Faculty of Engineering and Information Technology
University of Technology, Sydney

Volatility Modeling and Analysis via Coupled Wishart Process

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in Computing Sciences

by

Zhong She

November 2013

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

Production Note:
Signature removed prior to publication.

Acknowledgments

On having completed this thesis, I would like to give my greatest gratitude to my supervisors, Prof. Longbing Cao and Dr. Richard Xu. Only with Prof. Cao's generous financial support and guidance, I can make some progresses on my research topic and finish this thesis. Besides, his vision, working enthusiasm, diligence and ambition has made him a role model in the future of my way. My co-supervisor, Dr. Richard Xu offered me great help in my topic. He has a great interest in statistical learning with expert proficiency. I have benefited a lot from discussions with him.

I want to express my sincere thanks to my girlfriend Can Wang. With her reference, I get this opportunity to study abroad and come into a completely new research world. Also I want to thank all my colleagues in AAI, who has been offering me friendship and assistance during my period here. I can not forget the joys of our time together ever after.

My parents deserve my deepest thanks. During the past two years, they missed me so much and had to suffer from mine absence with them. Though more than 8,000 kilometers away, they cared about everything of me.

Lastly, I would like to thank myself. Honestly, these two years are one of the toughest periods of my life. I am glad I have survived with hope and determination. Hope luck and courage along with me on my future way.

Zhong She
August 2013 @ UTS

Contents

List of Figures

List of Tables

Abstract

Volatility refers to the measure for price fluctuation of specific financial instrument over time. It is a very important factor that can greatly influence investor's decisions and concerns every other participant in the stock market. High volatility implies great insatiability and will definitely increase liquidity whereas low volatility indicates poor activeness. Hence the research on volatility draws great attention and interest of researchers from different backgrounds. Including the methods from data mining and machine learning is essential to improve the quality of volatility analysis.

There are two main types of models on volatility analysis: the deterministic models and stochastic models. The deterministic models assume the volatility at particular time is a deterministic function of the past. The generalized autoregressive conditional heteroskedasticity (GARCH) model and its variations are in such category. The stochastic volatility (SV) models take the assumption that the volatility follows certain random process. Recent literature has shows that the stochastic models outperform the deterministic models to some extent. Among them, the Wishart process is a hot tool for modeling multivariate volatility.

However, the stock market is closely connected with the society and human behavior, which makes it difficult to model. Almost all the existing models assume independence between our target objects: prices or the hidden covariance matrices behind them. These assumption works well for rough research or when the relationship between objects is weak. For a more solid research, the coupling relationship must be taken into account.

In this thesis, we present two kinds of coupled Wishart process to model volatility: the homogenous coupled Wishart process and heterogenous coupled Wishart process. And corresponding algorithms are developed based on the models. The homogenous coupled Wishart process refers to model that our target objects belong to the same category. A two-chain coupled Wishart process is introduced in this thesis. Within such a model, the matrix in one chain is not only related with the past one from its own chain but also from its neighbors. After the derivation of its learning procedures, synthetic data are tested. Then, experiments are implemented with real data from two markets: U.S. and Hong Kong. In the two-chain coupled Wishart process, one chain indicates the volatility from U.S. stock market and the other the volatility indicates Hong Kong stock market.

The latter one is the heterogenous coupled Wishart process. Unlike the homogenous one, in such a model, the covariance matrices are coupled with vectors, scalars or even a system. We aim to model how the outside influence from other kinds of data affect the evolving of covariance matrices. For time limitation, we make a simplified setup to illustrate how the heterogeneous coupling works. Then we construct the learning algorithm based on the setups and test it on synthetic data.

To conclude, we include the thought of coupling into the analysis of volatility via Wishart process, with machine learning techniques. Sufficient experiments have proved the effectiveness of coupling in volatility analysis.

Chapter 1

Introduction

1.1 Preliminaries

In finance, volatility is the measure for price fluctuation of specific financial instrument over time (Roll 1984). It is a very important factor that can greatly influence investor's decisions: investors react emotionally to swings in market prices; when certain cash flows from selling a security are needed at a specific future date, higher volatility means a greater chance of a shortfall; higher volatility of returns while saving for retirement results in a wider distribution of possible final portfolio values; higher volatility of return when retired gives withdrawals a larger permanent impact on the portfolio's value; price volatility presents opportunities to buy assets cheaply and sell when overpriced.

As volatility is such an important indicator that the research on volatility is an issue that concerns every participant in the markets, not only the investors: government policy makers, market analysts and economists. Besides, given the rapid growth of financial markets and the continual development of new and increasingly complex financial instruments, the need for high quality volatility modeling methods is in great demand. Modeling and predicting volatility are of vital importance in various applications, such as asset pricing, portfolio selection, option pricing, hedging and risk management.

Besides, the burst of several bubbles aggravates the uncertainty of world market and economy. Especially after the economic crisis in 2008, modeling volatility is becoming a more and more critical task. Practitioners and academics are reassessing the adequacy of effective models to capture and predict financial volatility. As indicated in Figure. 1.1, every strike in volatility indicates a violate change in return prices: the great depression, the 1945 recession, and the 2008 financial crisis etc.

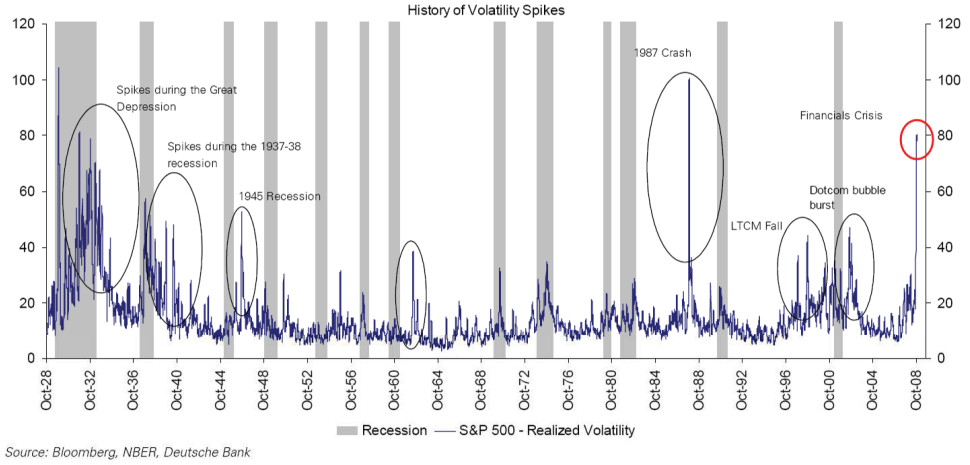

Figure 1.1: An example volatility of S&P 500 over decades.

Wishart process (Bru 1991, Gourieroux, Jasiak & Sufana 2009, Jacquier, Polson & Rossi 2002, Philipov & Glickman 2006) are a time series model that focuses on positive-definite matrix. It follows a Markov rule, at each time point t, the matrix variable Σ_t is only related to its previous one Σ_{t-1}. The transaction of covariance matrix follows a Wishart distribution. Due to its properties, Wishart process is adjusted and servers well as multivariate volatility model.

1.2 Background

Due to the importance of the study on volatility, there has been numerous research and various approaches on it. Among them, the two main approach-

es are deterministic models and stochastic models. For deterministic models, the GARCH model family are the most popular ones. The GARCH family models were largely extended after Engle's innovative work (Engle 1982), autoregressive conditional heteroscedasticity models (ARCH) are now commonly used to describe and forecast changes. Engle has won the Noble prize in economics for his innovative work on volatility. However, all these models are univariate models and focused on just one variable. It is intuitively acceptable that the volatilities of different assets and even markets are correlated in such aspects: the volatility of an asset could be transmitted to others, the correlations may change over time. Then, multivariate GARCH models are largely developed to capture such qualities.

In a GARCH model, a time series of continuously compounded returns (including dividends) is denoted $\{r_t\}_{t=1}^{T}$, and $\mathcal{F}_t$ denotes the information set available at t. The unobserved variance of returns conditional on $\mathcal{F}_t$ is defined as follows.

$$\sigma^2_{t+i|\mathcal{F}_t} = VAR(r_{t+i|\mathcal{F}_t}) \tag{1.2.1}$$

Variance predictions are obtained from a set of volatility models, formalized as $\mathcal{M} = \{m_1, m_2, \cdots, m_M\}$. Model m can generically be represented as

$$r_{t+1} = \epsilon_{t+1}\sqrt{h^{(m)}_{t+1}}, \tag{1.2.2}$$

where $h^{(m)}_{t+1}$ is an function and ϵ_{t+1} is an i.i.d zero mean/unit variance innovation. The specification of $h^{(m)}_{t+1}$ determines the conditional variance evolution and is typically a function of the history of returns as well as a vector of unknown parameters to be estimated from the data. The i-step ahead volatility forecast obtained by model m conditional on $\mathcal{F}_t$ is denoted $h^{(m)}_{t+i|t}$.

It is widely recognized that volatility of stock returns responds differently to bad news and good news. In particular, bad news tends to increase the future volatility while same-sized good news will only increase the future volatility by a smaller amount, or even cause decrease in the future volatility. The news impact function (NIF) (Engle & Ng 2012) has been developed as a powerful tool for analyzing the volatility asymmetry for GARCH-type mod-

els. The idea of the NIF is to examine the relationship between conditional volatility in period $t+1$ and the standardized shock to returns in period t in isolation.

The asymmetric effect in volatility is that the effects of positive returns on volatility are different from those of negative returns of a similar magnitude. On the other hand, leverage refers to the negative correlation between the current return and future volatility. Therefore leverage denotes asymmetry, but not all asymmetric effects display leverage. In the class of ARCH specifications that have been developed to capture asymmetric effects, the exponential GARCH (EGARCH) model of Nelson (Nelson 1991) and the GJR model of Glosten et al. (Glosten, Jagannathan & Runkle 2012) are widely used. Using the terminology given above, the EGARCH model can describe leverage, whereas the GJR model can capture asymmetric effects but not leverage (for further details, see (Asai & McAleer 2009)).

As stated above, a popular explanation for asymmetry is the leverage effect proposed by Christie (Christie 1982) . Other forms of asymmetry, such as the asymmetric V-shape, have to be explained by reasons other than the leverage effect. Alternative reasons for the volatility asymmetry that have been suggested in the literature include the volatility feedback effect (Campbell & Hentschel 1992).

The SV models family were proposed by several different different researchers via various models (Ahn & Wilmott 2003, Shephard 2005, Javaheri 2011, Kilin 2011). Among all these stochastic models, one kind of distribution and its process, the Wishart process with different construction has been proposed by several independent researchers, see (Gourieroux et al. 2009, Philipov & Glickman 2006, Asai & McAleer 2009). The basic assumption of Wishart process to capture is that returns follow a multivariate distribution, while the covariance matrix obey a one order autoregressive Markov chain. The notation is as follows.

$$\mathbf{y}_t|\Sigma_t \sim \mathcal{N}_k(\mathbf{0}, \Sigma_t), \tag{1.2.3}$$

$$\Sigma_t^{-1}|\nu, \Sigma_{t-1}^{-1} \sim Wishart_k(\nu, S_{t-1}), \tag{1.2.4}$$

where S_{t-1} is defined as follows:

$$S_{t-1} = \frac{1}{\nu}(A^{1/2})(\Sigma_{t-1}^{-1})^d(A^{1/2})', \tag{1.2.5}$$

where $\mathbf{y_t}$ is the return vector of k different assets at time t, and is set to obey a zero-mean multivariate Gaussian distribution. So it can be regarded as the return that "surprises" the expected. The actual returns are indicated by $\mathbf{y_t}$ together with the expected returns. Accordingly, during the experiment procedure, the raw return data will be filtered to be zero-mean. Σ_t is the latent covariance matrix of returns at time t. All the Σ_t and S_t are defined on $\mathcal{M}_+^k \subset \mathbb{R}^{k\times k}$, which is the set of all the real-valued symmetric and positive-definite matrices of dimension k. ν is the degree of freedom of this process and set to be invariant during the whole process. $t = 1, 2, \cdots, T$ is the indicator of time.

$A \in \mathcal{M}_+^k$ is a positive-definite symmetric matrix of parameters and A is decomposed through a Cholesky decomposition, denoted as $A = (A^{1/2})(A^{1/2})'$. This parameter matrix reveals how each element $\sigma_{ij}(t)$ of covariance matrix Σ_t at time t depends on the elements of covariance matrix Σ_{t-1} at time $t-1$. So A can be interpreted as a measure of intertemporal sensitivity (Philipov & Glickman 2006), (Casarin & Sartore 2008) . d is a scalar parameter to measure the overall strength of relationship between previous period and current period. As discussed in (Philipov & Glickman 2006), we set $d \in (0, 1)$.

1.3 Limitation and Motivation

However, our stock market have been in such a circumstance: with development of modern communication technology and computing power, information has been spreading with unprecedented speed. All the markets around

the world are connected closely and asset prices fluctuations are also coupled. The prices of different assets, different portfolios, the different market performance will interact with each other definitely.

Nevertheless, under the circumstance of globalization, only considering the influences from other assets is not enough. Outside influences often play an important role than one system itself. Like the world-wide economic crises in history, all of them start from one market and then spread to the rest of whole world. Imagine that a professional investor is managing a portfolio in Hong Kong stock market. He needs a good estimation on the covariance matrix of this portfolio. As the Hong Kong stock market is highly correlated with the U.S. market, taking into account the U.S. market will definitely help him make better estimation of the covariance matrix, see Figure 1.2. However, no work has explicitly and systematically address the coupling relationship across systems or markets for volatility due to its complexity with great challenges. The incomplete or local analysis of volatility will inevitably lead to tentative and less effective learning performance. Hence, modeling and analyzing volatility while considering the coupling relationship is a promising research topic.

1.4 Research Contribution

The contribution of this research is listed below:

1. We propose a framework for coupled volatility analysis. In this framework, we consider both the coupling relationships between assets and the coupling relationships among different systems.

2. We include the matrix variable Wishart process into our coupled volatility analysis framework. Besides, we applied our model in two main situations: homogenous coupling and heterogenous coupling. The former model focuses on parallel coupling between homogenous assets or portfolios, the later one models the outside heterogenous influence to target asset or portfolio.

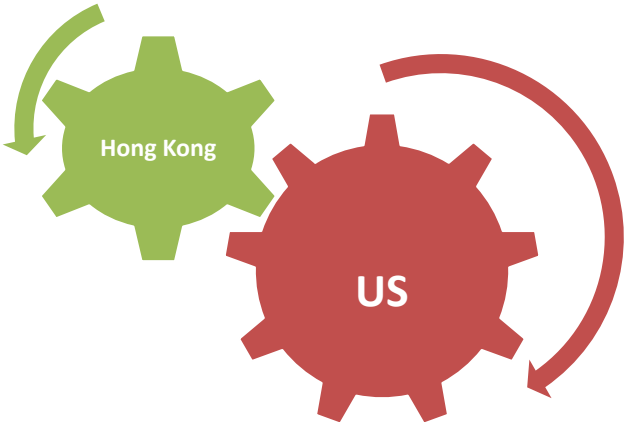

Figure 1.2: Coupling between markets.

3. Learning algorithms are developed based on MCMC, i.e. Gibbs sampling and Metropolis-Hasting sampling, to learn the coupled Wishart process. Detailed likelihood and conditional probability distributions are derived and presented. The parameters and hidden states are simultaneously learned via these algorithms.

1.5 Structure

This thesis focuses on the coupling relationship in the volatility modeling and tries to implement the relationship via coupled Wishart process. The main research issues can be found in Figure. 1.3.

In Chapter 1, we introduce the concepts of volatility and how the research on volatility analysis goes. The motivation and background of this research are presented.

Chapter 2 presents the review on current research available of volatility analysis, coupling methods, and Markov chain Monte Carlo methods respec-

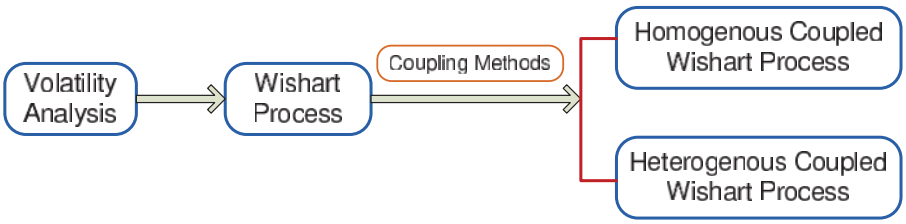

Figure 1.3: Research Issues.

tively.

In Chapter 3, the Wishart process is presented with details and the whole learning process is well presented. The Wishart process follows the Markov rule: the variable at certain time point only relates to its previous ones. And all the matrices are regarded as latent variables and they control the observations. As the model is a probabilistic one with clear graphic model, the learning procedures are based on statistical methods. Among those, a method based on sampling is proposed. Detailed procedures and setups are illustrated in this chapter.

Inspired by the thought that most of the financial activities are related with each other, a homogenous coupling Wishart process is proposed in Chapter 4 to make more precise prediction of the volatility.

To consider the outside influence, a heterogenous coupled Wishart process is proposed in Chapter 5. Unlike the homogenous coupling, heterogenous coupling models focus on the influences from different kinds of data.

Chapter 6 concludes the whole thesis and makes a comment of this research work.

Chapter 2

Literature Review

The literature review of this thesis are conducted in three aspects: volatility analysis, coupling methods, and Markov chain Monte Carlo methods.

2.1 Volatility Analysis

In this section, the related work on volatility modeling is presented. There are two main types of such models: the stochastic models and deterministic models.

2.1.1 GARCH Models

The GARCH family models were largely extended after Engle's innovative work (Engle 1982), autoregressive conditional heteroscedasticity models (ARCH) are now commonly used to describe and forecast changes. Engle has won the Noble prize in economics for his innovative work on volatility.

However, all these models are univariate models and focused on just one variable. It is intuitively acceptable that the volatilities of different assets and even markets are correlated in such aspects: the volatility of an asset could be transmitted to others, the correlations may change over time. Then, multivariate GARCH models are largely developed to capture such qualities.

A time series of continuously compounded returns (including dividends) is denoted $\{r_t\}_{t=1}^T$, and $\mathcal{F}_t$ denotes the information set available at t. The unobserved variance of returns conditional on $\mathcal{F}_t$ is defined as follows.

$$\sigma^2_{t+i|\mathcal{F}_t} = VAR(r_{t+i|\mathcal{F}_t}) \tag{2.1.1}$$

Variance predictions are obtained from a set of volatility models, formalized as $\mathcal{M} = \{m_1, m_2, \cdots, m_M\}$. Model m can generically be represented as

$$r_{t+1} = \epsilon_{t+1}\sqrt{h^{(m)}_{t+1}} \tag{2.1.2}$$

where $h^{(m)}_{t+1}$ is an function and ϵ_{t+1} is an iid zero mean/unit variance innovation. The specification of $h^{(m)}_{t+1}$ determines the conditional variance evolution and is typically a function of the history of returns as well as a vector of unknown parameters to be estimated from the data. The i-step ahead volatility forecast obtained by model m conditional on $\mathcal{F}_t$ is denoted $h^{(m)}_{t+i|t}$.

Here I introduce five models chosen from the vast literature on GARCH, due to their simplicity and demonstrated ability to forecast volatility over alternatives. The first, GARCH(1,1) (Engle 1982), is a natural starting point for model comparison due to its ubiquity and progenesis of alternative models. GARCH describes the volatility process as

$$h_{t+1} = \omega + \alpha r_t^2 + \beta h_t \tag{2.1.3}$$

Key features of this process are its mean reversion (imposed by the restriction $\alpha + \beta < 1$) and its symmetry (the magnitude of past returns, and not their sign, influences future volatility). I also include two asymmetric GARCH models, which are designed to capture the tendency for volatilities to increase more when past returns are negative. Threshold ARCH, or TARCH, appends a linear asymmetry adjustment. Specifically, we have

$$h_{t+1} = \omega + (\alpha + \gamma I_{r_t<c})r_t^2 + \beta h_t \tag{2.1.4}$$

where I is an indicator equaling one when the previous period's return is below some threshold c. The inclination of equity volatilities to rise more when past returns are negative leads to $\gamma > 0$.

Exponential GARCH, or EGARCH, models the log of variance,

$$\ln(h_{t+1}) = \omega + \alpha(|\epsilon_t| - E[|\epsilon_t|]) + \gamma\epsilon_t + \beta \ln(h_t) \tag{2.1.5}$$

where $\epsilon_t = r_t/\sqrt{h_t}$. The leverage effect is manifested in EGARCH as $\gamma < 0$.

The Nonlinear GARCH, or NGARCH, models asymmetry in the spirit of previous specifications using a different functional device. When $\gamma < 0$ the impact of negative news is amplified relative to positive news,

$$h_{t+1} = \omega + \alpha(\gamma + r_t)^2 + \beta h_t \tag{2.1.6}$$

Finally, asymmetric power ARCH (APARCH), devised by Ding et al. (Ding, Granger & Engle 1993), evolves according to

$$h_{t+1}^{\delta/2} = \omega + \alpha(|r_t| - \gamma r_t)^\delta + \beta h_t^{\delta/2} \tag{2.1.7}$$

Raising the left hand side to $2/\delta$ delivers the variance series. Ding et al. (Ding et al. 1993) show that serial correlation of absolute returns is stronger than squared returns. Hence, the free parameter δ can capture volatility dynamics more flexibly than other specifications, while asymmetries are incorporated via γ.

2.1.2 Stochastic Models

The SV models family were proposed by several different different researchers via various models (Ahn & Wilmott 2003, Shephard 2005, Javaheri 2011, Kilin 2011).

It is widely recognized that volatility of stock returns responds differently to bad news and good news. In particular, bad news tends to increase the future volatility while same-sized good news will only increase the future volatility by a smaller amount, or even cause decrease in the future volatility.

The news impact function (NIF) (Engle & Ng 2012) of is a powerful tool for analyzing the volatility asymmetry for GARCH-type models. The idea of the NIF is to examine the relationship between conditional volatility in period $t+1$ and the standardized shock to returns in period t in isolation.

In both the conditional and the stochastic volatility literature, there has been some confusion regarding the definitions of asymmetry and leverage. The asymmetric effect in volatility is that the effects of positive returns on volatility are different from those of negative returns of a similar magnitude. On the other hand, leverage refers to the negative correlation between the current return and future volatility. Therefore leverage denotes asymmetry, but not all asymmetric effects display leverage. In the class of ARCH specifications that have been developed to capture asymmetric effects, the exponential GARCH (EGARCH) model of Nelson (Nelson 1991) and the GJR model of Glosten et al. (Glosten et al. 2012) are widely used. Using the terminology given above, the EGARCH model can describe leverage, whereas the GJR model can capture asymmetric effects but not leverage (for further details, see (Asai & McAleer 2009)).

As stated above, a popular explanation for asymmetry is the leverage effect proposed by Christie (Christie 1982) . Other forms of asymmetry, such as the asymmetric V-shape, have to be explained by reasons other than the leverage effect. Alternative reasons for the volatility asymmetry that have been suggested in the literature include the volatility feedback effect (Campbell & Hentschel 1992). Most asymmetric MSV models are based on the basic SV specifications and hence the positive definiteness of H_t is ensured.

Among all these stochastic models, one kind of distribution and its process, the Wishart process with different construction has been proposed by several independent researchers, see (Gourieroux et al. 2009, Philipov & Glickman 2006, Asai & McAleer 2009). The basic assumption of Wishart process to capture is that returns follow a multivariate distribution, while the covariance matrix obey a one order autoregressive Markov chain. The

notation is as follows.

$$\mathbf{y}_t|\Sigma_t \sim \mathcal{N}_k(\mathbf{0}, \Sigma_t), \tag{2.1.8}$$

$$\Sigma_t^{-1}|\nu, \Sigma_{t-1}^{-1} \sim Wishart_k(\nu, S_{t-1}), \tag{2.1.9}$$

where S_{t-1} is defined as follows:

$$S_{t-1} = \frac{1}{\nu}(A^{1/2})(\Sigma_{t-1}^{-1})^d(A^{1/2})', \tag{2.1.10}$$

where $\mathbf{y_t}$ is the return vector of k different assets at time t, and is set to obey a zero-mean multivariate Gaussian distribution. So it can be regarded as the return that "surprises" the expected. The actual returns are indicated by $\mathbf{y_t}$ together with the expected returns. Accordingly, during the experiment procedure, the raw return data will be filtered to be zero-mean. Σ_t is the latent covariance matrix of returns at time t. All the Σ_t and S_t are defined on $\mathcal{M}_+^k \subset \mathbb{R}^{k\times k}$, which is the set of all the real-valued symmetric and positive-definite matrices of dimension k. ν is the degree of freedom of this process and set to be invariant during the whole process. $t = 1, 2, \cdots, T$ is the indicator of time.

$A \in \mathcal{M}_+^k$ is a positive-definite symmetric matrix of parameters and A is decomposed through a Cholesky decomposition, denoted as $A = (A^{1/2})(A^{1/2})'$. This parameter matrix reveals how each element $\sigma_{ij}(t)$ of covariance matrix Σ_t at time t depends on the elements of covariance matrix Σ_{t-1} at time $t-1$. So A can be interpreted as a measure of intertemporal sensitivity (Philipov & Glickman 2006), (Casarin & Sartore 2008) . d is a scalar parameter to measure the overall strength of relationship between previous period and current period. As discussed in (Philipov & Glickman 2006), we set $d \in (0, 1)$.

2.2 Coupling Methods

2.2.1 Coupling

In mechanic engineering, a coupling is a device used to connect two shafts together at their ends for the purpose of transmitting power. Here this term refers to the close relationship of two or several objects: the objects can be scalar, vectors, time series or systems. The introducing of the concept of coupling has its roots in the complex real world (Cao 2013). In various sceneries, especially the ones related with human activity, the complexity of our research target has reached such a level that the usual assumption of independency can not lead to a satisfactory solution. Behavioral and social applications are two main fields that such situation emerges as a huge challenge.

However most of the existing theories, tools and applications in statistics, data mining and machine learning are built on the assumption independence or conditional independence, which assumes the uncorrelation in the underlying objects, values or attributes. Based on a high level abstraction, it is assumed that objects are independent and identically distributed, with many corresponding algorithms proposed. This works well in simple business applications and abstract problems with weakened relations and heterogeneity, and serves as the foundation of classic theoretical systems and applied tools.

Thus, the introducing of coupling is intuitive from real world application and drawbacks from current research, and will definitely improve the learning of such complex models. There have been several researchers thinking about such a challenging issue and more and more work are being published. These works are from different aspects, theoretical exploration or application driven and they provide interesting and provoking thinking to the rest of us (Steinwart & Christmann 2009).

2.2.2 Coupling Analysis

Coupling in Categorical Data

Figure 2.1 (Topchy, Jain & Punch 2005) shows four possible base clusterings of 12 data objects into two clusters. Different partitions use different sets of labels. The target of clustering ensemble is to obtain a final clustering based on these four base clusterings.

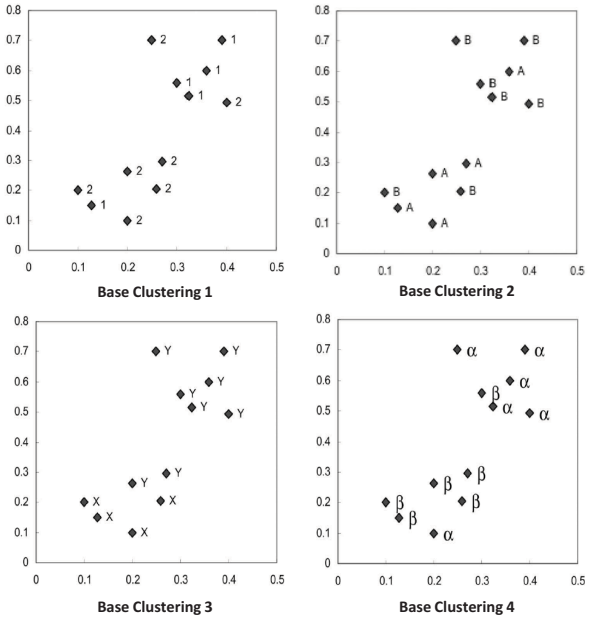

Figure 2.1: Coupled behaviors: clustering ensemble.

Here, we regard each object or observation as an entity of the quantitative behavior. Each base clustering is treated as a property of the quantitative behavior, and the clustering result of each base clustering is the corresponding property value or attribute value of the quantitative behavior. The base clusterings are expected to have interactions with one another, such as the co-occurrence of their cluster labels over the same set of objects, since they are all conducted on the same data objects. This kind of interactions embodies the coupling relationships between the properties of behaviors, which is one coupling aspect of the quantitative coupled behaviors.

In addition, each object has the neighborhood records as its environment. Thus, how this neighborhood impacts the clustering performance reflects another coupling aspect of the quantitative coupled behaviors in terms of the interactions among objects. Both these aspects deliver an example of the quantitative coupled behaviors. In particular, this example illustrates an application of the categorical coupled behaviors, since the associated property of base clustering is in essence a categorical attribute with cluster labels to be its values.

For this clustering ensemble problem, Wang et. al (Wang, She & Cao 2013*b*) introduce a coupled framework of clustering ensemble to formalize and learn both the coupling relationships between base clusterings (i.e, properties) and between data objects (i.e. entities), which are based on non-i.i.d. assumption.

Coupling in Numerical Data

Real-world data sets predominantly consist of quantitative attributes in diverse domains (Saria, Duchi & Koller 2011), such as finance and bioinformatics. The usual recommendation of numerical data is to deliver it as an information table (Kaytoue, Kuznetsov & Napoli 2011), which is a basic knowledge representation framework comprising a table with columns designating "attributes" and rows designating "objects". Each table cell therefore stands for the value of a particular attribute for a particular object. This traditional representation scheme only describes each object by associated variables and assumes the independent and identical distribution of them.

The fragment data of Iris (Table 2.1) is an example that six plant objects are characterized by four numerical attributes (i.e. "Sepal Length", "Sepal Width", "Petal Length", and "Petal Width"), and divided into three classes. For instance, the petal width of plant object u_1 is 0.2cm, which does not reflect any interaction with other attributes. Based on this classical representation, many data mining techniques and machine learning tasks (Plant 2012, Li & Liu 2012) including clustering and classification have been

Table 2.1: A Fragment Example of Iris Data Set

Iris	**Sepal.L** ($\mathbf{a_1}$)	**Sepal.W** ($\mathbf{a_2}$)	**Petal.L** ($\mathbf{a_3}$)	**Petal.W** ($\mathbf{a_4}$)	**Class**
u_1	5.5 cm	4.2 cm	1.4 cm	0.2 cm	Setosa
u_2	5.0 cm	3.4 cm	1.5 cm	0.2 cm	Setosa
u_3	6.1 cm	2.9 cm	4.7 cm	1.4 cm	Versicolor
u_4	6.2 cm	2.2 cm	4.5 cm	1.5 cm	Versicolor
u_5	6.3 cm	2.7 cm	4.9 cm	1.8 cm	Virginica
u_6	6.0 cm	2.2 cm	5.0 cm	1.5 cm	Virginica

performed. One of the critical parts in such applications is to study the pairwise distance between plant objects. A variety of distance metrics have been developed for numerical data, such as Euclidean and Minkowski metrics (Gan, Ma & Wu 2007). Since plant objects u_4 and u_6 have identical values of "Sepal.W" and "Petal.W", the normalized Euclidean distance between them is only 0.493, which is much smaller than that between u_4, u_3 (i.e. 0.950) and nearly half of that between u_6, u_5 (i.e. 0.982). It indicates that u_4 and u_6 stand a good chance to be clustered into the same group. However, in fact, u_4 and u_3 belong to "Versicolor", u_6 and u_5 are labeled as "Virginica". Similar cases can also be observed by the normalized Euclidean distance between u_3 and u_5 (i.e. 0.75), which is smaller than both the distances between u_3, u_4 and between u_5, u_6.

Both instances show that it is often problematic to analyze the numerical data by assuming all the continuous attributes are independent, while the traditional data representation schemes fail to capture the genuine couplings of attributes. In the real world, business and social applications such as investors in capital markets and members in social networking almost always see quantitative attributes coupled with each other (Cao, Ou & Yu 2012). It is very in demand from both practical and theoretical perspectives to develop effective representation method for analyzing continuous variables by considering the relationships among attributes (i.e. non-IIDness of numeri-

cal properties). A conventional way to explore the interaction of continuous attributes is to measure the agreement of shapes between variables via Pearson's correlation coefficient (Gan et al. 2007). Nevertheless, it only caters for the linear relationship between two variables. More often, numerical variables are associated with each other via nonlinear relationships, such as exponential and logarithmic functions. Our motivation is to consider both linear and nonlinear relationship functions, such couplings among variables are called global interactions or global dependency. In contrast, any method to study either the linear relationship or some specific nonlinear function only captures a local picture of the coupling relationships among variables, such as the Pearson's correlation. For Table 2.1, if we adopt the method in (Kalogeratos & Likas 2012) by treating each correlation as the pairwise similarity entry, we then obtain the normalized Euclidean distance between u_4 and u_6 as 0.223, which is still smaller than that between u_4 and u_3 (i.e. 0.329) but only a little larger than that between u_6 and u_5 (i.e. 0.218). It means the coupling relationships are only partially revealed with limited improvement on the distance.

A detailed review of the related work on numerical behavior analysis can be found in (Wang, She & Cao 2013*a*). Though most of the current strategies are based on the hypothesis of IIDness, great efforts have been made to reveal the implicit interactions between properties, such as Pearson's correlation (Gan et al. 2007), rank-correlated measure (Calders, Goethals & Jaroszewicz 2006), dependency clustering (Plant 2012). Nevertheless, no work that systematically and explicitly considers the global coupling relationships among continuous attributes has been reported .

Accordingly, (Wang et al. 2013*a*) proposes a framework of the coupled attribute analysis on numerical data to address the aforementioned research issues, based on the non-IIDness assumption. It considers both the intra-coupled interaction within an attribute, captured by the correlations between every attribute and its own powers; and the inter-coupled interaction among different attributes, quantified by the correlations between each attribute and

the powers of others. A coupled representation scheme is then introduced for quantitative objects to integrate the intra-coupled and inter-coupled interactions with the original information table representation via Taylor-like expansion in a global way.

Coupling in Time Series

The coupling between two or more time series is an interesting research topic and has been proved to be powerful model in some situations. Coupled hidden Markov models (CHMM) have been implemented in the application of speech recognition (Zhong & Ghosh 2002), fraud detection (Cao et al. 2012) and behavior analysis (Song & Cao 2012). The works of Ma (Ma 2007) and Jafari et. al (Jafari, Shirazi, Namaki & Raei 2011) on coupled time series focus on the behavior of financial markets. In these works, during the evolving of time series, the variables at certain time t are not only determined by the past of its own series but also by the neighbor series.

These research works do not take granted that these two time series evolve independently, but build some connections between the nodes based on particular setups. The coupling on time series, unlike the coupling in categorical data and numerical data, focus on the coupling between two systems and build the relationship on the higher level.

2.3 Markov Chain Monte Carlo

2.3.1 Monte Carlo Methods

Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. By running simulations many times according to same probability, just like actually playing and recording your results in a real casino situation, we can get large number of samples of target distribution. With these samples, many problems like optimization and integration can be conducted where ordinary analytic

solution is intractable.

One classic application of Monte Carlo methods is the calculation of π. Suppose we want to estimate π. We know that the area of a circle with radius r is πr^2, but it is also equal to the following definite integral:

$$I = \int_{-r}^{r} \int_{-r}^{r} \mathbb{I}(x^2 + y^2 \leq r^2) dx dy \tag{2.3.1}$$

Hence $\pi = I/(r^2)$. Let us approximate this by Monte Carlo integration. Let $f(x, y) = I(x^2 + y^2 \leq r^2)$ be an indicator function that is 1 for points inside the circle, and 0 outside, and let $p(x)$ and $p(y)$ be uniform distributions on $[-r, r]$, so $p(x) = p(y) = 1/(2r)$. Then

$$I = (2r)(2r) \int \int f(x, y) p(x) p(y) dx dy \tag{2.3.2}$$

$$= 4r^2 \int \int f(x, y) p(x) p(y) dx dy \tag{2.3.3}$$

$$\approx 4r^2 \frac{1}{S} \sum_{s=1}^{S} f(x_s, y_s) \tag{2.3.4}$$

With enough samples, we can plot the points that are accepted/ rejected as in Figure 2.2, with codes from (Murphy 2012).

2.3.2 Correlated Samples

There is one assumption that all these samples mentioned in Section 2.3.1 are i.i.d samples: identical, independent distribution. However, for most of the practical problems, the distributions of interest are usually not the standard ones we are familiar with. Mixture models are such kind of distributions that have wide practical applications. It is usually quite hard to get samples from mixture models directly. To address this issue, Markov chain Monte Carlo (MCMC) methods have been proposed to generated samples from identical distribution but correlated. The Metropolis algorithm, an instance of MCMC, is placed among the then algorithms that have had the greatest influence on the development and practice of science and engineering in the 20th century (Beichl & Sullivan 2000). The MCMC methods have played

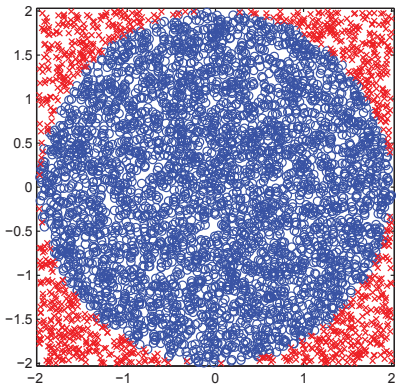

Figure 2.2: Estimating π by Monte Carlo integration, blue points are inside the circle, red crosses are outside.

a significant role in statistics, econometrics, physics, and computing science over the last two decades.

Bayesian inference (e.g., prediction or computation of posterior parameter estimates) relies on integration with respect to some potentially high-dimensional probability distribution. We will generically denote this distribution by π. Except in the simplest cases, such integrals cannot be computed in closed form.

Markov chain Monte Carlo (MCMC) methods provide a class of algorithms that produce estimates of the desired integral based on iterative sampling, combining Monte Carlo integration with samples from a specially constructed Markov chain. The key feature of these methods is that the sampling procedure does not rely on sampling from the distribution π, which is assumed to have an arbitrarily complex form.

The first step in understanding Monte Carlo integration involves formulating the desired integral as an expectation under the distribution π:

$$\int f(x)\pi(x)dx = \mathbb{E}_\pi[f(x)] \tag{2.3.5}$$

The Strong Law of Large Numbers (Gallager & Gallager 1996) informs us that the sample average based on a set of independent samples $x_i \sim \pi$, $i =$

$1, ..., n$, converges almost surely to the true expectation under π. Thus, we may consider the following approximation, which becomes arbitrarily precise for n sufficiently large:

$$\mathbb{E}_\pi[f(x)] \approx \sum_{i=1}^{n} \frac{1}{n} f(x_i). \tag{2.3.6}$$

The assumption of having i.i.d. samples x_i can be relaxed by examination of ergodic theory. The focus of MCMC methods is to develop an ergodic Markov chain with stationary distribution π, which we refer to as the target distribution, such that a sample path from this chain can be used to form the above estimate.

2.3.3 The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm provides a generic method for constructing an ergodic Markov chain, relying solely on defining a valid proposal distribution $q(\cdot|\cdot)$ and evaluation of the target distribution π up to a normalization constant. It is assumed evaluating $\pi(x)$ is easy, but sampling from this distribution is challenging. The Metropolis-Hastings algorithm is outlined in Algorithm 1.

Algorithm 1: Metropolis-Hastings algorithm

Initialize $x^{(0)}$;
for $i = 0$ **to** $N - 1$;
do Sample $u \sim \mathcal{U}_{[0,1]}$;
Sample $x^* \sim q(x^*|x^{(i)})$;
if $u \sim \mathcal{A}(x^*, x^{(i)}) = min\{1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\}$ **then**
  $x^{(i+1)} = x^*$;
else
  $x^{(i+1)} = x^{(i)}$;
end

The acceptance probability $\rho(y|x)$ is defined only when $\pi(x) > 0$. Howev-

er, as long as $\pi(x^0) > 0$, the chain defined in Algorithm 1 will have $\pi(x^t) > 0$ for all t. We use the convention that $\pi(y|x)$ is 0 if both $\pi(x)$ and $\pi(y)$ are zero. To analyze the properties of the Markov chain defined by the Metropolis-Hastings algorithm, it is useful to examine a condition known as detailed balance.

Let $K(y|x) = p(x_{n+1} = y|x_n = x)$ be the transition distribution or transition kernel for a given Markov chain. If $\mathcal{K}(y|x)$ satisfies detailed balance:

$$\mathcal{K}(y|x)\pi(x) = \mathcal{K}(x|y)\pi(y), \tag{2.3.7}$$

then the chain defined by this transition kernel has stationary distribution π. A Markov chain satisfying detailed balance is said to be reversible with respect to π.

Given a chain satisfying detailed balance,

$$\int \mathcal{K}(y|x)\pi(x)dx = \int \mathcal{K}(x|y)\pi(y)dx = \pi(y)\int \mathcal{K}(x|y)dx = \pi(y), \tag{2.3.8}$$

implying that π is indeed a stationary distribution of the Markov chain. It is straightforward to show that the transition kernel defined by Algorithm 1 satisfies detailed balance. With probability $\pi(y|x)$, the chain transitions from x to a sample $y \sim q(y|x)$; otherwise, the chain transitions back to x.

To prove that the Markov chain indeed converges to π (i.e., π is the unique invariant distribution for this chain and this distribution is reached from all initial states), we invoke some mild conditions under which the chain is both aperiodic and Harris recurrent (Robert & Casella 2004). Jointly, these conditions imply ergodicity.

Discussion on the rate of convergence to the stationary distribution can be found in (Gilks, Richardson & Spiegelhalter 1996, Robert & Casella 2004). In general, this burn-in period is challenging to quantify, except by conservative bounds, and is especially challenging to assess in high-dimensional spaces. Convergence can be greatly affected by the initialization of the Markov chain, and in practice, it is common to run multiple chains from different initializations (Gelman & Rubin 1992). Multimodal target distributions with

low valleys between the modes can lead to poorly mixing chains that stay in one region of the state space for long periods of time. Cleverly engineered proposal distributions, such as through tempering (Gelman & Rubin 1992), can play a significant role in the success of a sampling algorithm.

2.3.4 Gibbs Sampling

The Gibbs sampler (Gilks et al. 1996, Robert & Casella 2004) is a special case of the Metropolis-Hastings algorithm in which the proposed sample is always accepted. The Gibbs sampler for n-dimension random variables $(x_1, x_2, \cdots, x_n)$ is summarized in Algorithm 2, from which we see that in order to sample from the full joint distribution on n random variables, it is sufficient to iteratively sample from each of the possibly univariate conditional distributions. As discussed, a node in a directed graph is conditionally independent of all other nodes given its Markov property. Therefore, in the case of sparse graphs, the conditional density from which we are sampling is dependent only on a small subset of the other sampled nodes. We note that, as opposed to Metropolis-Hastings, the Gibbs sampler requires knowledge of the full conditional distributions and an ability to sample from them. Additionally, this algorithm is only applicable to models with at least two random variables.

Algorithm 2: Gibbs sampler

Initialize $x_0^{(1:n)}$;
for $i = 0$ **to** $N - 1$;
do Sample $x_1^{i+1} \sim p(x_1|x_2^{(i)}, x_3^{(i)}, \cdots, x_n^{(i)})$;
Sample $x_2^{i+1} \sim p(x_1|x_1^{(i+1)}, x_3^{(i)}, \cdots, x_n^{(i)})$;
$\cdots$;
Sample $x_j^{i+1} \sim p(x_j|x_1^{(i+1)}, x_{j-1}^{(i+1)}, \cdots, x_{j+1}^{(i)}, x_n^{(i)})$;
$\cdots$;
Sample $x_n^{i+1} \sim p(x_n|x_1^{(i+1)}, x_2^{(i+1)}, \cdots, x_{n-1}^{(i)})$;

To ensure a reversible chain, which leads to a Central Limit Theorem result for the estimator of (Gilks et al. 1996, Robert & Casella 2004), the reversible Gibbs sampler performs a sweep at every iteration from x_1 to x_n followed by a sweep in the reverse ordering back to x_1. Another variant on the standard Gibbs sampler, as proposed by Liu et al (Liu, Wong & Kong 1995), is to choose a random ordering for a single sweep, such an algorithm can lead to improved rates of convergence.

Chapter 3

Wishart Process and Its Learning

In this chapter, we present the Wishart process in detail, then show how to learn the hidden matrices and the parameters together with sampling methods.

3.1 Wishart Process

A Wishart process is a matrix valued continuous time stochastic process with a marginal Wishart distribution. The Wishart distribution is a matrix variate generalization of the chi-squared distribution. Since Wishart processes are defined as a solution to a stochastic differential equation, the existence and uniqueness of strong solutions will be discussed comprehensively. Wishart processes have the property of being symmetric positive definite and are therefore heavily used for modeling interest rates or the covariance matrix in stochastic volatility models.

3.1.1 Wishart Distribution

In statistics, the Wishart distribution is a generalization to multiple dimensions of the chi-squared distribution (Wishart 1928). The random variable

of this distribution is defined on the space $\mathcal{M}_+^k \subset \mathbb{R}^{k\times k}$, which is the set of all $k \times k$ positive definite matrix. These distributions are of great importance in the estimation of covariance matrices in multivariate statistics. In Bayesian statistics, the Wishart distribution is the conjugate prior of the inverse covariance-matrix of a multivariate-normal random-vector.

Suppose X is an $n \times p$ matrix, each row of which is independently drawn from a p-variate normal distribution with zero mean:

$$X_{(i)}=(x_i^1, \ldots, x_i^p) \sim N_p(0, V). \tag{3.1.1}$$

Then the Wishart distribution is the probability distribution of the $p \times p$ random matrix

$$S = X^T X \tag{3.1.2}$$

known as the scatter matrix. One indicates that S has that probability distribution by writing

$$S \sim W_p(V, n). \tag{3.1.3}$$

The positive integer n is the number of degrees of freedom. Sometimes this is written $W(V, p, n)$. If $p = 1$ and $V = 1$ then this distribution is a chi-squared distribution with n degrees of freedom. S^{-1} has the inverse Wishart distribution, $W_p^{-1}(V^{-1}, n)$, which is a conjugate prior for covariance matrices of zero-mean Gaussian distributions. This means that for data $\mathcal{D}$ if a prior $p(R)$ is Gaussian with zero mean, then the posterior $p(R|\mathcal{D})$ is also inverse Wishart.

3.1.2 Transition of Wishart Process

There have been several research from the filed of statistics and machine learning working on Wishart Process, see (Fox & West 2011, Wilson & Ghahramani 2011, Philipov & Glickman 2006). They defined different transition functions or distributions between time points. However, to model volatility we choose the transition from (Philipov & Glickman 2006) as this

model is designed specially to model volatility: every parameter has its unique effect in controlling the level of variance of asset prices.

Below is a graphic model indicating the transition, see Figure 3.1.

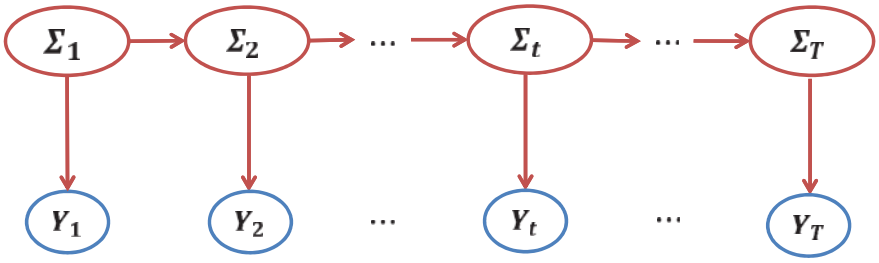

Figure 3.1: Graphic model of matrix transition.

In details, the transition relationship in (Philipov & Glickman 2006) is defined as follows:

$$\mathbf{y}_t|\Sigma_t \sim \mathcal{N}_k(\mathbf{0}, \Sigma_t), \tag{3.1.4}$$

$$\Sigma_t^{-1}|\nu, \Sigma_{t-1}^{-1} \sim \mathcal{W}_k(\nu, S_{t-1}), \tag{3.1.5}$$

where S_{t-1} is defined as follows,

$$S_{t-1} = \frac{1}{\nu}(A^{1/2})(\Sigma_{t-1}^{-1})^d(A^{1/2})'. \tag{3.1.6}$$

The parameter A can be interpreted as a measure of sensitivity on transition of time. This matrix reveals how each element of the current period covariances matrix depends on elements of the previous one. It is the parameter that in part determines mean reversion characteristics on a multivariate level. For example, without restrictions on this matrix parameter, each asset variance, σ_{ii}^2 will depend on the previous period variance of this asset's return as well as on its covariances with all other assets. Thus a change in the volatility of one asset will affect other assets' volatilities. The specification of the intertemporal variance relationship is actually presented in terms of A^{-1}, this can be explained in the conditional expectation of covariance matrix.

The scalar parameter d indicates the overall influence of these relationships. This parameter accounts for the presence of long memory or persistence. Such a phenomenon is quite universal when modeling financial data: today's return has a large effect on the forecast variance many periods in the future. Enormous evidence of long memory of volatility in univariate stochastic volatility models has been documented, see (Jacquier et al. 2002).

The persistence parameter d is the theoretically bound between -1 and 1. For practical financial data, the value for d is between 0 and 1. A value of d close to 0 indicates a weak overall effect of current volatility on future ones within a few subsequent periods. While a value close to 1 means high persistence, which implies that current volatility has a relatively strong effect on its future ones. The special case when $d = 0$ implies a constant volatility, which means the conditional expectation of the volatility in time t is a constant,

$$E(\Sigma_t) = (\nu - k - 1)^{-1} S_{t-1}^{-1} \tag{3.1.7}$$

$$= \frac{\nu}{\nu - k - 1} (A^{-1/2})(\Sigma_{t-1})^0 (A^{-1/2})' \tag{3.1.8}$$

$$= \frac{\nu}{\nu - k - 1} (A^{-1}) \tag{3.1.9}$$

The other special case when $d = 1$ and $A = I$, I is the k-dimension identity matrix, corresponds to a simple matrix-variate random walk,

$$E(\Sigma_t) = \nu S_{t-1} = \Sigma_{t-1} \tag{3.1.10}$$

As long as the parameter A is symmetric positive definite, the parameter d is bounded between 0 and 1, and the Wishart degrees of freedom, ν is greater than the dimension of the multi-variable in the model, then we have a autoregressive stochastic matrix model especially designed for volatility modeling.

3.2 Learning Procedures

In this section, we shall present how the simulation and learning procedures go for such a autoregressive model for matrix variable model. As we can see, it is a Bayesian model with hidden states: the observations are regarded to follow certain distributions.

First, we have to derive the full likelihood for such a model. Refer to the graphic model with parameters, easily, we have:

$$p(Y_{1:T}|\theta, \Sigma_{1:T}^{-1}) = \prod_{t=1}^{T} p(\Sigma_t^{-1}|\Sigma_{t-1}^{-1}, \theta) \cdot p(Y_t|\Sigma_t), \tag{3.2.1}$$

where $\theta = \{A^{-1}, d, \nu\}$.

Here, we have M independent observations for each Σ_t: $Y_t = \{y_{t1}, y_{t2}, \cdots, y_{tM}\}$. When $M = 1$, it is the original model with only one observation.

In this model, the hidden covariance matrices and parameters are learnt together. Following Bayes rule:

$$posterior \propto prior \times likelihood, \tag{3.2.2}$$

the posterior of hidden variables and parameters can be derived.

First, we set all our priors. A^{-1} follows a Wishart distribution $W(Q_0, \gamma_0)$, d is in a uniform distribution at the interval [0,1], ν follows a noninformative uniform distribution, we have:

$$p(A^{-1}) \sim \mathcal{W}(Q_0, \gamma_0) \propto |A^{-1}|^{(\gamma_0 - k - 1)/2} \cdot |Q_0|^{\gamma_0/2} \tag{3.2.3}$$

Besides, we have:

$$S_{t-1} = \frac{1}{\nu}(A^{1/2})(\Sigma_{t-1}^{-1})^d(A^{1/2})', \tag{3.2.4}$$

Then the posterior can be written as:

$$
\begin{aligned}
p(\theta, \Sigma_{1:T}^{-1}|Y_{1:T}) &\propto prior(\theta) \cdot \prod_{t=1}^{T} p(\Sigma_t^{-1}|\Sigma_{t-1}^{-1}, \theta) \cdot p(Y_t|\Sigma_t) \\
&\propto p(A^{-1}) \cdot p(d) \cdot p(\nu) \cdot \prod_{t=1}^{T} p(\Sigma_t^{-1}|\Sigma_{t-1}^{-1}, A^{-1}, d, \nu) \cdot p(Y_t|\Sigma_t) \\
&\propto |Q_0|^{\gamma_0/2} \cdot |A^{-1}|^{(\gamma_0-k-1)/2} \cdot \prod_{t=1}^{T} W(\Sigma_t^{-1}|S_{t-1}, \nu) \cdot \prod_{t=1}^{T}\prod_{m=1}^{M} \mathcal{N}(Y_{tm}|\Sigma_t) \\
&\propto |A^{-1}|^{(\gamma_0-k-1)/2} \cdot \prod_{t=1}^{T} [\frac{1}{2^{\nu k/2} \prod_{i=1}^{k} \Gamma(\frac{\nu+1-i}{2})} \cdot \\
&|S_{t-1}|^{-\nu/2} \cdot |\Sigma_t^{-1}|^{\nu-k-1} \cdot exp\left(-\frac{1}{2} tr(S_{t-1}^{-1}\Sigma_t^{-1})\right)] \cdot \\
&\prod_{t=1}^{T} \left[|\Sigma_t^{-1}|^{M/2} \cdot exp(-\frac{1}{2}\sum_{m=1}^{M} y_{tm}' \Sigma_t^{-1} y_{tm}) \right]
\end{aligned}
\tag{3.2.5}
$$

Now, we can deduce conditional probabilities of all the elements, by eliminating the irrelevant items.

For Σ_t^{-1}, when$t = 1 : T - 1$, we only preserve the items contains Σ_t^{-1}, including S_t, from Equation 5.5.5. (From step 2 to step 3, we use $Tr(ABC) = Tr(BCA) = Tr(CAB)$. And, for a scalar, $x = Tr(x)$.)

$$
\begin{aligned}
p(\Sigma_t^{-1}|\cdot) &\propto Wish(\Sigma_t^{-1}|\nu, S_{t-1}) \cdot Wish(\Sigma_{t+1}^{-1}|\nu, S_t) \cdot \mathcal{N}(Y_t|\Sigma_t) \cdot \\
&\propto |\Sigma_t^{-1}|^{(\nu-k-1)/2} \cdot exp(-\frac{1}{2} tr(S_{t-1}^{-1}\Sigma_t^{-1})) \cdot \\
&|S_t^{-1}|^{-\nu/2} exp(-\frac{1}{2} tr(S_t^{-1}\Sigma_{t+1}^{-1})) \cdot |\Sigma_t^{-1}|^{M/2} \cdot exp(-\frac{1}{2}\sum_{m=1}^{M} y_{tm}' \Sigma_t^{-1} y_{tm}) \\
&\propto |\Sigma_t^{-1}|^{(\nu-k-1)/2} \cdot exp(-\frac{1}{2} tr((S_{t-1}^{-1} + \sum_{m=1}^{M} y_{tm}' y_{tm})\Sigma_t^{-1})) \cdot |\Sigma_t^{-1}|^{(M-\nu d)/2} \\
&\cdot exp(-\frac{1}{2} tr(S_t^{-1}\Sigma_{t+1}^{-1})) \\
&\propto Wish(\Sigma_t^{-1}|\nu, \widetilde{S}_{t-1}) \cdot |\Sigma_t^{-1}|^{(M-\nu d)/2} \cdot exp(-\frac{1}{2} tr(S_t^{-1}\Sigma_{t+1}^{-1})),
\end{aligned}
\tag{3.2.6}
$$

where $\widetilde{S}_{t-1} = (S_{t-1}^{-1} + \sum_{m=1}^{M} y'_{tm} y_{tm})^{-1}$.

When $t = T$,

$$\begin{aligned} p(\Sigma_T^{-1}|\cdot) \propto & Wish(\Sigma_T^{-1}|\nu, S_{T-1}^{-1}) \cdot \mathcal{N}(Y_T|\Sigma_T) \\ \propto & |\Sigma_T^{-1}|^{(\nu-k-1)/2} \cdot exp(-\frac{1}{2} tr(S_{T-1}^{-1}\Sigma_T^{-1}))\cdot \\ & |\Sigma_T^{-1}|^{M/2} \cdot exp(-\frac{1}{2}\sum_{m=1}^{M} y'_{Tm}\Sigma_T^{-1} y_{Tm}) \\ \propto & Wish(\Sigma_t^{-1}|\nu + M, \widetilde{S}_{T-1}), \end{aligned} \tag{3.2.7}$$

where, $\widetilde{S}_{T-1} = (S_{T-1}^{-1} + \sum_{m=1}^{M} y'_{tm} y_{tm})^{-1}$.

For d, we find all the items with d are the S_t, so we only keep them in d's conditional posterior.

$$\begin{aligned} p(d|\cdot) \propto & \prod_{t=1}^{T} \left[|S_{t-1}|^{-\nu/2} exp(-\frac{1}{2} tr(S_{t-1}^{-1}\Sigma_t^{-1})) \right] \\ \propto & \prod_{t=1}^{T} \left[|\Sigma_{t-1}^{-1}|^{-\nu d/2} exp(-\frac{1}{2} tr(S_{t-1}^{-1}\Sigma_t^{-1})) \right] \end{aligned} \tag{3.2.8}$$

For ν, it is a bit more complicated than d,

$$\begin{aligned} p(\nu|\cdot) \propto & p(\nu) \cdot \prod_{t=1}^{T} Wish(\Sigma_t|\nu, S_{t-1}) \\ \propto & \prod_{t=1}^{T} \left[\frac{1}{2^{\nu k/2} \prod_{i=1}^{k} \Gamma(\frac{\nu+1-i}{2})} \cdot |S_{t-1}|^{-\nu/2} \cdot |\Sigma_t^{-1}|^{(\nu-k-1)/2} \cdot \right. \\ & \left. exp\left(-\frac{1}{2} \sum_{t=1}^{T} tr(S_{t-1}^{-1}\Sigma_t^{-1}) \right) \right] \\ \propto & (\frac{|\nu A^{-1}|^{\nu/2}}{2^{\nu k/2} \prod_{i=1}^{k} \Gamma(\frac{\nu+1-i}{2})})^T \prod_{t=1}^{T} |\Sigma_{t-1}^{-1}|^{-\nu d/2} \cdot |\Sigma_t^{-1}|^{(\nu-k-1)/2} \cdot \\ & exp\left(-\frac{1}{2} \sum_{t=1}^{T} tr(S_{t-1}^{-1}\Sigma_t^{-1}) \right) \end{aligned} \tag{3.2.9}$$

For A^{-1}, we keep the prior and all the S_t,

$$
\begin{aligned}
p(A^{-1}|\cdot) &\propto p(A^{-1}) \cdot \prod_{t=1}^{T} Wish(\Sigma_t|\nu, S_{t-1}) \\
&\propto Wish(A^{-1}|\gamma_0, Q_0) \cdot \prod_{t=1}^{T} Wish(\Sigma_t^{-1}|\nu, S_{t-1}) \\
&\propto |A^{-1}|^{\frac{\gamma_0-k-1}{2}} \cdot exp(-\frac{1}{2} \cdot tr(Q_0^{-1}A^{-1}))\cdot \\
&\prod_{t=1}^{T} |S_{t-1}^{-1}|^{\nu/2} exp(-\frac{1}{2} \cdot tr(S_{t-1}^{-1}\Sigma_t^{-1})) \\
&\propto |A^{-1}|^{\frac{\gamma_0-k-1}{2}} \cdot exp(-\frac{1}{2} \cdot tr(Q_0^{-1}A^{-1}))\cdot \\
&\prod_{t=1}^{T} |A^{-1}|^{\nu/2} exp(-\frac{1}{2} \cdot tr(S_{t-1}^{-1}\Sigma_t^{-1})) \\
&\propto |A^{-1}|^{\frac{T\nu+\gamma_0-k-1}{2}} \cdot exp(-\frac{1}{2} \cdot tr(Q_0^{-1}A^{-1}))\cdot \\
&\prod_{t=1}^{T} exp(-\frac{1}{2} \cdot tr(S_{t-1}^{-1}\Sigma_t^{-1})) \\
&\propto |A^{-1}|^{\frac{T\nu+\gamma_0-k-1}{2}} \cdot exp(-\frac{1}{2} \cdot tr((Q_0^{-1} + \nu \sum_{t=1}^{T} \Sigma_{t-1}^{d/2}\Sigma_t^{-1}\Sigma_{t-1}^{d/2})A^{-1})) \\
&\propto Wish(A^{-1}|T\nu + \gamma_0, (Q_0^{-1} + \nu \sum_{t=1}^{T} \Sigma_{t-1}^{d/2}\Sigma_t^{-1}\Sigma_{t-1}^{d/2})^{-1})
\end{aligned}
\tag{3.2.10}
$$

With above posteriors, the procedures of learning hidden states and parameters are as follows, based on Gibbs sampling.

1. Initialize $\theta = \{A, d, \nu\}$ and $\{\Sigma_t\}$.
2. Sample Σ_t from $\Sigma_t|\Sigma_{\setminus t}, A, d, \nu$ for $t = 1, \cdots, T$.
3. Sample A from $A|\{\Sigma_t\}, d, \nu$.
4. Sample d from $d|\{\Sigma_t\}, A, \nu$.
5. Go to Step 2 until enough iteration times.

Please note that for every step, the new samples must be updated in next step. For $t = 1, \cdots, T-1$, the posterior of Σ_t is not a standard distribution

that we can directly sample from. We can adopt the Metropolis Hastings (MH) sampling method, and use a Wishart distribution as the proposal density. When $t = T$, we can sample Σ_T directly from a Wishart distribution with specific parameters. For ν and d, their posteriors are not common distributions either, we can get it by a discrediting method.

3.3 Results Analysis

We set the time series length $T = 100$, and $\nu = 20$, $d = 0.7$, $A = [130, 50; 50, 130]$. Besides, at each time t, M independent observations are generated from Σ_t and we have $M = 20$. Then, we ran 25,000 times of iterations with a Gibbs sampler. The first 5,000 samples are discarded and the rest are recorded. In this section, an analysis is conducted on the obtained samples.

3.3.1 Parameters

In Figure 3.2, the red line indicates the real value, the blue lines are samples. We can see that A and ν are biased estimated, while d suffers less.

In related literature, within the synthetic experiment of Philipov et. al (Philipov & Glickman 2006), there also exists severe bias in parameter estimation. In Wolfgang et. al's report (Rinnergschwentner, Tappeiner & Walde 2011) on the same topic, which has corrected many errors of Philipov et al's original paper, the authors present a very beautiful results in Table 1, however, it is not so convincing, as it is one result from one data sample. In Table 2 of this paper, the mean results of 100 data sets are shown, we can also see strong bias from this.

3.3.2 Determinant Diagnose

Despite the bias, we do not focus on the parameters much. What we want to know is the quality of estimation of covariance matrices. Below in Figure

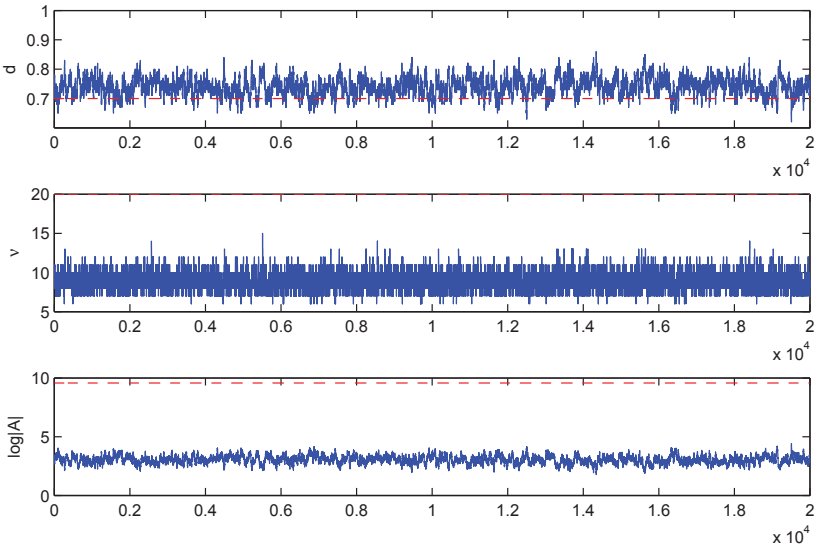

Figure 3.2: Samples of the parameters: d, ν, A.

3.3, we plot the determinant of the true covariance matrices, $\widehat{E(\Sigma_t^{-1}|\cdot)}$, and $\widehat{\widehat{E(\Sigma_t^{-1}|\cdot)}}$, where,

$$\widehat{E(\Sigma_t^{-1}|\cdot)} = (\widehat{A}^{1/2})(\Sigma_{t-1}^{-1})^d(\widehat{A}^{1/2})' \tag{3.3.1}$$

$$\widehat{\widehat{E(\Sigma_t^{-1}|\cdot)}} = (\widehat{A}^{1/2})(\widehat{\Sigma}_{t-1}^{-1})^d(\widehat{A}^{1/2})' \tag{3.3.2}$$

The red line indicates true value, while blue indicates $\widehat{E(\Sigma_t^{-1}|\cdot)}$, and green indicates $\widehat{\widehat{E(\Sigma_t^{-1}|\cdot)}}$.

We can see from the figure that there exists some bias in the estimation of the parameters. However, the parameters are not our focus, but the hidden matrices.

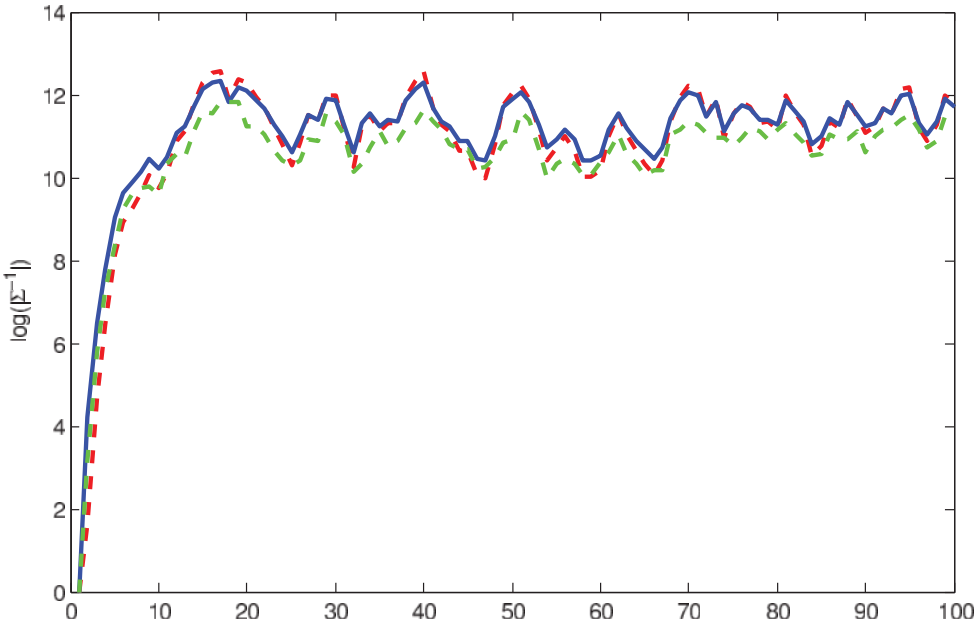

Figure 3.3: Plot of determinant.

3.3.3 Element Diagnose

In this subsection, we plot each element of the covariance matrices, from true value, $\widehat{E(\Sigma_t^{-1}|\cdot)}$, and $\widehat{\widehat{E(\Sigma_t^{-1}|\cdot)}}$ respectively, see Figure 3.4.

3.3.4 Some Issues in Implementing with Singularity

In the Gibbs sampling framework of this project, one important task is to get matrix samples for Σ_t^{-1} and A^{-1}. To produce such matrices, we use a Wishart distribution sampler (for A^{-1} and Σ_T) or use Wishart distribution sampler to get a sample candidate in a Metropolis-Hastings algorithm. In Matlab, this procedure is implemented by a function *wishrnd(sigma,df)*. This function requires one matrix parameter 'sigma' to be positive-definite. However, due to some properties of the model and computing inaccuracy, the intermediate variable $\widetilde{S}$ (to be in *wishrnd* as *sigma*) easily fall into errors. There are two types of errors:

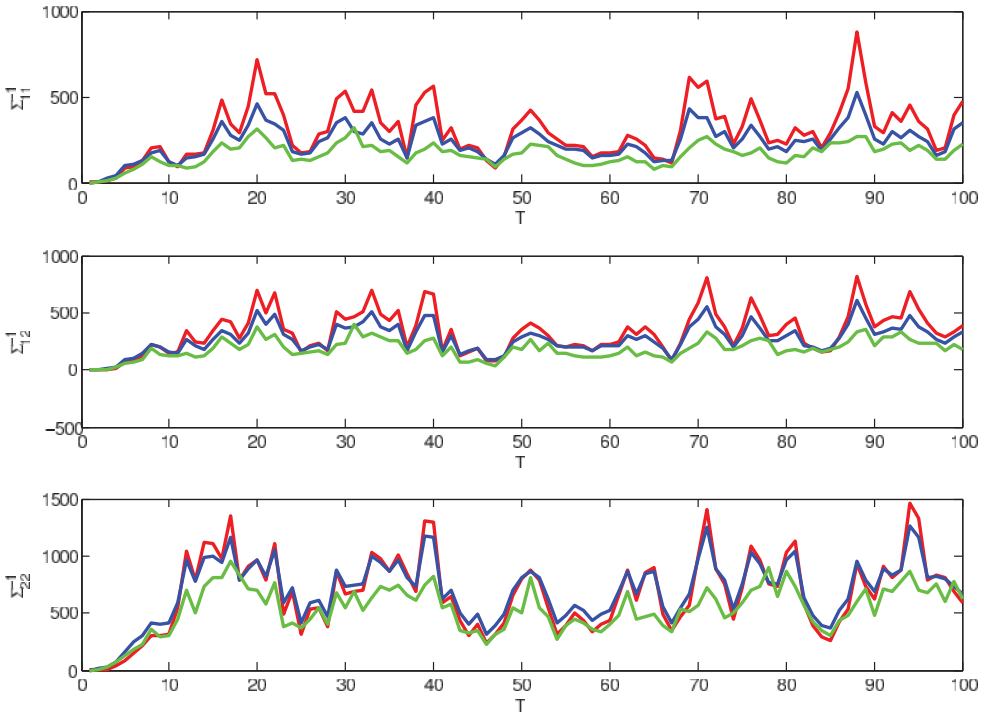

Figure 3.4: Plot of each element.

1. The $\widetilde{S}$ is not so "well" positive-definite, we add a small enough identity matrix to make sure all its eigenvalues larger than 0.

2. The $\widetilde{S}$ is NaN or with negative eigenvalues. This is because the S_{t-1} 'close to singular or badly scaled'.

How does such errors come from? According to the definition of S_t, when Σ_{t-1}^{-1} is positive-definite, S_t should also be positive-definite out of question. However, when Σ_{t-1}^{-1} is close to singular, S_{t-1}^{-1} is close to singular. Then $\widetilde{S}_{t-1}$ will be quite 'sick' and leads to error when calling *wishrnd*.

As to the bad $\Sigma_t^{(-1)}$, one way is to add a small enough identity matrix to it as not to change its property. However, this solution works for some matrices but not for some another matrices, make it not a reliable method in large number of iterations.

Another solution is to make a filter when doing the sampling. Here we use the indicator *rcond* (reciprocal conditional number). When the *rcond* of sampled matrix less than a very small number: eps, we just drop it and resample a new one. If we consider the space of matrix ($rcond < eps$) is small ($rcond(eye(k)) = 1$), it is acceptable to do such a dropping. And the results of sampling are not affected after dropping those whose $rcond < eps$.

3.3.5 Summary

Through the analysis of this sampling results of this data set, and from both previous experience and results from other literature. There commonly exists bias on parameters. However, diagnose on covariance matrices is more soothing, at least from figures: Figure 3.3 and Figure 3.4. Of course, this analysis is superficial, further quantity analysis is needed after more results are obtained.

Chapter 4

Homogenous Coupling Wishart Process

In this chapter, we construct the first type of coupling Wishart process: the homogenous coupling Wishart process. To begin with, we present the background and motivation, then the framework and concrete implement are demonstrated, the following part are the leaning process and evaluation results.

4.1 Background

Volatility indicates the price variations of a specific financial instrument (Engle 1982). The research on volatility concerns every participant in the financial markets: government policy makers, market analysts, fund managers, and economists. Along with the rapid growth of financial markets and the continual development of new and increasingly complex financial instruments, the need for high-quality volatility modeling methods is in great demand. Modeling and predicting volatility is also of vital importance in various applications (de Mattos Neto, Ferreira & Cavalcanti 2011, Philipov & Glickman 2006, Abdullah & Zeng 2010), such as asset pricing, portfolio selection, hedging and risk management. In addition, the burst of several

bubbles aggravates the uncertainty of world market and economy. Especially after the economic crisis in 2008, modeling and analyzing volatility is becoming a more and more critical task.

There has been a great number of research (Jacquier et al. 2002) on the univariate volatility models, which consider the volatility of just one variable. However, studies on the volatility of a certain financial instrument should also include the influences from other assets to capture the dependency and spill-over effect. For this purpose, multivariate volatility models (Asai & McAleer 2009) have been proposed based on the univariate ones. Multivariate volatility models focus on the covariance matrix of a portfolio, and regard that the volatilities evolve together over time, according to specific model setups.

To model the dynamic covariance matrix evolvement in multivariate volatility analysis, Wishart process has been adopted in several literatures (Philipov & Glickman 2006, Gourieroux et al. 2009). As a probabilistic model, Wishart process demonstrates its flexibility and simplicity in modeling the covariance matrix (Asai & McAleer 2009). In a Wishart process, the return of one asset is regarded as an observation of the latent covariance matrix, and conditional independent with others. The latent covariance matrix evolves in a one-order Markov way: only related to the state of its last time.

Nevertheless, under the circumstance of globalization, only considering the influences from other assets is not enough. Outside influences often play an important role than one system itself. Like the world-wide economic crises in history, all of them start from one market and then spread to the rest of whole world. Imagine that a professional investor is managing a portfolio in Hong Kong stock market. He needs a good estimation on the covariance matrix of this portfolio. As the Hong Kong stock market is highly correlated with the U.S. market, taking into account the U.S. market will definitely help him make better estimation of the covariance matrix. However, no work has explicitly and systematically address the coupling relationship across systems or markets for volatility due to its complexity with great challenges. The

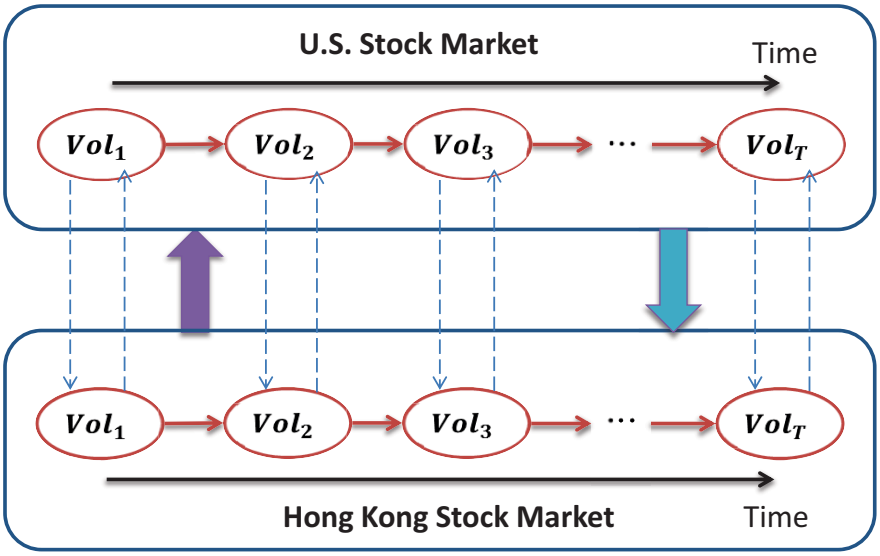

Figure 4.1: The coupling relationship of volatility.

incomplete or local analysis of volatility will inevitably lead to tentative and less effective learning performance.

Here, we introduce the thought of coupling and include the outside influences from other markets. Both in nature and society, there exist some kinds of coupling relationships between individuals and systems. The research on such couplings and interactions is of great significance in diverse fields, like the coupled behavior analysis (Cao et al. 2012), the coupled anomaly detection (Song, Cao, Wu, Wei, Ye & Ding 2012), and the coupled attribute analysis (Wang, Cao, Wang, Li, Wei & Ou 2011). In the example of the Hong Kong investor, a coupled volatility analysis framework can be constructed as in Figure. 4.1. Within this framework, at each time point, the covariance matrix of a portfolio from Hong Kong market is evolving with that of a similar portfolio from U.S. market.

In this chapter, we propose the coupled volatility analysis with linear coupling relationship between the covariance matrices. This relationship is dynamic with time, providing much more flexibility and better accuracy for models. In summary, the key contributions are listed as follows:

– We propose a coupled Wishart process by considering the interactions between different systems, in which we construct linear coupling relationship with properly designed weights.

– An estimation method based on MCMC, i.e. Gibbs sampling and Metropolis-Hasting sampling, is adopted to learn all the parameters and hidden covariance matrices.

– Rigorous experiments have been conducted to show the superiority of our proposed coupled model.

In Section 4.2, we briefly review the related work. Basic knowledge on Wishart distribution and Wishart process is presented in Section 4.3. Section 4.4 proposes the coupled volatility analysis model, a linear coupling approach is introduced in Section 4.5. The estimation method for this model is then specified in Section 4.7. Experiment results are shown in Section 4.9, which prove the effectiveness of our proposed model in capturing the coupling relationship. Finally, the conclusion and future work are presented in Section 4.10.

4.2 Related Work

Volatility analysis is an important research issue that has been broadly studied in econometrics and other communities (Bollerslev 1986). The univariate volatility models deal with the volatility of just one variable (Jacquier et al. 2002). However, such methods lack the consideration of the influences from other variables. Thus, univariate volatility analysis has been extended to the multivariate cases by several researchers (Harvey, Ruiz & Shephard 1994). Two main approaches for volatility analysis are GARCH models and stochastic models. The GARCH models assume that the volatility is a deterministic function of the past (Engle 1982, Bollerslev 1986), while the stochastic models suppose the volatility follows a random process (Harvey et al. 1994). In addition, considering the properties of stock

market, the leverage effect, the asymmetric influence between different stocks and markets, numerous research works have been carried out (Harvey & Shephard 1996). The stochastic models have gained increasing popularity for their flexibility and capturing power (Philipov & Glickman 2006, Asai & McAleer 2009, Gourieroux et al. 2009). Among the methods via stochastic models, several researchers adopt the Wishart process to simulate the fluctuations of a portfolio or the whole market (Asai & McAleer 2009), which provides a probabilistic model and is easy to simulate. The approaches based on the Wishart process regard the volatility as hidden variables and the observations following a normal distribution (Philipov & Glickman 2006). Unlike our focus, these models all deal with the variables within only one system.

Coupling is another aspect of this research, the thought of coupling has been included in a variety of areas and to different levels, as the coupling relationship exists everywhere in real world. Several works have been done on the coupling of discrete data and its applications (Wang et al. 2011), while the coupling between sequences and time series is also studied (Cao et al. 2012, Boronowski & Frangakis 1998). During the evolving of time series, the variables at certain time t are not only determined by the past of its own series but also by the neighbor series. Coupled hidden Markov models (CHMM) are such approaches with extensive studies and applications (Zhong & Ghosh 2002). These research works, in particular for the coupling of time series, however, do not take into account the interactions between different systems. No work has been reported to explicitly and systematically consider the coupling relationship between systems or markets for the volatility analysis, whereas our proposed model addresses this promising research issue.

4.3 Preliminary Knowledge

In this section, we briefly review the Wishart distribution and Wishart process, which have been included to model multivariate volatilities in sever-

al works (Philipov & Glickman 2006, Asai & McAleer 2009, Gourieroux et al. 2009).

4.3.1 Wishart Distribution

The Wishart distribution (Bishop & Nasrabadi 2006) is defined on the space $\mathcal{M}_+^k \subset \mathbb{R}^{k\times k}$, which is the set of $k \times k$ symmetric positive definite matrices. The probability density function of a Wishart distribution is formalized as follows,

$$\mathcal{W}(\mathbf{\Sigma}|\nu, \mathbf{W}) = \frac{1}{2^{\frac{\nu k}{2}} |\mathbf{W}|^{\frac{\nu}{2}} \Gamma_k(\frac{\nu}{2})} |\mathbf{\Sigma}|^{\frac{\nu-k-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\mathbf{W}^{-1}\mathbf{\Sigma})}, \tag{4.3.1}$$

where $\mathbf{\Sigma}$ is the matrix random variable, ν is a scalar parameter larger than k, $\mathbf{W}$ is a $k \times k$ symmetric positive definite matrix parameter, and $\Gamma_k(\cdot)$ is the multivariate gamma function defined as

$$\Gamma_k(\nu/2) = \pi^{k(k-1)/4} \prod_{j=1}^{k} \Gamma\left[(\nu + 1 - j)/2\right]. \tag{4.3.2}$$

4.3.2 Wishart Process

The Wishart process (Philipov & Glickman 2006, Gourieroux et al. 2009, Fox & West 2011) is proposed to model the dynamic time series of covariance matrices. The approach introduced by Philipov and Glickman (Philipov & Glickman 2006) is adopted in this paper. The graphic model for a Wishart process is presented in Figure. 4.2. The dynamic covariance structure is modeled on Wishart distribution:

$$\mathbf{y}_t|\Sigma_t \sim \mathcal{N}_k(\mathbf{0}, \Sigma_t), \tag{4.3.3}$$

$$\Sigma_t^{-1}|\nu, \Sigma_{t-1}^{-1} \sim \mathcal{W}_k(\nu, S_{t-1}), \tag{4.3.4}$$

where S_{t-1} is defined as follows,

$$S_{t-1} = \frac{1}{\nu}(A^{1/2})(\Sigma_{t-1}^{-1})^d(A^{1/2})'. \tag{4.3.5}$$

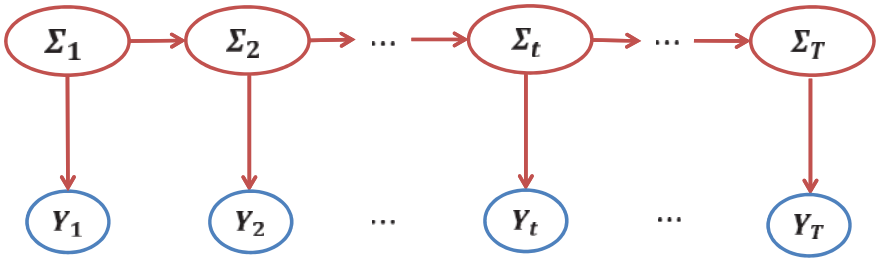

Figure 4.2: A single Wishart process.

The variables and parameters in the above equations are explained here. $\mathbf{y_t}$ is the return vector of k different assets at time t, and is set to obey a zero-mean multivariate Gaussian distribution (denoted as $\mathcal{N}_k$). So it can be regarded as the return that "surprises" the expected. Σ_t is the latent covariance matrix of returns at time t, and treated as the volatility for a portfolio. Σ_t and S_t are symmetric positive definite matrices, defined on $\mathcal{M}_+^k \subset \mathbb{R}^{k\times k}$. ν is the degree of freedom and set to be invariant during the whole process. $A \in \mathcal{M}_+^k$ is also a symmetric positive definite matrix parameter, and can be decomposed through a Cholesky decomposition, denoted as $A = (A^{1/2})(A^{1/2})'$. This parameter matrix reveals how each entry of covariance matrix Σ_t at time t depends on the entries of covariance matrix Σ_{t-1} at time $t-1$. So A can be interpreted as a measure of intertemporal sensitivity (Philipov & Glickman 2006). d is a scalar parameter to measure the overall strength of relationship between previous period and current period. As discussed in (Philipov & Glickman 2006), we set $d \in (0, 1)$. Besides, $t = 1, 2, \cdots, T$ is the time indicator.

4.4 Coupled Volatility Analysis

Considering the interaction of world financial markets, the volatility of one market is greatly influenced by that of other markets. Therefore, constructing a coupled volatility model to analyze the volatility between several different markets is quite essential. Inspired by the research (Cao et al. 2012) based

on the Hidden Markov Model (HMM) and some works on the coupled time series (Qiu, Lu, Cao & He 2011), a coupled model on volatility analysis is proposed.

Below, we present a coupled Wishart process to capture the interactive volatilities upon multiple Wishart processes. Within a coupled framework, it is assumed that the covariance matrix is conditioned with not only the previous state of its own Wishart process, but also its "neighbor" covariance matrix at last step from other Wishart processes.

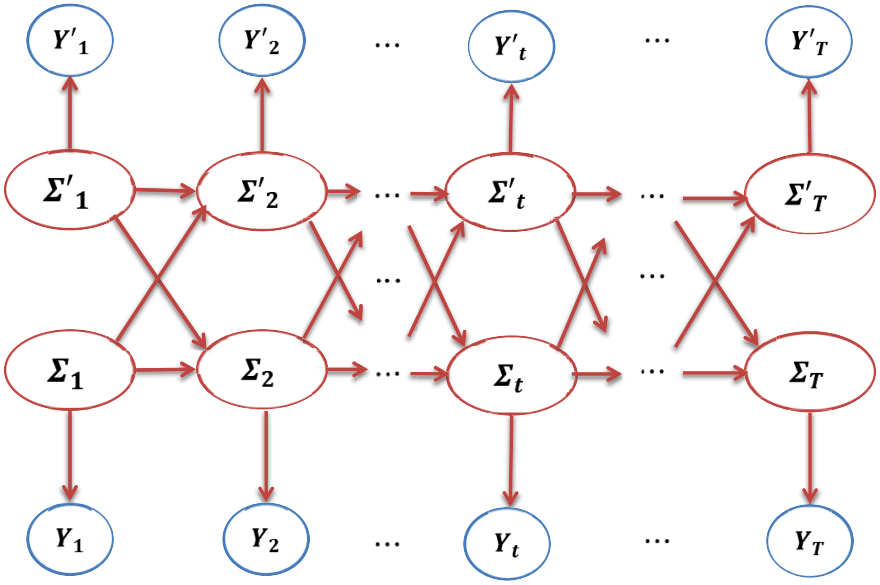

Figure 4.3: An example of the coupled Wishart process.

Take a two-chain coupled Wishart process for example, the graphic model is exhibited in Figure. 4.3. The two Wishart processes (i.e. W and W') are coupled during the procedure of evolving. As can be observed from Figure. 4.3, at each time t, the covariance matrix Σ_t is not only determined by its previous one Σ_{t-1}, but also depends on Σ'_{t-1}. Formally, we have the

formalization as below,

$$\mathbf{y_t}|\Sigma_t \sim \mathcal{N}_k(\mathbf{0}, \Sigma_t) \tag{4.4.1}$$

$$\Sigma_t^{-1}|\nu, \Sigma_{t-1}^{-1}, (\Sigma'_{t-1})^{-1} \sim \mathcal{W}_k(\nu, g(\Sigma_{t-1}, \Sigma'_{t-1})) \tag{4.4.2}$$

$$\mathbf{y_t}'|\Sigma'_t \sim \mathcal{N}_k(\mathbf{0}, \Sigma'_t) \tag{4.4.3}$$

$$\Sigma_t'^{-1}|\nu', \Sigma_{t-1}^{-1}, (\Sigma'_{t-1})^{-1} \sim \mathcal{W}_k(\nu', g'(\Sigma_{t-1}, \Sigma'_{t-1})) \tag{4.4.4}$$

As well presented in Equation (4.4.2) and Equation (4.4.4), at a new time t, the coupled evolvement is passed from last stage $t-1$ by a coupling of covariance matrices, i.e. $g(\Sigma_{t-1}, \Sigma'_{t-1})$ and $g'(\Sigma_{t-1}, \Sigma'_{t-1})$. Here, $g(\cdot)$ and $g(\cdot)'$ are the coupling functions of covariance matrices. How to specifically construct $g(\cdot)$ and $g(\cdot)'$ is another issue, which will be addressed in the next section.

4.5 A Linear Coupling Approach

In this section, we present a linear coupling model for the coupled volatility analysis and its weights setting, together with their theoretical supports.

4.5.1 Linear Model

Based on several coupled frameworks on HMM (Zhong & Ghosh 2002) and linear dynamic systems (Qiu et al. 2011), which consider both inter-coupling and intra-coupling in a linear way, a similar coupled model is constructed accordingly. Below is a coupled system with Q Wishart processes, for each process W_q $(q = 1, 2, \cdots, Q)$ at every time point t, the evolution of the corresponding covariance matrix is coupled with the other $Q-1$ covariance

matrices.

$$\begin{cases} W^{(1)} : \Sigma_t^{(1)} \sim \mathcal{W}_k(\nu^{(1)}, g_{t-1}^{(1)}) \\ W^{(2)} : \Sigma_t^{(2)} \sim \mathcal{W}_k(\nu^{(2)}, g_{t-1}^{(2)}) \\ \qquad \cdots \\ W^{(q)} : \Sigma_t^{(q)} \sim \mathcal{W}_k(\nu^{(q)}, g_{t-1}^{(q)}) \\ \qquad \cdots \\ W^{(Q)} : \Sigma_t^{Q} \sim \mathcal{W}_k(\nu^{(Q)}, g_{t-1}^{(Q)}) \end{cases} \tag{4.5.1}$$

In detail, for the qth Wishart process $W^{(q)}$ at time t, the evolution of covariance is determined by two parameters: $\nu^{(q)}$ and $g_{t-1}^{(q)}$. $\nu^{(q)}$ is the parameter that indicates its own property, and set to be invariant for each process. $g_{t-1}^{(q)}$ is a function of the linear combination of all the covariance matrices at time $t-1$, note that it is a linear specification of $g(\cdot)$ in Equation (4.4.2) and Equation (4.4.4).

The linear combination function $g_{t-1}^{(q)}$ is defined as follows. Equation (4.5.2) is based on the property of Wishart process, and its structure is similar to Equation (4.3.5). Equation (4.5.3) illustrates how exactly the coupling goes. The weights $\{\omega_{t-1}^{(j)}\}$ represents the extent of influences at time t from the jth Wishart process. α is a weight that indicates the self-influence and set to be invariant at our current stage.

$$g_{t-1}^{(q)} = \frac{1}{\nu^{(q)}}(A_q^{1/2})\widehat{\Sigma_{t-1}^{(q)}}^{-d^{(q)}}(A_q^{1/2})' \tag{4.5.2}$$

$$\widehat{\Sigma_{t-1}^{(q)}} = \alpha \cdot \Sigma_{t-1}^{(q)} + (1-\alpha) \cdot \sum_{j=1}^{Q} \omega_{t-1}^{(j)} \cdot \Sigma_{t-1}^{(j)} \tag{4.5.3}$$

The linear structure of coupling relationship on covariance matrices has several advantages. First, linear relationship leads to a clear and easy-to-inference structure, the coupling influences from other systems are modeled with weights. Second, under the linear structure, the self-evolution can be easily modeled, which is an very important part in the evolvement (Qiu et al. 2011).

Here we take the example from Section 4.1 to demonstrate how linear coupling of matrices makes sense. Assume Σ_t and Σ'_t are the covariance matrices of two identical assets portfolio (i.e. manufacturing and consuming) from U.S. and Hong Kong stock markets respectively, we make a linear combination C of them. The linear combination of these two matrices with weights, corresponds to the linear combination of every entry in the matrices. Any element of the combined matrix, like $C(1,1)$ indicates the corresponding linear combination of element $\Sigma_t(1,1)$ from covariance matrix Σ_t and $\Sigma'_t(1,1)$ from covariance matrix Σ'_t.

Here naturally arises another problem: how to acquire a set of reasonable weights? As the coupling objects are not ordinary variables, but positive definite matrices, some related issues are addressed in the next part.

4.5.2 Weights Setting

How to define proper weights for those covariance matrices is an important issue. We define the weights as follows

$$\omega_{t-1}^{(j)} = \frac{||y_t^{(j)} - y_{t-1}^{(j)}||_2}{\sum_{j=1}^{Q} ||y_t^{(j)} - y_{t-1}^{(j)}||_2} \tag{4.5.4}$$

This selection of weights is based on the thought that the violent system tends to exert a greater influence on the quieter ones (Asai & McAleer 2009). So the extent of changes in observations at time $t-1$ determines the weight of evolvement to next time stage.

4.5.3 Theoretical Support

First, we easily prove that a linear combination of covariance matrices is still a symmetric positive definite matrix, which means that the linear structure is a reasonable candidate to capture the coupling relationship.

Theorem 4.5.1 *Suppose A and B are two symmetric positive definite matrices, $\omega_1, \omega_2 \in (0,1)$ are two weights. A linear combination of A and B is*

still a symmetric positive definite matrix.

In fact, for any non-zero vector x, we have

$$x(\omega_1 A + \omega_2 B)x' = \omega_1 \cdot xAx' + \omega_2 \cdot xBx' > 0 \tag{4.5.5}$$

According to the definition of symmetric positive definite matrix, the linear combination of them is still a symmetric positive definite matrix.

Next, we demosntrate that a linear combination of covariance matrices A and B, denoted as C, is still between A and B on the space of covariance matrices $\mathcal{M}_+^k \subset \mathbb{R}^{k\times k}$. The reason is that we accordingly have $d(A,C) \leq d(A,B)$ and $d(B,C) \leq d(A,B)$, detailed proof is available by request. As our target objects are not a vector or a scalar, we apply a distance metric on covariance matrix (Förstner & Moonen 1999).

$$d(A,B) = \sqrt{\sum_{i=1}^{n} \ln^2 \lambda_i(A,B)}, \tag{4.5.6}$$

where $\lambda_i(A,B)$ is the eigenvalue from $|\lambda A - B| = 0$. Positivity, symmetry, and triangle inequality of this metric can be found in (Förstner & Moonen 1999).

4.6 Matrix Metric of Linear coupling

As our target objects are covariances matrices, and we try to come to a linear coupling of them, one metric on the distance between is needed. W. Förstner proposed a metric for covariance matrices in one of his technic report.

The distance measure between covariance matrix A and B is defined as follows,

$$d(A,B) = \sqrt{\sum_{i=1}^{n} \ln^2 \lambda_i(A,B)} \tag{4.6.1}$$

where $\lambda_i(A,B)$ are the eigenvalues from $|\lambda A - B| = 0$. The authors have clarify their properties of positivity, symmetry, and triangle inequality.

As to the linear coupling, we take a simple example, $C = f(A,B,\omega) = (1-\omega)A + \omega B$. We can prove that it satisfies the followings:

1. When $\omega = 1$, the distance between C and one of the border B is 0. That is $d(C, B) = 0$.

2. When $\omega = 0$, the distance between C and another border A is 0. That is $d(C, A) = 0$.

3. When $0 < \omega < 1$, the distance between C and B, and the distance between C and A, are within the distance between A and B. That is,

$$max\{d(C, B), d(C, A)\} < d(A, B) \tag{4.6.2}$$

We prove it to be true. If we can prove $d(C, A) < d(A, B)$ and $d(C, B) < d(A, B)$ to be true respectively. Next, we try to prove $d^2(C, A) < d^2(A, B)$.

For the n eigenvalues of $|\lambda^0 A - B| = 0$, we denote them as $\lambda_1^0, \lambda_2^0, \cdots, \lambda_n^0$. For the n eigenvalues of $|\lambda^1 A - C| = 0$, we denote them as $\lambda_1^1, \lambda_2^1, \cdots, \lambda_n^1$. According to the definition, $d^2(C, A) = \sum_{i=1}^n \ln^2 \lambda_i^1(C, A)$, and $d^2(A, B) = \sum_{i=1}^n \ln^2 \lambda_i^0(A, B)$.

$$|\lambda^1 A - C| = |\lambda^1 A - [(1 - \omega)A + \omega B]| = |(\lambda^1 + \omega - 1)A - \omega B| \tag{4.6.3}$$

From above equation, we can find

$$\lambda_i^0 = \frac{\lambda_i^1 + \omega - 1}{\omega} \tag{4.6.4}$$

then,

$$\sum_{i=1}^n \ln^2 \lambda_i^0 - \sum_{i=1}^n \ln^2 \lambda_i^1 = \sum_{i=1}^n [\ln^2 \lambda_i^0 - \ln^2 \lambda_i^1] \tag{4.6.5}$$

then,

$$\ln^2 \lambda_i^0 - \ln^2 \lambda_i^1 = \ln \frac{\lambda_i^0}{\omega\lambda_i^0 - \omega + 1} \ln \lambda_i^0(\omega\lambda_i^0 - \omega + 1) \tag{4.6.6}$$

We can further prove Eq. (4.6.6) is larger than 0, on matter λ_i is larger or small than 1.

1. When $\lambda_i = 1$, we can get $\ln^2 \lambda_i^0 - \ln^2 \lambda_i^1 = 0$.

2. When $\lambda_i < 1$, we can get $\ln^2 \lambda_i^0 - \ln^2 \lambda_i^1 > 0$

3. When $\lambda_i > 1$, we can get $\ln^2 \lambda_i^0 - \ln^2 \lambda_i^1 > 0$

Thus, we have $d^2(C, A) < d^2(A, B)$. Similarly, we can have $d^2(C, B) < d^2(A, B)$. Then, Eq. (4.6.2) holds.

4.7 Estimation Methods

For such a probabilistic framework with a clear graphic model, Markov chain Monte Carlo (MCMC) (Andrieu, De Freitas, Doucet & Jordan 2003) methods are promising ways to simulate the time series, as the parameters and latent variables can be estimated together. In this section, we propose an MCMC technique to estimate the coupled Wishart process.

As a Bayesian approach, the idea behind MCMC methods is to produce values for specific variables from a known distribution of interest (usually multivariate distribution) by sampling a Markov chain, whose invariant transition distribution is just the target distribution. Unlike the usual problem of acquiring the maximum likelihood of parameter Θ, in this model, the parameter space is augmented to include all the latent variables. After the Markov chain converges to the target distribution, these draws are treated as the samples from marginal posterior densities. Therefore, several statistics such as mean and moment can be approximately calculated.

From the construction of our model, we find that our target posterior distribution has a high-dimension density (see Equation (4.7.1)), with both scalar and matrix variables. For such a high-dimension posterior, Gibbs sampling (Andrieu et al. 2003) is a good choice, which produces samples of every element from its conditional probability with all others, then repeats this process while updating the samples according to the conditional probability density function. The most advantage of adopting Gibbs sampling is that it produces high-dimension samples via its own procedures. Besides, a Gibbs sampler accepts all the candidate draws (Andrieu et al. 2003), which avoids the trouble of choosing a proper proposal distribution in practice.

To implement the Gibbs sampling, we deduce the joint posterior of pa-

rameters and hidden variables first. The joint posterior function in a coupled Wishart process is proportional to the priors multiplied by the likelihood. Take a two-chain coupled Wishart process as an example, below is the posterior distribution:

$$\begin{aligned} & p(\nu^{(1)}, \nu^{(2)}, A^{(1)}, A^{(2)}, d^{(1)}, d^{(2)}, \Sigma_{1:T}^{(1)}, \Sigma_{1:T}^{(2)} | \mathbf{y}_{1:T}^{(1)}, \mathbf{y}_{1:T}^{(2)}) \\ \propto\ & p(\nu^{(1)}) \cdot p(\nu^{(2)}) \cdot p(A^{(1)}) \cdot p(A^{(2)}) \cdot p(d^{(1)}) \cdot p(d^{(2)}) \cdot \\ & \prod_{i=1}^{T} p(\Sigma_t^{(1)} | \cdot) \prod_{i=1}^{T} p(\Sigma_t^{(2)} | \cdot) \prod_{i=1}^{T} p(\mathbf{y}_t^{(1)} | \cdot) \prod_{i=1}^{T} p(\mathbf{y}_t^{(2)} | \cdot) \end{aligned} \tag{4.7.1}$$

For the space limit, we do not give a completely detailed result here. Based on the above posterior function, we derive the distribution density of every item conditioned with the rest and conduct Gibbs sampling. After the parameters have been estimated, predictions can be made. The conditional posterior density of each element is specified in Appendix.

The structure of the Gibbs sampler on the estimation of parameters and hidden variables is as follows:

---

(1) Initialize parameters $A^{(1)}$, $d^{(1)}$, $A^{(2)}$, $d^{(2)}$ and latent variables $\{\Sigma_t^{(1)}\}$, $\{\Sigma_t^{(2)}\}$.

(2) Sample $\Sigma_t^{(1)}$ from $\Sigma_t^{(1)} | \{\Sigma_{\setminus t}^{(1)}, A^{(1)}, d^{(1)}, A^{(2)}, d^{(2)}\}$, for $t = 1, \cdots, T$.

(3) Sequentially, sample each parameter or latent matrix from its conditional posterior distribution density. Before every drawing within one cycle, we update the known samples when producing draws for other elements.

(4) Go to Step (2) until predefined interations.

Cycling through steps (2)-(6) is a complete flow path of MCMC sampler for the coupled Wishart process. Actually, in each step of sampling, other techniques like Metropolis sampling is also applied (Andrieu et al. 2003), as most of the posterior densities are not the common distributions that we can directly make samples from.

4.8 Likelihood and Conditional Posteriors

For a single Wishart process proposed by (Philipov & Glickman 2006), Gibbs sampling has been proposed for its simulation. However, for the coupled Wishart process, the Gibbs sampler is much more complicated than the single one. Here we deduce all the conditional probabilities in the Gibbs sampling framework.

We can refer to the graphic model in Figure 4.3. There are two chains $W^{(1)}$ and $W^{(2)}$, the parameters of each chain is denoted as $\theta^{(1)} = A^{(1)}, d^{(1)}, \nu^{(1)}$ and $\theta^{(2)} = A^{(2)}, d^{(2)}, \nu^{(2)}$. If they are coupled with each other, the joint posterior can be written as follows,

$$p(\Sigma_{1:T}^{(1)}, \Sigma_{1:T}^{(2)}, \theta^{(1)}, \theta^{(2)} | y_{1:T}^{(1)}, y_{1:T}^{(2)}) = p(\theta^{(1)}) \cdot p(\theta^{(2)}) \cdot \tag{4.8.1}$$

$$\prod_{t=1}^{T} p(\Sigma_t^{(1)} | \Sigma_{t-1}^{(1)}, \Sigma_{t-1}^{(2)}) \cdot p(y_t^{(1)} | \cdot) \cdot \tag{4.8.2}$$

$$\prod_{t=1}^{T} p(\Sigma_t^{(2)} | \Sigma_{t-1}^{(1)}, \Sigma_{t-1}^{(2)}) \cdot p(y_t^{(2)} | \cdot) \tag{4.8.3}$$

$$= p(A^{(1)}) \cdot p(d^{(1)}) \cdot p(\nu^{(1)}) \cdot p(A^{(2)}) \cdot p(d^{(2)}) \cdot p(\nu^{(2)}) \cdot \tag{4.8.4}$$

$$\prod_{t=1}^{T} Wish(\Sigma_t^{(1)} | \nu^{(1)}, S_{t-1}^{(1)}) \cdot N(y_t^{(1)} | \Sigma_t^{(1)}) \cdot \tag{4.8.5}$$

$$\prod_{t=1}^{T} Wish(\Sigma_t^{(2)} | \nu^{(2)}, S_{t-1}^{(2)}) \cdot N(y_t^{(2)} | \Sigma_t^{(2)}) \tag{4.8.6}$$

where $S_{t-1}^{(1)}$ and $S_{t-1}^{(2)}$ are functions of the combination of the covariance matrices.

$$S_{t-1}^{(1)} = \frac{1}{\nu^{(1)}} \cdot A(1)^{1/2} (\Sigma_{t-1}^{-1}(c1))^{d^{(1)}} (A(1)^{1/2})' \tag{4.8.7}$$

$$S_{t-1}^{(2)} = \frac{1}{\nu^{(2)}} \cdot A(2)^{1/2} (\Sigma_{t-1}^{-1}(c2))^{d^{(2)}} (A(2)^{1/2})' \tag{4.8.8}$$

and,

$$\Sigma_{t-1}^{-1}(c1) = \omega_t^{(1)}\Sigma_{t-1}^{-1}(1) + (1-\omega_t^{(1)})\Sigma_{t-1}^{-1}(2) \tag{4.8.9}$$

$$\Sigma_{t-1}^{-1}(c2) = \omega_t^{(2)}\Sigma_{t-1}^{-1}(1) + (1-\omega_t^{(2)})\Sigma_{t-1}^{-1}(2) \tag{4.8.10}$$

the ω are weights of the covariance matrices.

From the joint posterior distribution above, we can deduce all the conditional distribution of each element sequentially.

Conditional posterior for Σ

$$\begin{aligned} p(\Sigma_t^{-1}(1)|\cdot) \propto\ & Wish(\Sigma_t^{-1}(1)|\nu^{(1)}, S_{t-1}^{(1)})\cdot \\ & Wish(\Sigma_{t+1}^{-1}(1)|\nu^{(1)}, S_t^{(1)})\cdot \\ & Wish(\Sigma_{t+1}^{-1}(2)|\nu^{(2)}, S_t^{(2)}) \cdot N(y_t^{(1)}|\Sigma_t^{(1)}) \\ \propto\ & |\Sigma_t^{-1}(1)|^{(\nu^{(1)}-k-1)/2} \cdot exp(-\frac{1}{2}tr(S_{t-1}^{(1)}\Sigma_t^{-1}(1)))\cdot \\ & |S_t^{(1)}|^{-\nu^{(1)}/2}exp(-\frac{1}{2}tr(S_t^{(1)}\Sigma_{t+1}^{-1}(1)))\cdot \\ & |S_t^{(2)}|^{-\nu^{(2)}/2}exp(-\frac{1}{2}tr(S_t^{(2)}\Sigma_{t+1}^{-1}(2)))\cdot \\ & |\Sigma_t(1)|^{-1/2}exp(-\frac{1}{2}(y_t^{(1)})'\Sigma_t^{-1}(1)y_t^{(1)}) \end{aligned} \tag{4.8.11}$$

And,

$$\begin{aligned} p(\Sigma_t^{-1}(2)|\cdot) \propto\ & Wish(\Sigma_t^{-1}(2)|\nu^{(2)}, S_{t-1}^{(2)})\cdot \\ & Wish(\Sigma_{t+1}^{-1}(2)|\nu^{(2)}, S_t^{(2)})\cdot \\ & Wish(\Sigma_{t+1}^{-1}(1)|\nu^{(1)}, S_t^{(1)}) \cdot N(y_t^{(2)}|\Sigma_t^{(2)}) \\ \propto\ & |\Sigma_t^{-1}(2)|^{(\nu^{(2)}-k-1)/2} \cdot exp(-\frac{1}{2}tr(S_{t-1}^{(2)}\Sigma_t^{-1}(2)))\cdot \\ & |S_t^{(2)}|^{-\nu^{(1)}/2}exp(-\frac{1}{2}tr(S_t^{(2)}\Sigma_{t+1}^{-1}(2)))\cdot \\ & |S_t^{(1)}|^{-\nu^{(1)}/2}exp(-\frac{1}{2}tr(S_t^{(1)}\Sigma_{t+1}^{-1}(1)))\cdot \\ & |\Sigma_t(2)|^{-1/2)}exp(-\frac{1}{2}(y_t^{(2)})'\Sigma_t^{-1}(2)y_t^{(2)}) \end{aligned} \tag{4.8.12}$$

when $t = T$,

$$\begin{aligned} p(\Sigma_T^{-1}(1)|\cdot) \propto & Wish(\Sigma_T^{-1}(1)|\nu^{(1)}, S_{T-1}^{(1)}) \cdot N(y_T^{(1)}|\Sigma_T^{(1)}) \\ \propto & |\Sigma_t^{-1}(1)|^{(\nu^{(1)}-k-1)/2)} \cdot exp(-\frac{1}{2}tr(S_{t-1}^{(1)}\Sigma_t^{-1}(1))) \cdot \\ & |\Sigma_t(1)|^{-1/2} exp(-\frac{1}{2}(y_t^{(1)})'\Sigma_t^{-1}(1)y_t^{(1)}) \end{aligned} \tag{4.8.13}$$

and,

$$\begin{aligned} p(\Sigma_T^{-1}(2)|\cdot) \propto & Wish(\Sigma_T^{-1}(2)|\nu^{(2)}, S_{T-1}^{(2)}) \cdot N(y_T^{(2)}|\Sigma_T^{(2)}) \\ \propto & |\Sigma_t^{-1}(2)|^{(\nu^{(2)}-k-1)/2)} \cdot exp(-\frac{1}{2}tr(S_{t-1}^{(2)}\Sigma_t^{-1}(2))) \cdot \\ & |\Sigma_t(2)|^{-1/2} exp(-\frac{1}{2}(y_t^{(2)})'\Sigma_t^{-1}(2)y_t^{(2)}) \end{aligned} \tag{4.8.14}$$

Conditional posterior for d

$$\begin{aligned} p(d^{(1)}|\cdot) \propto & \prod_{t=1}^{T} \left[|S_{t-1}^{(1)}|^{-\nu^{(1)}/2} exp(-\frac{1}{2}tr(S_{t-1}^{(1)}\Sigma_t^{-1}(1))) \right] \\ \propto & \prod_{t=1}^{T} \left[|\Sigma_{t-1}^{-1}(1)|^{-\nu^{(1)}d^{(1)}/2} exp(-\frac{1}{2}tr(S_{t-1}^{(1)}\Sigma_t^{-1}(1))) \right] \end{aligned} \tag{4.8.15}$$

And,

$$\begin{aligned} p(d^{(2)}|\cdot) \propto & \prod_{t=1}^{T} \left[|S_{t-1}^{(1)}|^{-\nu^{(1)}/2} exp(-\frac{1}{2}tr(S_{t-1}^{(1)}\Sigma_t^{-1}(1))) \right] \\ \propto & \prod_{t=1}^{T} \left[|\Sigma_{t-1}^{-1}(1)|^{-\nu^{(1)}d^{(1)}/2} exp(-\frac{1}{2}tr(S_{t-1}^{(1)}\Sigma_t^{-1}(1))) \right] \end{aligned} \tag{4.8.16}$$

Conditional posterior for ν

$$\begin{aligned} p(\nu^{(1)}|\cdot) \propto & \, p(\nu^{(1)}) \cdot \prod_{t=1}^{T} Wish(\Sigma_t^{(1)}|\nu^{(1)}, S_{t-1}^{(1)}) \\ \propto & \, (\frac{|\nu^{(1)}A(1)^{-1}|^{\nu^{(1)}/2}}{2^{\nu^{(1)}k/2}\prod_{i=1}^{k}\Gamma(\frac{\nu^{(1)}+1-i}{2})})^T \prod_{t=1}^{T} |\Sigma_{t-1}(1)|^{\nu^{(1)}d^{(1)}/2} \cdot \\ & |\Sigma_t^{-1}(1)|^{\nu^{(1)}/2} \cdot exp\left(-\frac{1}{2}\sum_{t=1}^{T} tr(S_{t-1}^{-1}(1)\Sigma_t^{-1}(1)) \right) \end{aligned} \tag{4.8.17}$$

And,

$$
\begin{aligned}
p(\nu^{(2)}|\cdot) &\propto p(\nu^{(2)}) \cdot \prod_{t=1}^{T} Wish(\Sigma_t^{(2)}|\nu^{(2)}, S_{t-1}^{(2)}) \\
&\propto \left(\frac{|\nu^{(2)} A(2)^{-1}|^{\nu^{(2)}/2}}{2^{\nu^{(2)}k/2} \prod_{i=1}^{k} \Gamma(\frac{\nu^{(1)}+1-i}{2})}\right)^T \prod_{t=1}^{T} |\Sigma_{t-1}(2)|^{\nu^{(2)} d^{(2)}/2} \cdot \\
&\quad |\Sigma_t^{-1}(2)|^{\nu^{(2)}/2} \cdot exp\left(-\frac{1}{2}\sum_{t=1}^{T} tr(S_{t-1}^{-1}(2)\Sigma_t^{-1}(2))\right)
\end{aligned} \tag{4.8.18}
$$

Conditional posterior for A

$$
\begin{aligned}
p(A^{-1}(1)|\cdot) &\propto p(A^{-1}(1)) \cdot \prod_{t=1}^{T} Wish(\Sigma_t^{(1)}|\nu^{(1)}, S_{t-1}^{(1)}) \\
&\propto Wish(A^{-1}(1)|\gamma_0, Q_0) \cdot \prod_{t=1}^{T} Wish(\Sigma_t^{(1)}|\nu^{(1)}, S_{t-1}^{(1)}) \\
&\propto |A^{-1}(1)|^{\frac{\gamma_0-k-1}{2}} \cdot exp(-1/2 \cdot tr(Q_0^{-1}A^{-1}(1))) \cdot \\
&\prod_{t=1}^{T} |\Sigma_{t-1}^{-1}(1)|^{\nu^{(1)}/2} exp(-1/2 \cdot tr(S_{t-1}^{-1}(1)\Sigma_t^{-1}(1)))
\end{aligned} \tag{4.8.19}
$$

And,

$$
\begin{aligned}
p(A^{-1}(2)|\cdot) &\propto p(A^{-1}(2)) \cdot \prod_{t=1}^{T} Wish(\Sigma_t^{(2)}|\nu^{(2)}, S_{t-1}^{(2)}) \\
&\propto Wish(A^{-1}(2)|\gamma_0, Q_0) \cdot \prod_{t=1}^{T} Wish(\Sigma_t^{(2)}|\nu^{(2)}, S_{t-1}^{(2)}) \\
&\propto |A^{-1}(2)|^{\frac{\gamma_0-k-1}{2}} \cdot exp(-1/2 \cdot tr(Q_0^{-1}A^{-1}(2))) \cdot \\
&\prod_{t=1}^{T} |\Sigma_{t-1}^{-1}(2)|^{\nu^{(1)}/2} exp(-1/2 \cdot tr(S_{t-1}^{-1}(1)\Sigma_t^{-1}(1)))
\end{aligned} \tag{4.8.20}
$$

4.9 Experiment and Evaluation

Several experiments are performed on synthetic and real-life data sets to show the effectiveness of our proposed coupled Wishart process. The experiments

for this research are designed into two stages. The first part is the simulation conducted on two sets of coupled synthetic data to test the capacity of our proposed model in capturing the coupling relationship. In the second stage, we implement this model and corresponding learning methods to a real-life data set.

4.9.1 Evaluation Measures

In the following experiments, three evaluation measures, mean absolute percentage error (MAPE) (Philipov & Glickman 2006), mean squared error (MSE) (Wilson & Ghahramani 2011), and determinant error (Det_error) (Philipov & Glickman 2006) are included to assess the quality of our method.

Mean Absolute Percentage Error

To evaluate the prediction quality, we use the MAPE between the predicted and true covariance matrices. The MAPE for the (i, j) element is calculated as

$$MAPE_{ij} = \frac{1}{T}\sum_{t=1}^{T}\frac{|\Sigma_t^{-1}(i,j) - \widehat{E(\Sigma_t^{-1}|\cdot)}|(i,j)}{|\Sigma_t^{-1}(i,j)|}, \tag{4.9.1}$$

where Σ_t^{-1} is the inverse of covariance matrix at time t. $\widehat{E(\Sigma_t^{-1}|\cdot)}$ is the estimated inverse covariance matrix mean, defined as $\widehat{E(\Sigma_t^{-1}|\cdot)} = (\widehat{A}^{1/2})(\Sigma_{t-1}^{-1})^d(\widehat{A}^{1/2})'$.

Mean Squared Error

Another evaluation metric is the MSE, usually adopted to measure the fitted time series values in statistics. With simulated data, we can directly calculate the MSE as the true value is known, see Equation (4.9.2). Here, we do not use the inverse of covariance matrices, but themselves.

$$MSE_{ij} = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\Sigma_t}(i,j) - \Sigma_t(i,j))^2, \tag{4.9.2}$$

where $\widehat{\Sigma_t}$ is the estimated covariance matrix mean.

However, in real applications, we never observe the true covariance matrix Σ_t. When the ground truth is not known, we alternatively use the proxy $S_t(i,j) = y_t(i)y_t(j)$ where y_i is the ith component of the multivariate observation y_t. This is because $E[y_t(i)y_t(j)] = \Sigma_t(i,j)$, assuming y(t) has a zero mean, and a brief proof is provided in Equation (4.9.3).

$$\begin{aligned}\Sigma_t(i,j) &= \mathbf{Cov}(y_t(i), y_t(j)) \\ &= \mathbf{E}(y_t(i) - E(y_t(i)))(y_t(j) - E(y_t(j))) \\ &= \mathbf{E}(y_t(i)y_t(j))\end{aligned} \tag{4.9.3}$$

In a thorough empirical study, Brownlees et al. (Brownlees, Engle & Kelly n.d.) use the univariate analogue of this proxy. The MSE error then can be calculated by

$$MSE_{ij} = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\Sigma_t}(i,j) - y_t(i)y_t(j))^2 \tag{4.9.4}$$

Determinant Error

Besides, we compute the logarithmic determinants of the estimated covariance matrix and the true value, and calculate the Det_error as

$$Det\_error = \frac{1}{T}\sum_{t=1}^{T}(\log|\Sigma_t| - \log|\widehat{E(\Sigma_t)}|)^2, \tag{4.9.5}$$

where $\widehat{\Sigma(t)}$ is the estimated covariance matrix mean.

For all these three measures, the smaller ones indicate closer estimated values with the ground truth, corresponding to better models.

4.9.2 Synthetic Data Analysis

In the first part, we set all the parameters as follows to generate two time series $\{Y_t^{(1)}\}$ and $\{Y_t^{(2)}\}$:

- Process 1: $\nu^{(1)} = 30$, $d^{(1)} = 0.7$, $A^{(1)} = (130\ 50; 50\ 130)$, and $\Sigma_0^{(1)} = (3\ 0.8; 0.8\ 3)$.

– Process 2: $\nu^{(2)} = 20$, $d^{(2)} = 0.3$, $A^{(2)} = (13\ 5; 5\ 13)$, and $\Sigma_0^{(2)} = (3\ 0.8; 0.8\ 3)$.

The coupling setup is: $\alpha = 0.5$, $w_1 = 0.7$ and $w_2 = 0.3$, which means the coupling relationship in this simulation is not symmetric.

Based on the above settings, we produce two sets of the coupled time series: $\{Y_t^{(1)}\}$ and $\{Y_t^{(2)}\}$. In order to estimate these parameters, the MCMC simulation is conducted with 100000 iterations. The first 20000 draws are discarded and the remains are kept. By averaging the selected samples, we obtain the estimation of both parameters and latent covariance matrices. Then we conduct this group of experiments for ten times and analyze the statistical property of the results.

Firstly, we present the estimation of parameters in box plots. As shown in Figure. 4.4 and Figure. 4.5, some of the parameters are well estimated like $d^{(1)}$ and $d^{(2)}$ (i.e., the estimated values for $d^{(1)}$ and $d^{(2)}$ are 0.68 and 0.23 respectively, while the true values are 0.7 and 0.3). However, some other parameters, like $\nu^{(1)}$, $\nu^{(2)}$ are underestimated due to its sensitivity to the appropriation of covariance matrix (Philipov & Glickman 2006). Just like the learning of original Wishart process in (Asai & McAleer 2009), some parameters are estimated with bias. As our ultimate goal is to model the covariance matrices, not the parameters, now we want to know whether this underestimation would affect the learning for covariance matrices.

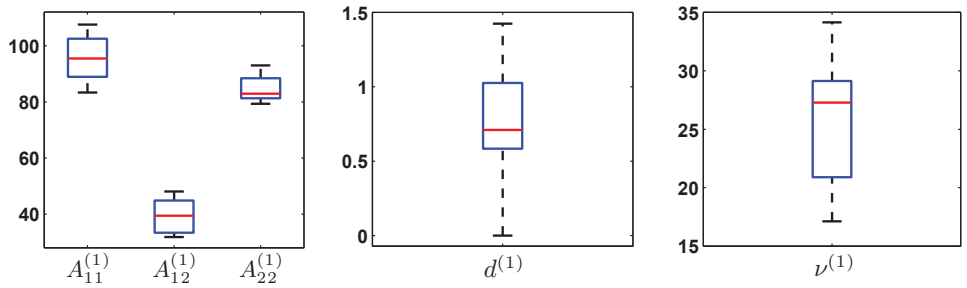

Figure 4.4: Box plots of the samples of parameters in chain 1.

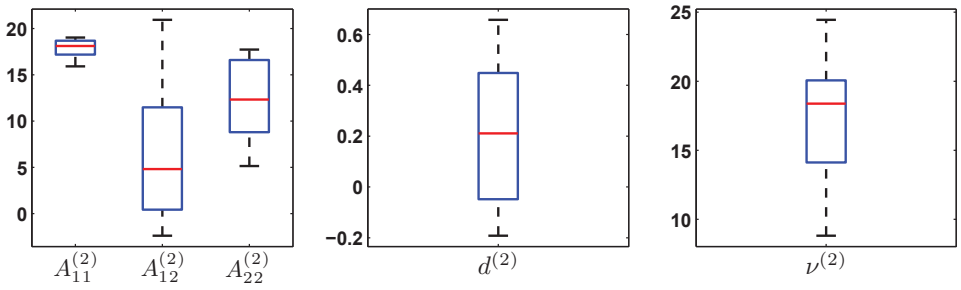

Figure 4.5: Box plots of the samples of parameters in chain 2.

Figure. 4.6 exhibits the estimation of covariance matrices with statistical property. The blue lines denote the true values, the red lines indicate the expectation of learnt values, with 95% highest posterior density interval (HPDI) in dotted lines. All the estimated values and true values fall into the 95% HPDI, which indicates the quality of estimation is quite good, also supported by the evaluation values in Table 4.1. Here, we conclude that the slight bias of parameters do not effect the learning quality of covariance matrices, which is also consistent with the conclusion in (Philipov & Glickman 2006).

In Figure 4.7, we present the estimated results from the coupled Wishart process and single Wishart process, compared with true values. We observe that the red lines (i.e. coupled results) mostly fit the ground truth (i.e. true values in blue) better than the uncoupled values (in green), indicating that our proposed coupled Wishart process simulates the truth more effectively. The above conclusion is also supported by Table 4.1, in which the bold refers to better results under the three evaluation metrics: MAPE, MSE and Det_error. In Table 4.1, we compute the MAPE and MSE of every element of the two-dimension covariance matrices from both chains, and Det_errors of each chain. All these evaluation values demonstrate that our proposed model outperforms the single Wishart process.

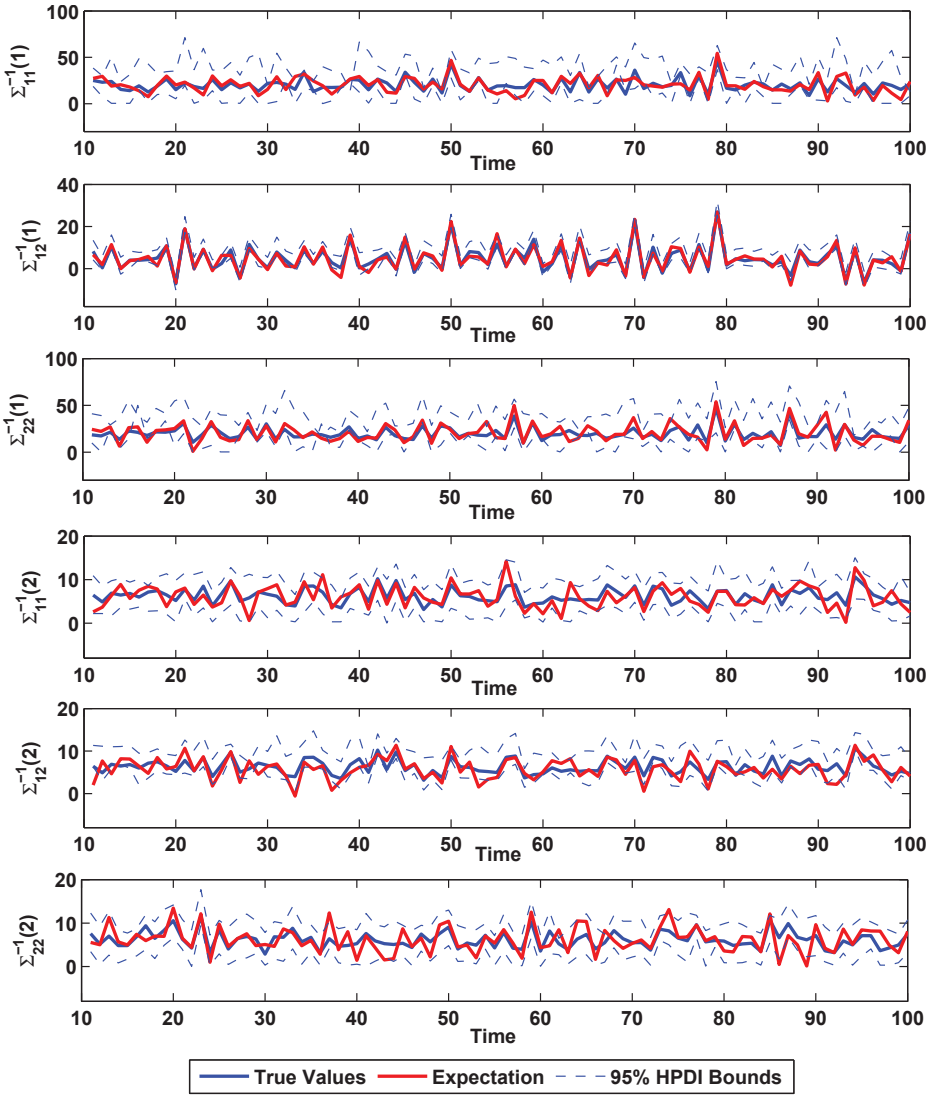

Figure 4.6: Learnt expectations and true values: the subscripts indicate the corresponding elements of matrices, (1) and (2) refer to the chains.

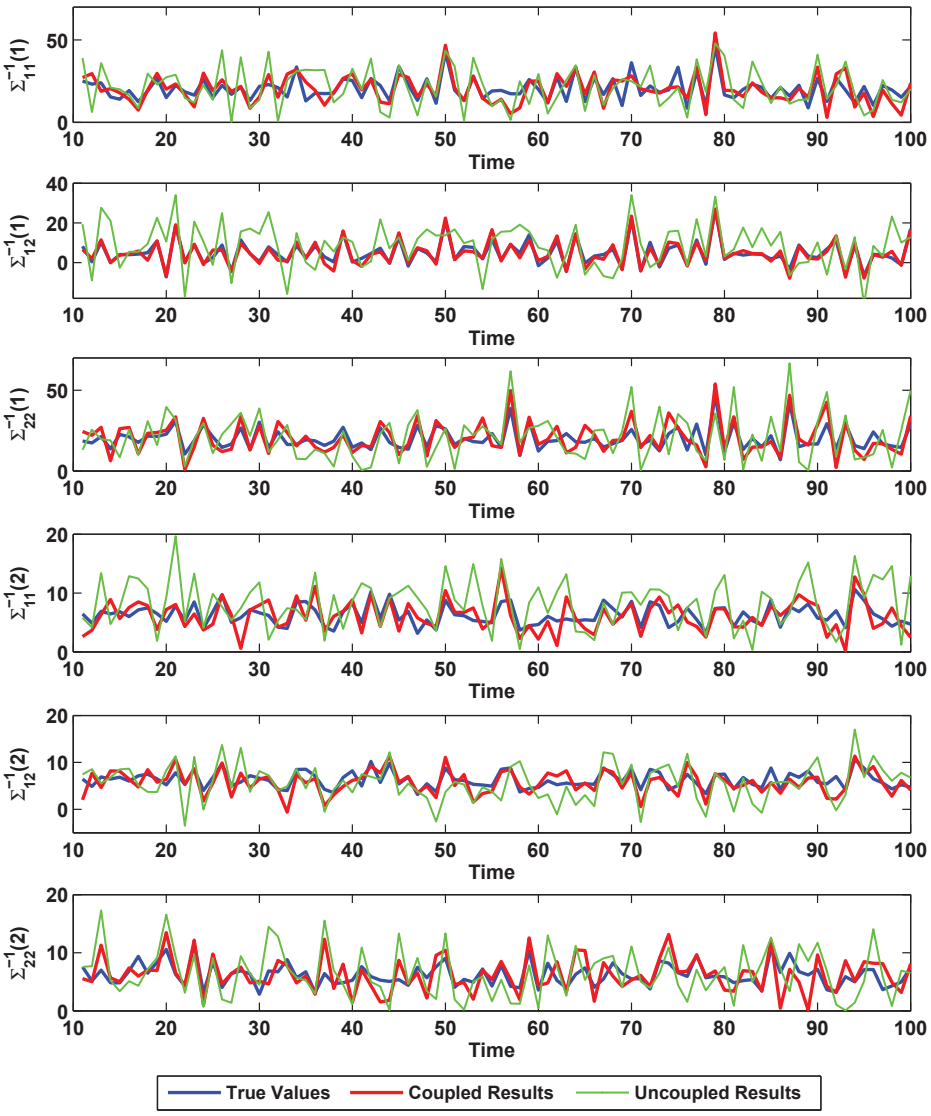

Figure 4.7: Coupled learnt v.s. uncoupled learnt: the subscripts indicate the corresponding elements of matrices, (1) and (2) refer to the chains.

Table 4.1: Evaluations of Estimated Variables for Synthetic Data

		MAPE	MSE	Det_error
Uncoupled	*Chain 1*			
	$\Sigma_{11}^{-1}(1)$	0.54	104.84	
	$\Sigma_{12}^{-1}(1)$	35.17	99.95	13.43
	$\Sigma_{22}^{-1}(1)$	0.64	129.45	
	Chain 2			
	$\Sigma_{11}^{-1}(2)$	0.62	18.18	
	$\Sigma_{12}^{-1}(2)$	14.80	31.29	9.11
	$\Sigma_{22}^{-1}(2)$	0.54	12.96	
Coupled	*Chain 1*			
	$\Sigma_{11}^{-1}(1)$	**0.27**	**32.61**	
	$\Sigma_{12}^{-1}(1)$	**1.04**	**3.36**	**13.22**
	$\Sigma_{22}^{-1}(1)$	**0.25**	**33.44**	
	Chain 2			
	$\Sigma_{11}^{-1}(2)$	**0.27**	**3.96**	
	$\Sigma_{12}^{-1}(2)$	**12.05**	**22.03**	**1.34**
	$\Sigma_{22}^{-1}(2)$	**0.29**	**5.02**	

Table 4.2: Descriptive Statistics of Real-life Data

	US		**Hong Kong**	
	Manu	Consm	Manu	Consm
Mean	1.174	1.0495	0.8506	0.6434
SD	6.6842	6.5721	5.3678	4.3881
Median	1.255	1.21	1.355	1.135
Minimum	-25.02	-24.43	-20.8	-15.16
Maximum	23.95	33.33	17.51	12.13

4.9.3 Real-life Data Analysis

For the experiments on real data, we collect the monthly returns of two industries: manufacturing (Manu for short) and consuming (Consm for short) from US stock market and Hong Kong stock market. The return series include 132 months, from January 2002 to December 2012. The descriptive statistics are provided in Table 4.2. Our research goal is to find the volatility of these two asset portfolios, considering the interaction between these markets under the circumstance of globalization.

Before implementation with our proposed model, we first prefilter the data with an AR(1) model (Jacquier et al. 2002), which makes the data consistent with the model assumption: $\mathbf{y}_t|\Sigma_t \sim \mathcal{N}_k(\mathbf{0}, \Sigma_t)$, or more specifically, $E(\mathbf{y}_t) = 0$. After such a preprocess, we implement the coupled Wishart process on these two data sets, under the linear coupling setup in Section 4.5. After 20000 times of sampling, we discard the first 2000 samples, and record the rest. By averaging them, we get the estimation of all parameters and hidden covariance matrices.

Just like the synthetic data, we concern the hidden covariance matrices most. After averaging the selected samples, we compare the three evaluation metrics from both the coupled Wishart process and single Wishart process, see Figure. 4.8, Figure. 4.9, and Figure. 4.10. Note that for the real-life data, as we do not know the true covariance matrix, a proxy proposed in Equation (4.9.4) is used in the calculation of MSE, while for the synthetic data, we refer to Equation (4.9.2).

We can see that under such three evaluation metrics, the results of coupled Wishart process are always better than those of the single Wishart process.

In summary, in the first stage of experiments, we simulate two sets of coupled time series and then conduct the experiments, results show that our proposed learning method can properly capture the coupling relationship and learn the hidden covariance matrices. Based on the conclusion obtained from stage 1, we continue to implement our proposed model on the real-life data. With comparisons, the coupled Wishart process is again demonstrated to

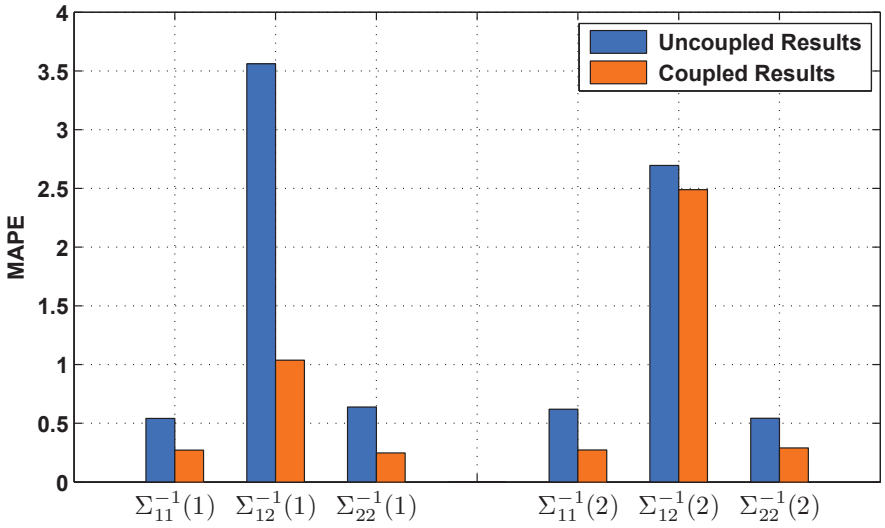

Figure 4.8: Comparison of MAPE results: the subscripts indicate the corresponding elements of matrices, (1) and (2) refer to the chains.

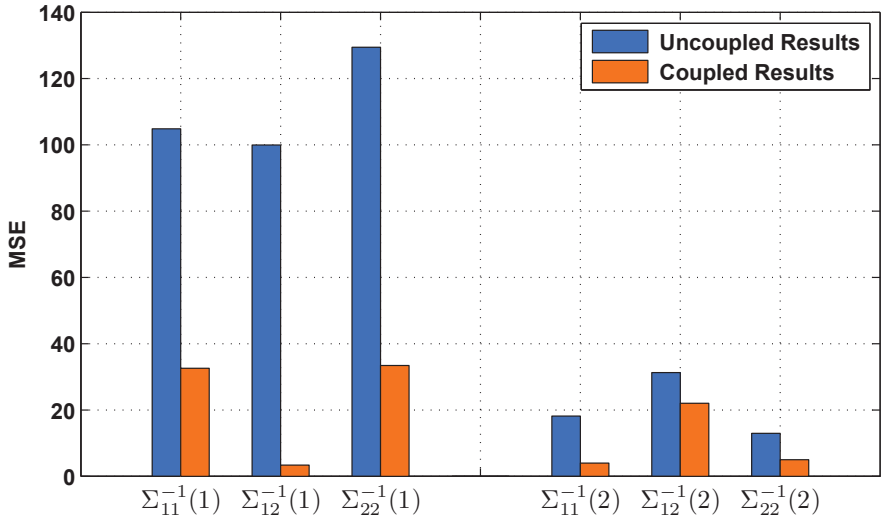

Figure 4.9: Comparison of MSE results: the subscripts indicate the corresponding elements of matrices, (1) and (2) refer to the chains.

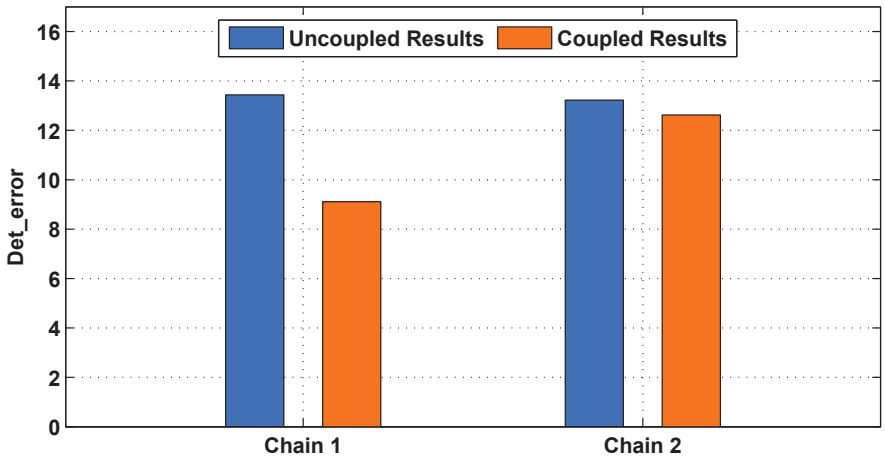

Figure 4.10: Comparison of Det_error results.

outperform the single Wishart process.

4.10 Summary

Wishart process has been proposed to model the volatility as one effective approach with great flexibility and capturing power. In this paper, we propose a coupled volatility analysis model to capture the interaction between different systems. Based on several synchronous Wishart processes, a linear coupled Wishart process is put forward. At time t, the covariance matrices exert their influences according to a learnt weight. After the model has been set up, based on the observations, we can deduce the posterior distribution of both parameters and latent variables. After that, we develop an algorithm to learn the parameters and latent covariance matrices, mainly with Gibbs sampling and and Metropolis-Hasting sampling. Experiments on synthetic data and real data show that our proposed coupled Wishart process performs better than the single Wishart process.

Although our model can effectively capture the coupling relationship when modeling the volatility of several synchronic markets, the complexity of models and simulation methods put high demand on the computation ability. When the dimension of observations or the number of Wishart pro-

cesses grows, we need to re-deduce the posterior distribution and the computation cost highly increases. There is great demand for the simplification of models and more efficient learning methods. In the future, we will improve our proposed model on such directions. In addition, this coupled model is currently used on financial data, we can also apply it to other fields, such as neurological science (Fox & West 2011).

Chapter 5

Heterogeneous Coupling Wishart Process

In last Chapter, a coupled model focusing on two parallel matrix sequences, as the target of these two sequences are both matrices, it can be categorized as homogenous coupling Wishart process. However in this Chapter, another aspect of coupled Wishart process is explored: heterogenous coupling Wishart process.

5.1 Background

We have presented the coupling between two Wishart processes, together with learning procedures and applications. The homogenous coupled Wishart processes focus on the interactions between time series of matrices. However, the real world is so complex that coupling relationships may exist in different kinds of data or data with various structures. Homogenous coupling models can not handle such data sets. Hence we introduce the heterogenous coupling Wishart process. Think of the situation that our target is to model the affect from other kinds data, but not parallel covariance matrices. In such cases, the time series are heterogenous, which means their interactions are not so obvious and we have to build the latent connections.

In this chapter, we aim to model the volatilities considering the heterogenous influences. First, a general framework is proposed for such heterogenous coupled time series. Then, a specific model is setup with details with the domain knowledge from finance. Next, we implement the models on synthetic and real-life data set.

5.2 Heterogenous Coupling Framework

In this part, a general coupling framework is proposed for parallel time series presented. In real life, there more or less exists coupling relationship between two or several time series.

These latent affects between variables or time series is of vital importance if we want to build a concise and reliable model.

Considering the volatility models, though the stochastic models have been introduced and developed for their flexibility and capturing power (Gourieroux et al. 2009), they lack the capability to capture outside influence. We have $\{\Sigma_t\}$ as the covariance matrix at each time point t, and $\{M_t\}$ as the outside influential source, M_t can be any kind of variable like matrix, vector or scalar.

A brief graphic model of such heterogenous coupling is illustrated in Figure 5.1.

5.3 Problem Statement

In such a heterogenous coupling model, the observations still follow a Gaussian distribution 5.3.1:

$$\mathbf{y_t}|\Sigma_t \sim \mathcal{N}_k(\mathbf{0}, \Sigma_t) \tag{5.3.1}$$

The coupling relationship lies in the latent part: at each time point t, the outside influential source M_t, will affect the covariance matrix transition. While at the mean time, covariance matrix also affect outside influential source, at a low level. Such coupling relationship can be formalized in below

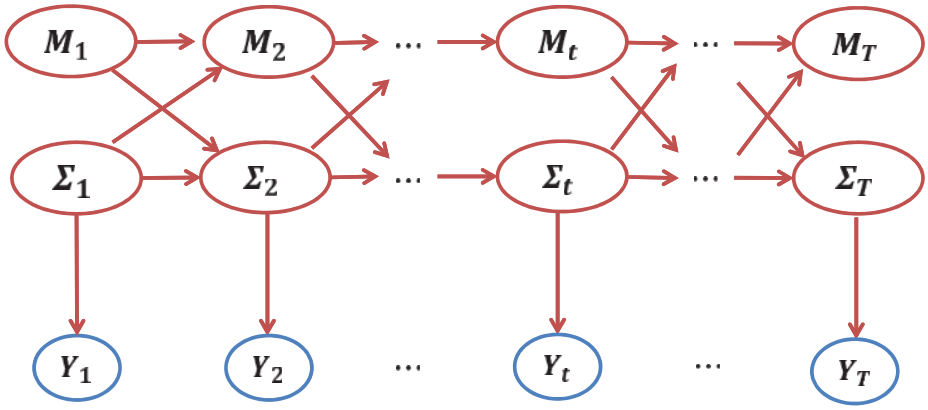

Figure 5.1: Graphic model for a heterogeneous coupled Wishart process.

equations:

$$\Sigma_{t+1} = f(\Sigma_t, M_t) \tag{5.3.2}$$

$$M_{t+1} = g(\Sigma_t, M_t) \tag{5.3.3}$$

To formulate a full heterogenous coupled Wishart process, there are two issues need to be addressed:

1. The motivation seems strong, however, find such a sequence of certain data is a challenging issue. Do the covariance matrices really correlated with the sequence of vectors: $\{M_t\}$ and $\{\Sigma_t\}$?

2. How do we transfer the information from a vector at time t to a matrix at time $t+1$, from M_t and Σ_{t+1}, and vice versa? As the our objects in front of us are vectors and covariance matrices respectively, the information transferring is a challenging issue.

Let's go back to the parameter settings of Wishart process in Chapter 3, the matrix parameter A is interpreted as a measure of intertemporal sensitivity while d denotes the overall influence strength from previous time to next. It is intuitive that we put the outside violence into A. To make such a influence distinct, we shall make specific setup in following sections.

5.4 Model Setups

As the limit of time, for this topic: heterogenous Wishart process, we can only develop a simplified version of it. Through a simplified version, we can still present the value of this model and the motivation behind it. And also, the learning procedures are given with details.

First, we set matrix A as a diagonal matrix $A = diag\{(a_1, \ldots, a_k)\}$. This setup can help model the influence from outside to the Wishart process.

$$\alpha_t = \frac{||M_t - M_{t+1}|| - \overline{\Delta M}}{\overline{\Delta M}} \tag{5.4.1}$$

and,

$$A_{t+1} = A_t(1 + \alpha_t) \tag{5.4.2}$$

As to the matrix transition,

$$\Sigma_t^{-1}|\nu, \Sigma_{t-1}^{-1} \sim \mathcal{W}_k(\nu, S_{t-1}), \tag{5.4.3}$$

where

$$S_{t-1} = \frac{1}{\nu}(A_{t-1}^{1/2})(\Sigma_{t-1}^{-1})^d(A_{t-1}^{1/2})'. \tag{5.4.4}$$

We can see from the setup that the fluctuation in M_t exert positive or negative influence on the transition of matrices. As M_t here is either scalar or vector, then $||\cdot||$ refers to the norm of the changes. $\overline{\Delta M}$ indicates the average fluctuation norm, when from time t to $t+1$, the fluctuation is below the average change, then the diagnose elements A_t will be less in next time step, if higher than the average change, then opposite. The diagonal matrix setting (Casarin & Sartore 2008) is to make such fluctuations on outside factors more direct to the matrices.

5.5 Learning Procedures

Based on the Bayesian structure of the above model, we also take the Gibbs sampling as the learning method. First, we have to derive the full likelihood

for such a model. Refer to the graphic model with parameters, easily, we have:

$$p(Y_{1:T}|M_{1:T},\theta,\Sigma_{1:T}^{-1}) = \prod_{t=1}^{T} p(\Sigma_t^{-1}|\Sigma_{t-1}^{-1},\theta)\cdot p(Y_t|\Sigma_t), \tag{5.5.1}$$

where $\theta = \{A^{-1}, d, \nu\}$.

Following Bayes rule:

$$posterior \propto prior \times likelihood, \tag{5.5.2}$$

the posterior of hidden variables and parameters can be derived.

First, we set all our priors. A^{-1} follows a Wishart distribution $W(Q_0,\gamma_0)$, d is in a uniform distribution at the interval [0,1], ν follows a noninformative uniform distribution, we have:

$$p(A^{-1}) \sim \mathcal{W}(Q_0,\gamma_0) \propto |A^{-1}|^{(\gamma_0-k-1)/2}\cdot|Q_0|^{\gamma_0/2} \tag{5.5.3}$$

Unlike the definition in Chapter 3, we have:

$$S_{t-1} = \frac{1}{\nu}(A_{t-1}^{1/2})(\Sigma_{t-1}^{-1})^d(A_{t-1}^{1/2})', \tag{5.5.4}$$

the A_{t-1} here fluctuates according to the variation of $\{M_t\}$.

Then the posterior can be written as:

$$
\begin{aligned}
p(\theta, \Sigma_{1:T}^{-1} | Y_{1:T}, M_{1:T}) &\propto prior(\theta) \cdot \prod_{t=1}^{T} p(\Sigma_t^{-1} | \Sigma_{t-1}^{-1}, \theta) \cdot p(Y_t | \Sigma_t) \\
&\propto p(A^{-1}) \cdot p(d) \cdot p(\nu) \cdot \prod_{t=1}^{T} p(\Sigma_t^{-1} | \Sigma_{t-1}^{-1}, A_{t-1}^{-1}, d, \nu) \cdot p(Y_t | \Sigma_t) \\
&\propto |Q_0|^{\gamma_0/2} \cdot |A^{-1}|^{(\gamma_0-k-1)/2} \cdot \prod_{t=1}^{T} W(\Sigma_t^{-1} | S_{t-1}, \nu) \cdot \prod_{t=1}^{T} \mathcal{N}(Y_t | \Sigma_t) \\
&\propto |A^{-1}|^{(\gamma_0-k-1)/2} \cdot \prod_{t=1}^{T} [\frac{1}{2^{\nu k/2} \prod_{i=1}^{k} \Gamma(\frac{\nu+1-i}{2})} \cdot \\
&|S_{t-1}|^{-\nu/2} \cdot |\Sigma_t^{-1}|^{\nu-k-1} \cdot exp\left(-\frac{1}{2} tr(S_{t-1}^{-1} \Sigma_t^{-1})\right)] \cdot \\
&\prod_{t=1}^{T} \left[|\Sigma_t^{-1}|^{1/2} \cdot exp(-\frac{1}{2} y_t' \Sigma_t^{-1} y_t) \right]
\end{aligned}
\tag{5.5.5}
$$

Now, we can deduce conditional probabilities of all the elements, by eliminating the irrelevant items.

For Σ_t^{-1}, when $t = 1 : T-1$, we only preserve the items contains Σ_t^{-1}, including S_t, from Equation (5.5.5). (From the second line to the third line, we use $Tr(ABC) = Tr(BCA) = Tr(CAB)$. And, for a scalar, $x = Tr(x)$.)

$$
\begin{aligned}
p(\Sigma_t^{-1} | \cdot) &\propto Wish(\Sigma_t^{-1} | \nu, S_{t-1}) \cdot Wish(\Sigma_{t+1}^{-1} | \nu, S_t) \cdot \mathcal{N}(Y_t | \Sigma_t) \cdot \\
&\propto |\Sigma_t^{-1}|^{(\nu-k-1)/2} \cdot exp(-\frac{1}{2} tr(S_{t-1}^{-1} \Sigma_t^{-1})) \cdot \\
&|S_t^{-1}|^{-\nu/2} exp(-\frac{1}{2} tr(S_t^{-1} \Sigma_{t+1}^{-1})) \cdot |\Sigma_t^{-1}|^{1/2} \cdot exp(-\frac{1}{2} y_t' \Sigma_t^{-1} y_t) \\
&\propto |\Sigma_t^{-1}|^{(\nu-k-1)/2} \cdot exp(-\frac{1}{2} tr((S_{t-1}^{-1} + y_t' y_t) \Sigma_t^{-1})) \cdot |\Sigma_t^{-1}|^{(1-\nu d)/2} \cdot \\
&exp(-\frac{1}{2} tr(S_t^{-1} \Sigma_{t+1}^{-1})) \\
&\propto Wish(\Sigma_t^{-1} | \nu, \widetilde{S}_{t-1}) \cdot |\Sigma_t^{-1}|^{(1-\nu d)/2} \cdot exp(-\frac{1}{2} tr(S_t^{-1} \Sigma_{t+1}^{-1})),
\end{aligned}
\tag{5.5.6}
$$

where $\widetilde{S}_{t-1} = (S_{t-1}^{-1} + y_t' y_t)^{-1}$.

When $t = T$,

$$\begin{aligned} p(\Sigma_T^{-1}|\cdot) \propto & Wish(\Sigma_T^{-1}|\nu, S_{T-1}^{-1}) \cdot \mathcal{N}(Y_T|\Sigma_T) \\ \propto & |\Sigma_T^{-1}|^{(\nu-k-1)/2} \cdot exp(-\frac{1}{2}tr(S_{T-1}^{-1}\Sigma_T^{-1}))\cdot \\ & |\Sigma_T^{-1}|^{M/2} \cdot exp(-\frac{1}{2}y_T'\Sigma_T^{-1}y_T) \\ \propto & Wish(\Sigma_t^{-1}|\nu+1, \widetilde{S}_{T-1}), \end{aligned} \tag{5.5.7}$$

where, $\widetilde{S}_{T-1} = (S_{T-1}^{-1} + y_t'y_t)^{-1}$.

For d, we find all the items with d are the S_t, so we only keep them in d's conditional probability.

$$\begin{aligned} p(d|\cdot) \propto & \prod_{t=1}^{T}\left[|S_{t-1}|^{-\nu/2}exp(-\frac{1}{2}tr(S_{t-1}^{-1}\Sigma_t^{-1}))\right] \\ \propto & \prod_{t=1}^{T}\left[|\Sigma_{t-1}^{-1}|^{-\nu d/2}exp(-\frac{1}{2}tr(S_{t-1}^{-1}\Sigma_t^{-1}))\right] \end{aligned} \tag{5.5.8}$$

For ν, it is a bit more complicated than d,

$$\begin{aligned} p(\nu|\cdot) \propto & p(\nu)\cdot\prod_{t=1}^{T}Wish(\Sigma_t|\nu, S_{t-1}) \\ \propto & \prod_{t=1}^{T}\left[\frac{1}{2^{\nu k/2}\prod_{i=1}^{k}\Gamma(\frac{\nu+1-i}{2})}\cdot|S_{t-1}|^{-\nu/2}\cdot|\Sigma_t^{-1}|^{(\nu-k-1)/2}\cdot\right. \\ & \left. exp\left(-\frac{1}{2}\sum_{t=1}^{T}tr(S_{t-1}^{-1}\Sigma_t^{-1})\right)\right] \\ \propto & (\frac{|\nu A^{-1}|^{\nu/2}}{2^{\nu k/2}\prod_{i=1}^{k}\Gamma(\frac{\nu+1-i}{2})})^T\prod_{t=1}^{T}|\Sigma_{t-1}^{-1}|^{-\nu d/2}\cdot|\Sigma_t^{-1}|^{(\nu-k-1)/2}\cdot \\ & exp\left(-\frac{1}{2}\sum_{t=1}^{T}tr(S_{t-1}^{-1}\Sigma_t^{-1})\right) \end{aligned} \tag{5.5.9}$$

For A^{-1}, we keep the prior and all the S_t,

$$
\begin{aligned}
p(A^{-1}|\cdot) &\propto p(A^{-1}) \cdot \prod_{t=1}^{T} Wish(\Sigma_t|\nu, S_{t-1}) \\
&\propto Wish(A^{-1}|\gamma_0, Q_0) \cdot \prod_{t=1}^{T} Wish(\Sigma_t^{-1}|\nu, S_{t-1}) \\
&\propto |A^{-1}|^{\frac{\gamma_0-k-1}{2}} \cdot exp(-\frac{1}{2} \cdot tr(Q_0^{-1}A^{-1}))\cdot \\
&\prod_{t=1}^{T} |S_{t-1}^{-1}|^{\nu/2} exp(-\frac{1}{2} \cdot tr(S_{t-1}^{-1}\Sigma_t^{-1})) \\
&\propto |A^{-1}|^{\frac{\gamma_0-k-1}{2}} \cdot exp(-\frac{1}{2} \cdot tr(Q_0^{-1}A^{-1}))\cdot \\
&\prod_{t=1}^{T} |A^{-1}|^{\nu/2} (1 + \sum_{1}^{t-1} \alpha_t) exp(-\frac{1}{2} \cdot tr(S_{t-1}^{-1}\Sigma_t^{-1}))
\end{aligned}
\tag{5.5.10}
$$

With above posteriors, the procedures of learning hidden states and parameters are as follows, according to Gibbs sampling.

1. Initialize $\theta = \{A, d, \nu\}$ and $\{\Sigma_t\}$.
2. Sample Σ_t from $\Sigma_t|\Sigma_{\setminus t}, A, d, \nu$ for $t = 1, \cdots, T$.
3. Sample A from $A|\{\Sigma_t\}, d, \nu$.
4. Sample d from $d|\{\Sigma_t\}, A, \nu$.
5. Go to 2 until enough iteration times.

Please note that for every step, the new samples must be updated in next step. For $t = 1, \cdots, T-1$, the posterior of Σ_t is not a standard distribution that we can directly sample from. We can adopt the Metropolis Hastings (MH) sampling method, and use a Wishart distribution as the proposal density. When $t = T$, we can sample Σ_T directly from a Wishart distribution with specific parameters. For ν and d, their posteriors are not common distributions either, we can get it by a discrediting method.

5.6 Synthetic Data Experiment

Based on the above setups, we can generate a set of matrices (volatilities). These matrices $\{\Sigma_t\}$ are generated based on Wishart process, and importing the influences from $\{M_t\}$.

First, we set our simulation model parameters: $k = 5$, $d = 0.3$, $\nu = 19$ and $A^{-1} = 0.0125 \cdot diag\{1, \cdots, 1\}$. Then Figure 5.2 and Figure 5.3 exhibits the time evolution of the observation and latent volatilities. In Figure 5.2, σ_{iit} denotes the volatility of the i_{th} variable.

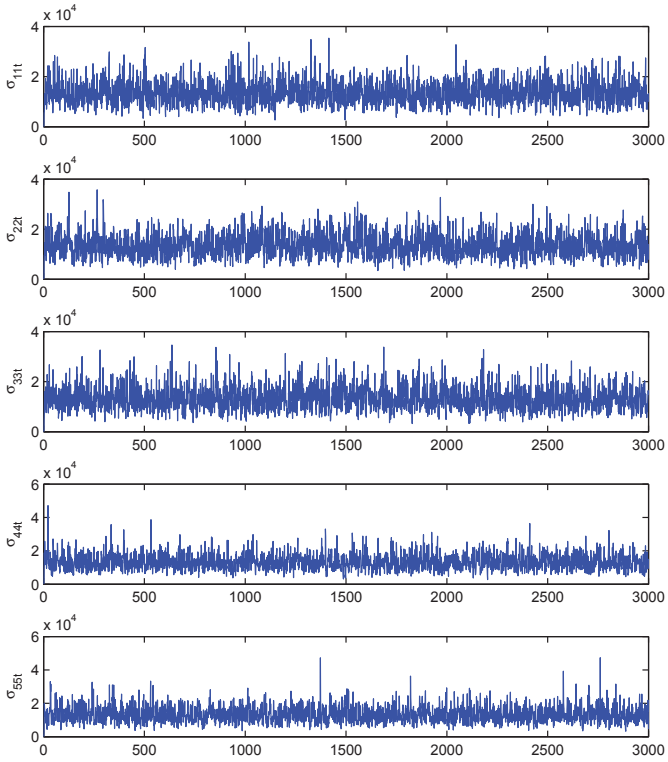

Figure 5.2: Simulated latent volatilities.

Now, we include the outside influence to the evolving of volatilities, ac-

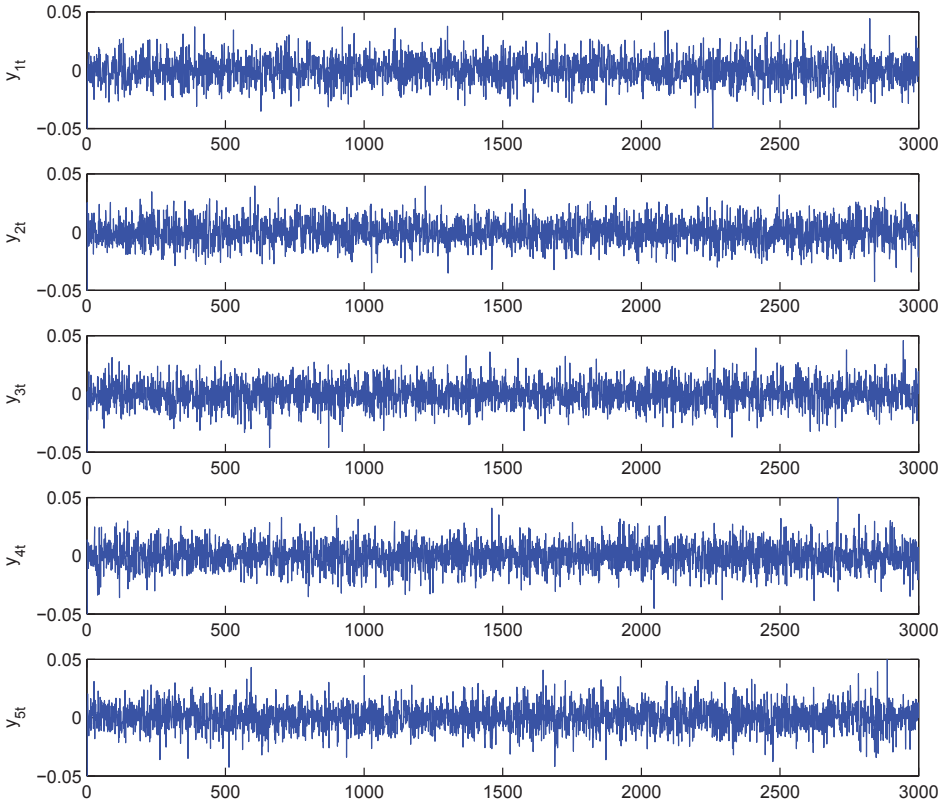

Figure 5.3: Simulated observations.

cording to our model setup in Section 5.4. The daily S&P 500 index from 7/08/2001 to 12/07/2013 (*Source: Yahoo Finance*) are imported as $\{M_t\}$. Then we want to see whether the fluctuation of index can have influence on target volatilities.

Figure 5.4 shows the history record of S&P 500 index, and we can see violent variations during the 2008 financial crisis. After including it into our model setup, the volatilities of the five variables shows the great impact from S&P data, as see in Figure 5.5.

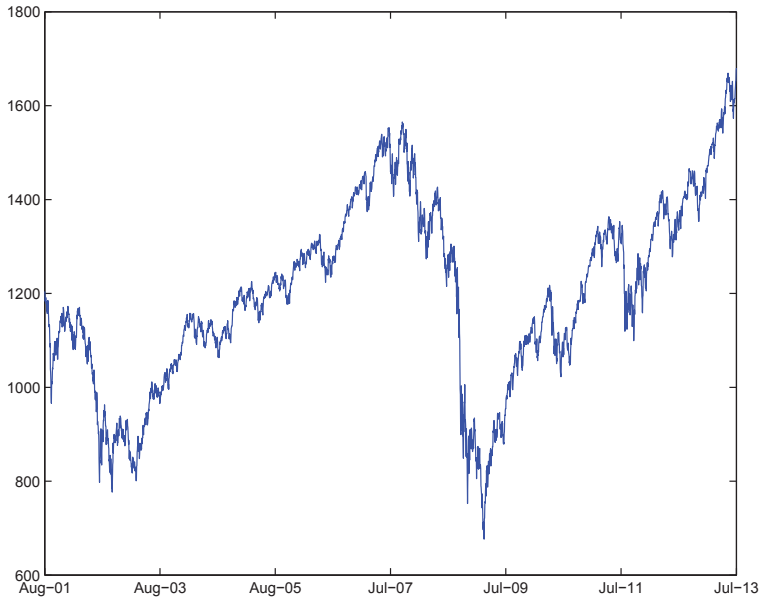

Figure 5.4: History daily S&P 500.

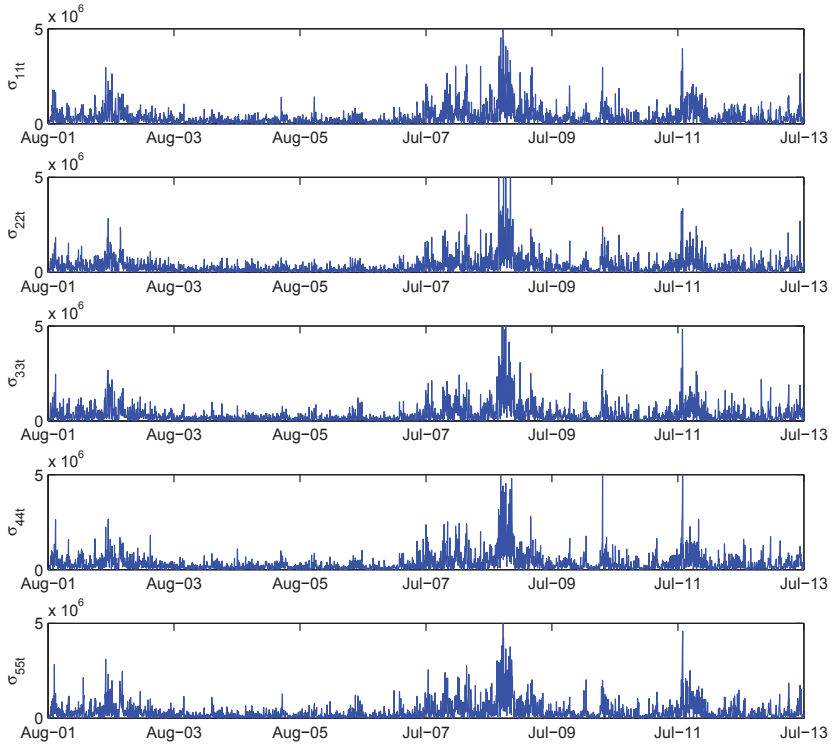

Figure 5.5: Simulated latent volatilities with outside influences.

Chapter 6

Conclusions

6.1 Summarization

The analysis of volatility is a hot topic for people from different backgrounds for its vital importance. However the complexity of a stock market and the everywhere relationship between same or different objects. Hence comes the need of coupling methods. In thesis, we make some explorations on introducing the coupling methods into volatility analysis via Wishart process: an autoregressive time series model. Two kinds of coupled Wishart processes are presented: homogenous and heterogenous.

The homogenous coupled Wishart process focus on chains in which same kinds of data interact with each other. Its structure makes it especially suitable for modeling the time-lag effect and the highly correlated international markets. After model construction and test on synthetic data, the homogenous coupled Wishart process has proved to model the volatility better when coupling relationship is considered than that not considered. All the evaluation indicators support this conclusion.

For the heterogenous coupled Wishart process, an interesting and simplified setup is made to demonstrate the model clearly. This model mainly depicts outside influences' effect on the evolving. As the outside indicator often differs from our target data, the coupling relationship is no longer ho-

mogenous but heterogenous. Based on the synthetic data, the heterogenous coupling is learnt with strong evidence.

6.2 Future Work

One challenge for coupling relationship learning is that how to assure the existence of coupling relationship. The previous models' independent assumption lacks the ability to capture explicit or latent coupling relationship in complex real world data sets. But the overfitting problem may arise when the coupling relationship is not so strong and result in inaccuracy in modeling. So how to construct a coupling model that can adjust to the relationship level is a challenging issue.

Another main challenge is how to connect the target objects, especially the heterogenous situation. Despite a general framework in Chapter 5 , a specific problem demands appropriate setups for an effective modeling. No plausible suggestion has been given in current literature. In other words, a lot of interesting can be done in this field.

Appendix A

Appendix: List of Publications

- **Zhong She**, Can Wang (2013). "Volatility Analysis via Coupled Wishart Process". The 2013 International Joint Conference on Neural Networks (**IJCNN 2013**), full paper accepted.
- **Zhong She**, Can Wang, Longbing Cao (2012), "CCE: A Coupled Framework of Clustering Ensembles". The 26th Conference on Artificial Intelligence (**AAAI 2012**), pp. 2455-2456, (*poster*).
- Can Wang, **Zhong She**, Longbing Cao (2013), "Coupled Attribute Analysis on Numerical Data". The 23rd International Joint Conference on Artificial Intelligence (**IJCAI 2013**), full paper accepted.
- Can Wang, **Zhong She**, Longbing Cao (2013), "Coupled Clustering Ensemble: Incorporating Coupling Relationships Both between Base Clusterings and Objects". The 29th IEEE International Conference on Data Engineering (**ICDE 2013**), pp. 374-385.
- Can Wang, Mingchun Wang, **Zhong She**, Longbing Cao (2012), "CD: A Coupled Discretization Algorithm". The 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (**PAKDD 2012**), pp. 407-418.

Appendix B

Appendix: List of Symbols

The following list is neither exhaustive nor exclusive, but may be helpful.

Σ_t	Covariance at time t.
ν	The degree of freedom in a Wishart process.
S_t	An intermediate variable for covariance matrix transition.
θ	The whole parameters.
Y_t	Observation at time t.
A	Parameter that stands for intertemporal sensitivity.
M_t	Outside indicators, can be vectors or scalars.
d	Parameter, a scalar.
ω	Weights.
S_t	Intermediate variable.
Q_0	One of the prior arguments.
$\mathcal{M}_+^k$	Space of real-valued symmetric and positive-definite matrices of dimension k.

Bibliography

Abdullah, S. & Zeng, X. (2010), Machine learning approach for crude oil price prediction with artificial neural networks-quantitative (ann-q) model, *in* 'Proceedings of the 2010 International Joint Conference on Neural Networks', pp. 1–8.

Ahn, H. & Wilmott, P. (2003), 'Stochastic volatility and mean-variance analysis', *Wilmott Magazine* p. 421.

Andrieu, C., De Freitas, N., Doucet, A. & Jordan, M. (2003), 'An introduction to MCMC for machine learning', *Machine Learning* **50**(1), 5–43.

Asai, M. & McAleer, M. (2009), 'The structure of dynamic correlations in multivariate stochastic volatility models', *Journal of Econometrics* **150**(2), 182–192.

Beichl, I. & Sullivan, F. (2000), 'The metropolis algorithm', *Computing in Science & Engineering* **2**(1), 65–69.

Bishop, C. M. & Nasrabadi, N. M. (2006), *Pattern Recognition and Machine Learning*, Vol. 1, Springer New York.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **31**(3), 307–327.

Boronowski, D. & Frangakis, A. (1998), Neural networks for nonlinear mutual prediction of coupled chaotic time series, *in* 'Proceedings of the 1998

International Joint Conference on Neural Networks', Vol. 3, pp. 1937–1942.

Brownlees, C., Engle, R. & Kelly, B. (n.d.), A practical guide to volatility forecasting through calm and storm, 2009, *in* 'URL http://ssrn.com/abstract'.

Bru, M. (1991), 'Wishart processes', *Journal of Theoretical Probability* **4**(4), 725–751.

Calders, T., Goethals, B. & Jaroszewicz, S. (2006), Mining rank-correlated sets of numerical attributes, *in* 'Proceedings of the 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining', ACM, pp. 96–105.

Campbell, J. & Hentschel, L. (1992), 'No news is good news: An asymmetric model of changing volatility in stock returns', *Journal of financial Economics* **31**(3), 281–318.

Cao, L. (2013), 'Non-iidness learning: an overview', *The Computer Journal* pp. 1–18.

Cao, L., Ou, Y. & Yu, P. S. (2012), 'Coupled behavior analysis with applications', *IEEE Transactions on Knowledge and Data Engineering* **24**, 1378–1392.

Casarin, R. & Sartore, D. (2008), 'Matrix-state particle filter for wishart stochastic volatility processes', *Working Papers* .

Christie, A. (1982), 'The stochastic behavior of common stock variances: Value, leverage and interest rate effects', *Journal of Financial Economics* **10**(4), 407–432.

de Mattos Neto, P., Ferreira, T. & Cavalcanti, G. (2011), A simulation environment for volatility analysis of developed and in development market-

s, *in* 'Proceedings of the 2011 International Joint Conference on Neural Networks', pp. 2450–2456.

Ding, Z., Granger, C. & Engle, R. (1993), 'A long memory property of stock market returns and a new model', *Journal of Empirical Finance* **1**(1), 83–106.

Engle, R. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation', *Econometrica: Journal of the Econometric Society* pp. 987–1007.

Engle, R. & Ng, V. (2012), 'Measuring and testing the impact of news on volatility', *The Journal of Finance* **48**(5), 1749–1778.

Förstner, W. & Moonen, B. (1999), 'A metric for covariance matrices', *Quo vadis geodesia* pp. 113–128.

Fox, E. & West, M. (2011), 'Autoregressive models for variance matrices: Stationary inverse wishart processes', *Arxiv preprint arXiv:1107.5239* .

Gallager, R. G. & Gallager, R. G. (1996), *Discrete stochastic processes*, Vol. 101, Kluwer Academic Publishers Boston.

Gan, G., Ma, C. & Wu, J. (2007), *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, Philadelphia.

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* pp. 457–472.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996), *Markov chain Monte Carlo in Practice*, Vol. 2, Chapman & Hall/CRC.

Glosten, L., Jagannathan, R. & Runkle, D. (2012), 'On the relation between the expected value and the volatility of the nominal excess return on stocks', *The Journal of Finance* **48**(5), 1779–1801.

Gourieroux, C., Jasiak, J. & Sufana, R. (2009), 'The wishart autoregressive process of multivariate stochastic volatility', *Journal of Econometrics* **150**(2), 167–181.

Harvey, A. C. & Shephard, N. (1996), 'Estimation of an asymmetric stochastic volatility model for asset returns', *Journal of Business & Economic Statistics* **14**(4), 429–434.

Harvey, A., Ruiz, E. & Shephard, N. (1994), 'Multivariate stochastic variance models', *The Review of Economic Studies* **61**(2), 247–264.

Jacquier, E., Polson, N. & Rossi, P. (2002), 'Bayesian analysis of stochastic volatility models', *Journal of Business & Economic Statistics* **20**(1), 69–87.

Jafari, G., Shirazi, A. H., Namaki, A. & Raei, R. (2011), 'Coupled time series analysis: Methods and applications', *Computing in Science & Engineering* pp. 84–89.

Javaheri, A. (2011), *Inside Volatility Arbitrage: the Secrets of Skewness*, Vol. 317, Wiley.

Kalogeratos, A. & Likas, A. (2012), 'Text document clustering using global term context vectors', *Knowledge and Information Systems* **31**(3), 455–474.

Kaytoue, M., Kuznetsov, S. & Napoli, A. (2011), Revisiting numerical pattern mining with formal concept analysis, *in* 'Proceedings of the 22nd International Joint Conference on Artificial Intelligence', pp. 1342–1347.

Kilin, F. (2011), 'Accelerating the calibration of stochastic volatility models', *The Journal of Derivatives* **18**(3), 7–16.

Li, D. & Liu, C. (2012), 'Extending attribute information for small data set classification', *IEEE Transactions on Knowledge and Data Engineering* **24**(3), 452–464.

Liu, J. S., Wong, W. H. & Kong, A. (1995), 'Covariance structure and convergence rate of the gibbs sampler with various scans', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 157–169.

Ma, W.-J. (2007), 'Coupled random walk approach to complex financial time series', *AAPPS Bulletin* **17**(2).

Murphy, K. P. (2012), *Machine Learning: a Probabilistic Perspective*, The MIT Press.

Nelson, D. (1991), 'Conditional heteroskedasticity in asset returns: A new approach', *Econometrica: Journal of the Econometric Society* pp. 347–370.

Philipov, A. & Glickman, M. (2006), 'Multivariate stochastic volatility via wishart processes', *Journal of Business and Economic Statistics* **24**(3), 313–328.

Plant, C. (2012), Dependency clustering across measurement scales, *in* 'Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining', pp. 361–369.

Qiu, J., Lu, J., Cao, J. & He, H. (2011), 'Tracking analysis for general linearly coupled dynamical systems', *Communications in Nonlinear Science and Numerical Simulation* **16**(4), 2072–2085.

Rinnergschwentner, W., Tappeiner, G. & Walde, J. (2011), Multivariate stochastic volatility via wishart processes - a continuation, Working Papers 2011-19, Faculty of Economics and Statistics, University of Innsbruck.

Robert, C. P. & Casella, G. (2004), *Monte Carlo Statistical Methods*, Vol. 319, Citeseer.

Roll, R. (1984), 'A simple implicit measure of the effective bid-ask spread in an efficient market', *The Journal of Finance* **39**(4), 1127–1139.

Saria, S., Duchi, A. & Koller, D. (2011), Discovering deformable motifs in continuous time series data, *in* 'Proceedings of the 22nd International Joint Conference on Artificial Intelligence', pp. 1465–1471.

Shephard, N. (2005), 'Stochastic volatility.', *Economics Group, Nuffield College, University of Oxford, Economics Papers* .

Song, Y. & Cao, L. (2012), Graph-based coupled behavior analysis: A case study on detecting collaborative manipulations in stock markets, *in* 'Neural Networks (IJCNN), The 2012 International Joint Conference on', IEEE, pp. 1–8.

Song, Y., Cao, L., Wu, X., Wei, G., Ye, W. & Ding, W. (2012), Coupled behavior analysis for capturing coupling relationships in group-based market manipulations, *in* 'Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 976–984.

Steinwart, I. & Christmann, A. (2009), Fast learning from non-iid observations, *in* 'Advances in Neural Information Processing Systems', pp. 1768–1776.

Topchy, A., Jain, A. & Punch, W. (2005), 'Clustering ensembles: models of consensus and weak partitions', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(12), 1866–1881.

Wang, C., Cao, L., Wang, M., Li, J., Wei, W. & Ou, Y. (2011), Coupled nominal similarity in unsupervised learning, *in* 'Proceedings of the 20th ACM International Conference on Information and Knowledge Management', ACM, pp. 973–978.

Wang, C., She, Z. & Cao, L. (2013*a*), Coupled attribute analysis on numerical data, *in* 'The 23rd International Joint Conference on Artificial Intelligence'.

Wang, C., She, Z. & Cao, L. (2013*b*), Coupled clustering ensemble: incorporating coupling relationships both between base clusterings and objects, *in* 'Proceedings of the 29th IEEE International Conference on Data Engineering', p. accepted.

Wilson, A. & Ghahramani, Z. (2011), Generalised wishart processes, *in* 'Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)', AUAI Press, Corvallis, Oregon, pp. 736–744.

Wishart, J. (1928), 'The generalised product moment distribution in samples from a normal multivariate population', *Biometrika* **20**(1/2), 32–52.

Zhong, S. & Ghosh, J. (2002), Hmms and coupled hmms for multi-channel eeg classification, *in* 'Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on', Vol. 2, IEEE, pp. 1154–1159.