms of spillover effects in information arrivals are supported by different degrees. The contemporary correlations in shocks to

information arrival processes are well documented in all estimated models, but only part of the stock pairs or stock triplets show cross-security historical dependencies in information arrival processes. In Tables 5.10 and 5.11, I have shown the results of jointly testing the off-diagonal entries of $\boldsymbol{B}$ and $\boldsymbol{V}$. These test results are consistent with those shown above. The off-diagonal entries of $\widehat{\boldsymbol{V}}$ are significantly different from zero for all estimated models, while the off-diagonal entries of $\widehat{\boldsymbol{B}}$ are most significant in these models: *BA* and *GE*, *KO* and *PEP*, *BAC*, *C* and *WFC*, and *CVX*, *RDSB* and *XOM*.

Table 5.5: Estimation Results Using Stock Pairs Data

Asset 1	AAPL	AXP	BA	COST	DIS	DOW	GOOGL	GS	HD	KO	T
Asset 2	PG	V	GE	WMT	FOXA	DD	MSFT	JPM	LOW	PEP	VZ
μ_{11}	0.094	0.103	0.089	0.091	0.133	0.077	0.07	0.111	0.136	0.039	0.037
	(0.054)	(0.044)	(0.045)	(0.035)	(0.041)	(0.054)	(0.046)	(0.051)	(0.039)	(0.033)	(0.033)
μ_{12}	1.003	0.991	1.034	1.06	1.004	1.007	1.024	1.008	1.035	1.062	1.048
	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)
μ_{21}	0.056	0.133	0.065	0.063	0.127	0.082	0.092	0.101	0.144	0.064	0.041
	(0.03)	(0.045)	(0.04)	(0.031)	(0.05)	(0.04)	(0.045)	(0.048)	(0.049)	(0.029)	(0.038)
μ_{22}	1.038	1.008	1.016	1.042	1.005	1.025	1.002	1.017	1.006	1.068	1.07
	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)
$\Sigma_{1,11}$	2.214	1.339	1.427	0.846	1.181	2.135	1.526	1.89	1.079	0.76	0.767
	(0.114)	(0.069)	(0.073)	(0.044)	(0.061)	(0.11)	(0.079)	(0.097)	(0.055)	(0.039)	(0.039)
$\Sigma_{1,12}$	-0.004	-0.008	-0.014	0.013	-0.009	-0.001	0.005	-0.003	-0.008	-0.004	-0.015
	(0.008)	(0.006)	(0.007)	(0.005)	(0.006)	(0.009)	(0.007)	(0.008)	(0.006)	(0.005)	(0.005)
$\Sigma_{1,22}$	0.026	0.025	0.028	0.029	0.026	0.03	0.026	0.026	0.026	0.028	0.025
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\Sigma_{2,11}$	0.641	1.46	1.084	0.678	1.783	1.124	1.506	1.669	1.754	0.585	1.031
	(0.033)	(0.075)	(0.056)	(0.035)	(0.092)	(0.058)	(0.077)	(0.086)	(0.09)	(0.03)	(0.053)
$\Sigma_{2,12}$	0.012	0.006	-0.005	-0.01	0.003	-0.02	0.004	-0.02	-0.007	0.011	0.001
	(0.005)	(0.007)	(0.006)	(0.005)	(0.008)	(0.006)	(0.007)	(0.007)	(0.008)	(0.004)	(0.005)
$\Sigma_{2,22}$	0.028	0.028	0.026	0.028	0.033	0.029	0.026	0.024	0.027	0.027	0.024
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
B_{11}	0.722	0.7	0.604	0.684	0.57	0.659	0.738	0.522	0.619	0.559	0.435
	(0.026)	(0.03)	(0.031)	(0.031)	(0.035)	(0.03)	(0.029)	(0.038)	(0.037)	(0.034)	(0.042)
B_{12}	-0.02	-0.03	0.145	0.044	0.026	0.007	-0.013	0.154	-0.012	0.124	0.166
	(0.029)	(0.029)	(0.035)	(0.034)	(0.027)	(0.028)	(0.033)	(0.036)	(0.036)	(0.034)	(0.03)
B_{21}	-0.002	-0.025	0.065	0.059	0.065	-0.026	0.099	0.01	0.226	0.131	-0.16
	(0.027)	(0.034)	(0.028)	(0.03)	(0.04)	(0.029)	(0.03)	(0.038)	(0.039)	(0.031)	(0.047)
B_{22}	0.627	0.62	0.641	0.608	0.637	0.719	0.556	0.695	0.441	0.623	0.865
	(0.029)	(0.033)	(0.031)	(0.033)	(0.032)	(0.028)	(0.034)	(0.036)	(0.038)	(0.031)	(0.033)
V_{11}	0.069	0.074	0.084	0.084	0.072	0.096	0.082	0.072	0.068	0.076	0.087
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)
V_{12}	0.012	0.04	0.019	0.034	0.036	0.036	0.034	0.036	0.034	0.021	0.062
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
V_{22}	0.073	0.096	0.066	0.08	0.099	0.093	0.088	0.07	0.076	0.061	0.108
	(0.004)	(0.005)	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.003)	(0.004)	(0.003)	(0.006)

Table 5.6: Estimation Results Using Stock Triplets Data

Asset 1	AET	BAC	CAT	CSCO	CVX	JNJ	MCD
Asset 2	ANTM	C	MMM	INTC	RDSB	MRK	SBUX
Asset 3	UNH	WFC	UTX	IBM	XOM	PFE	YUM
μ_{11}	0.108	0.168	0.012	0.071	0.023	0.081	0.004
	(0.05)	(0.065)	(0.049)	(0.046)	(0.039)	(0.029)	(0.029)
μ_{12}	1.015	1.035	1.017	1.008	1.044	1.063	1.047
	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)	(0.007)
μ_{21}	0.094	0.104	0.111	0.071	0.009	0.074	0.086
	(0.05)	(0.062)	(0.033)	(0.047)	(0.043)	(0.039)	(0.048)
μ_{22}	1.024	1.041	1.047	1.018	1.012	1.055	1.042
	(0.007)	(0.006)	(0.006)	(0.006)	(0.008)	(0.007)	(0.006)
μ_{31}	0.102	0.11	0.073	-0.011	0.025	0.069	0.037
	(0.044)	(0.04)	(0.04)	(0.037)	(0.036)	(0.037)	(0.046)
μ_{32}	1.017	1.027	1.035	1.052	1.061	1.07	1.007
	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)	(0.007)
$\Sigma_{1,11}$	1.841	2.926	1.682	1.502	1.051	0.592	0.595
	(0.095)	(0.149)	(0.087)	(0.077)	(0.054)	(0.03)	(0.03)
$\Sigma_{1,12}$	0.001	-0.001	-0.009	-0.018	-0.015	0	-0.011
	(0.008)	(0.009)	(0.007)	(0.007)	(0.005)	(0.004)	(0.004)
$\Sigma_{1,22}$	0.027	0.025	0.027	0.026	0.024	0.024	0.028
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\Sigma_{2,11}$	1.826	2.702	0.768	1.551	1.272	1.052	1.641
	(0.094)	(0.139)	(0.039)	(0.08)	(0.066)	(0.054)	(0.085)
$\Sigma_{2,12}$	0.014	-0.01	-0.001	-0.009	-0.026	0.004	0
	(0.008)	(0.009)	(0.005)	(0.007)	(0.008)	(0.006)	(0.007)
$\Sigma_{2,22}$	0.029	0.024	0.025	0.027	0.04	0.028	0.027
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
$\Sigma_{3,11}$	1.427	1.141	1.129	0.961	0.863	0.941	1.521
	(0.073)	(0.059)	(0.058)	(0.049)	(0.044)	(0.049)	(0.078)
$\Sigma_{3,12}$	-0.011	-0.007	0	-0.014	0.007	-0.003	-0.01
	(0.007)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.007)
$\Sigma_{3,22}$	0.026	0.024	0.027	0.026	0.022	0.024	0.03
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
B_{11}	0.623	0.613	0.518	0.613	0.455	0.675	0.549
	(0.039)	(0.039)	(0.037)	(0.032)	(0.042)	(0.032)	(0.034)
B_{12}	0.022	0.104	0.104	-0.045	0.045	0.012	0.069
	(0.032)	(0.047)	(0.042)	(0.032)	(0.019)	(0.029)	(0.029)
B_{13}	-0.032	0.026	0.043	0.045	0.229	0.019	0.035
	(0.041)	(0.05)	(0.04)	(0.033)	(0.044)	(0.027)	(0.024)
B_{21}	0.014	0.021	0.13	-0.021	0.091	0.024	0.051
	(0.043)	(0.037)	(0.033)	(0.031)	(0.064)	(0.037)	(0.034)
B_{22}	0.656	0.558	0.531	0.669	0.688	0.608	0.719
	(0.035)	(0.044)	(0.038)	(0.031)	(0.029)	(0.034)	(0.028)
B_{23}	0.001	0.179	0.065	0.01	0.083	0.092	0.005
	(0.046)	(0.046)	(0.036)	(0.032)	(0.068)	(0.032)	(0.024)
B_{31}	0.046	-0.022	0.061	0.002	0.075	-0.003	0.031
	(0.037)	(0.034)	(0.034)	(0.032)	(0.041)	(0.033)	(0.042)
B_{32}	0.051	0.155	-0.036	0.136	0.079	0.019	0.003
	(0.031)	(0.04)	(0.039)	(0.032)	(0.018)	(0.03)	(0.035)
B_{33}	0.545	0.517	0.61	0.618	0.559	0.79	0.664
	(0.04)	(0.042)	(0.037)	(0.033)	(0.043)	(0.028)	(0.03)
V_{11}	0.097	0.09	0.084	0.086	0.057	0.064	0.084
	(0.005)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)
V_{12}	0.058	0.051	0.03	0.022	0.021	0.03	0.026
	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
V_{13}	0.053	0.04	0.029	0.034	0.03	0.026	0.033
	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)
V_{22}	0.118	0.079	0.068	0.083	0.142	0.09	0.081
	(0.006)	(0.004)	(0.003)	(0.004)	(0.007)	(0.005)	(0.004)
V_{23}	0.051	0.038	0.034	0.029	0.025	0.032	0.029
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
V_{33}	0.088	0.064	0.071	0.085	0.054	0.069	0.13
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.007)

Table 5.7: Conditional Correlations between Returns and Volumes

AAPL	AET	ANTM	AXP	BA	BAC	C
-0.016	0.004	0.06	-0.042	-0.069	-0.005	-0.039
(0.001)	(0)	(0.003)	(0.002)	(0.004)	(0)	(0.002)
CAT	COST	CSCO	CVX	DD	DIS	DOW
-0.04	0.081	-0.092	-0.098	-0.109	-0.05	-0.004
(0.002)	(0.005)	(0.005)	(0.005)	(0.006)	(0.003)	(0)
FOXA	GE	GOOGL	GS	HD	IBM	INTC
0.013	-0.029	0.024	-0.016	-0.045	-0.091	-0.043
(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.005)	(0.002)
JNJ	JPM	KO	LOW	MCD	MMM	MRK
-0.004	-0.102	-0.028	-0.032	-0.081	-0.009	0.022
(0)	(0.006)	(0.002)	(0.002)	(0.005)	(0.001)	(0.001)
MSFT	PEP	PFE	PG	RDSB	SBUX	T
0.02	0.087	-0.022	0.091	-0.117	0.001	-0.111
(0.001)	(0.005)	(0.001)	(0.005)	(0.007)	(0)	(0.006)
UNH	UTX	V	VZ	WFC	WMT	XOM
-0.058	0.003	0.032	0.007	-0.043	-0.075	0.049
(0.003)	(0)	(0.002)	(0)	(0.002)	(0.004)	(0.003)
YUM						
-0.046						
(0.003)						

Table 5.8: Correlations between Shocks to Stocks' Information Arrival Processes

AAPL &	AXP &	BA &	COST &	DIS &	DOW &	GOOGL &	GS &
PG	V	GE	WMT	FOXA	DD	MSFT	JPM
0.17	0.469	0.262	0.413	0.423	0.385	0.404	0.51
(0.034)	(0.027)	(0.033)	(0.029)	(0.029)	(0.03)	(0.029)	(0.025)
HD &	KO &	T &	AET &	AET &	ANTM &	BAC &	BAC &
LOW	PEP	VZ	ANTM	UNH	UNH	C	WFC
0.476	0.314	0.643	0.537	0.571	0.501	0.606	0.523
(0.027)	(0.031)	(0.02)	(0.025)	(0.023)	(0.026)	(0.022)	(0.025)
C &	CAT &	CAT &	MMM &	CSCO &	CSCO &	INTC &	CVX &
WFC	MMM	UTX	UTX	INTC	IBM	IBM	RDSB
0.531	0.397	0.372	0.49	0.255	0.394	0.349	0.235
(0.024)	(0.029)	(0.03)	(0.026)	(0.033)	(0.03)	(0.031)	(0.034)
CVX &	RDSB &	JNJ &	JNJ &	MRK &	MCD &	MCD &	SBUX &
XOM	XOM	MRK	PFE	PFE	SBUX	YUM	YUM
0.539	0.281	0.397	0.392	0.409	0.313	0.319	0.277
(0.023)	(0.033)	(0.029)	(0.029)	(0.029)	(0.032)	(0.032)	(0.033)

Table 5.9: Comparison between Sample and Estimated Moments

Ticker	Desc. Ret. Mean	Desc. Vol. Mean	Desc. Ret. Std. Dev	Desc. Vol. Std. Dev	Est. Ret. Mean	Est. Vol. Mean	Est. Ret. Std. Dev	Est. Vol. Std. Dev.
AAPL	0.093	0.993	1.690	0.422	0.101	1.079	1.544	0.459
AET	0.104	0.979	1.437	0.449	0.116	1.092	1.409	0.468
ANTM	0.092	1.000	1.475	0.612	0.107	1.163	1.442	0.654
AXP	0.095	0.918	1.178	0.369	0.111	1.069	1.203	0.463
BA	0.085	0.979	1.243	0.535	0.096	1.111	1.239	0.471
BAC	0.157	0.968	1.828	0.421	0.181	1.118	1.780	0.486
C	0.096	0.963	1.726	0.401	0.113	1.131	1.714	0.507
CAT	0.011	0.950	1.363	0.385	0.013	1.089	1.341	0.448
COST	0.084	0.981	0.950	0.471	0.098	1.145	0.957	0.500
CVX	0.020	0.943	1.041	0.313	0.024	1.095	1.049	0.379
DD	0.076	0.947	1.097	0.504	0.091	1.129	1.114	0.552
DIS	0.127	0.959	1.126	0.381	0.141	1.063	1.119	0.406
DOW	0.075	0.985	1.559	0.505	0.083	1.094	1.523	0.498
FOXA	0.121	0.955	1.387	0.499	0.139	1.095	1.395	0.509
GE	0.059	0.935	1.068	0.363	0.069	1.083	1.075	0.434
GOOGL	0.066	0.966	1.383	0.508	0.076	1.121	1.293	0.527
GS	0.107	0.966	1.441	0.394	0.117	1.060	1.410	0.381
HD	0.130	0.990	1.104	0.371	0.144	1.095	1.070	0.413
JNJ	0.074	0.977	0.783	0.449	0.085	1.117	0.789	0.396
JPM	0.095	0.961	1.411	0.423	0.109	1.101	1.345	0.485
KO	0.036	0.988	0.920	0.425	0.042	1.132	0.900	0.450
LOW	0.139	0.973	1.447	0.398	0.151	1.053	1.356	0.367
MCD	0.004	0.979	0.814	0.422	0.004	1.127	0.800	0.482
MMM	0.102	0.961	0.907	0.372	0.117	1.097	0.898	0.382
MRK	0.069	0.972	1.072	0.476	0.084	1.195	1.093	0.660
MSFT	0.088	0.968	1.370	0.480	0.097	1.066	1.266	0.422
PEP	0.059	0.973	0.778	0.377	0.068	1.134	0.789	0.439
PFE	0.062	0.966	0.953	0.578	0.074	1.141	1.002	0.451
PG	0.054	0.998	0.866	0.434	0.060	1.102	0.825	0.430
RDSB	0.008	0.909	1.111	0.561	0.009	1.114	1.183	0.553
SBUX	0.082	0.998	1.415	0.492	0.091	1.103	1.318	0.419
T	0.035	0.978	0.911	0.440	0.039	1.112	0.903	0.429
UNH	0.098	0.975	1.274	0.452	0.108	1.076	1.230	0.408
UTX	0.069	0.977	1.060	0.367	0.078	1.095	1.093	0.415
V	0.129	0.974	1.330	0.527	0.145	1.092	1.259	0.486
VZ	0.038	0.997	0.973	1.390	0.051	1.342	1.138	1.032
WFC	0.102	0.958	1.104	0.324	0.114	1.072	1.092	0.358
WMT	0.058	0.956	0.885	0.404	0.067	1.107	0.850	0.435
XOM	0.022	0.952	0.937	0.332	0.026	1.134	0.961	0.454
YUM	0.036	0.976	1.454	0.639	0.042	1.151	1.318	0.662

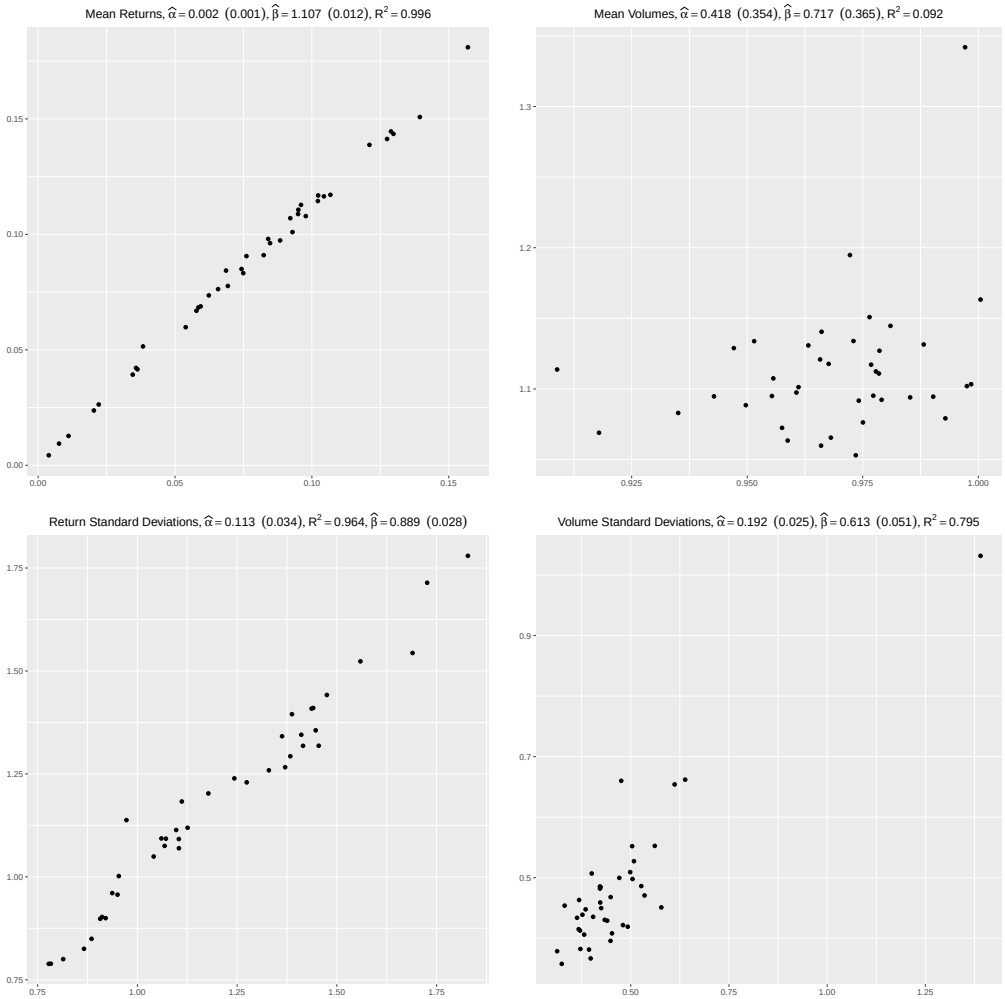


Figure 5.6: Scatter Plots of Sample Moments against Estimated Moments for Returns and Volumes

Table 5.10: Results of Joint Hypothesis Tests for Stock Pairs

Test	AAPL PG	AXP V	BA GE	COST WMT	DIS FOXA	DOW DD
$B_{ij} = 0$	0.490	1.950	25.530	6.789	4.272	0.778
p-value (χ^2_2)	0.783	0.377	0.000	0.034	0.118	0.678
$B_{ij} = 0$ and $V_{ij} = 0$	23.382	139.480	76.881	117.258	119.756	99.209
p-value (χ^2_3)	0.000	0.000	0.000	0.000	0.000	0.000
Test	GOOGL MSFT	GS JPM	HD LOW	KO PEP	T VZ	
$B_{ij} = 0$	10.776	20.823	34.752	36.512	31.583	
p-value (χ^2_2)	0.005	0.000	0.000	0.000	0.000	
$B_{ij} = 0$ and $V_{ij} = 0$	119.136	178.003	177.061	108.184	247.893	
p-value (χ^2_3)	0.000	0.000	0.000	0.000	0.000	

Table 5.11: Results of Joint Hypothesis Tests for Stock Triplets

Test	AET ANTM UNH	BAC C WFC	CAT MMM UTX	CSCO INTC IBM	CVX RDSB XOM	JNJ MRK PFE	MCD SBUX YUM
$B_{ij} = 0$	7.184	57.261	41.691	30.003	91.243	12.755	14.131
p-value (χ^2_6)	0.304	0.000	0.000	0.000	0.000	0.047	0.028
$V_{ij} = 0$	219.984	225.442	172.931	136.687	185.766	154.900	109.657
p-value (χ^2_3)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$B_{ij} = 0$ and $V_{ij} = 0$	227.041	281.779	213.812	166.985	277.262	167.510	123.505
p-value (χ^2_9)	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Chapter 6

Conclusion

In this essay, by extending the mixture distribution hypothesis to a multi-security framework, I have found evidence supporting the existence of spillover effects in securities' information arrival processes. This research has defined spillover effects in two forms based on assumptions on how news may occur and spread: One is the contemporary correlation in the shocks to different securities' information arrival processes, and the other is the cross-security historical dependency of different securities' information arrival processes. The estimation results shown in Chapter 5 gives strong supports for the existence of the first form of spillover effects, as all estimated models show that the contemporary correlations in shocks are statistically significant. In the mean time, the models estimated in Chapter 5 show different degrees of support for spillover effects of the second form. Among the 11 bi-stock and 22 tri-stock models estimated, *BA* and *GE*, *KO* and *PEP*, *BAC*, *C* and *WFC*, and *CVX*, *RDSB* and *XOM* have shown the strongest supports for cross-security historical dependencies. In addition, the cross-security historical dependency is not necessarily symmetric, as most evidently shown by 6 bi-stock models. The information arrival process of the stock with larger market capitalization may have big impacts on that of a smaller-capitalized stock, but the contrary does not always hold.

The biggest challenge for the estimation procedure is the invisibility of securities' information arrival processes, which introduces a high-dimensional and intractable integral to the likelihood function. In order to solve the likelihood maximization problem, I have applied the technique devised by Gu and Kong (1998), which transforms the MLE problem into an equation solving problem involving a conditional expectation, and then uses the stochastic approximation algorithm to approximate the conditional expectation and update parameter estimates. This estimation technique not only enables to search for the MLE point estimate for parameters, but also gives an approximation for the Fisher information matrix as a byproduct, allowing the implementation of inferences.

However, prior to this research, the practical performance of the above estimation technique has not been evaluated in any real estimation problems as complicated as the one studied in this research. Thus before applying the method to real data estimations, I simulated 300 datasets for four combinations of different parameter settings to test the performance of the estimation technique. In these four scenarios, the values that the parameters take cover well the ranges in which real model's parameter values may take, as implied by previous studies.

Then I applied the above estimation technique to these simulated datasets, obtained the parameter estimates, and calculated the their confidence intervals. The simulation results show that the estimation technique works in general. The intervals spanned by simulated information variables can cover the true information variables with probabilities close to one; the fi-

nal parameter estimates it produced are close to the true parameter values; and the confidence intervals it produced for most parameters can cover the true values of the corresponding parameters by probabilities that are close to the confidence levels. However, there are exceptions to the last case. The estimates of the conditional mean parameter for trading volumes are most volatile as compared to the standard errors such that The confidence intervals are often too narrow to cover the true parameter values. I checked the distributions of the estimated parameters, which are close to normal distributions. Therefore, the remaining explanation for the bad performance of the standard error of this parameter is that the estimation algorithm has not been iterated long enough. Due to constraints in computation costs, I have iterated each simulation by 30 parameter update steps.

Based on the simulation results, I elongated the number of parameter update steps to 100 in real data estimation. The first and second-order moments of return and trading volumes, calculated from parameter estimates match up well with the corresponding sample moments, which suggests that the parameter estimates are reliable. The signs and magnitudes of the parameter estimates conform well with implications of the multi-security mixture distribution hypothesis. For example, the off-diagonal entries of $\widehat{\boldsymbol{V}}$ are all positive, as well as most off-diagonal entries of $\widehat{\boldsymbol{B}}$. The negative off-diagonal entries of $\widehat{\boldsymbol{B}}$ are insignificant, except for one case. Then I did hypothesis tests to search for evidence supporting the existence of information arrival spillover effects, up which I drew the conclusions stated in the beginning of this chapter. However, there is something that could potentially be improved in the hypothesis testing procedure. I tested a series of single inequality constraints and multiple equality constraints using Wald tests, but the multiple-security mixture distribution hypothesis actually gives multiple inequality constraints. It would be more appropriate to test these inequality constraints simultaneously, which requires the incorporations of multiple inequality constraints in the stochastic approximation process.

Part II

Information Spillover Effects across Asset Classes and International Markets

Chapter 7

Introduction

In the first part of the dissertation, I extended the mixture distribution hypothesis to a multi-security setting, and studied the information spillover effects among different stocks using the multi-security a framework. The empirical results show that at least some stocks pairs and triplets in the energy and finance industries exhibit strong evidence supporting the existence of spillover effects among multiple stocks (please see 5 for details). While the assumption for conducting such a study is that information associated with one company may have implications on another, it is natural to extend such an assumption to assets other than individual stocks. For example, information associated with the category of large market capitalization stocks may have important implications for stocks in the mid- and small-cap categories. Information associated with equity assets may have important implications for bond assets, and vice versa. And information associated with equity assets in one country may also have important implications for those in other countries.

In the second part of my dissertation, I will study the information spillover effects in the three scenarios listed above. Since the objects investigated in this study cannot be represented by individual assets (for example, when studying the spillover effects among international markets, a single stock cannot represent a countries' stock market well), market indexes that are representative for the studied objects are more suitable for examining the spillover effects among these objects. It then requires to collect the indexes' price and volume trading data sampled at fixed intervals.

While price data for market indexes are readily available, trading volumes are not well defined for market indexes, unless the market indexes are directly traded. The primary data sources I used, i.e. Google Finance (`www.google.com/finance`) and Yahoo Finance (`finance.yahoo.com`), provide data for some market indexes' trading volumes, but such data seem inconsistent with each other. For example, Figure 7.1 plots the time series observations of S&P 500 index's trading volumes provided by the two data sources between Jan. 3, 2012 and Dec. 31, 2014, and Figure 7.2 does a scatter plot between the two series of trading volumes. Though the two series are highly correlated with a correlation coefficient of 0.804, the correlation is far from perfect.

For comparison, the index levels of S&P 500 provided by the two data sources are also plotted, which is shown in Figure 7.3. The differences between the two data sources are very small that the two series are indistinguishable from each other. Thus the index level observations are reliable, but the volume observations seem not.

The more severe problem is the lack of proper definition for market indexes' trading volumes. If we define a market index's trade volume as the average trading volume of its component assets, then assets with more outstanding shares and low per share prices will impact too much the average

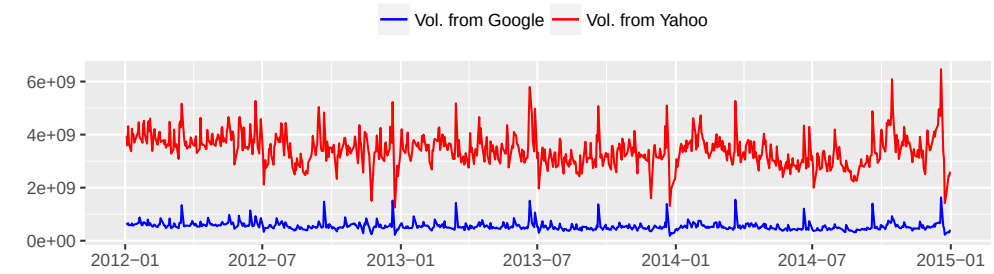


Figure 7.1: Time Series Plots of S&P 500 Index Trading Volumes from Google Finance and Yahoo Finance

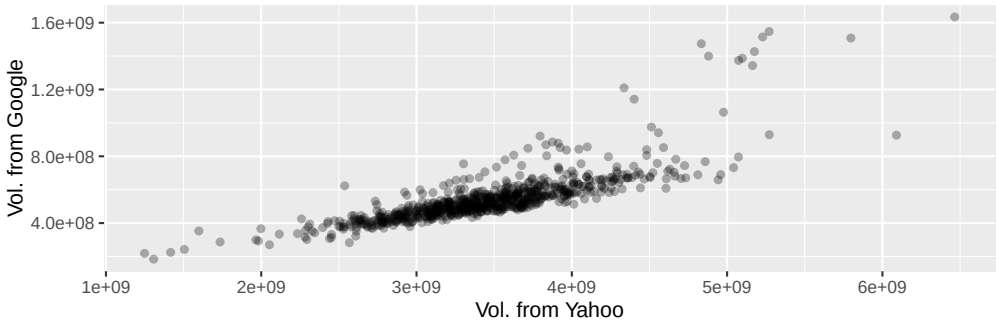


Figure 7.2: Scatter Plot of S&P 500 Index Trading Volumes from Google Finance and Yahoo Finance

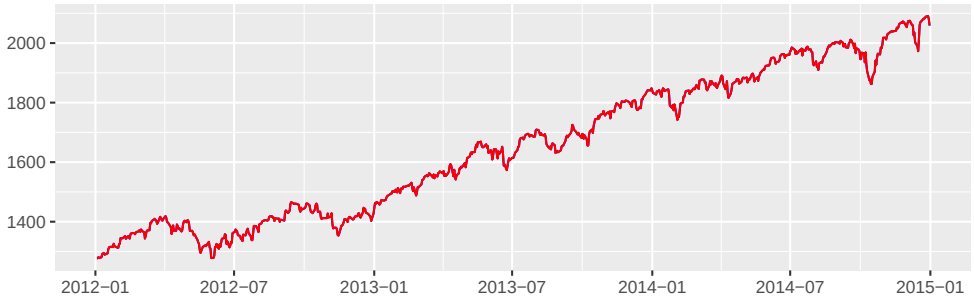


Figure 7.3: S&P 500 Index Values from Google Finance and Yahoo Finance

trading volume. If we use some variable such as price and market capitalization to weight component assets' trading volumes, then the resulting trading volume will contain price and return information, which we would like to keep separate from trading volumes.

The key to solve the problem is to find market indexes that are directly traded. It turned out that ETF (exchange traded funds) may be a good choice for approximating market indexes. An ETF usually tracks a market index, and is traded in the same way as stocks on public exchanges. An arbitrage free mechanism known as *Creation and Redemption* insures

that the discrepancy between the returns of indexes and the corresponding ETF's should be close to zero.[1]

For example, Figure 7.4 has plotted the returns of both the S&P 500 index and SPY (which is an ETF tracking S&P 500), showing that the two return series are very highly correlated with each other (the correlation coefficient is 0.998), such that the two series are barely distinguishable from each other. Figure 7.5 shows the scatter plot of the returns of S&P 500 and SPY, as well as the coefficient estimates for an ordinary least squares regression between the two return series. The plots and OLS coefficient estimates demonstrate that the ETFs are good approximations of the tracked market indexes. Therefore, for studying the spillover effects among market indexes, I decided to use ETF trading data. Since ETFs are traded in the same way as stocks, trading volumes are naturally defined for them.

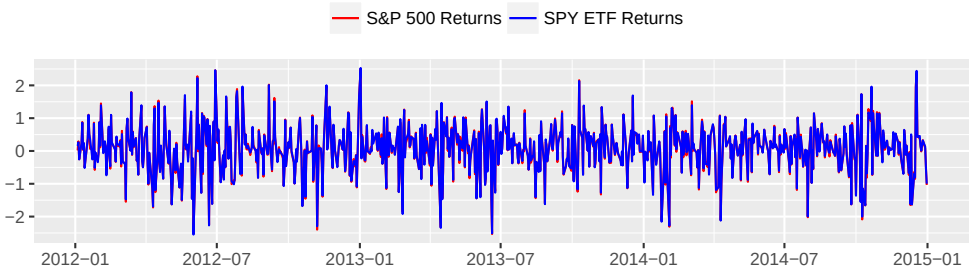


Figure 7.4: The S&P 500 Index Values

Table 7.1 shows the models I estimated using ETF data, corresponding to the three scenarios that I introduced at the beginning of this chapter. Table 7.2 lists some information about the used ETFs.

Table 7.1: ETF Triplets Used in Read Data Estimation

Varying market caps	SPY	IJH	IJR
Varying countries	SPY	EFA	VWO
Varying asset classes	SPY	AGG	GLD

[1]When the price of an ETF is too high to be in line with the tracked market index's value, an authorized participant (usually a market maker) can buy the component assets of the index and deliver them to the ETF issuer, get ETF shares, and then sell them in the market to get risk free profits, and vice versa.

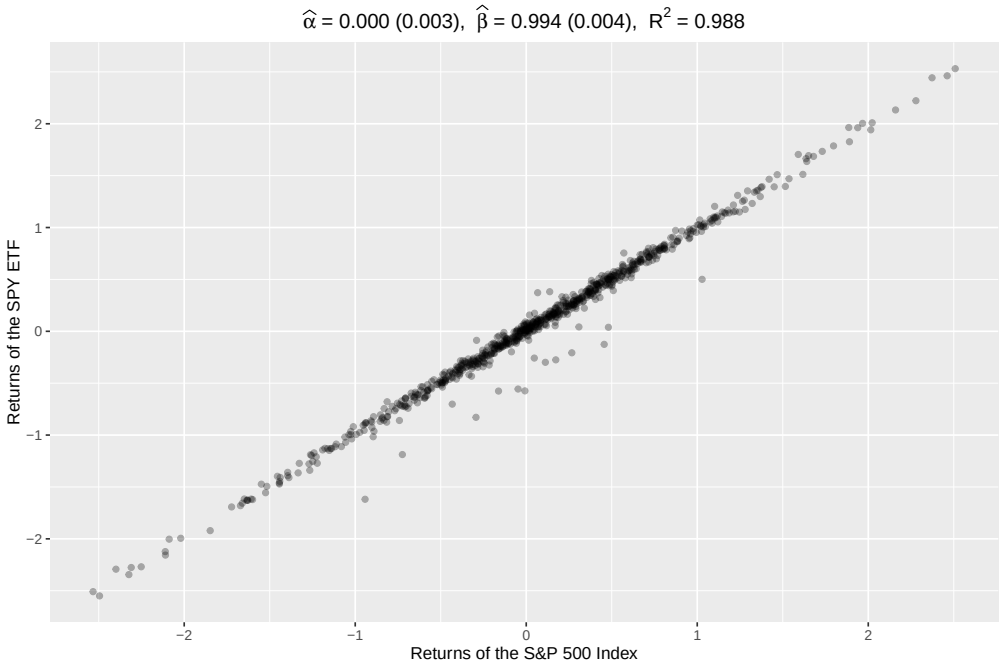


Figure 7.5: The S&P 500 Index Values

Table 7.2: The Names, Issuers and Targeted Indexes of the ETF's used in Real Data Estimation

ETF Name	Issuer	Target Index
SPDR S&P 500 ETF (SPY)	State Street Corporation	S&P 500 Index (large-cap stocks)
iShares Core S&P Mid-Cap ETF (IJH)	BlackRock, Inc.	S&P MidCap 400 (mid-cap stocks)
iShares Core S&P Small-Cap ETF (IJR)	BlackRock, Inc.	S&P SmallCap 600 (small-cap stocks)
iShares MSCI EAFE ETF (EFA)	BlackRock, Inc.	MSCI EAFE Index (large- and mid-cap stocks in non US, non Canadian developed countries)
Vanguard FTSE Emerging Markets ETF (VWO)	The Vanguard Group	FTSE Emerging Markets All Cap China A Transition Index (large-, mid- and small-cap stocks in emerging markets)
iShares Barclays Aggregate Bond ETF (AGG)	BlackRock, Inc.	Barclays US Aggregate Bond Index (US government and investment grade corporate bonds)
SPDR Gold Shares ETF (GLD)	State Street Corporation	The price of one share is based on 1/10 ounce of gold

Chapter 8

Literature Review

In this chapter I will review some previous researches in the area of spillover effects.

Hamao et al. (1990) studied the Tokyo, London and New York stock markets' daily opening and closing levels of market indexes using the ARCH family of models. They found evidence of price volatility spillovers in directions from New York to Tokyo, from London to Tokyo, and and from New York to London.

Theodossiou and Lee (1993) investigated the interdependence among the stocks markets of US, Japan, UK, Canada and Germany with a GARCH-M model. He found out US market's mean spilled over to the UK, Canadian and German markets, and Japan market's mean spilled over to the Germany market. In terms of volatilities, the US market spills over to all other four markets, the UK market spills over to the Canadian market, and the German market spills over to the Japanese market.

Booth et al. (1997) used a multivariate EGARCH model to study the four Scandinavian stock markets, i.e. Danish, Norwegian, Swedish, and Finnish stock markets. They found asymmetric volatility spillovers, in that bad news tends to spread more than good news.

Y. Tse (1999) studied the DJIA spot and future markets minute level prices using Hasbrouck's (1995) cointegration model, and found out the most price-discovery took place in the future market. Furthermore, he found out that volatility spillovers exist in both directions, though the future market impact the spot market more. Asymmetries between good and bad news also exist, with bad news tending to spillover more. Zhong et al. (2004) obtained similar results studying the Mexican stock market.

Kanas (2000) used a bivariate EGARCH model to study to spillover effects between stock and foreign exchange markets in six developed countries: US, UK, Japan, Germany, France, and Canada, and found evidence for spillover effects from stock markets to foreign exchange markets in countries other than Germany. In addition, the spillover effects are symmetric in the sense that the spillover effects by positive and negative shocks are of the same magnitudes.

Hong (2001) developed a class of $\mathcal{N}(0,1)$ tests defined as weighted sums of squared sample cross-correlations between squared standardized residuals for examining volatility spillovers between two time series that are conditionally heteroscedastic. Applying the test to studying two exchange rates, US Dollar-Deutschemark and US Dollar-Japanese Yen, the author finds evidence for simultaneous interactions between the two exchange rates in both means and volatilities. However, volatility spillover in the form of Granger causality only exists uni-directionally from Deutschemark's exchange rate to Japanese Yen's.

Baele (2005) used a regime switching model to study volatility spillovers from the aggregated European market and US market to 13 local European

markets, and found out the spillover intensity increased substantially from 1980s to 1990s, due to increased trade integration, equity market development, and low inflation.

Diebold and Yilmaz (2009) studied 19 global equity markets from the early 1990s, and using the separate measures of return spillover and volatility spillover, they find out the dynamics of return and volatility spillovers are very different: return spillovers have a gently increasing trend without bursts, while volatility spillovers have bursts but no trends.

Wu et al. (2011) studied volatility spillovers from crude oil prices to corn spot and future prices, and found out the time-varying impacts from the energy market to corn market were strengthened after 2005. The impacts are also stronger when the ratio in ethanol and gasoline consumptions exceeds a certain threshold. They also devised a cross-hedging strategy that uses oil market to hedge against risks in the corn markets, but found out the improvement is marginal compared to the traditional hedging strategy using corn markets alone.

All the current researches uses multivariate GARCH or Stochastic Volatility based methods to study the spillover effects. Thus, my current research of using a multi-security MDH framework will contribute to the volatility spillover literature by providing a new perspective.

Chapter 9

ETF Data Estimation

In this chapter, I will show the descriptive statistics of the ETF data I used, as well as showing the estimation results.

9.1 Data Description

Table 9.1: Return Descriptive Statistics of ETF Data Used

	AGG	EFA	GLD	IJH	IJR	SPY	VWO
Mean	0.01	0.04	-0.04	0.07	0.07	0.07	0.02
Std.Dev	0.19	0.93	1.11	0.87	0.95	0.74	1.13
Skewness	-0.46	-0.20	-0.99	-0.41	-0.22	-0.24	0.02
Kurtosis	2.54	1.10	8.42	1.11	0.47	1.19	0.88
Maximum	0.84	3.54	4.26	2.85	2.98	2.53	4.54
Minimum	-1.10	-3.74	-9.19	-3.39	-3.47	-2.55	-4.44
p-Value	0.05	0.04	0.76	0.02	0.15	0.50	0.10

The daily observations on returns and volumes of the 7 ETFs cover the period between Jan 1, 2012 and Dec 31, 2014, with 754 observations in total for each ETF. Returns are calculated as the natural logarithms of two consecutive close prices, which have been adjusted for stock splits and dividends by the data source, Yahoo Finance. The p-Value lines report the p-Values of Ljung-Box tests with 30 lags. 3 of 7 ETF have p-Values smaller than 0.1.

Table 9.1 shows some descriptive statistics of the 7 ETFs' returns. The statistics are qualitatively close to those of individual stocks. That is, the mean returns are close to zero, while the standard deviations of returns are relatively large. All return series have excessive kurtosis, and 3 out of 7 have the p-value for Ljung-Box test smaller than 0.1. The autocorrelation functions are also plotted in Figure 9.1, which shows that absolute and squared returns tend to have higher levels of autocorrelation than return series.

Figure 9.2 shows the histograms of the 7 ETF's return series. We can see that compared to normal distributions, the "bell shapes" of the returns series are thinner and taller, which is similar to the cases of individual stocks.

Table 9.2 shows the descriptive statistics of the ETF's de-trended trading volumes. The de-trending method is first calculating two-year moving averages as trend values of trading volumes, then dividing the raw trading volumes to their trend values. The detrending process causes the mean trading volumes to be close to 1, as shown in Table 9.2. The standard deviations of trading volumes are much smaller compared to return series. Since trading volumes are positive, it should be non-surprising to see their skewness are all positive. The p-value of the Ljung-Box tests for zero autocorrelation for lags from 1 to 30 are all zero, contradicting to the case of

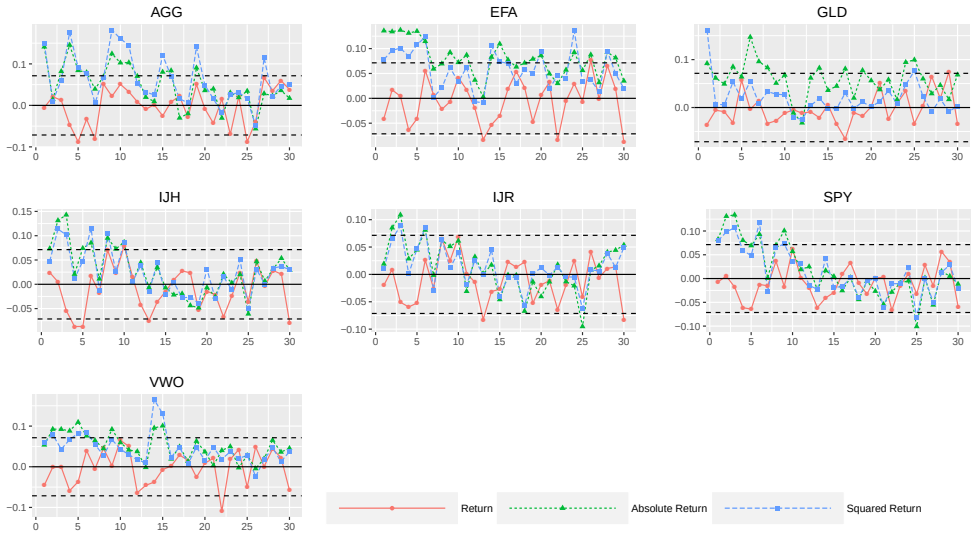


Figure 9.1: The Autocorrelation Functions of ETFs' Returns, Absolute Returns and Squared Returns

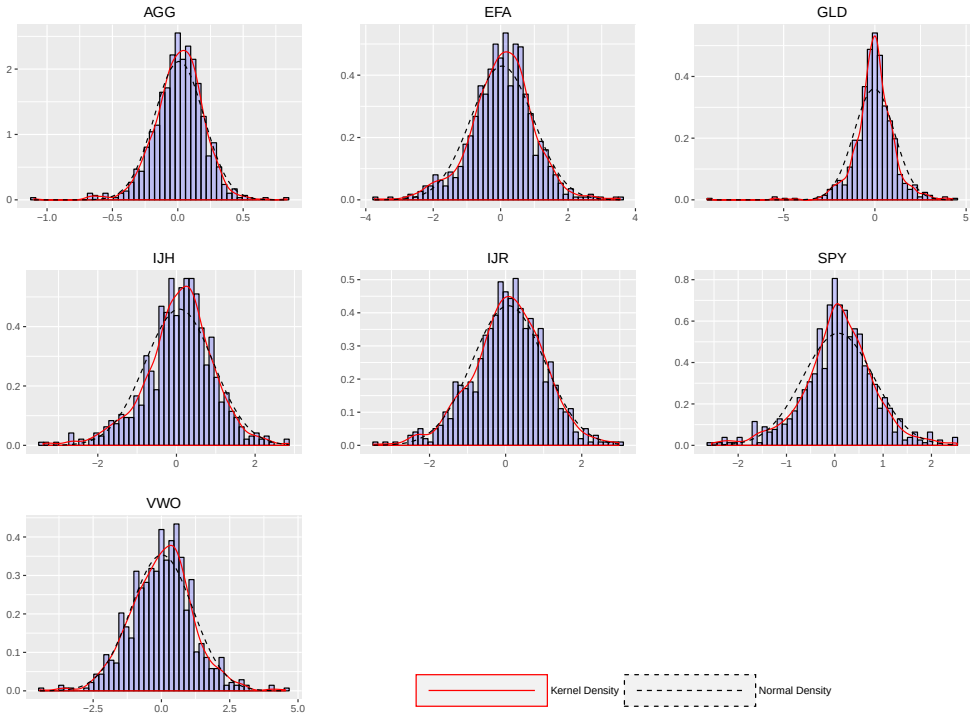


Figure 9.2: The Histograms of ETFs' Returns

returns. The last line gives the correlation between trading volumes and squared returns, supporting the mixture distribution hypothesis.

Table 9.2: Volume Descriptive Statistics of ETF Data Used

	AGG	EFA	GLD	IJH	IJR	SPY	VWO
Mean	0.96	0.96	0.96	1.00	0.96	0.94	0.98
Std.Dev	0.54	0.33	0.60	1.99	0.50	0.34	0.43
Skewness	2.02	1.44	5.63	7.46	1.69	1.67	1.97
Kurtosis	6.12	3.53	62.19	63.93	4.35	4.88	7.38
Maximum	4.40	2.72	9.47	25.17	4.00	3.29	4.24
Minimum	0.21	0.35	0.23	0.19	0.26	0.37	0.19
p-Value	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cor.	0.10	0.38	0.74	0.21	0.16	0.51	0.41

The first six lines of the table has self-explanatory line names. Lines with name p-Value give the p-Values of Ljung-Box tests for no autocorrelation. The lines with name Cor. give the correlation coefficients between squared returns and de-trended volumes. The raw volume data is downloaded from Yahoo Finance, which is then fed to a de-trending procedure. Specifically, the de-trending procedure first finds two-year centered moving averages as trend values, then de-trend the raw volumes by taking the ratios between raw volumes and corresponding trend values.

Figure 9.3 plots the autocorrelation functions of detrended trading volumes, showing that detrended trading volumes show high levels of autocorrelation. Figure 9.4 plots the histograms of logarithmic detrended trading volumes. The distribution of detrended trading volumes are actually close to logarithmic normals, except for the case of IJH (Blackrock's ETF tracking the mid-cap stock index S&P 400).

9.2 Estimation Results

Tables 9.3 and 9.4 show present the estimation results from applying the multi-security mixture distribution hypothesis to ETF data, with 9.3 focusing on the conditional distribution parameters and 9.4 on the information process parameters. Using the parameter estimates in these two tables, I calculated the unconditional means and variances of returns and detrended trading volumes, which are shown in Table 9.5. Table 9.5 does not contain *AGG* and *GLD*, because the model containing these two and *SPY* has complex eigenvalues for $\widehat{\boldsymbol{B}}$. And my method of computing unconditional moments cannot be applied.

Since we are most interested in identifying the spillover effects among information arrival processes, let us move on to examining the estimates for

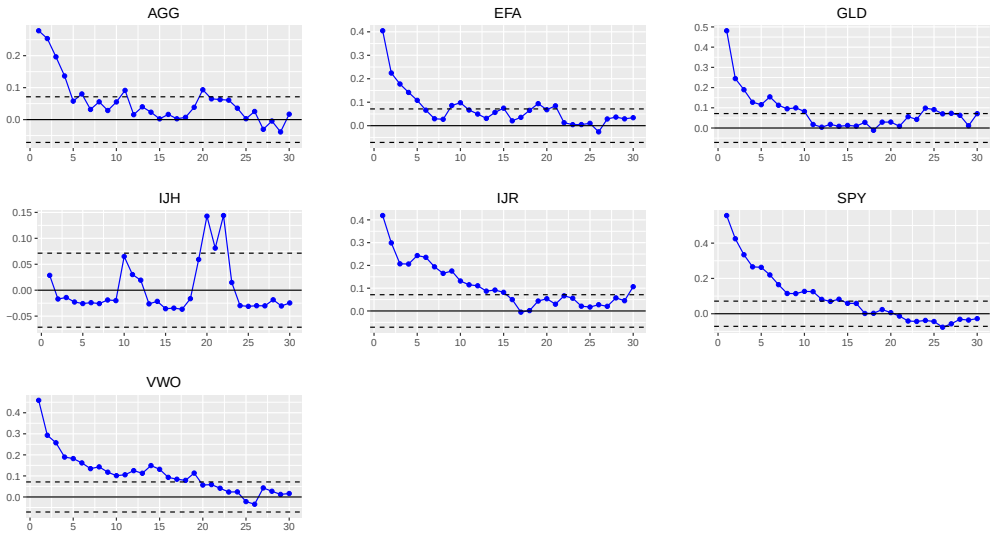


Figure 9.3: The Autocorrelation Functions of ETFs' Detrended Trading Volumes

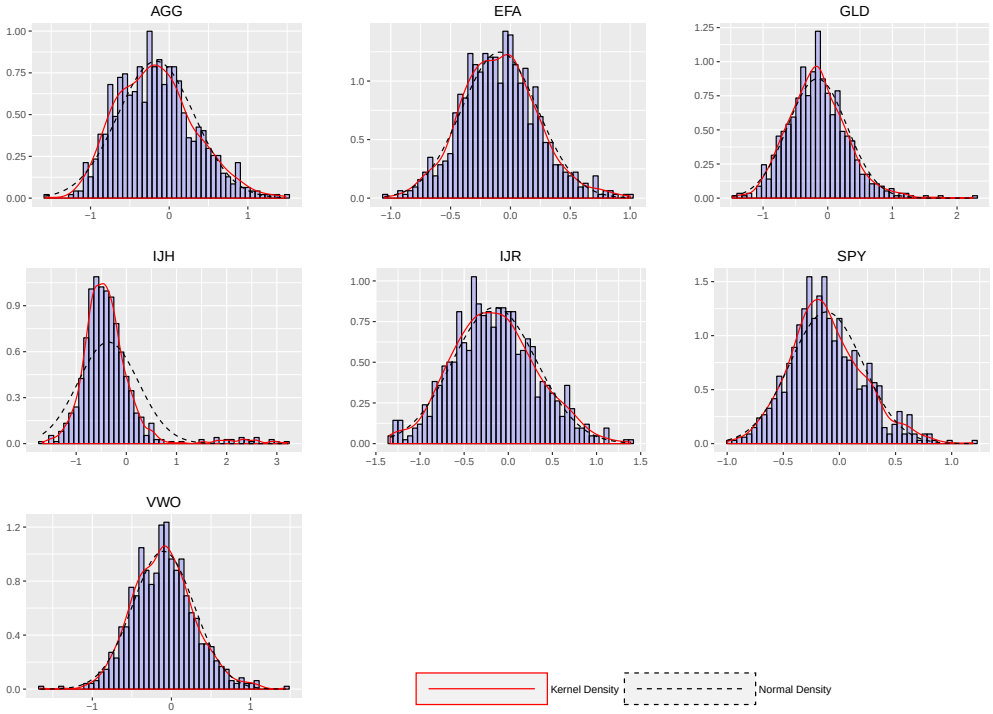


Figure 9.4: The Histograms of ETFs' Logarithmic Detrended Trading Volumes

entries in $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{V}}$. We expect the off-diagonal entries of both $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{V}}$ to be non-negative, which is largely confirmed by Table 9.4. There are three negative off-diagonal entries of $\widehat{\boldsymbol{B}}$, i.e. $\widehat{B}_{12}$ and $\widehat{B}_{32}$ in Column 2, and $\widehat{B}_{21}$ in Column 4, with only the last one being statistically significant, though its magnitude is still small (-0.074). The off-diagonal entries of $\widehat{V}$ are all positive.

Table 9.3: Estimation Results Using ETF Data

Asset 1	AGG	EFA	IJH
Asset 2	GLD	SPY	IJR
Asset 3	SPY	VWO	SPY
μ_{11}	0.012	0.044	0.071
	(0.012)	(0.035)	(0.035)
μ_{12}	1.188	1.054	0.975
	(0.009)	(0.007)	(0.007)
μ_{21}	-0.042	0.084	0.081
	(0.037)	(0.028)	(0.038)
μ_{22}	1.032	1.074	1.056
	(0.007)	(0.007)	(0.008)
μ_{31}	0.087	0.018	0.087
	(0.029)	(0.043)	(0.029)
μ_{32}	1.111	1.062	1.113
	(0.007)	(0.007)	(0.007)
$\Sigma_{1,11}$	0.073	0.85	0.967
	(0.003)	(0.044)	(0.05)
$\Sigma_{1,12}$	0	-0.021	0.018
	(0.002)	(0.005)	(0.007)
$\Sigma_{1,22}$	0.051	0.028	0.039
	(0.002)	(0.001)	(0.002)
$\Sigma_{2,11}$	0.953	0.53	0.996
	(0.049)	(0.027)	(0.051)
$\Sigma_{2,12}$	-0.03	-0.024	-0.014
	(0.006)	(0.004)	(0.007)
$\Sigma_{2,22}$	0.036	0.026	0.044
	(0.002)	(0.001)	(0.002)
$\Sigma_{3,11}$	0.55	1.278	0.552
	(0.028)	(0.066)	(0.028)
$\Sigma_{3,12}$	-0.028	-0.005	-0.028
	(0.004)	(0.007)	(0.004)
$\Sigma_{3,22}$	0.027	0.034	0.028
	(0.001)	(0.002)	(0.001)

The significantly positive off-diagonal entries of $\widehat{\boldsymbol{B}}$ include $\widehat{B}_{13}$, $\widehat{B}_{21}$, $\widehat{B}_{31}$ in Column 1, $\widehat{B}_{12}$, $\widehat{B}_{13}$, $\widehat{B}_{32}$ in Column 2, and $\widehat{B}_{13}$ and $\widehat{B}_{23}$ in Column 3. These significant estimates indicate that spillover effects in the form of cross-security historical dependency does exist among ETFs. Furthermore, similar to the case of stocks in which we observed asymmetric spillover effects, the spillover effects among ETFs are also asymmetric. Take the model of *IJH* (mid-cap stocks), *IJR* (small-cap stocks) and *SPY* (large-cap stocks) for an example, though *SPY* has significant spillover effects

Table 9.4: Estimation Results Using ETF Data

Asset 1	AGG	EFA	IJH
Asset 2	GLD	SPY	IJR
Asset 3	SPY	VWO	SPY
B_{11}	0.567	0.278	0.27
	(0.032)	(0.045)	(0.039)
B_{12}	-0.065	0.268	0.106
	(0.039)	(0.041)	(0.058)
B_{13}	0.318	0.097	0.516
	(0.052)	(0.034)	(0.08)
B_{21}	0.096	0.02	-0.074
	(0.03)	(0.044)	(0.022)
B_{22}	0.486	0.683	0.557
	(0.037)	(0.041)	(0.033)
B_{23}	0.033	0.057	0.325
	(0.049)	(0.033)	(0.046)
B_{31}	0.088	0.067	0.007
	(0.019)	(0.052)	(0.016)
B_{32}	-0.02	0.17	0.035
	(0.024)	(0.048)	(0.024)
B_{33}	0.712	0.499	0.747
	(0.032)	(0.039)	(0.032)
V_{11}	0.17	0.065	0.37
	(0.009)	(0.003)	(0.019)
V_{12}	0.045	0.034	0.066
	(0.006)	(0.003)	(0.008)
V_{13}	0.021	0.034	0.053
	(0.004)	(0.003)	(0.005)
V_{22}	0.153	0.061	0.121
	(0.008)	(0.003)	(0.006)
V_{23}	0.048	0.036	0.027
	(0.004)	(0.003)	(0.003)
V_{33}	0.059	0.088	0.056
	(0.003)	(0.004)	(0.003)

Table 9.5: The Moments of ETFs' Returns and Detrended Traindg Volumes Calculated from Parameter Estimates

	EFA	VWO	IJH	IJR	SPY
$\mathrm{E}(r_t)$	0.05	0.02	0.09	0.09	0.10
$\mathrm{E}(V_t)$	1.10	1.15	1.20	1.15	1.26
$\mathrm{Var}(r_t)$	0.88	1.39	1.19	1.09	0.63
$\mathrm{Cov}(r_t, V_t)$	-0.02	-0.00	0.07	0.00	0.00
$\mathrm{Var}(V_t)$	0.13	0.27	0.77	0.29	0.48

on *IJH* ($\widehat{B}_{13} = 0.516$), the effect does not exist in the reverse direction ($\widehat{B}_{31} = 0.007$). This is aligned with my finding in Chapter 5 that assets with large market capitalizations tend to impact more those with smaller ones.

The off-diagonal entries of $\widehat{\boldsymbol{V}}$ strongly support the existence of spillover effects in the form of contemporary correlation in shocks. Estimates for all such parameters are significantly positive. In Table 9.6 I have calculated the contemporary correlations, which are of both statistical and practical significance.

Therefore, we can conclude that the existence of spillover effects among ETFs are supported by empirical evidence. Spillover effects in the form of contemporary correlations in shocks widely exist among all models I estimated, while cross-security historical dependency only exists between certain ETFs, and such dependency is asymmetric that assets with large market capitalizations tend to give more impacts on smaller ones.

Table 9.6: Contemporaneous Correlations between Shocks to Information Processes

AGG & GLD	AGG & SPY	GLD & SPY	EFA & SPY	EFA & VWO
0.278	0.214	0.507	0.532	0.454
(0.034)	(0.034)	(0.026)	(0.024)	(0.028)
SPY & VWO	IJH & IJR	IJH & SPY	IJR & SPY	
0.487	0.311	0.367	0.332	
(0.026)	(0.033)	(0.03)	(0.031)	

Chapter 10

Conclusion

In the second part of my dissertation, I applied the Multi-security Mixture Distribution framework developed in the first part to studying the spillover effects in information arrivals among different market indexes. I took three perspectives, i.e. spillovers among large-cap, mid-cap and small-caps stocks, spillovers among US, other developed countries' and developing countries' markets, and spillovers among stock, bond and gold markets. Due to the difficulty in defining trading volumes for market indexes, I used ETFs to proxy the corresponding market indexes.

The empirical results support the existence of spillover effects among studied markets in both forms. For spillovers in the form of cross-security historical dependency in information arrivals, large-cap stocks exhibit strong spillovers to mid-cap and small-cap stocks, but is not affected much by the two latter markets. The US market exhibits strong spillovers to other developed and developing countries' markets, and again the latter two markets' spillover effects are not as strong. And lastly, the US stock market spills over heavily to the bond market, the bond market spills over slightly to both the stock and gold market, and the gold market does not exhibit spillover effects to either of the other two markets.

For spillover effects in the form of contemporaneous correlations in shocks to information arrival processes, all studied models show strong evidence supporting this type of spillovers.

Bibliography

Alexander Philipov, Mark E. Glickman (2006). "Multivariate Stochastic Volatility via Wishart Processes". In: *Journal of Business & Economic Statistics* 24.3, pp. 313–328.

Andersen, Torben G. (1996). "Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility". In: *The Journal of Finance* 51.1, pp. 169–204.

Andrieu, Christophe et al. (2003). "An introduction to MCMC for machine learning". In: *Machine learning* 50.1-2, pp. 5–43.

Asai, Manabu and Michael McAleer (2006). "Asymmetric Multivariate Stochastic Volatility". In: *Econometric Reviews* 25.2-3, pp. 453–473.

Asai, Manabu, Michael McAleer, and Jun Yu (2006). "Multivariate Stochastic Volatility: A Review". In: *Econometric Reviews* 25.2-3, pp. 145–175.

Baele, Lieven (2005). "Volatility Spillover Effects in European Equity Markets". In: *Journal of Financial and Quantitative Analysis* 40 (02), pp. 373–401.

Bauwens, Luc, Sébastien Laurent, and Jeroen V.K. Rombouts (2006). "Multivariate GARCH models: a survey". In: *Journal of Applied Econometrics* 21.1, pp. 79–109.

Blum, Julius R. (1954). "Multidimensional Stochastic Approximation Methods". In: *The Annals of Mathematical Statistics* 25.4, pp. 737–744.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of Econometrics* 31.3, pp. 307–327.

— (1990). "Modelling the Coherence in Short-run Nominal Exchange Rates: A Multivariate Generalized ARCH Model". In: *The Review of Economics and Statistics* 72.3, pp. 498–505.

Bollerslev, Tim, Robert F Engle, and Jeffrey M Wooldridge (1988). "A Capital Asset Pricing Model with Time-Varying Covariances". In: *Journal of Political Economy* 96.1, pp. 116–31.

Booth, G.Geoffrey, Teppo Martikainen, and Yiuman Tse (1997). "Price and volatility spillovers in Scandinavian stock markets". In: *Journal of Banking & Finance* 21.6, pp. 811–823.

Brooks, S. et al. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Celeux, Gilles and Jean Diebolt (1985). "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem". In: *Computational Statistics Quarterly* 2.1, pp. 73–82.

Chan, David, Robert Kohn, and Chris Kirby (2006). "Multivariate Stochastic Volatility Models with Correlated Errors". In: *Econometric Reviews* 25.2-3, pp. 245–274.

Chan, Kung Sik and Charles J Geyer (1994). "Discussion: Markov chains for exploring posterior distributions". In: *The Annals of Statistics* 22.4, pp. 1747–1758.

Chib, Siddhartha, Yasuhiro Omori, and Manabu Asai (2009). "Multivariate Stochastic Volatility". In: *Handbook of Financial Time Series*. Ed. by Thomas Mikosch et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 365–400.

Clark, Peter K. (1973). "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices". In: *Econometrica* 41.1, pp. 135–155.

Danìelsson, Jòn (1998). "Multivariate stochastic volatility models: Estimation and a comparison with VGARCH models". In: *Journal of Empirical Finance* 5.2, pp. 155–173.

Delyon, Bernard, Marc Lavielle, and Eric Moulines (1999). "Convergence of a stochastic approximation version of the EM algorithm". In: *Ann. Statist.* 27.1, pp. 94–128.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38.

Diebold, Francis X. and Kamil Yilmaz (2009). "Measuring Financial Asset Return and Volatility Spillovers, with Application to Global Equity Markets". In: *The Economic Journal* 119.534, pp. 158–171.

Engle, Robert F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". In: *Econometrica* 50.4, pp. 987–1007.

Engle, Robert F. and Kenneth F. Kroner (1995). "Multivariate Simultaneous Generalized Arch". In: *Econometric Theory* 11.1, pp. 122–150.

Engle, Robert, Victor K. Ng, and Michael Rothschild (1990). "Asset pricing with a factor-arch covariance structure: Empirical estimates for treasury bills". In: *Journal of Econometrics* 45.1-2, pp. 213–237.

Epps, Thomas W. and Mary Lee Epps (1976). "The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distributions Hypothesis". In: *Econometrica* 44.2, pp. 305–321.

Geman, Stuart and Donald Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 6, pp. 721–741.

Geyer, Charles (2011). "Introduction to Markov Chain Monte Carlo". In: *Handbook of Markov Chain Monte Carlo*, pp. 3–48.

Glosten, Lawrence R. and Paul R. Milgrom (1985). "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders". In: *Journal of Financial Economics* 14.1, pp. 71–100.

Gu, Ming Gao and Fan Hui Kong (1998). "A Stochastic Approximation Algorithm with Markov Chain Monte-Carlo Method for Incomplete Data Estimation Problems". In: *Proceedings of the National Academy of Sciences of the United States of America* 95.13, pp. 7270–7274.

Gu, Ming Gao and Hong-Tu Zhu (2001). "Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 339–355.

Hamao, Yasushi, Ronald W. Masulis, and Victor Ng (1990). "Correlations in Price Changes and Volatility across International Stock Markets". In: *Review of Financial Studies* 3.2, pp. 281–307.

Han, Yufeng (2006). "Asset Allocation with a High Dimensional Latent Factor Stochastic Volatility Model". In: *Review of Financial Studies* 19.1, pp. 237–271.

Harris, Lawrence (1986). "Cross-Security Tests of the Mixture of Distributions Hypothesis". In: *The Journal of Financial and Quantitative Analysis* 21.1, pp. 39–46.

— (1987). "Transaction Data Tests of the Mixture of Distributions Hypothesis". In: *The Journal of Financial and Quantitative Analysis* 22.2, pp. 127–141.

Harvey, Andrew, Esther Ruiz, and Neil Shephard (1994). "Multivariate Stochastic Variance Models". In: *The Review of Economic Studies* 61.2, pp. 247–264.

Hasbrouck, Joel (1995). "One Security, Many Markets: Determining the Contributions to Price Discovery". In: *The Journal of Finance* 50.4, pp. 1175–1199.

Hastings, W Keith (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1, pp. 97–109.

Hong, Yongmiao (2001). "A test for volatility spillover with application to exchange rates". In: *Journal of Econometrics* 103, pp. 183–224.

Jeantheau, Thierry (1998). "Strong Consistency of Estimators for Multivariate ARCH Models". In: *Econometric Theory* 14.01, pp. 70–86.

K. Pitt, Michael and Neil Shephard (1999). "Filtering via Simulation: Auxiliary Particle Filters". In: *Journal of the American Statistical Association* 94.446, pp. 590–599.

Kanas, Angelos (2000). "Volatility Spillovers Between Stock Returns and Exchange Rate Changes: International Evidence". In: *Journal of Business Finance & Accounting* 27.3-4, pp. 447–467.

Kawakatsu, Hiroyuki (2006). "Matrix exponential GARCH". In: *Journal of Econometrics* 134.1, pp. 95–128.

Kesten, Harry (1958). "Accelerated Stochastic Approximation". In: *Ann. Math. Statist.* 29.1, pp. 41–59.

Kiefer, J. and J. Wolfowitz (1952). "Stochastic Estimation of the Maximum of a Regression Function". In: *Ann. Math. Statist.* 23.3, pp. 462–466.

Lai, Tze Leung (2003). "Stochastic approximation: invited paper". In: *Ann. Statist.* 31.2, pp. 391–406.

Lange, Kenneth (1995). "A Gradient Algorithm Locally Equivalent to the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.2, pp. 425–437.

Liesenfeld, Roman (1998). "Dynamic Bivariate Mixture Models: Modeling the Behavior of Prices and Trading Volume". In: *Journal of Business & Economic Statistics* 16.1, pp. 101–109.

— (2001). "A generalized bivariate mixture model for stock price volatility and trading volume". In: *Journal of Econometrics* 104.1, pp. 141–178.

Liu, Chuanhai and Donald B. Rubin (1994). "The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence". In: *Biometrika* 81.4, pp. 633–648.

Louis, Thomas A. (1982). "Finding the Observed Information Matrix when Using the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2, pp. 226–233.

Mandelbrot, Benoit (1963). "The Variation of Certain Speculative Prices". In: *The Journal of Business* 36, p. 394.

Meng, Xiao-Li and Donald B. Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework". In: *Biometrika* 80.2, pp. 267–278.

Metropolis, Nicholas et al. (1953). "Equation of state calculations by fast computing machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.

Park, Beum-Jo (2010). "Surprising information, the MDH, and the relationship between volatility and trading volume". In: *Journal of Financial Markets* 13.3, pp. 344–366.

— (2011). "Forecasting Volatility in Financial Markets Using a Bivariate Stochastic Volatility Model with Surprising Information". In: *Journal for Economic Forecasting* 3, pp. 37–58.

Pelletier, Denis (2006). "Regime switching for dynamic correlations". In: *Journal of Econometrics* 131.1-2, pp. 445–473.

Peskun, Peter H (1973). "Optimum monte-carlo sampling using markov chains". In: *Biometrika* 60.3, pp. 607–612.

Powell, Warren B. (2007). "Stochastic Approximation Methods". In: *Approximate Dynamic Programming*. John Wiley & Sons, Inc. Chap. 6, pp. 179–224.

Robbins, Herbert and Sutton Monro (1951). "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407.

Roberts, Gareth O and Adrian FM Smith (1994). "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms". In: *Stochastic processes and their applications* 49.2, pp. 207–216.

Shephard, Neil and Torben G. Andersen (2008). *Stochastic Volatility: Origins and Overview*. Economics Series Working Papers 389. University of Oxford, Department of Economics.

Silvennoinen, Annastiina and Timo Teräsvirta (2005). *Multivariate Autoregressive Conditional Heteroskedasticity with Smooth Transitions in Conditional Correlations*. SSE/EFI Working Paper Series in Economics and Finance 577. Stockholm School of Economics.

— (2009). "Multivariate GARCH Models". In: *Handbook of Financial Time Series*. Ed. by Thomas Mikosch et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 201–229.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. London ; New York : Chapman and Hall, 1986.

Simon, Carl P. and Lawrence Blume (1994). *Mathematics for economists*. New York : Norton.

Spall, J. C. (1987). "A Stochastic Approximation Technique for Generating Maximum Likelihood Parameter Estimates". In: *American Control Conference, 1987*, pp. 1161–1167.

— (1992). "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation". In: *IEEE Transactions on Automatic Control* 37.3, pp. 332–341.

— (2000). "Adaptive stochastic approximation by the simultaneous perturbation method". In: *IEEE Transactions on Automatic Control* 45.10, pp. 1839–1853.

Tanner, Martin A (1996). *Tools for statistical inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 3rd ed. Springer Series in Statistics. Springer-Verlag New York.

Tauchen, George E. and Mark Pitts (1983). "The Price Variability-Volume Relationship on Speculative Markets". In: *Econometrica* 51.2, pp. 485–505.

Taylor, Stephen J. (1982). "Financial returns modelled by the product of two stochastic processes : a study of daily sugar prices, 1961-79". In: *Time Series Analysis : Theory and Practice*. Ed. by O. D. Anderson. Vol. 1. Amsterdam: North-Holland, pp. 203–226.

Theodossiou, Panayiotis and Unro Lee (1993). "Mean and Volatility Spillovers Across Major National Stock Markets: Further Empirical Evidence". In: *Journal of Financial Research* 16.4, pp. 337–350.

Tierney, Luke (1994). "Markov chains for exploring posterior distributions". In: *the Annals of Statistics*, pp. 1701–1728.

Tsay, Ruey S. (2002). *Analysis of financial time series. [electronic resource].* Wiley series in probability and statistics. New York : Wiley, c2002.

Tsay, Ruey S (2005). *Analysis of financial time series*. Vol. 543. John Wiley & Sons.

Tse, Y. K and Albert K. C Tsui (2002). "A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model With Time-Varying Correlations". In: *Journal of Business & Economic Statistics* 20.3, pp. 351–362.

Tse, Yiuman (1999). "Price discovery and volatility spillovers in the DJIA index and futures markets". In: *Journal of Futures Markets* 19.8, pp. 911–930.

Wei, Greg C. G. and Martin A. Tanner (1990). "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms". In: *Journal of the American Statistical Association* 85.411, pp. 699–704.

Weide, Roy van der (2002). "GO-GARCH: a multivariate generalized orthogonal GARCH model". In: *Journal of Applied Econometrics* 17.5, pp. 549–564.

Wu, Feng, Zhengfei Guan, and Robert J. Myers (2011). "Volatility spillover effects and cross hedging in corn and crude oil futures". In: *Journal of Futures Markets* 31.11, pp. 1052–1075.

Ying, Charles C. (1966). "Stock Market Prices and Volumes of Sales". In: *Econometrica* 34.3, pp. 676–685.

Yu, Jun and Renate Meyer (2006). "Multivariate Stochastic Volatility Models: Bayesian Estimation and Model Comparison". In: *Econometric Reviews* 25.2-3, pp. 361–384.

Zhong, Maosen, Ali F. Darrat, and Rafael Otero (2004). "Price discovery and volatility spillovers in index futures markets: Some evidence from Mexico". In: *Journal of Banking & Finance* 28.12, pp. 3037–3054.