ABSTRACT

LIU, PENG. A Stochastic Volatility Model and Inference for the Term Structure of Interest Rates. (Under the direction of Professor Peter Bloomfield).

This thesis builds a stochastic volatility model for the term structure of interest rates, which is also known as the dynamics of the yield curve. The main purpose of the model is to propose a parsimonious and plausible approach to capture some characteristics that conform to some empirical evidence and conventions. Eventually, the development reaches a class of multivariate stochastic volatility models, which is flexible, extensible, providing the existence of an inexpensive inference approach.

The thesis points out some inconsistency among conventions and practice. First, yield curves and their related curves are conventionally smooth. But in the literature these curves are modeled as random functions, and the co-movement of points on the curve are usually assumed to be governed by some covariance structures that do not generate smooth random curves. Second, it is commonly agreed that constant volatility is not a sound assumption, but stochastic volatilities have not been commonly considered in related studies.

Regarding the above problems, we propose a multiplicative factor stochastic volatility model, which has a relatively simple structure. Though it is apparently simple, the inference is not, because of the presence of stochastic volatilities. We first study the sequential-Monte-Carlo-based maximum likelihood approach, which extends the perspectives of Gaussian linear state-space modeling. We propose a systematic procedure that guides the inference based on this approach. In addition, we also propose a saddlepoint approximation approach, which integrates out states. Then the state propagates by an exact Gaussian approximation. The approximation works reasonably well for univariate models. Moreover, it works even better for the multivariate model that we propose, because we can enjoy the asymptotic property of the saddlepoint approximation.

A Stochastic Volatility Model and Inference for the Term Structure of Interest Rates

by

Peng Liu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2007

Approved By:

Dr. William H. Swallow

Dr. A. Ronald Gallant

Dr. Denis Pelletier

Dr. Peter Bloomfield
Chair of Advisory Committee

Dr. David Dickey

Biography

Peng Liu obtained his B.E. in Industrial Foreign Trade and B.E. in Computer Science from Northeastern University, People's Republic of China in 1998. He obtained his M.S. in Applied Probability and Statistics from Northern Illinois University, United States in 2003. He joined the Ph.D program in the Department of Statistics at North Carolina State University in the autumn of 2003.

Acknowledgements

I would like to thank my adviser, Professor Peter Bloomfield. I sincerely appreciate that he leads me into this fruitful area. It is also his rigorous critiques, plenteous knowledge and experience, and heuristic guidance that helps me pave the way for the research. Two year research experience with him brings me the confidence to overcome any fear about future research challenges.

I also would like to thank other four members in the advisory committee. Professor William Swallow had been supportive since I applied for the program, and recommended me for the long term SAS fellowship support. The remarkable research works by Professor David Dickey and Professor Ronald Gallant have as well inspired my researches on time series, optimization, and coding. Professor Denis Pelletier has shown me useful warnings, stop signs, and directions when I touched unfamiliar areas.

Special thanks to Mr. John Sall at the SAS Institute. He generously granted me a three-year fellowship. He also offered me two internship opportunities of working for the JMP group at SAS. I have gained precious experience while working with the group of the most talented statistical software developers.

I also would like to thank Professor Mohsen Pourahmadi at Northern Illinois University. He pointed out the glorious history of time series analysis, and encouraged me to work with the famous time series specialists at NC State. Meanwhile, his work brings an important corner stone to this thesis.

I must thank Dr. Christopher Gotwalt at SAS. He advises my work at SAS. His expertise in optimization and inference has provided importance guidance to the methodology development of this thesis.

Finally, I would like to thank my parents Zhonghua and Shoupei Liu, and my wife Bing. Their encouragement is the most invaluable support in my life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Overview

Term structure of interest rate modeling, started probably by Vasicek (1977), is an active topic at least in two disciplines: econometrics and financial mathematics. The major concern of an econometrician is to understand the dynamics of a collection of interest rates. The concern of a financial mathematician is to price derivatives whose underlying assets are interest rates. The topic may also be of interest to physicists, electrical engineers, and others, because theoretical or practical tools from those disciplines might be applicable. The topic is rather fascinating to statisticians, especially those who are working on time series, because it has roots in classical time series analysis, and it leads to current study on nonlinear and non-Gaussian systems.

The object to be modeled is a curve evolving over time, as illustrated in Figure (1.1). This curve itself is not naturally analytical. Only a finite number of points on the curve can be observed, so the curve is not observable completely. There are different ways of describing *the* curve. For example, the forward rate curve is a type of derived curve. Moreover, this curve and its dynamics have been modeled under two measures, physical and risk-neutral. In another word, modeling under the risk-neutral measure must neutralize the differences, which exist under the physical measure, among expected returns due to risks. Among empirical studies, we find that two stylized facts are interesting to us. The first stylized fact is that the first difference of a time series of a yield with a fixed time-to-

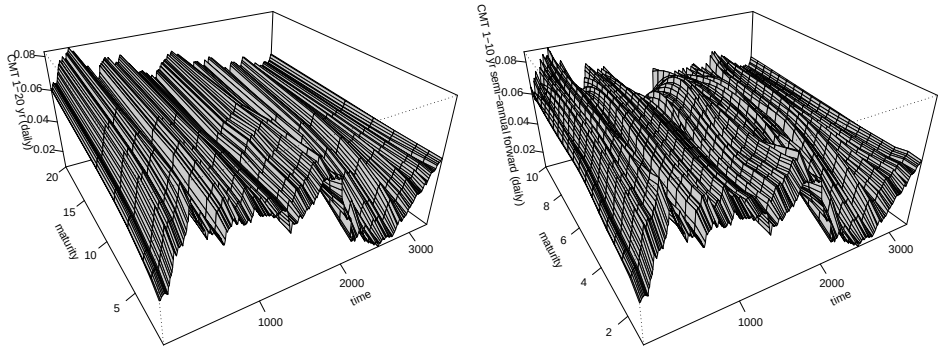

Figure 1.1: Dynamics of Static Yield Curves. The left plots the Constant Maturity Rates, sampled every ten observations, from October 1st, 1993, to May 10th, 2006. The right plots the corresponding semi-annual forward rates. Source: http://research.stlouisfed.org

maturity presents the so called volatility clustering phenomenon. The second is that those differenced series with different time-to-maturity present correlations among them. The correlation deceases if their time-to-maturity apart further from each other.

Unfortunately those stylized facts are sometimes ignored or misused in modeling and inferential practice. For example volatility clustering is sometimes ignored, such that complicated variance modeling can be avoided. Moreover, correlation structures are often picked due to their convenience in computation and presentation, which often does not reveal what the stylized fact attempts to tell. These are two mismatches that we have seen in the literature. Regarding such problems, we are interested in the following questions. First, what are the characteristics of the curve? Is the curve itself random, or static but observed with errors? Is the curve smooth? And so on. Second, by allowing stochastic volatilities along the physical time axis, how complicated is the process? There are many other important issues, such as mean reversion and unit roots. But it is unrealistic to do many things at the same time. We focus on those two issues, because by such, we may decompose a complex system in two orthogonal directions, and the remaining issues will fit naturally into the framework. This perspective may be introduced by Santa-Clara (2001) in

terms of solving a stochastic partial differential equation. The perspective is itself inherited from the string model perspective, which was initiated by Kennedy (1994; 1997).

We consider term structure modeling as a motivation, and propose a multivariate stochastic volatility model that is of our interest, which is partially inspired by the string model proposed by Kimmel (2004). We will ultimately focus on making inference, based on observed time series, about the unknown parameters in the model. Therefore, our model is a discrete time multivariate stochastic volatility model. The model is different from existing ones. And we will reveal the appealing properties of the model that will lead to a non-simulation-based inference.

The rest of the thesis consists of five chapters. Chapter 2 serves as a short presentation, about canonical forms of term structure models, and how they are used in practice. In this chapter, we bring up the discrete time multivariate stochastic volatility model that we will propose. Chapter 3 refers to some empirical results and practice, in order to point out the disagreement among conventions and practice, from the string model perspective. By such, we pursue a more specific form of the models that we propose in Chapter 2. More specifically, the covariance functions that have been specified in the string model literature usually violate the smoothness convention on static yield curves. In addition, effects from stochastic volatilities are often ignored during the estimation of the covariance functions. Those two issues motivate our proposal of two types of covariance structures, and incorporate stochastic volatilities. Two covariance (or correlation) structures are also introduced in the chapter. By such, we assume that increments of static yield curves are Gaussian, conditionally on the source of stochastic volatilities.

Chapter 4 starts with discussing the inference of the canonical stochastic volatility model. We follow the maximum likelihood approach and the likelihood is constructed from sequential Monte Carlo filtering. We will present the difficulty and computational expense of this approach. In addition, we will explain why we do not feel confident on making inference by following this approach. We propose a patch to this practice, which however will cause even more computational expenses.

The above approach does bring deep insight about how inference can be improved dramatically for some systems, e.g. the one that we propose. For our model, we still construct likelihood through filtering. However, we take advantage of the linear Gaussian structure of the state transition. The new approach is not simulation-based, which is different from all canonical inference that we have been aware of. Briefly, our proposal constructs the likelihood by integrating out the state variables, using the saddlepoint approximation, which also approximates the predicted state distribution as a Gaussian. We will present the details in Chapter 5.

Besides applications in quantitative finance, Chapter 5 develops a general purpose methodology for situations, in which state variables can be estimated consistently. When state variables can be estimated consistently, parameter inference may be conducted without expensive simulations. We have shown several rather simple models in Chapter 5, and the last chapter extends discussions and points out future work.

Chapter 2

Term Structure Modeling in Financial Mathematics

Term structure of interest rates in derivative pricing is a large topic. Though an overall review deserves a separate study, we provide a short presentation about what our model and inference will serve. Also we present our model later in this chapter.

2.1 Motivation

The seminal paper of Black and Scholes (1973) is one milestone in the complete story. Besides the elegant pricing formula, their work provides a probabilistic model that prescribes the dynamics of price movements, based on which a *fair* price of an option can be derived. Mathematical forms of those dynamics are known as Itô processes. Since then, continuous time model framework (Merton 1990) has become the foundation of modern quantitative finance. And prescribing the dynamics of underlying assets is one major task for pricing their derivatives.

The initial development of the above work started with such a problem. Suppose an investor has been given a right, at time t, to purchase an asset at time T, at a fixed price K whatever the actual price P_T will be at T. The pricing question is what the fair cost is

at t for such a right. This right is known as a European call option in finance. The option is known as a derivative from the asset. The famous Black-Scholes formula gave an explicit solution to such a problem in early seventies, and brought quantitative finance into a new era. While stocks are assets in the stock market, there is a much larger market that is less familiar to the public. It is known as the fixed-income security market. In this market, the trading instruments are related to interest rates. Derivatives from interest rate instruments enlarge the notional value of the market rapidly. In some sense, similar pricing problems in the fixed income security market are even more important.

While pricing a derivative, it is crucial to know the dynamics of the underlying asset. For example in Black-Scholes, the stock price is assumed to have a geometric Brownian motion with a constant volatility. The stochastic differential is of the form

$$dS_t = rS_t + \sigma S_t dW_t, \tag{2.1}$$

where S_t stands for the stock price at time t, r for the risk-free short rate, and W_t for a standard Wiener process. The corresponding European call option price is

$$\begin{aligned} C(S,T) &= S \times \Phi\left(\frac{\ln(S/K) + (r+\sigma^2/2)T}{\sigma\sqrt{T}}\right) \\ &\quad -K \times e^{-rT} \times \Phi\left(\frac{\ln(S/K) + (r+\sigma^2/2)T}{\sigma\sqrt{T}} - \sigma\sqrt{T}\right), \end{aligned}$$

where $\Phi(\cdot)$ stands for the standard Gaussian cumulative density function (CDF), and S is the current stock price.

From the above formula, we can see volatility σ and interest rate r play important roles in pricing, where both were assumed constant. It has been recognized that those assumptions are not realistic in some situations, such that the geometric Brownian motion is not a good approximation. A well known empirical evidence to statisticians or econometricians is the differenced log series, or the return series, does not have apparently constant variance, or volatility in terms of financial mathematics. In addition, the pattern has be summarized as *volatility clustering*.

More general assumptions are that the volatility and interest rate(s) are stochastic processes. The dynamics of interest rates are stochastic differential by themselves, with their

own volatilities, which might be stochastic as well. The dynamics of interest rates are then used for pricing their own derivatives. Pricing under stochastic volatilities is an emerging research area; see Fouque et al. (2000). In addition, interest rates are observed as vector time series. Therefore, we are inevitably interested in multivariate stochastic volatility models. For pricing, the importance of inference on stochastic volatility models is obvious. Later, inference will be developed for more general problems, and possible applications can be largely extended. We first briefly present the development of modeling term structure of interest rates.

2.2 Term Structure Models

The area started to benefit from natural science after Vasicek (1977), which came after Black-Scholes. More specifically, in the Vasicek model, the short rate (r_t) – instantaneous borrowing and lending rate – was prescribed by an Ornstein-Uhlenbeck process:

$$dr_t = (\mu - \alpha r_t)dt + \sigma dW_t,$$

where μ, α, and σ were constant, and W_t was a standard Wiener process. Over the next decade, many researches were on adding dynamics of state variables to the Vasicek model, which drove r_t, μ, and σ. In many situations, the static yield curve was then constructed as a deterministic function of those state variables.

The Vasicek model and its successors, which are known as short rate models, are under the umbrella of continuous time modeling. According to the full treatment in Merton (1990), we can see that continuous time models are not simply extensions or generalizations to their discrete counterparts, because price is always generated discretely over time. To our understanding, such models are continuous time approximations to discrete time stochastic processes that are observable, such that algebraic simplicity in calculation can be used in analysis. However, inference about unknown parameters is not simplified. Inference of univariate continuous time models with deterministic drift and volatility functions is a branch of active research. Meanwhile, models with stochastic drift and volatility are of

great interest. The inference about this stem is known as stochastic volatility models (SVM), initiated probably by Taylor (1980; 1982) to model the volatility clustering phenomenon. In parallel, autoregressive conditional heteroscedasticity (ARCH) and generalized ARCH (GARCH) belong to the other approach.

After Heath et al. (1992) proposed their HJM framework of modeling the characteristics of the entire instantaneous forward rate curve f_t, the difficulty of specifying the functional that links the short rate and yield curve is circumvented. The dynamics are expressed in the following stochastic differential:

$$df_{t,s} = \mu_{t,s}dt + \vec{\sigma}_{t,s}^T d\vec{W}_t, \text{ where } t \leq s < T^*, \tag{2.2}$$

where T^* is a finite constant, $\vec{W}_t$ is a standard d-dimensional Brownian motion, $\mu_{t,s}$ and $\vec{\sigma}_{t,s}$ are drift and diffusion functions in respective scalar and vector forms, the superscript T for $\vec{\sigma}_{t,s}$ stands for vector transpose, and $\vec{\sigma}_{t,s}^T d\vec{W}_t$ is an inner product of two vectors. This form prescribes infinitely many processes for $s \in (t, T^*)$, which are driven by d-dimensional innovations $d\vec{W}_t$. The co-movement between two forward rates at s_1 and s_2 can be described by the covariance $\text{Cov}(df_{t,s_1}, df_{t,s_2}) = \vec{\sigma}_{t,s_1}^T \vec{\sigma}_{t,s_2} dt$. The crucial benefit of the framework is that it suffices to specify $\vec{\sigma}_{t,s}$ to determine the stochastic differential (2.2). To statisticians, the task is similar to specifying a covariance function for the driving process $\vec{\sigma}_{t,s}^T d\vec{W}_t$, $s \in (t, T^*)$. For a finite collection of forward rates, $f_{t,s_1}, \ldots, f_{t,s_k}$, the form (2.2) is represented by the corresponding discrete form:

$$\begin{aligned}
df_{t,s_1} &= \mu_{t,s_1}dt + \sum_{j=1}^{d} \sigma_{t,s_1,j} dW_{t,j}, \text{ where } t \leq s_1 < T^*, && (2.3)\\
&\vdots && (2.4)\\
df_{t,s_i} &= \mu_{t,s_i}dt + \sum_{j=1}^{d} \sigma_{t,s_i,j} dW_{t,j}, \text{ where } t \leq s_i < T^*, && (2.5)\\
&\vdots && (2.6)\\
df_{t,s_k} &= \mu_{t,s_k}dt + \sum_{j=1}^{d} \sigma_{t,s_k,j} dW_{t,j}, \text{ where } t \leq s_k < T^*. && (2.7)
\end{aligned}$$

Therefore, the covariance matrix of $df_{t,s_1}, \ldots, df_{t,s_k}$ must be singular, if $k > d$. Specifying

a covariance in this way is very similar to what the factor analysis does, i.e. the matrix $[\sigma_{t,s_i,j}]$ is similar to a factor loading matrix. Thus, the singularity issue is easy to follow.

Several years after that, using the similar strategy, market models were popularized, and took observed rates as direct inputs; see Brace et al. (1997) and Rutkowski (1998). Such models may still be considered to be in the HJM family. Along this line, most of the efforts are put upon specifying either $\vec{\sigma}_{t,s}$ or $d\vec{W}_t$, such that the covariance structure of their product is parsimonious, but still rich enough to represent a large class of structures. Such models include factor models, string models (or field models), and so on.

The term structure plays an important role in fixed-income asset-related pricing problems, which are related to a much larger market than the equity market. There have been several comprehensive treatments on existing models; see James and Webber (2000), Brigo and Mercurio (2001), and Musiela and Rutkowski (1997) among other. Sundaresan (2000) surveys a broad range of research areas, including term structure modeling.

We have reviewed the literature and textbooks in the financial mathematics discipline. We can see two perspectives on making inference. One is under the so-called risk-neutral measure. The other is under the physical measure. The first perspective is taken by financial mathematicians for risk-neutral pricing. The goal is to guarantee that, given the parameters, the model is able to recover market prices, by using a conventional formula such as the Black-Scholes. Such approaches are known as calibrations. The second perspective is taken by econometricians and probably statisticians. The inference then relies on historical time series. Parameters obtained by two approaches are often different from each other. A theoretical offset is known as the risk premium, which is the drift difference by change of measures. In addition, to our understanding, disagreement due to model misspecification has not been addressed. Before any uncertainties besides the risk premium can be excluded, two approaches are complements to each other. We still estimate parameters from time series, in discrete time, in particular.

In addition, by allowing volatilities to be stochastic, inference about the forward rate models calls for multivariate versions of either SVM or GARCH. There are existing

models. However, we will propose a suitable one for the term structure of interest rates, which can be used in more general situations as well. Our model is inspired by the string model perspective.

2.3 String Model Perspective

We have mentioned that, chronologically, short rate models, forward rate models, market models were developed during the past two decades. Our proposal is related to models that are categorized by driving processes. In the literature, the driving processes are either processes of state variables, which have economic interpretations, or purely Wiener processes. We will focus on the second category. Among the category, some are known as finite factor models, e.g. $\vec{W}_t$ in Equation (2.2) is a standard d-dimensional Wiener processes, which is basically a collection of independent standard Wiener processes.

The random field models (see Kennedy 1994, among others) or the string models, which inspire our study, extend the finite factor models in a way such that the number of factors is infinitely many, while parsimony is still preserved. In string models, $\vec{W}_t$ is replaced by infinitely many correlated Wiener processes $Z_{t,s}$, for $s \in (t, T^*)$. As a whole piece, $Z_{t,s}$ is treated as a *field*, or a *string*, indexed by s. In the terminology of analysis of covariance functions, this object is known as a continuous one-dimensional Gaussian random function, or Gaussian process. We adopt a form similar to the one from Goldstein (2000), in the following, to facilitate the discussion:

$$df_{t,s} = \mu_{t,s}dt + \sigma_{t,s}dZ_{t,s}, \text{ where } t < s < T^*, \tag{2.8}$$

where $\mu_{t,s}$ and $\sigma_{t,s}$ are both scalar functions, $\text{Var}(dZ_{t,s}) = dt$, and $\text{Cov}(dZ_{t,s_1}, dZ_{t,s_2}) = \rho(s_1, s_2)dt$, where $s_1, s_2 \in (t, T^*)$. Therefore, for a finite collection of k forward rates, the

corresponding form for the stochastic differential (2.8) is as follows:

$$\begin{aligned} df_{t,s_1} &= \mu_{t,s_1}dt + \sigma_{t,s_1}dZ_{t,s_1}, \text{ where } t < s < T^*, \\ \ldots &\quad \ldots \quad \ldots \\ df_{t,s_k} &= \mu_{t,s_k}dt + \sigma_{t,s_k}dZ_{t,s_k}, \text{ where } t < s < T^*. \end{aligned}$$

The covariance of $df_{t,s_1}, \ldots, df_{t,s_k}$ is DRD, where D is a diagonal matrix with $D^2_{i,i} = \sigma^2_{t,s_i}dt$, and R is a correlation matrix with $R_{i,j} = \rho(s_i, s_j)$.

By properly prespecifying the correlation structure, any number of Wiener processes can be used as driving processes. Thus, the task of covariance modeling through a factor analysis style approach in (2.2) has been changed into working with covariance directly in (2.8). To financial mathematicians, it is important that the argument in the HJM framework remains the same in string models, that is it suffices to specify the covariance to prescribe the stochastic differential. Beyond seeking a more parsimonious and flexible parameterization for the driving processes, this class of models has been blended with the market model framework, which leads to simpler valuation practice; see Longstaff and Schwartz (2001). To statisticians, the task remains the same – covariance modeling, but more attractive. In practice, to statisticians, the factor analysis approach is sometimes harder, in terms of both inference and interpretation. The string model approach opens the door to many other directions.

A brief history of the development along this line is as follows. Kennedy (1994) pioneered the development of this line. The development is made more clear by Kennedy (1997), and calibration was first implemented in Pang (1999). Goldstein (2000) extends the driving process to non-Gaussian random field. Santa-Clara (2001) connects the perspective to stochastic partial differential equations. Kimmel (2004) adds conditional volatility. Albeverio et al. (2004) works on Lévy field assumptions. Baaquie (2001; 2002), and Baaquie and Srikant (2004) connect the problem to quantum field theory. Bester (2004) makes some empirical comparisons between random field model approach and affine model approach. Gall et al. (2004; 2006) work on discrete versions.

Practical advantages of string models have been discussed by their inventors. These models might be promoted because they overcome the HJM drawback that the number of factors must be much smaller than the number of interest rates. In practice, the inner product of the $k \times d$ matrix $[\sigma_{t,s_i,j}]$ in equations (2.3) through (2.7) must be calibrated to an $k \times k$ covariance matrix, which is implied from market prices. In order to have a unique solution, the condition $k + 1 - 2d \geq 0$ must be satisfied. While $k + 1 - 2d >> 0$, parsimony is achieved, which is similar to the factor analysis approach in statistics literature. String models do not have such a restriction; this approach still involves a covariance matrix, whose entries are modeled simultaneously, which are sometimes hardly modeled parsimoniously. To statisticians, parsimony may be achieved by structured parameterizations, among which we will choose two particular ones for some special cases in this research.

Some string model approaches decompose the term structure in two directions, because some authors suggest to solve a stochastic partial differential equation (SPDE), rather than a system of ordinary stochastic differential equations; see Santa-Clara (2001). Therefore, instead of considering term structure as a functional process or vector process with one index t, it has been considered as a scalar process with two indices. The first index is commonly agreed, which is the physical time horizon, denoted by t. The second index used in literature is either maturity date or maturity tenor. It is very important to distinguish them in financial mathematics literature, because the no-arbitrage condition is different under two choices. In this research, as we will address later, our interest is a covariance matrix of fixed dimension that evolves over physical time horizon. Therefore, the choice of the second index is not crucial. We ignore possible meaning of the index, and denote it by s, along which a second stochastic process has been specified in the SPDE approach.

It has been tried, in the literature, to specify a Gaussian process along the second index. It is well known to statisticians that it is equivalent to specify a valid covariance function to define a Gaussian process, and it has been mostly discussed in string model literature. In the next chapter, we point out that some covariance structures proposed for

$\sigma_{t,s}dZ_{t,s}$ do not match the convention that the yield curve is smooth. We then present two structures that generate either non-smooth or smooth random curves, and discuss their properties for inference. The discussion is based on the modified Cholesky decomposition proposed by Pourahmadi (1999).

Even more fundamental, the curve along the second index must be continuous, so that pure arbitrage does not exist. The proof is not hard. First, a zero-coupon bond curve must be continuous, otherwise a pair of instantaneous short and long position at the discontinuity point will create a riskless profit. Second, continuous transformation of a continuous curve preserves the continuity. Fortunately, representations for the yield curve that we are aware of are all continuous transformations of the zero-coupon bond curve. During our model construction, if we assume there are latent volatility processes that drive the curve, the continuity cannot be broken. This concern finally leads to our multiplicative factor model. In the next section, we briefly present the development of discrete time stochastic volatility models.

2.4 Discrete Time Stochastic Volatilities Models

There are two promising approaches in the literature, regarding making inference about volatility processes. The first is the autoregressive and conditional heteroscedasticity (ARCH), started by Engle (1982). The second one is the stochastic volatility model (SVM), started by Taylor (1980). The canonical form of ARCH is:

$$Y_t = H_t^{1/2}\epsilon_t, \tag{2.9}$$

$$H_t = \alpha_0 + \alpha_1 Y_{t-1}^2 \tag{2.10}$$

where ϵ_t is usually assumed to be a Guassian white noise with $\mathrm{Var}(\epsilon_t) = 1$. Generalized ARCH (GARCH), by Bollerslev (1986), has the following minimal structure.

$$Y_t = H_t^{1/2}\epsilon_t, \tag{2.11}$$

$$H_t = \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 H_{t-1}, \tag{2.12}$$

which is known as GARCH(1,1). The canonical form of SVM is:

$$Y_t = \beta \exp(X_t/2)\epsilon_t \tag{2.13}$$

$$X_t = \phi X_{t-1} + \sigma u_t, \tag{2.14}$$

where $\beta > 0$, u_t and ϵ_t are usually assumed to be orthogonal Gaussian white noises with $\mathrm{Var}(u_t) = \mathrm{Var}(\epsilon_t) = 1$. By letting $H_t = \exp(X_t + \log \beta^2)$ and putting back into the canonical SVM, we get

$$Y_t = H_t^{1/2}\epsilon_t \tag{2.15}$$

$$\log(H_t) = (1-\phi)\log\beta^2 + \phi\log(H_{t-1}) + \sigma u_t. \tag{2.16}$$

Thus, we can tell the difference between two approaches, which is that SVM has innovations in both equations, and the second equation does not have a positivity constraint on the designated process. It is SVM that is more attractive to us, because it is a special case of non-linear state-space models. We usually call (2.13) the observation equation and (2.14) the state equation. To financial mathematicians, Equation (2.14) in SVM corresponds to a suitable stochastic differential, which may be desirable under some circumstances. Further comparison is beyond the scope of this research.

The statistical inference of ARCH or GARCH will be straight forward, by maximizing the likelihood. The GARCH(1,1) likelihood is:

$$\begin{aligned} p(Y_1, \ldots, Y_n) &= p(Y_1)\prod_{t=2}^{n} p(Y_t|Y_{t-1}) \\ &= p_N(Y_1; 0, H_1)\prod_{t=2}^{n} p_N(Y_t; 0, \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 H_{t-1}), \end{aligned}$$

where $p_N(\ \cdot\ ; \mu, \sigma^2)$ stands for a Gaussian density with mean μ and variance σ^2. H_1 is either provided or not, but it does not bring critical difficulty to inference. Meanwhile, SVM does not have such a convenience. We postpone our discussion on SVM inference to later chapters.

2.4.1 Multivariate Extensions of Stochastic Volatility Models

Though inference is postponed, in order to facilitate the discussion in the following chapters, we first present the form of the model that we are interested in. Because we are handling multiple time series, we are interested in multivariate extensions of SVM. Suppose now the observed time series $\vec{Y}_t$ is of dimension m, and the unobservable process $\vec{X}_t$ is of dimension k. In addition, we denote the components of $\vec{Y}_t$ by $\{Y_{1t}, \ldots, Y_{mt}\}$.

Harvey et al. (1994) propose the following model, denoted by **M1**, in which $m = k$.

$$\vec{Y}_t = \text{diag}\{\exp(\vec{X}_t/2)\}\vec{\epsilon}_t, \tag{2.17}$$

$$\vec{X}_t = \vec{\mu} + \Phi\vec{X}_{t-1} + \vec{\eta}_t, \tag{2.18}$$

where $\vec{\mu}$ is a mean vector, Φ is a $m \times m$ matrix, $\vec{\epsilon}_t \sim N_m(0, \Sigma_\epsilon)$, and $\vec{\eta}_t \sim N_m(0, \Sigma_\eta)$. A concern of this model is that instantaneous correlation of $\vec{Y}_t$ is constant. In addition, from an inferential point of view about the number of unknown quantities, neither is this system very much different from m independent univariate SVM, because it has as many unobservable processes as m independent SVMs. The major difference is that conditional means, for example $E(Y_{1t}|Y_{2t}, \ldots, Y_{mt}, \vec{X}_t)$, are not necessarily zeros.

Pitt and Shephard (1999b) propose the following factor SVM, denoted by **M2**, with $m < k = k_1 + m$, in which $(\vec{F}_t^T, \vec{X}_t^T)^T$ is the unobservable process:

$$\vec{Y}_t = B \cdot \text{dot}\{\exp(\vec{F}_t/2)\vec{\epsilon}_{t1}\} + \text{diag}\{\exp(\vec{X}_t/2)\}\vec{\epsilon}_{t2}, \tag{2.19}$$

$$\vec{F}_t = \Phi_1\vec{F}_{t-1} + \vec{\eta}_{t1}, \tag{2.20}$$

$$\vec{X}_t = \vec{\mu} + \Phi_2\vec{X}_{t-1} + \vec{\eta}_{t2}, \tag{2.21}$$

where $\text{dot}\{\cdot\}$ is a dot product operator, $\vec{Y}_t$ is m-dimensional, $\vec{F}_t$ is k_1-dimensional, $\vec{X}_t$ is m-dimensional, $\vec{\epsilon}_{t1} \sim N_{k_1}(0, I)$, $\vec{\epsilon}_{t2} \sim N_m(0, \Lambda_1)$, $\vec{\eta}_{t1} \sim N_{k_1}(0, \Lambda_2)$, $\vec{\eta}_{t2} \sim N_m(0, \Lambda_3)$, $\vec{\mu}$ is a mean vector, $\Lambda_{1,2,3}$ and $\Phi_{1,2}$ are diagonal matrices. This model contradicts the common usage of factors, which is to reduce the number of unknown quantities. This system has even more unobservable processes than observation processes. The reason might be that $\text{Var}(\vec{Y}_t)$ will be singular, if **M2** does not have Equation (2.21) and the second term in the

right-hand side of Equation (2.19). Suppose we ignore $\vec{X}_t$, and compare the new form with what we will propose in the next paragraph. The new form of above, denoted by **M3**, is:

$$\vec{Y}_t = B \cdot \text{vec}\{\exp(\vec{F}_t/2)\vec{\epsilon}_t\}, \tag{2.22}$$

$$\vec{F}_t = \Phi\vec{F}_{t-1} + \vec{\eta}_t, \tag{2.23}$$

in which we suppose B is $m \times k$, and Φ is $k \times k$.

Different from the above two approaches, we are interested in the following setting, denoted by **M4**:

$$\vec{Y}_t = \sum_{i=1}^{k} \exp(X_{it}/2)\vec{Z}_{it}, \tag{2.24}$$

$$\vec{X}_t = \Phi\vec{X}_{t-1} + \vec{\eta}_t, \tag{2.25}$$

where $\vec{Z}_{it} \sim N_m(0, \Sigma_i)$ for $i = 1, \ldots, k$, $\vec{\eta}_t \sim N_k(0, \Sigma_\eta)$, and Φ is a $k \times k$ matrix. The simplest version of the above, denoted by **M5**, is:

$$\vec{Y}_t = \exp(X_t/2)\vec{Z}_t, \tag{2.26}$$

$$X_t = \phi X_{t-1} + \eta_t, \tag{2.27}$$

with $k = 1$, $\vec{Z}_t \sim N(0, \Sigma)$, and conditionally $\vec{Y}_t | X_t \sim N(0, \exp(X_t)\Sigma)$. This simple version is directly corresponding to Equation (8) in Kimmel (2004). The following shows that **M5** is a special case of Kimmel's model. Adopting our notation, we rewrite Kimmel's Equation (8) as follows:

$$df_{t,s} = \mu(X_t, t, s)dt + \sum_{i=1}^{m} \sigma_{wi}(X_t, t, s)dW_{ti} + \sigma_z(X_t, t, s)dZ_t, \tag{2.28}$$

where X_t is a latent process, W_{ti}'s are independent standard Brownian motions, Z_t is the random field as in Equation (2.8), σ_{wi}'s and σ_z are scalar functions. Simplifying the process by defining $\sigma_{wi} = 0$, we get the following:

$$df_{t,s} = \mu(X_t, t, s)dt + \sigma_z(X_t, t, s)dZ_t. \tag{2.29}$$

This simplification is to ignore additive factors. Recall the HJM framework, we can see that $\sigma_z(X_t, t, s)$ also determines $\mu(X_t, t, s)$ under the risk-neutral measure. Here, in order

to focus discussion on inference related to volatilities, we over-simplify the problem, ignore the drift, and focus on the following process:

$$df^*_{t,s} = \sigma_z(X_t, t, s)dZ_t. \tag{2.30}$$

For a finite collection, the processes are:

$$\begin{aligned} df^*_{t,s_1} &= \sigma_z(X_t, t, s_1)dZ_{t,s_1}, \\ \cdots \quad &\cdots \quad \cdots \\ df^*_{t,s_k} &= \sigma_z(X_t, t, s_k)dZ_{t,s_k}, \end{aligned}$$

with $\mathrm{Cov}(df^*_{t,s_i}, df^*_{t,s_j}) = \sigma_z(X_t, t, s_i)\sigma_z(X_t, t, s_j)\rho(s_i, s_j)dt$. Thus, in discrete time, we may get Equation (2.26). And the $(i,j)^{th}$ entries of the matrix $\exp(X_t)\Sigma$ are mapped by the following equations:

$$[\exp(X_t)\Sigma]_{i,j} = [\sigma_z(X_t, t, s_i)\sigma_z(X_t, t, s_j)\rho(s_i, s_j)]\,.$$

Our task is to propose an inferential methodology for the multivariate SVM that we have presented.

We consider our approach **M4** different from **M3**, in which each common volatility factor F_{ti} is realized only once at t, by attaching its transformation to a scalar innovation $\epsilon_{t,i}$. The realized volatility factor then enters into the observation vector, by linear combination. In contrast, a common volatility factor X_{it}, in **M4**, is realized multiple times at t by attaching its transformation to multiple innovations as components of a vector $\vec{Z}_{it}$, the components of which may not be necessarily independent. Though **M3** and **M4** are different, it will be very interesting to see how they are related. By defining perfectly correlated components in each $\vec{Z}_{it}$ in Equation (2.24), such that the corresponding covariance matrix is rank one, we can see that **M3** is a special case of **M4**. Therefore, it reveals that we control the perfectness of co-movements through controlling the correlation of $\vec{Z}_{it}$ in **M4**, rather than controlling the loading matrix B in **M3**. In addition, as we mentioned before, by such, the latent volatility process does not break the continuity of the *field* or *string* easily, and achieve a nonsingular dynamic covariance.

M4 brings both simplicity and complexity. The simplicity and flexibility of our model is appealing. First, the number of latent processes can be small, i.e. $k < m$, while the other two models have at least as many latent processes as observed time series. Second, our model does not need to balance common factors and singularity in covariance: as long as vector innovations have non-singular covariance, $\vec{Y}_t$ will have non-singular covariance. Third, dynamic correlation of $\vec{Y}_t$ is immediately obtained, when $k > 1$. The complexity is, for k-factors in **M4**, k random vectors $\vec{Z}_{it}$ for $i = 1, \ldots, k$, introduce k covariance matrices, each of which is $m \times m$. Although bringing one more common factor increases unknown parameters by at least $O(m^2)$, reducing one common factor will decreases unknown quantities by at least $O(N)$, where N is the time series length. For long time series, when $N >> m^2$, our model still has less unknown quantities than others. In addition, we suggest imposing structures on these covariance matrices, which is a very common approach in large covariance modeling, in order to reduce the number of unknown quantities in covariance matrices. From the next chapter, we start to focus on the inference and work with **M5**.

Chapter 3

Conventions, Practice, and Mismatches

We have mentioned that we may impose covariance structure on $\vec{Z}_t$ in Equation (2.26). The reason is both from achieving the parsimony and, more important, providing a chance of revealing the true nature of the term structure. This chapter points out that lack of understanding about implications of covariance structures may lead to possible misunderstanding about the nature of the term structure. Providing evidence that supports any existing covariance structures is beyond the scope of this study. However, we hope the material presented here can bring a fresh view regarding covariance modeling for the term structure. In order to facilitate the discussion, we model covariance by separating variance and correlation, and structures are embedded in the correlation. Regarding only structures, we use the term covariance structure and correlation structure indistinguishably in the discussion.

3.1 Smooth Yield Curves

A convention is that static yield curves are smooth. In empirical study on the term structure, the first thing is to construct static yield curves, or curves of other rates.

Under short rate models and HJM framework, short rate $r(t)$ and forward rate $f(t,T)$ are not observable, and must be computed from other quantities, such as bond prices, coupon rates, swap rates, and so on. A conventional method is known as the *bootstrap* method, which is not related to what was developed by Efron (1979). We refer to Hull (2002) and Fabozzi (1997) for the relationships among bond prices, coupon rates, bond yields, par rates, spot rates, and forward rates. Because bond prices and coupon rates are typically observable, other rates can be computed based on the relationships. The bootstrap procedure starts with zero-coupon securities that mature in a year, whose spot rates and par rates can be determined fairly easily. The remaining rates can be determined iteratively one by one with longer and longer maturities. The procedure might be plausible, if the observable curve can be observed at finer and finer grid resolution. In practice, that is impossible, especially for securities with long maturities which are much less traded. A conventional solution is to interpolate the observable curve by smoothing, e.g. cubic spline. Corresponding documentations can be easily found in finance literature and the Federal Reserve files on constructing constant maturity rates.

The convention that we are following in this study is that static yield curves are random functions. Therefore, forward rate curves that are derived from yield curves are also stochastic. By re-constructing such curves through interpolation techniques such as splines, the resulting curves may have a smoothness property that the underlying process does not have. We notice that such an issue leads to disagreement between empirical studies and mathematical modeling on the term structure as a random function.

3.2 Covariance Structure in Mathematical Modeling

The exponential correlation function $\rho_{ij} = \exp(-\beta|T_i - T_j|)$ has been used in Rebonato (2002) for modeling forward rate curves. This correlation function, and many of its variations have appeared in the pricing literature. This correlation function is known to characterise linear Gaussian Markov process, which is known as autoregressive process

of order one in discrete time, or Ornstein-Uhlenbeck process in continuous time. It is a known result that such a correlation function cannot govern a smooth random function in the mean square sense. References on this result and further on random functions and covariance functions include Yaglom (1987), Stein (1999), among many time series and spatial process analysis.

3.3 Covariance Structure in Empirical Studies

Simple historical estimation was described in Jarrow (1996), and used by string model practice (Longstaff et al. 2001). Interesting patterns are discovered, such as the decay pattern in correlations while two series are far apart, and historical correlations are larger than implied correlations that are calibrated to market prices. Their inference is based on a very crucial assumption that, for instance HJM model under the physical measure,

$$df(t,T) = \mu(t,T,f(t,T))dt + \sum \sigma_j(t,T,f(t,T))dW_j(t),\ t \leq T,$$

drift and volatilities are functions of only $(T-t)$, which is the difference of T and t. Briefly speaking, the procedure utilizes the discrete version of the term structure dynamics and constructs series of increments which have constant means specified by the drift function. Since the drift and volatilities are functions of $T-t$, then the series can be approximated from constant tenor curves. The increments are then treated as independent Gaussian vectors. Covariance is then easy to compute. There have been many discussions on the implications of stylized facts that are obtained by the simple estimation method and restrictive assumptions. We feel such discussions should be interpreted with caution and that stochastic volatilities or serial dependence has been ignored.

Though the stylized facts are obtained under rather restrictive conditions, some still deliver very useful information. For example, the correlation matrix shows a nearly Toeplitz structure, and correlation decays off the diagonals. Figure (3.1) graphically represents the correlation matrix for forward rates reported in Table I in Longstaff et al. (2001). This graph presents a smooth ridge, which implies that the corresponding random function

may be differentiable. Therefore, we are interested in the "Gaussian" correlation function $\rho_{ij} = \exp(-\beta(T_i - T_j)^2)$, or its variations. This correlation function can characterize a smooth Gaussian process, which is infinitely differentiable in the mean square sense. We call its discrete representation a Gaussian correlation matrix.

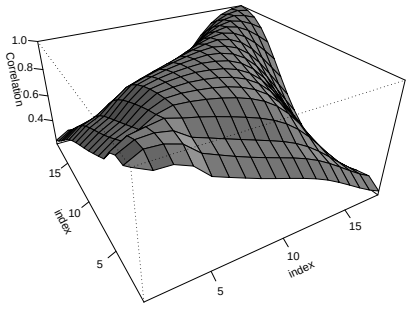

Figure 3.1: Simple Historical Correlation Estimate. The height of points on the surface represents the value in the correlation matrix. Matrix indices are along the other two horizontal axises.

To us, for inferential purposes, one attractive feature about those two correlation functions is that they have explicit forms of Cholesky decomposition, which is very useful in constructing Gaussian likelihood function. The decomposition of an exponential correlation matrix is well known. The decomposition of a Gaussian correlation matrix has been shown by Loh and Lam (2000).

Besides the computational usefulness of the Cholesky decomposition, one of its variations can interpret the decomposition in a natural way. The variation is even more convenient for likelihood construction. We present this variation in the next section.

3.4 Modified Cholesky and Covariance Modeling

Suppose V is symmetric, positive definite matrix. The Cholesky decomposition of V can be expressed as $V = LL^T$, where L is a lower triangle matrix, and L^T is the transpose of L. The modified Cholesky decomposition by Pourahmadi (1999) refers to

$$TVT^T = D, \tag{3.1}$$

where T is a lower triangle with ones on its diagonal, and D is a diagonal matrix. The relationship is $LL^T = T^{-1}D^{1/2}D^{1/2}(T^T)^{-1}$. Moreover, matrix inversion is easy to compute, by $V^{-1} = T^T D^{-1} T$.

T and D have very meaningful interpretations. Suppose V is the covariance matrix of a vector of zero mean Gaussian variables $Y_1, \ldots, Y_p$. Run regressions

$$Y_i = \sum_{j=1}^{i-1} \phi_{i,j} Y_j + \epsilon_i, \text{ for } i = 2, \ldots, p.$$

Then, $-\phi_{i,j}$ is the $(i,j)^{th}$ element of T. $Var(\epsilon_i)$ is the i^{th} element on D's diagonal.

In recent studies, the modified Cholesky decomposition has been used in longitudinal data analysis and large scale covariance modeling. The former utilizes the natural order among longitudinal data, while the later utilizes the computational simplicity.

We are interested in two types of covariance matrices. The first is the exponential covariance matrix, which is the discrete representation of an exponential covariance function. A typical matrix is as follows:

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\theta & \sigma^2\theta^2 & \ldots & \ldots & \sigma^2\theta^{m-1} \\ \sigma^2\theta & \sigma^2 & \sigma^2\theta & \ldots & \ldots & \ldots \\ \sigma^2\theta^2 & \sigma^2\theta & \sigma^2 & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \sigma^2\theta & \sigma^2\theta \\ \sigma^2\theta^{m-1} & \ldots & \ldots & \ldots & \sigma^2\theta & \sigma^2 \end{bmatrix},$$

where $|\theta| < 1$. The corresponding T and D matrices are as follows:

$$T = \begin{bmatrix} 1 & 0 & 0 & \dots & \dots & 0 \\ -\theta & 1 & 0 & \dots & \dots & \dots \\ 0 & -\theta & 1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 1 & 0 \\ 0 & \dots & \dots & \dots & -\theta & 1 \end{bmatrix},$$

$$D = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & \dots & 0 \\ 0 & (1-\theta^2)\sigma^2 & 0 & \dots & \dots & \dots \\ 0 & 0 & (1-\theta^2)\sigma^2 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & (1-\theta^2)\sigma^2 & 0 \\ 0 & \dots & \dots & \dots & 0 & (1-\theta^2)\sigma^2 \end{bmatrix}.$$

It is more complicated for the Gaussian covariance function and its corresponding matrix, which are presented in the following as functions of row and column indices. The notation L indicates a d-dimensional lower triangle matrix. For $1 \le i, j \le d$,

$$\begin{aligned} L_{i,j} &= (-w)^{i-j} G(j-1, i-1; w^2), \\ T_{i,j} &= \frac{L_{i,j}}{L_{i,i}}, \\ D_{i,i} &= \frac{1}{L_{i,i}^2}, \end{aligned}$$

where

$$G(m, n; q) = \begin{cases} \frac{(1-q^{n-m+1})(1-q^{n-m+2})\dots(1-q^n)}{(1-q)(1-q^2)\dots(1-q^m)} & \text{, if } 0 \le m \le n \\ 0 & \text{, otherwise,} \end{cases}$$

The function $G(m, n; q)$ is known as the Gaussian polynomial; see Loh and Lam (2000). Their inverse also have explicit forms. Therefore, the Gaussian likelihood does not require numerical inversion. We use the above structures as special cases in the rest of our study.

Chapter 4

Stochastic Volatility Model and Inference

We have presented several forms of SVMs in Chapter 2. The simplest form of our model is described by Equation (2.26) and (2.27). In Chapter 3, We have discussed about modeling covariance structures for the covariance of $\vec{Z}_t$ in Equation (2.26). The next question is about inference, after injecting a stochastic volatility process. In this chapter, we focus on the inference about the canonical SVM.

The chapter is arranged as follows. The first section will revisit the canonical SVM. A review on three canonical approaches follows. Then, we study the approach that uses particle filtering to construct the likelihood, and make inference about unknown parameters, which are assumed fixed. As we will point out, the quality of such a likelihood construction does not have a commonly agreed assessment, we suggest a procedure for making inference that is based on this approach.

4.1 The Canonical Form

The canonical form (Taylor 1982) of a stochastic volatility model is as follows:

$$Y_t = \beta \exp(X_t/2)\epsilon_t \tag{4.1}$$

$$X_t = \phi X_{t-1} + \sigma u_t, \tag{4.2}$$

where we may assume that u_t and ϵ_t are identical and independent Gaussian white noises. The last equation is known as the autoregressive of order one (AR1) process. This is corresponding to the following stochastic differential:

$$\begin{aligned} dY_t &= \gamma_t dW_{1t} \\ \gamma_t &= \beta \exp(X_t/2) \\ dX_t &= -\theta X_t dt + \sigma_\omega dW_{2t}, \end{aligned}$$

where W_{1t} and W_{2t} are independent Wiener processes. The last stochastic differential is known as a zero mean Ornstein-Uhlenbeck (OU) process. A discrete series sampled from an OU process every Δt unit has the covariance function $\gamma(\Delta t) = \frac{\sigma_\omega^2}{2\theta} e^{-\theta \cdot \Delta t}$. Moreover, relationship between OU and AR1 is as follows.

$$\phi = e^{-\theta} \tag{4.3}$$

$$\sigma_\epsilon^2 = \frac{1 - e^{-2\theta}}{2\theta} \sigma_\omega^2. \tag{4.4}$$

The derivation of OU covariance is given in Appendix A.1. We have omitted the drift term, and probably over-simplified the problem. But for situations with constant drift or deterministic drift, the generalization is trivial. More complicated generalization includes mean reverting, etc. We will probably postpone such extensions to future work.

By reparameterization, the canonical form can be expressed by the following equivalent form as well:

$$Y_t = \exp(X_t/2)\epsilon_t \tag{4.5}$$

$$X_t = \alpha + \phi X_{t-1} + \sigma u_t, \tag{4.6}$$

in which the state process has a mean, which is not necessarily zero. Two parameterizations are connected by the relationship

$$\beta = \exp\left\{\frac{\alpha}{2(1-\phi)}\right\}. \tag{4.7}$$

This parameterization has been used in some literature, and we may as well compare our results with those that use this parameterization.

In literature, autoregressive conditional heteroscadesticity (ARCH) and generalized ARCH (GARCH) are also well known for modeling volatility clustering, due to Engle (1982) and Bollerslev (1986), respectively. It is known that pricing under GARCH-type volatilities, the market completeness assumption is not broken. However, pricing under Taylor's stochastic volatility will break the assumption. Details are in Fouque et al. (2000). We will not discuss issues that are not related to inference. We focus on Taylor's structure, because it leads to more general problems in other areas.

More general problems include state-space modeling. Here, in particular, the problems are either non-linear or non-Gaussian, or both. Linear and Gaussian state-space models have been studied by many authors. See Harvey (1989) as a textbook treatment. The efficiency of the Kalman filter (Kalman 1960) is well known. Its success in autoregressive and moving average (ARMA) modeling (Jones 1980) is also well noted.

A typical state-space problem usually has the Markovian property, which is often represented by two conditional densities $p(X_{t+1}|X_t,\theta)$ and $p(Y_t|X_t,\theta)$ for a certain parameter θ, where $\{X_t\}_{t=1}^T$ denote unobservable state variables, and $\{Y_t\}_{t=1}^T$ are observed time series. Usually, we assume that $p(X_1)$ is known. The first conditional density prescribes how state variables propagate. The second conditional density prescribes the transition from unobservable variables to observable variables. By the Markovian property, conditionally on X_t, two variables X_{t+1} and Y_t are independent of $X_{1:t-1}$ and $Y_{1:t-1}$, where $X_{1:t-1}$ is the state variable sequence from 1 to $t-1$, and $Y_{1:t-1}$ is the observed variable sequence from 1 to $t-1$.

In the Kalman filter, the iteration starts with a known density $p(X_t|Y_{1:t-1},\theta)$ at

t, with a known parameter set θ. We may sometimes drop θ to make the notation simpler. The task is to compute $p(X_{t+1}|Y_{1:t}, \theta)$ when Y_t is observed. Because $p(Y_t|X_t, \theta)$ is given, we can derive

$$p(X_t|Y_{1:t}, \theta) \propto p(Y_t|X_t, \theta)p(X_t|Y_{1:t-1}, \theta). \tag{4.8}$$

Next, we can derive

$$p(X_{t+1}|Y_{1:t}, \theta) = \int p(X_{t+1}|X_t, \theta)p(X_t|Y_{1:t}, \theta)dX_t. \tag{4.9}$$

Fortunately, for linear Gaussian transitions, $p(X_t|Y_{1:t}, \theta)$ and $p(X_{t+1}|Y_{1:t}, \theta)$ have closed forms. And it suffices to compute mean and variance of $X_t|Y_{1:t}$, which is Gaussian distributed. The following is a graph that demonstrates the iteration.

$$\begin{array}{ccccccc} & & Y_{t-1} & & Y_t & & Y_{t+1} \\ & & \uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow \\ \rightarrow & X_{t-1} & \rightarrow & X_t & \rightarrow & X_{t+1} & \rightarrow \end{array}$$

In the above graph, right and upward arrows denote known state and observation transitions. Downward arrows denote computing the posterior.

For non-linear and non-Gaussian state-space problems, computing $p(X_t|Y_{1:t}, \theta)$ and $p(X_{t+1}|Y_{1:t}, \theta)$ is difficult, because closed forms usually do not exist. The recently developed particle filter (Gordon 1993) is an *exact* generalization to the Kalman filter. The *exact* means that all dependent structures are not changed or broken, which is different from the Extended Kalman filter (EKF). EKF linearizes the state-space equations around the filtered state at each iteration. Particle filters compute exact densities, although not analytically. We postpone the presentation on particle filtering to a later section.

4.2 Canonical Inferences

The state-space structure brings flexibility to describing dynamics. However, it brings difficulties to inference about parameters. That is because the state sequence is introduced and unobservable, and it has the same length as the observed series. We are

aware of three canonical inferences for stochastic volatility models; see Shephard (2005). They include Markov chain Monte Carlo (MCMC), efficient method of moments (EMM), and particle filtering. In the following sub-sections, we briefly present the first two, and focus on the third one. The first two involve general purpose inference methods, which deserve detailed review and study, which are beyond the scope of this review.

4.2.1 MCMC

From a logical perspective, MCMC approaches are straight forward, if the joint density of parameters, unobservable variables, and observable variables, can be expressed in a closed ring of conditional densities. Thus, a sampler can explore the space of random variables by following the rules that are designated by the conditional densities. There are a variety of implementations to guide the sampler's movement. An accessible and comprehensive reference on both theory and practice is Gilks et al. (1996). In stochastic volatility literature, Jacquier et al. (1994) pioneered this approach.

4.2.2 Efficient Method of Moments

EMM approach for stochastic volatilities, by Gallant et al. (1997), also utilizes a general inference methodology, which belongs to the family of indirect inference. The basic idea of indirect inference is the following deduction: data from the generator with the same parameter set must have similar characteristics in terms of some summary statistics, by fitting the same auxiliary model. In addition, in order to make inductive conclusion reasonable, the auxiliary model must be a close approximation to the data generator. A preferred approximation is known as semi-nonparametric density, and the feature statistic is the score function. See details in Gallant and Tauchen (1996) and references therein.

4.2.3 Particle Filtering and State Inference

Actually, particle filtering does not make inference about fixed parameters. The filtering can only make inference about unknown states, given all other parameters. More

specifically, the filtering consecutively computes $p(X_t|Y_{1:t},\theta)$ and $p(X_{t+1}|Y_{1:t},\theta)$ in Equation (4.8) and (4.9), when θ is known. Related names include sequential Monte Carlo, sequential importance sampling, sequential sampling importance resampling, etc. An accessible and comprehensive reference is Doucet et al. (2001). There have been modifications and enhancement to the first successful filter by Gordon (1993). We only review the original version in this research. The foundations include basic sampling importance resampling, state prediction, and state filtering. We first discuss sampling importance resampling (SIR).

Sampling from a Posterior and SIR

Suppose we know densities $p(X)$ and $p(Y|X)$ for random variables Y and X. The task is to compute $p(X|Y) \propto p(Y|X)p(X)$. From simulation point of view, being able to compute the density $p(X|Y)$ is equivalent to being able to generate a sample from $p(X|Y=y)$, for any y, if it is possible. There are apparently two approaches to generate such a sample; see Smith and Gelfand (1992, and references therein). The first is known as the rejection sampling; see von Neumann (1951). The second is SIR.

To facilitate the discussion regarding sample points from certain distributions, we adopt the following notations. We use $\widehat{x}^{(i)}$ to denote the i^{th} sample point from $p(X)$. The notation will be intuitive for a "predicted" or "proposed" sample point. We use $\{\widehat{x}^{(i)}\}_{i=1}^{M}$ to denote such a sample of size M. We use $x^{(i)}$ to denote the i^{th} sample point from $p(X|Y=y)$, for some y. We use upper cases for variables, and lower cases for realized values. We often simply use $p(X|y)$ for $p(X|Y=y)$. $\{x^{(i)}\}_{i=1}^{M}$ denotes a sample of size M from $p(X|y)$. The curly brackets notation sometimes also stands for a set of indexed numbers.

In order to use the rejection sampling, first we assume $p(y|X)$ is maximized at X^{max}. Then $p(y|X^{max})$ serves as the constant C in the following inequality:

$$p(X,y) = p(y|X)\cdot p(X) \leq p(y|X^{max})\cdot p(X) = C\cdot p(X).$$

Then for each sample point $\widehat{x}^{(i)}$ from $p(X)$, accept the point with probability

$$\frac{p(\widehat{x}^{(i)},y)}{C\cdot p(\widehat{x}^{(i)})},$$

until M sample points are accepted. Then $\{x^{(i)}\}_{i=1}^{M}$ represent a sample of size M from the density $p(X|y) = p(X,y)/p(y)$; see Ripley (1987) for the proof.

In order to use SIR, we must find a weight function that is determined by the ratio $p(X|y)/p(X)$, in which $p(X|y)$ is the target distribution, and $p(X)$ is the proposal distribution. In the current situation,

$$p(X|y)/p(X) \propto p(y|X)p(X)/p(X) = p(y|X).$$

Therefore, $p(y|\widehat{x}^{(i)})$ serves as the importance weight for the sample $\widehat{x}^{(i)}$ from $p(X)$. An implementation of SIR is as follows:

Procedure 4.2.1. *An SIR Implementation*

(a) generate $\{\widehat{x}^{(i)}\}_{i=1}^{M}$ *from* $p(X)$*;*

(b) compute $\widehat{w}^{(i)} = p(y|\widehat{x}^{(i)})$*, for* $i = 1, \ldots, M$*;*

(c) compute normalized weights $w^{(i)} = \frac{\widehat{w}^{(i)}}{\sum_{i=1}^{M} \widehat{w}^{(i)}}$*, for* $i = 1, \ldots, M$*;*

(d) resample from $\{\widehat{x}^{(i)}\}_{i=1}^{M}$*, regarding weights* $\{w^{(i)}\}_{i=1}^{M}$*, and get a sample* $\{x^{(i)}\}_{i=1}^{M}$.

Thus, we obtain a sample that is from $p(X|y)$. We then apply this technique to filtering.

SIR and Filtering

Recall the general Markov state-space structure with known transitions. We assume that $p(X_t|Y_{1:t-1})$ is known. Due to the Markovian property, it is equivalent to know $p(X_t|Y_{t-1})$. Sampling from $p(X_t|Y_t)$, which is the first task, is achieved by using SIR. This accomplishes the task of Equation (4.8).

The notation for the following discussion is similar to those discussed previously, with extra subscripts. We use $\widehat{x}_t^{(i)}$ for the i^{th} sample from $p(X_t|Y_{1:t-1})$, which stands for a sample of the predicted state at t. Use $x_t^{(i)}$ for the i^{th} sample from $p(X_t|Y_{1:t})$, which stands for a sample of the filtered state at t. Finally, use $\widetilde{x}_t^{(i)}$ for the i^{th} sample from $p(X_t|Y_{1:T})$,

which stands for a sample of the smoothed state at t, where $Y_{1:T}$ stands for the entire observed series.

Suppose we generate a sample $\{\widehat{x}_t^{(i)}\}_{i=1}^M$ of size M from $p(X_t|Y_{t-1} = y_{t-1})$. Use $p(y_t|X_t)$ as the weight function. Adopt SIR, we obtain $\{x_t^{(i)}\}_{i=1}^M$ from $p(X_t|y_t)$. The ultimate goals is to compute $p(X_{t+1}|Y_{1:t})$ as in Equation (4.9). Due to the Markovian property,

$$p(X_{t+1}|Y_{1:t}) = p(X_{t+1}|Y_t) = \int p(X_{t+1}|X_t)p(X_t|Y_t)dX_t = \int p(X_{t+1}, X_t|Y_t)dX_t.$$

Therefore, generating a sample from $p(X_{t+1}|Y_{1:t})$ is equivalent to first generate a sample from $p(X_{t+1}, X_t|Y_t)$, then marginalize it with respect to X_t. For any y_t, generating a pair $(\widehat{x}_{t+1}^{(i)}, x_t^{(i)})$ from $p(X_{t+1}, X_t|Y_t = y_t)$ is achieved by generating $x_t^{(i)}$ from $p(X_t|y_t)$ first, then generate $\widehat{x}_{t+1}^{(i)}$ from $p(X_{t+1}|X_t = x_t^{(i)})$. By such, we can implement the filtering, at step t, as follows:

Procedure 4.2.2. *Adopt SIR in Filtering*

(a) sample $\widehat{x}_t^{(i)}$ *from* $p(X_t|X_{t-1} = x_{t-1}^{(i)})$, *for* $i = 1, \ldots, M$;

(b) compute $\widehat{w}_t^{(i)} = p(y_t|\widehat{x}_t^{(i)})$, *for* $i = 1, \ldots, M$;

(c) compute normalized weights $w_t^{(i)} = \frac{\widehat{w}_t^{(i)}}{\sum_{i=1}^M \widehat{w}_t^{(i)}}$, *for* $i = 1, \ldots, M$;

(d) resample from $\{\widehat{x}_t^{(i)}\}_{i=1}^M$, *regarding weights* $\{w_t^{(i)}\}_{i=1}^M$, *and get a sample* $\{x_t^{(i)}\}_{i=1}^M$.

The simulation runs from $t = 1$ through T where the series terminate, given all other parameters. To initiate the simulation, $\{\widehat{x}_1^{(i)}\}_{i=1}^M$ is a sample generated from a known distribution. This procedure serves the basis of particle filtering, where *particle* is a vivid name for sample points. As a convention, state estimation is simply $E(X_t|Y_t)$, which is approximated by a sample average $\frac{1}{M}\sum_{i=1}^M x_t^{(i)}$.

We have mentioned that there are at least two approaches to generate a sample from a posterior distribution. In this filter, SIR is generally adopted. A reason is that finding a constant for the rejection sampling is not required for SIR. Thus it is more general and hopefully easy to implement.

4.2.4 Particle Filtering and Parameter Inference

Though particle filtering provides state inference, parameters must be given separately. Hürzeler and Künsch (2001) give a comprehensive outline on likelihood based inference. However, likelihood construction is based on simulation. The following are three likelihood construction methods, which have been discussed in the above reference.

Pointwise Approximation

The following computes the likelihood at a fixed parameter set θ. In order to avoid unnecessary confusion and make notations more clear, we allow the subscript $t-1$ to be 0, in which case the conditioning drops off. For example, $p(y_1|y_0)$ is the same as $p(y_1)$. This simplified notation applies to the rest of discussions. The likelihood is:

$$L(\theta|y_{1:T}) \quad = \quad p(y_{1:T}|\theta) = \prod_{t=1}^{T} p(y_t|y_{1:t-1},\theta) = \prod_{t=1}^{T} \int p(y_t|x_t,\theta)p(x_t|y_{1:t-1},\theta)dx_t,$$

in which the integration term $\int p(y_t|x_t,\theta)p(x_t|y_{1:t-1},\theta)dx_t$ is approximated by sample average over particles as follows:

$$\frac{1}{M}\sum_{i=1}^{M} p(y_t|\widehat{x}_t^{\{i\}},\theta).$$

The maximum likelihood approach can optimize over the parameter space, until reaching a local or global optimum, depending on the method. For every different parameter set, filtering must be conducted to obtain trajectories $\{\widehat{x}_{1:T}^{(i)}\}_{i=1}^{M}$.

Function Approximation

Pointwise approximation requires re-filtering when parameters change. The likelihood surface is also noisy. The following constructs a smooth likelihood function about an arbitrary θ, after a filtering has been done at θ_0. The approximation utilizes the importance

sampling trick multiple times. The formulation is as follows:

$$\begin{aligned} & L(\theta|y_{1:T},\theta_0) = p(y_{1:T}|\theta) = \prod_{t=1}^{T} p(y_t|y_{1:t-1},\theta) \\ = & \prod_{t=1}^{T} \int p(y_t|x_t,\theta) p(x_t|y_{1:t-1},\theta) dx_t \\ = & \prod_{t=1}^{T} \int p(y_t|x_t,\theta) \frac{p(x_t|y_{1:t-1},\theta)}{p(x_t|y_{1:t-1},\theta_0)} p(x_t|y_{1:t-1},\theta_0) dx_t, \end{aligned}$$

where

$$\begin{aligned} & p(x_t|y_{1:t-1},\theta) \\ = & \int p(x_t|x_{t-1},\theta) p(x_{t-1}|y_{1:t-1},\theta) dx_{t-1} \\ = & \int p(x_t|x_{t-1},\theta) \frac{p(x_{t-1}|y_{1:t-1},\theta)}{p(x_{t-1}|y_{1:t-1},\theta_0)} p(x_{t-1}|y_{1:t-1},\theta_0) dx_{t-1} \end{aligned}$$

and

$$\begin{aligned} & \frac{p(x_{t-1}|y_{1:t-1},\theta)}{p(x_{t-1}|y_{1:t-1},\theta_0)} \\ = & \frac{p(y_{t-1}|x_{t-1},\theta)}{p(y_{t-1}|x_{t-1},\theta_0)} \times \frac{p(x_{t-1}|y_{1:t-2},\theta)}{p(x_{t-1}|y_{1:t-2},\theta_0)} \times \frac{p(y_{t-1}|y_{1:t-2},\theta_0)}{p(y_{t-1}|y_{1:t-2},\theta)}. \end{aligned}$$

Two starting densities $p(x_1|\theta)$ and $p(x_1|\theta_0)$ are known. Among above quantities, $p(\cdot|\cdot,\theta_0)$ are directly feasible from the filtered sample at θ_0. Meanwhile, $p(y_t|x_t,\ \cdot)$ and $p(x_{t+1}|x_t,\ \cdot)$ are available from known transitions for any parameter. $p(x_t|y_{1:t-1},\theta)$ and $p(y_t|y_{1:t-1},\theta)$ are computed recursively, by knowing $p(x_1|\theta)$ and $p(x_1|\theta_0)$ first.

The major step among the above expressions is to approximate $p(y_t|y_{1:t-1},\theta)$, which is sample average as follows:

$$\begin{aligned} A_{0,t} = p(y_t|y_{1:t-1},\theta) & \approx \frac{1}{M} \sum_{i=1}^{M} p(y_t|\widehat{x}_t^{(i)},\theta) \frac{p(\widehat{x}_t^{(i)}|y_{t-1},\theta)}{p(\widehat{x}_t^{(i)}|y_{t-1},\theta_0)} \qquad (4.10) \\ & = \frac{1}{M} \sum_{i=1}^{M} p(y_t|\widehat{x}_t^{(i)},\theta) \frac{A_1(X_t=\widehat{x}_t^{(i)})}{B_1(X_t=\widehat{x}_t^{(i)})}, \qquad (4.11) \end{aligned}$$

among which the term $p(y_t|\widehat{x}_t^{(i)},\theta)$ is easy to compute with the known transition from state to observation. Two functions $A_1(X_t=x_t)$ and $B_1(X_t=x_t)$, which take $X_t=x_t$ as the

argument, are computed as follows:

$$A_1(X_t = x_t) = p(x_t|y_{t-1}, \theta) \approx \frac{1}{M}\sum_{j=1}^{M} p(x_t|x_{t-1}^{(j)}, \theta)\frac{p(x_{t-1}^{(j)}|y_{t-1}, \theta)}{p(x_{t-1}^{(j)}|y_{t-1}, \theta_0)} \quad (4.12)$$

$$= \frac{1}{M}\sum_{j=1}^{M} p(x_t|x_{t-1}^{(j)}, \theta)\frac{A_{2,t}}{B_{2,t}}, \quad (4.13)$$

$$B_1(X_t = x_t) = p(x_t|y_{t-1}, \theta_0) \approx \frac{1}{M}\sum_{j=1}^{M} p(x_t|x_{t-1}^{(j)}, \theta_0), \quad (4.14)$$

where the term $p(x_t|x_{t-1}^{(j)}, \theta_0)$ and $p(x_t|x_{t-1}^{(j)}, \theta)$ are easy to compute with the known state transition. Two other terms $A_{2,t}$ and $B_{2,t}$, are computed as follows:

$$A_{2,t} = p(x_{t-1}^{(j)}|y_{t-1}, \theta) = \frac{p(y_{t-1}|x_{t-1}^{(j)}, \theta)p(x_{t-1}^{(j)}|y_{t-2}, \theta)}{p(y_{t-1}|y_{t-2}, \theta)} \quad (4.15)$$

$$= p(y_{t-1}|x_{t-1}^{(j)}, \theta)\frac{A_1(X_{t-1} = x_{t-1}^{(j)})}{A_{0,t-1}}, \quad (4.16)$$

$$B_{2,t} = p(x_{t-1}^{(j)}|y_{t-1}, \theta_0) = \frac{p(y_{t-1}|x_{t-1}^{(j)}, \theta_0)p(x_{t-1}^{(j)}|y_{t-2}, \theta_0)}{p(y_{t-1}|y_{t-2}, \theta_0)} \quad (4.17)$$

$$= p(y_{t-1}|x_{t-1}^{(j)}, \theta_0)\frac{B_1(X_{t-1} = x_{t-1}^{(j)})}{B_{0,t-1}}, \quad (4.18)$$

where

$$B_{0,t} = p(y_t|y_{t-1}, \theta_0) \approx \frac{1}{M}\sum_{j=1}^{M} p(y_t|\widehat{x}_t^{(j)}, \theta_0). \quad (4.19)$$

Moreover, in order to initiate the propagation, $A_1(X_1 = x)$ and $B_1(X_1 = x)$ are computed as following:

$$A_1(X_1 = x) = p(x|\theta), \quad (4.20)$$

$$B_1(X_1 = x) = p(x|\theta_0). \quad (4.21)$$

Therefore, by maximizing over θ, the inference then runs an approximation maximization iterative approach, which is

$$\theta_{k+1} = \arg\max_{\theta} \log L(\theta|y_{1:T}, \theta_k).$$

Though the function approximation is mathematically elegant, the complexity of the approximation, compared to the pointwise approximation, is $O(M^2)$, which is sometimes too large. The complexity of the pointwise approximation is $O(M)$. The major benefit of the function approximation is the possible smoothness. However, the analytic gradient for such a complicated function is not necessarily available in general. In this case the numerical gradient or Hessian must be used in derivative-based optimization routines, and the gain might be questionable. On the contrary, by using non-derivative-based optimization routines, e.g. Nelder-Mead, the optimizer may converge by using a certain number of pointwise function evaluations which is still cheap, compared to $O(M^2)$, even though the likelihood is not smooth. A question arises about how we can draw such an inference from a noisy surface. Recall, the function approximation also involves particle randomness effects, although the surface is smoothing. After running filtering the second time at $\theta_1 \equiv \theta_0$, by using a different random seed, the new likelihood function $L(\theta|y_{1:T}, \theta_1)$ does not necessarily agree with $L(\theta|y_{1:T}, \theta_0)$. Therefore, smoothness is not crucial in comparing pointwise and function approximations.

Expectation-Maximization Algorithm

Regarding states as missing observations, the E-M algorithm iterates as follows

$$\theta_{k+1} = \arg\max_{\theta} \int \log p(y, x|\theta)\mu(dx|y, \theta_k),$$

where the integral denotes the expectation of the log-likelihood of the complete data with respect to the probability measure on the missing quantity. The expectation is approximated by the average over a smoothed samples from $p(X_{1:T}|Y_{1:T}, \theta_k)$, or $p(X_{1:T}|Y_{1:T})$ for simplicity. The smoothed sample can be obtained by backward smoothing on filtered samples as follows:

$$p(x_t|x_{t+1}, y_{1:T}, \theta_k) = p(x_t|x_{t+1}, y_{1:t}, \theta_k) \propto p(x_t|y_{1:t}, \theta_k)p(x_{t+1}|x_t, \theta_k),$$

which is again obtained by SIR, where $p(x_{t+1}|x_t, \theta_k)$ is the weight function. The following is an algorithm that obtains a smooth sample $\{\widetilde{x}_{1:T}^{(i)}\}_{i=1}^{M}$. At the k^{th} iteration, the algorithm propagates backwards from $t+1$ to t by following these three steps for each i:

Procedure 4.2.3. *Adopt SIR in Smoothing*

(a) compute $w_t^{(i,j)} = p(\widetilde{x}_{t+1}^{(i)}|x_t^{(j)})$, *for* $j = 1, \ldots, M$;

(b) compute $\widetilde{w}_t^{(i,j)} = \frac{w_t^{(i,j)}}{\sum_{j=1}^{M} w_t^{(i,j)}}$, *for* $j = 1, \ldots, M$;

(c) sample $\widetilde{x}_t^{(i)}$ *from* $\{x_t^{(j)}\}_{j=1}^{M}$, *regarding weights* $\{\widetilde{w}_t^{(i,j)}\}_{j=1}^{M}$.

In addition, the smooth sample at T coincides with the filtered sample at T. Thus, in terms of particles, the E-M algorithm steps from k to $k+1$ as follows:

$$\theta_{k+1} = \arg\max_{\theta} \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \log\left(p(y_t|\widetilde{x}_t^{(i)}, \theta)\right).$$

In addition, it would not be difficult to see that the complexity of the algorithm is $O(M^2)$ as well.

4.3 Quality of Likelihood and a Systematic Framework

Regarding the three approaches of constructing the likelihood, we want to address that there are no formal inferential procedures that consider both estimates and the quality of likelihood approximation. All likelihood constructions mentioned in the previous section are based on a finite set of state trajectories. The question arises when maximum likelihood estimates and their covariance matrix are reported. This is because they may be affected by not only the length of time series, but also the number of particles or state trajectories. Our opinion is that following the simulated likelihood approach, the estimates can be trusted only if the randomness caused by the number of state trajectories is negligible. We do not have a theoretical justification about how much randomness is negligible. However, because a certain amount of difference in information criteria will affect decision making, we would like to suggest that the randomness caused by the number of state trajectories should be controlled within the critical range that affects decision making. For example, if a difference by 3 between Akaike information criteria (AIC) will call for a decision making between two different models, we might like to control the 90% span of the log-likelihood

uncertainty due to particles within, say 1.5, because AIC requires twice of the log-likelihood. Otherwise, if two models have difference in information criteria by 3, it would be difficult to tell whether two models are really different, or just due to the particle randomness. However, balancing among the accuracy of the simulated likelihood, saving computing expense on function evaluations, and effectively finding the optimal is rather subjective. We suggest the following iterative procedure:

Procedure 4.3.1. *A Systematic Framework for Particle Filtering-based MLE*

(a) at step k, set the number of particles to M, quit if M exceeds a limit;

(b) utilize an optimization routine to find the maximum of the log likelihood, initiated at θ_{k-1} that is the estimate obtained at step $k-1$, using any kind of likelihood approximation;

(c) in case optimization fails to satisfy individual stopping criteria, go to step $k+1$ and increase M;

(d) at the optimum θ_k, bootstrap the likelihood evaluated at θ_k;

(e) if the variation of the bootstrapped likelihood is larger than a prespecified threshold, go to step $k+1$ and increase M;

(f) compute statistics necessary for reports.

For step (b), we suggest using pointwise approximation and a fixed random seed each time the function is evaluated. Thus, we obtain static approximation to the entire likelihood surface by fixing the random seed, and reduce the roughness by increasing the number of particles. Fixing random seed is an effective practice; see Durham et al. (2002). More efficiency can be gained by spending more on pre-storing random numbers. The static approximation is still non-smooth, due to the sampling step. Fortunately, robust local optimization routines exist. For example, the non-gradient based simplex method

(Nelder and Mead 1965) is such a well known routine. More routines can be found in Kelley (1999), among others.

The tricks about increasing number of particles and fixing random seeds are not new. We implement the procedure more systematically, and add a necessary step of assessing the quality of likelihood approximation. The procedure provides an exploratory approach at the beginning when the number of particles is small and filtering is relatively cheap. Based on such very rough surface, the Nelder-Mead optimization is still capable of approaching towards an optimum. By increasing M, we help the Nelder-Mead approach the optimum under finer and finer resolutions, and assess the quality of likelihood approximation in addition.

In step (d), bootstrap can be done by repeating the likelihood evaluation with different random seeds. The size of bootstrap sample can be subjective. In case of regarding the span of the log-likelihood, the sample size should be relatively large, if the coverage is large. Step (f) can also be subjective. For example, we can keep several independent runs to get estimates, with different random seeds but the same number of particles. Then we summarize the estimates. We can also adopt the function approximation to the likelihood once to obtain a set of estimates. Finally, how to increase M is also subjective.

The following studies the result from a simulation to illustrate how the procedure works. The simulated data is generated from the canonical SVM with parameters $\{\beta = 0.6, \phi = 0.95, \sigma = 0.3\}$. This setting generates the observed time series with a clear volatility clustering pattern. States are slowly mean reverting with relatively small noise. Figure (4.1) gives the plots of observed time series, and the *true* state series.

Table (4.1) illustrates our procedure. The initial guess about the parameters is arbitrary but realistic, which is $\{\beta_0 = 0.3, \phi_0 = 0.9, \sigma_0 = 0.5\}$. The stopping criterion The first column in the table indicates the number of particles that are used at that stage. Using that amount of particles, the optimization routine, which is Nelder-Mead in our study, finds an optimum at the values under columns $\widehat{\beta}$, $\widehat{\phi}$, and $\widehat{\sigma}$. The corresponding log-likelihood values are also listed in the third column. The second column tells that there

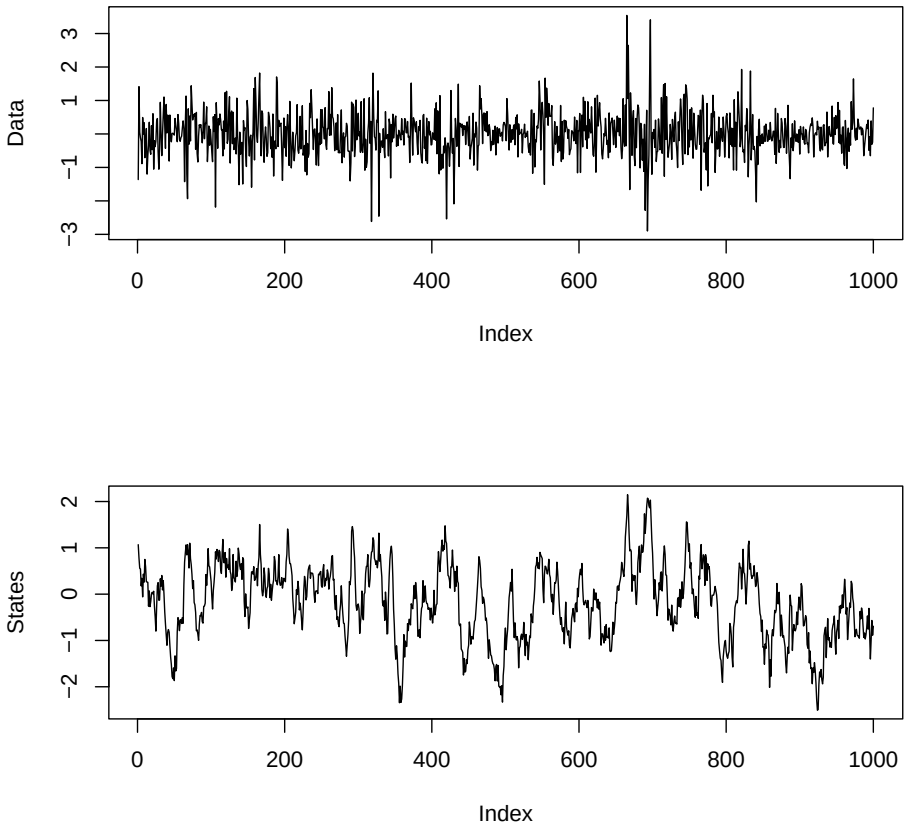

Figure 4.1: Simulated Stochastic Volatility Process with $\{\beta = 0.6, \phi = 0.95, \sigma = 0.3\}$

have been no convergence failure reported by Nelder-Mead, in this simulation. From this

Table 4.1: History of the Estimation

M	converge	log-likelihood	$\widehat{\beta}$	$\widehat{\phi}$	$\widehat{\sigma}$
100	yes	-923.94	0.58	0.90	0.34
200	yes	-925.93	0.57	0.89	0.42
400	yes	-926.17	0.55	0.94	0.27
800	yes	-925.12	0.56	0.90	0.34
1600	yes	-926.15	0.55	0.92	0.33
3200	yes	-926.39	0.54	0.93	0.32

table, regardless of the first column, we do not know about the quality of the estimates and realized log-likelihood. Because, by changing random seeds, estimates and log-likelihood will change. Therefore, at each stage with a particular M, we bootstrap the log-likelihood at the corresponding estimates. Figure (4.2) provides the bootstrapped log-likelihood at each stage with increasing number of particles. The graph is presented in consecutive boxplots (Tukey 1977). In order to bootstrap log-likelihood, we simply use a different random seed

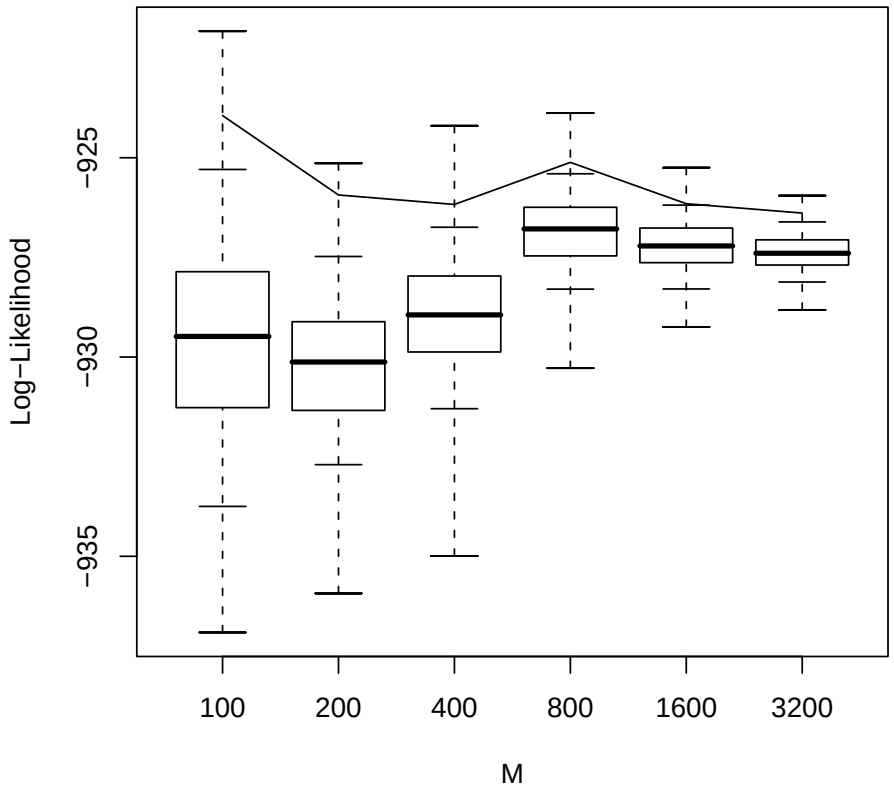

Figure 4.2: Boxplots for Bootstrapped Log-likelihood. Bootstrap sample size is 1000. Two additional bars to each boxplot indicate 5% and 95% percentiles. The line runs through boxplots indicates optima that the Nelder-Mead optimizer finds under six scenario with different number of particles.

to get simulated log-likelihood, which is computed by the pointwise approximation formula.

Meanwhile, we should notice that even the optimizer adopts the function approximation to the likelihood, bootstrapping the log-likelihood at the estimates gives 1 to all importance weights. So function approximation in log-likelihood bootstrapping reduces to pointwise approximation. At this point, we can see the smoothness of the approximated likelihood is not essential for assessing the quality of the likelihood approximation. With a fixed number of particles, the pointwise approximation and the function approximation, with the same set of parameters, will vary over the same range. We want to assess the range, which indicates the quality.

From Figure (4.2) we can see log-likelihood varies less as M increases. Bootstrapping terminates when $M = 3200$, at the moment that the 90% of the twice of the bootstrapped log-likelihood are in a region of width 3, which is a predetermined threshold. In Figure (4.2), there is a line through the boxplots. This line connects the log-likelihood in Table (4.1). It is clear, the optima that the optimizer finds are at the upper tail of the bootstrapped log-likelihood samples. Recall, we use pointwise approximation for function evaluation in the optimizer. The pattern confirms that the Nelder-Mead is capable to find an optimum on a noisy surface.

In order to make a connection between the boxplots and the actual static log-likelihood surface that we have generated with a fixed random seed, we profile the log-likelihood that uses $M = 100$ particles, around the obtained estimates. Each profile log-likelihood is evaluated at 1000 equally spaced points. In order to avoid the delusion that those 1000 function values present the curve that passes through all the points on the profile log-likelihood, we do not concatenate those points. The profile log-likelihood plots confirm the position of the boxplot for $M = 100$. These plots also help to imagine a scenario that an arbitrary number of particles are used to construct the simulated likelihood, and estimates are reported without assessing the quality of the likelihood approximation. A question that we are trying to answer here is that after an optimum is found, how well the simulated likelihood can represent the expected likelihood. Using the suggested procedure, we think the question is answered.

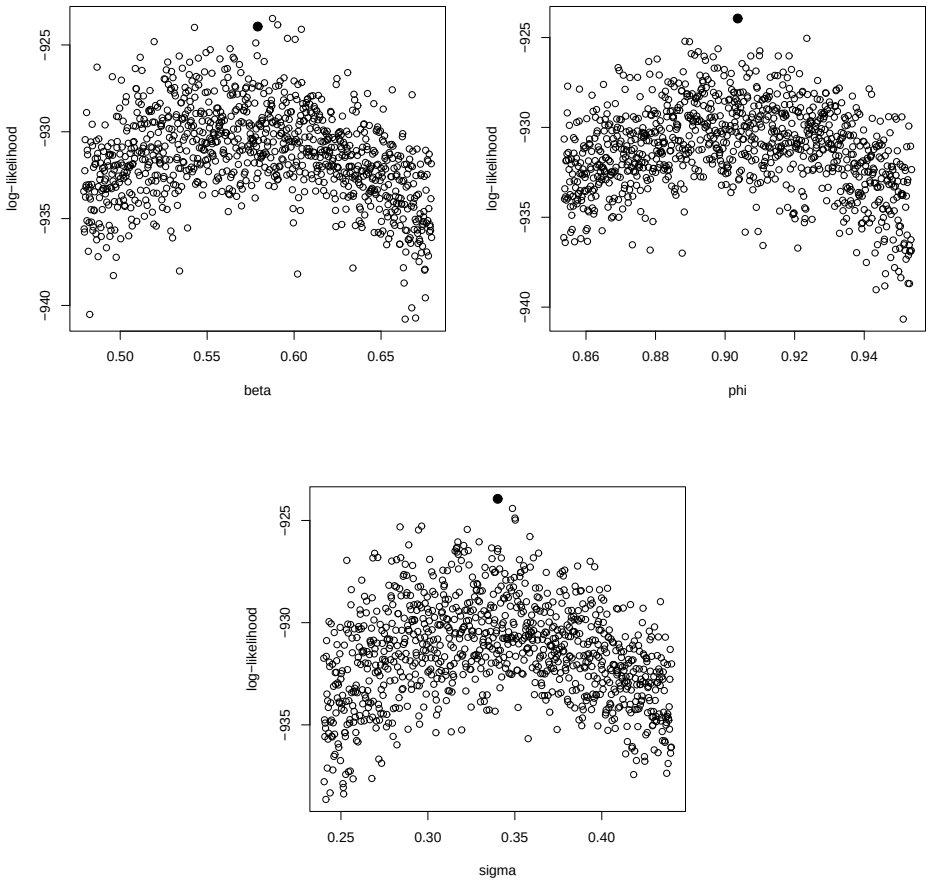

Figure 4.3: Static profile log-likelihood against β, ϕ, and σ. The number of particles $M = 100$. The solid dots are the optimal found by the Nelder-Mead optimizer.

4.4 Summary

In this chapter, we mainly study the particle filtering-based maximum-likelihood for the inference about the canonical stochastic volatility model. The major difficulty is to integrate over a huge space of the unobserved state trajectories. The space is huge, because the trajectory estimates can never be consistent under the current situation. All these simulated approaches try to generate samples from the space effectively. Based on their work, we suggest a systematic procedure to guide the parameter estimation. Briefly, we use

a strategy that is relatively cheap to find a region that might hold the parameters. This strategy does not require a rigorous initial guess, which is not generally feasible. Meanwhile, a fair number of particles is suggested later in the procedure. Then a variety of published strategies can be used to obtain estimates and statistics that can be reported.

Chapter 5

A Multivariate Stochastic Volatility Model and Inference

Based on our experience, the particle filtering approach to the inference is not as practical as the approach that Box et al. (1994) take to make inference about autoregressive and moving average (ARMA) models. Computationally, the method is too expensive. Fortunately, the particle filtering approach does setup connections, so that other inferential methods can be applied. This chapter pursues a saddlepoint approximation to integrate out state variables.

The rest of chapter is arranged as follows. First, we study the approach on the canonical SVM. Next, we study a multivariate extension described by Equation (2.26) and (2.27). After that, we report simulation results.

5.1 Saddlepoint Approximation Approach for the Univariate Model

The canonical form of stochastic volatility models is linear and Gaussian in the state transition. That gives some hope for an analytic approximation for some calculations that were done by the sequential Monte Carlo. This situation has been discussed as early as

in Masreliez (1975), in which approximating $p(Y_t|Y_{t-1})$ is not an easy task. In this section, we use the saddlepoint approximation (Daniels 1954) for all levels of approximation.

In this section, we focus on the canonical SVM described by Equation (2.13) and (2.14). Suppose $p(X_t|Y_{t-1}) \sim N(\widehat{\mu}_t, \widehat{\sigma}_t^2)$ and $p(Y_t|X_t)$ are known. We are interested in computing $p(Y_t|Y_{t-1})$, $p(X_t|Y_t)$, and $p(X_{t+1}|Y_t)$, such that filtering can propagate. In non-linear and non-Gaussian state-space, $p(X_t|Y_t)$ is not Gaussian, neither is $p(X_{t+1}|Y_t)$. However, if it is plausible to approximate $p(X_t|Y_t)$ by Gaussian, then $p(X_{t+1}|Y_t)$ is Gaussian naturally for the canonical SVM. And iteration continues. The saddlepoint approximation can be used for computing the posterior $p(X_t|Y_t)$ and the density $p(Y_t|Y_{t-1}) = \int p(Y_t|X_t)p(X_t|Y_{t-1})dX_t$.

Recall that the saddlepoint approximation is usually but not necessarily stated as an integration problem; see Goutis and Casella (1999). Suppose we are interested in the integral

$$f(x) = \int m(x,\theta)d\theta.$$

By defining $k(x,\theta) = \log m(x,\theta)$, the saddlepoint approximation to $f(x)$ is

$$f(x) \approx \exp\left\{k(x,\hat{\theta}(x))\right\}\left(-\frac{2\pi}{\frac{\partial^2 k(x,\theta)}{\partial\theta^2}\Big|_{\hat{\theta}(x)}}\right)^{1/2},$$

where $\hat{\theta}(x)$ maximizes $m(x,\theta)$. Therefore, we may apply this to computing $p(Y_t|Y_{t-1})$. The approximation is often stated as well as a Bayesian problem of computing the expectation of a function with respect to a posterior distribution, which is

$$E[g(\theta)|X=x] = \frac{\int g(\theta)f(x|\theta)\pi(\theta)d\theta}{\int f(x|\theta)\pi(\theta)d\theta},$$

where $g(\theta)$ is the function of interest, $\pi(\theta)$ is the prior distribution, and $f(x|\theta)$ is the likelihood. The approximation is

$$E[g(\theta)|X=x] \approx \exp\{k_n(x,\hat{\theta}(x)) - k_d(x,\theta^*(x))\}\left(\frac{\frac{\partial^2 k_d(x,\theta)}{\partial\theta^2}\Big|_{\theta^*(x)}}{\frac{\partial^2 k_n(x,\theta)}{\partial\theta^2}\Big|_{\hat{\theta}(x)}}\right)^{1/2},$$

where $k_n(x,\theta) = \log[g(\theta)f(x|\theta)\pi(\theta)]$ and $k_d(x,\theta) = \log[f(x|\theta)\pi(\theta)]$. Two functions $k_n(x,\theta)$ and $k_d(x,\theta)$ are maximized at $\hat{\theta}(x)$ and $\theta^*(x)$ respectively. It is required that $g(\theta)$ is a positive function, such that the saddlepoint approximation can be applied to $E[g(\theta)]$. In case that $g(\theta)$ is not positive, the moment generating function $M(t) = E[e^{tg(\theta)}]$ is computed first. After that, $E[g(\theta)]$ can be obtained. More general discussions can be found in Tierney et al. (1989). We may use this to obtain $E[X_t|Y_t]$ and $E[X_t^2|Y_t]$. Having these, we are ready to compute $\prod_{t=1}^{T} p(Y_t|Y_{t-1})$ for the canonical SVM.

5.1.1 Likelihood Computation

The computation is still under the filtering framework. Two steps involve the saddlepoint approximation. First is the saddlepoint approximation of the density $p(Y_t|Y_{t-1}) = \int p(Y_t|X_t, Y_{t-1})p_N(X_t|Y_{t-1})dX_t$, where $X_t|Y_{t-1}$ is Gaussian with known mean and variance. The second is the saddlepoint approximation of filtering, which computes $E[X_t|Y_t]$ and $E[X_t^2|Y_t]$. Then $X_{t+1}|X_t$ is approximated by Gaussian with mean $\phi E[X_t|Y_t]$ and variance $\phi^2 Var[X_t|Y_t] + \sigma^2$. The iteration then moves forward.

Theoretical and practical details about the saddlepoint approximation can be found in Schervish (1995). Several key steps and derivations are put in Appendix (A.2), as a self-contained reference. Our implementation can be summarized as follows, where the subscript N denotes a Gaussian density indicated with or without mean and variance. The likelihood components include

$$\begin{aligned} p(y_t|y_{t-1}) &= \int p(y_t|x_t)p(x_t|y_{t-1})dx_t \approx \int p(y_t|x_t)p_N(x_t|y_{t-1})dx_t \\ &= \int p(y_t|x_t)p_N(x_t|y_{t-1}; \phi \mathrm{E}[x_{t-1}|y_{t-1}], \phi^2 \mathrm{Var}[x_{t-1}|y_{t-1}] + \sigma^2)dx_t \\ &= \tilde{p}(y_t|y_{t-1}), \end{aligned}$$

where the filtering computes

$$p(x_{t-1}|y_{t-1}) \approx \frac{p(y_{t-1}|x_{t-1})p_N(x_{t-1}|y_{t-2})}{\int p(y_{t-1}|x_{t-1})p_N(x_{t-1}|y_{t-2})dx_{t-1}} = \frac{p(y_{t-1}|x_{t-1})p_N(x_{t-1}|y_{t-2})}{p(y_{t-1}|y_{t-2})}.$$

We compute $p(y_t|y_{t-1})$ and $\mathrm{M}_{x_{t-1}|y_{t-1}}(k) = \int e^{kx_{t-1}} p(x_{t-1}|y_{t-1})dx_{t-1}$, using the saddlepoint technique. Derivations are put in Appendix (A.3). The maximum likelihood estimator is

$$\arg\max_{\theta} \sum_{t=1}^{T} \log \tilde{p}(y_t|y_{t-1}, \theta).$$

5.1.2 Criticism

During our study, we found that Shimada and Tsukuda (2005) have studied the saddlepoint approach already. However, instead of computing posterior mean and variance rigorously, their implementation plugs in the Maximize-a-Posteriori (MAP) estimator. By such, their method removes the computation demand on the posterior mean and variance. Thus, their method is much faster. From the parameter inference point of view, that may cause little difference, since neither theirs nor ours enjoy the large sample property of the saddlepoint approximation, because we only have one observation at t. Therefore, a formal justification will be impossible, in comparing two approximations, or with simulated likelihood approaches. In the next section, we reveal a situation in which we can enjoy the asymptotic property.

5.2 A Factor Stochastic Volatility Model

In Chapter 3, we have discussed the reason that we are interested in the model described by Equation (2.26) and (2.27), which is as follows to facilitate discussions:

$$\begin{aligned} \vec{Y}_t &= \exp(X_t/2)\vec{Z}_t, \\ X_t &= \phi X_{t-1} + \eta_t, \end{aligned}$$

where $\vec{Z}_t \sim N_m(\vec{0}, \Sigma)$ and $\eta_t \sim N_1(0, \sigma_\eta^2)$ are mutually independent Gaussian white noises in vector and scaler forms. In order to make a maximum likelihood inference for the mul-

tivariate SVM, we need to compute the likelihood

$$\begin{aligned}
L(\theta|\vec{y}_{1:T}) &= p(\vec{y}_{1:T}|\theta) = \prod_{t=1}^{T} p(\vec{y}_t|\vec{y}_{t-1},\theta) \\
&= \prod_{t=1}^{T} \int p(\vec{y}_t|x_t,\theta) p(x_t|\vec{y}_{t-1},\theta) dx_t \\
&\approx \prod_{t=1}^{T} \int p(\vec{y}_t|x_t) p_N(x_t|\vec{y}_{t-1}) dx_t \\
&= \prod_{t=1}^{T} \int p(\vec{y}_t|x_t) p_N(x_t|\vec{y}_{t-1}, \phi \mathrm{E}[X_{t-1}|\vec{y}_{t-1}], \phi^2 \mathrm{Var}[X_{t-1}|\vec{y}_{t-1}] + \sigma_\eta^2) dx_t \\
&= \prod_{t=1}^{T} \tilde{p}(\vec{y}_t|\vec{y}_{t-1}).
\end{aligned}$$

Among above steps, assuming $X_t|\vec{y}_{t-1} \sim N(\mu, \sigma^2)$, we approximate $\mathrm{E}[X_t|\vec{y}_t]$ and $\mathrm{Var}[X_t|\vec{y}_t]$ as follows:

$$\begin{aligned}
\mathrm{E}[X_t|\vec{y}_t] \approx \widehat{\mu}_t &= \arg\max_{x_t} \left\{ -\frac{1+m}{2}\log(2\pi) - \frac{1}{2}\log\|\Sigma\| - \frac{1}{2}\log\sigma^2 \right. \\
&\qquad \left. -\frac{m x_t}{2} - \frac{\exp(-x_t)}{2} \vec{y}_t\,'\Sigma^{-1}\vec{y}_t - \frac{1}{2\sigma^2}(x_t-\mu)^2 \right\}, \\
\mathrm{Var}[X_t|\vec{y}_t] \approx \widehat{\sigma}_t^2 &= \left(\frac{\exp(-x_t)}{2} \vec{y}_t\,'\Sigma^{-1}\vec{y}_t + \frac{1}{\sigma^2} \right)^{-1} \Bigg|_{x_t=\widehat{\mu}},
\end{aligned}$$

for which derivations are in Appendix (A.4). They are indeed MAP estimators, which are more easy to obtain in general situations. The quality of this approximation, instead of rigorously computing the moment generating function, is reasonably good, which has been discussed in Schervish (1995). Using such a construction, we make inference about $(\Sigma, \phi, \sigma_\eta)$, by maximizing the approximated likelihood $\prod_{t=1}^{T-1} \tilde{p}(\vec{y}_t|\vec{y}_{t-1})$.

For maximum likelihood inference, it would always be wise to provide a reasonable initial guess about the parameter. It would be ideal if the initial guess is consistent. For the current situation, we adopt the following procedure to obtain initial values.

Procedure 5.2.1. *Obtain Initial Values*

(a) regard $\{\vec{y}_t\}_{t=1}^{T}$ *identical and independent, compute the sample covariance* Σ_y,

(b) provide an arbitrary, but reasonable, guess about (ϕ, σ_η),

(c) compute $\sigma_{ex}^2 = (E[\exp(X_t/2)|\phi, \sigma_\eta])^2$.

(d) compute $\Sigma = \Sigma_y/\sigma_{ex}^2$ *to approximate* $Var[\vec{Z}_t]$,

(e) get state estimates by filtering, using $(\Sigma, \phi, \sigma_\eta)$,

(f) update (ϕ, σ_η) *by fitting an AR1 model to the filtered states,*

(g) goto the step (c), and repeat this loop several times,

(h) use $(\Sigma, \phi, \sigma_\eta)$ *from the last updating step as an initial guess.*

Basically, the above procedure is a cheap analogue to the E-M iteration. The step (e) above is an analogue to the E-step in an E-M iteration, and the step (f) is an analogue to the M-step in an E-M iteration. The procedure is cheaper than a formal E-M, because we conduct filtering in step (e), rather than smoothing. In addition, $E[\exp(X_t/2)|\phi, \sigma_\eta]$ in step (c) is easy to compute due to the fact that $\exp(X_t/2)$ is log-normal.

5.3 Simulation Studies

In order to illustrate the performance of the proposed inferential method for this class of models. We conduct five simulation studies.

5.3.1 Simulation Study - 1

The purpose of this simulation study is to set up a similar setting to that in Jacquier et al. (1994). In their simulation study, the model is the canonical SVM, but in the equivalent form, which is described by Equation (4.5) and (4.6). We use **parameterization 1** to denote that parameterization. We actually conduct simulation by using the form (4.1) and (4.2), the parameterization of which is denoted by **parameterization 2**. Pitt and Shephard (1999a) discussed the convergence performance from a purely Bayesian

inference point view, and prefer **parameterization 1**. However, parameterization preference is considered differently here, because the methodology is different. More often, the preference here is only related to optimization performance, and possibly convergence rate. However, the convergence rate in optimization will heavily depend on individual optimization routines. For example, many local optimization routines perform well if the surface around the optimum is approximately quadratic, and perform poorly if the surface is close to singular. In our simulation, we adopt **parameterization 2**, which has more economic meaning; see Kim et al. (1998).

In the following, we briefly present how parameter values are chosen for the simulation. Details are in the original reference. First ϕ is predetermined to be 0.9, 0.95, or 0.98. The other two values are chosen, such that following equalities hold.

$$\begin{aligned} \exp\left\{2\log\beta + \frac{\sigma^2}{2(1-\phi^2)}\right\} &= \exp\left\{\frac{\alpha}{1-\phi} + \frac{\sigma^2}{2(1-\phi^2)}\right\} = 0.0009 \\ \exp\left\{\frac{\sigma^2}{1-\phi^2}\right\} - 1 &= 10, \text{ or } 1, \text{ or } 0.1. \end{aligned}$$

It can be recognized that the above quantities are related to a log-normal random variable with location parameter $\frac{\alpha}{1-\phi}$ and scale parameter $\frac{\sigma^2}{(1-\phi^2)}$, which are unconditional mean and variance of an AR1 process with intercept α, autocorrelation coefficient ϕ, and innovation variance σ^2.

The simulation in the original reference is for a univariate model, while we are interested in a multivariate model. Therefore, for each simulated state trajectory, we independently simulate m observed series. That means $\Sigma = \beta^2 I_m$ for $\vec{Z}_t$ in Equation (2.26), where I_m is an m-dimensional identity matrix. We choose m to be either 5 or 20. We also increase the number of simulated samples for each setting. The number is 1000 in our study, which was 500 in Jacquier et al. (1994). We also choose the length of series to be either $T = 500$ or $T = 1000$, in order to study the effects from the time series length.

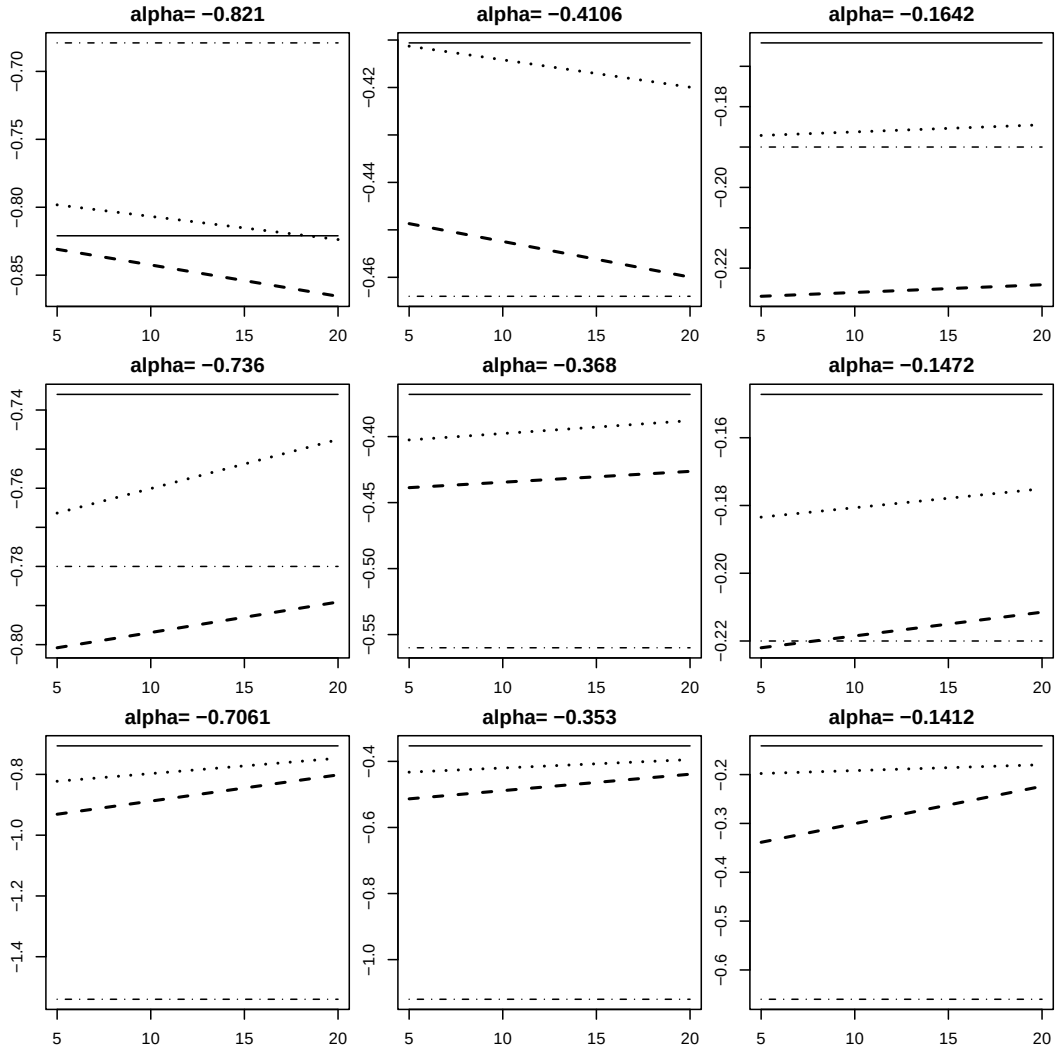

Figure 5.1: Estimates of α in Nine Scenarios. Solid horizontal lines indicate true values. Dash-dotted lines indicate Bayesian estimates. Dashed lines indicate $T = 500$. Dotted lines indicate $T = 1000$. The trend of dashed and dotted lines indicates the trend by increasing m from 5 to 20.

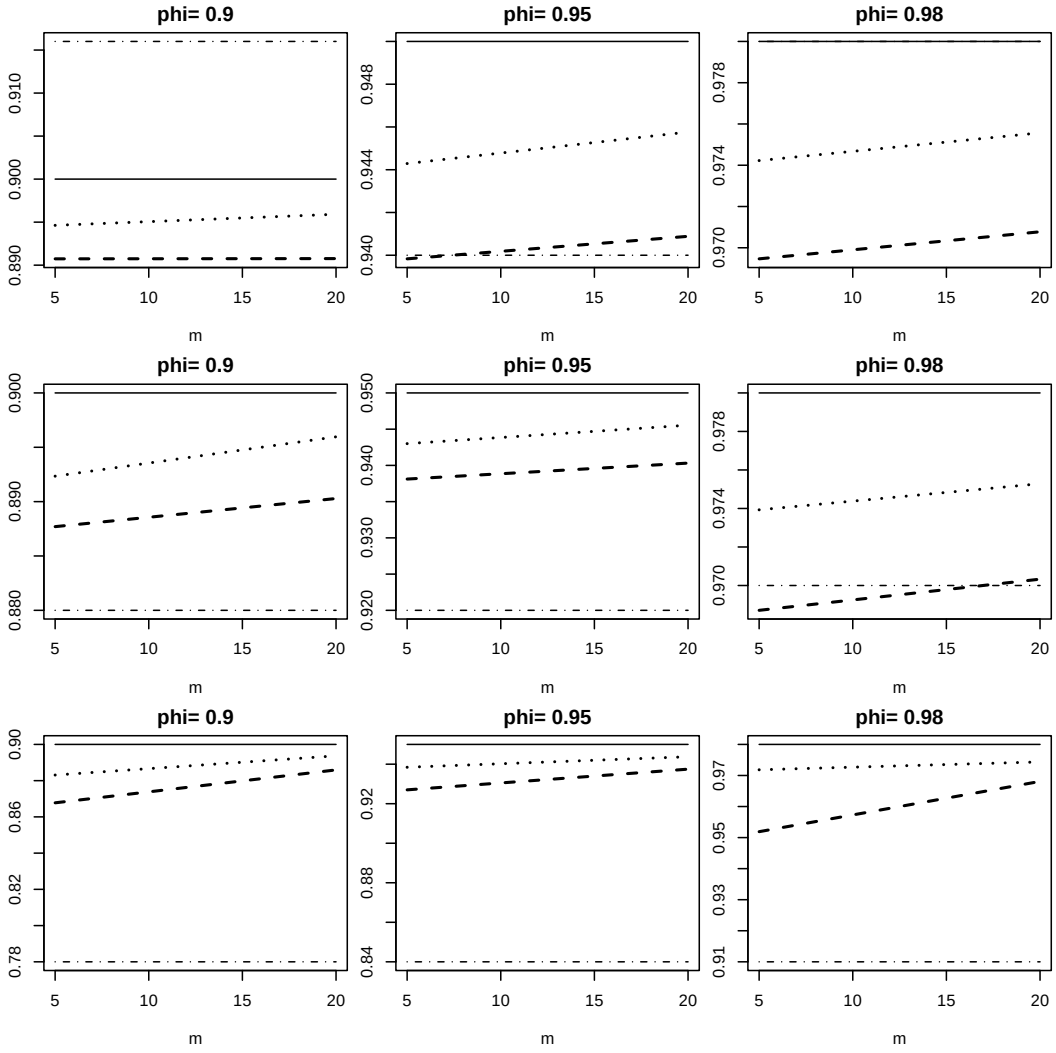

Figure 5.2: Estimates of ϕ in Nine Scenarios. Solid horizontal lines indicate true values. Dash-dotted lines indicate Bayesian estimates. Dashed lines indicate $T = 500$. Dotted lines indicate $T = 1000$. The trend of dashed and dotted lines indicates the trend by increasing m from 5 to 20.

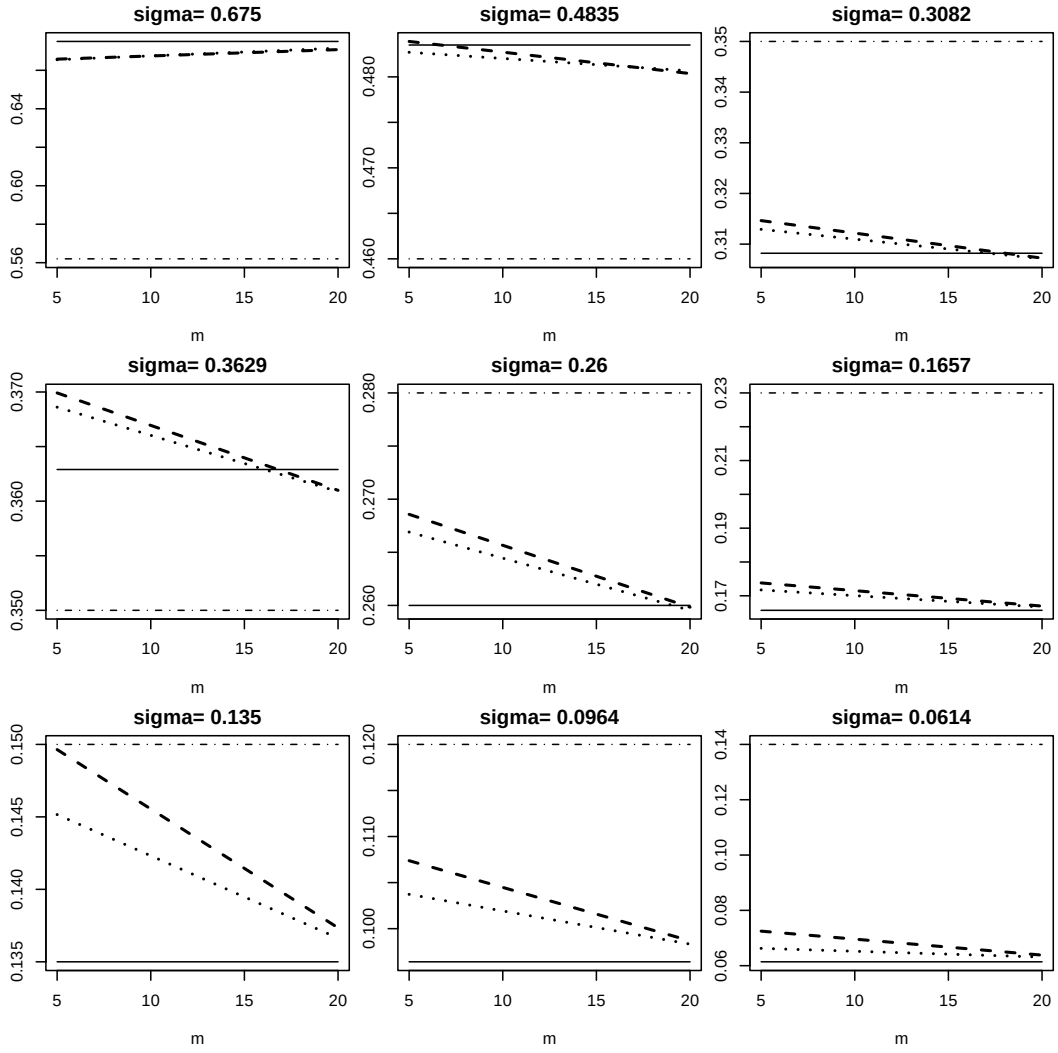

Figure 5.3: Estimates of σ in Nine Scenarios. Solid horizontal lines indicate true values. Dash-dotted lines indicate Bayesian estimates. Dashed lines indicate $T = 500$. Dotted lines indicate $T = 1000$. The trend of dashed and dotted lines indicates the trend by increasing m from 5 to 20.

Table 5.1: Simulation Results. **Parameterization 1**.

α	ϕ	σ	α	ϕ	σ	α	ϕ	σ
−0.821	0.900	0.675	−0.411	0.950	0.483	−0.164	0.980	0.308
−0.736	0.900	0.363	−0.368	0.950	0.260	−0.147	0.980	0.166
−0.706	0.900	0.135	−0.353	0.950	0.096	−0.141	0.980	0.061
				Estimates				
T=500	m=5							
−0.831	0.891	0.666	−0.449	0.940	0.484	−0.227	0.969	0.315
(0.197)	(0.025)	(0.044)	(0.151)	(0.020)	(0.036)	(0.127)	(0.017)	(0.030)
−0.801	0.888	0.370	−0.439	0.938	0.269	−0.222	0.969	0.174
(0.217)	(0.031)	(0.037)	(0.166)	(0.024)	(0.031)	(0.136)	(0.019)	(0.025)
−0.931	0.868	0.150	−0.513	0.927	0.107	−0.339	0.952	0.072
(0.628)	(0.089)	(0.045)	(0.460)	(0.065)	(0.035)	(0.724)	(0.103)	(0.032)
T=500	m=20							
−0.866	0.891	0.671	−0.460	0.941	0.480	−0.224	0.971	0.307
(0.190)	(0.024)	(0.029)	(0.148)	(0.019)	(0.024)	(0.118)	(0.015)	(0.020)
−0.789	0.890	0.361	−0.426	0.940	0.260	−0.211	0.970	0.167
(0.183)	(0.026)	(0.022)	(0.144)	(0.020)	(0.019)	(0.116)	(0.016)	(0.016)
−0.802	0.886	0.137	−0.439	0.937	0.099	−0.224	0.968	0.064
(0.254)	(0.036)	(0.019)	(0.190)	(0.027)	(0.016)	(0.152)	(0.022)	(0.012)
T=1000	m=5							
−0.798	0.895	0.666	−0.411	0.944	0.483	−0.187	0.974	0.313
(0.141)	(0.017)	(0.031)	(0.099)	(0.013)	(0.025)	(0.068)	(0.010)	(0.021)
−0.766	0.892	0.369	−0.403	0.943	0.267	−0.183	0.974	0.172
(0.149)	(0.021)	(0.026)	(0.106)	(0.015)	(0.022)	(0.074)	(0.011)	(0.018)
−0.823	0.883	0.145	−0.433	0.938	0.104	−0.198	0.972	0.066
(0.306)	(0.043)	(0.030)	(0.186)	(0.027)	(0.022)	(0.114)	(0.016)	(0.015)
T=1000	m=20							
−0.824	0.896	0.672	−0.420	0.946	0.481	−0.184	0.976	0.307
(0.126)	(0.016)	(0.021)	(0.093)	(0.012)	(0.018)	(0.066)	(0.009)	(0.014)
−0.748	0.896	0.361	−0.388	0.946	0.260	−0.175	0.975	0.167
(0.120)	(0.017)	(0.016)	(0.089)	(0.013)	(0.014)	(0.066)	(0.009)	(0.011)
−0.747	0.894	0.137	−0.395	0.944	0.098	−0.180	0.974	0.063
(0.161)	(0.023)	(0.013)	(0.115)	(0.016)	(0.010)	(0.080)	(0.011)	(0.008)

Table 5.2: Simulation Results. **Parameterization 2**.

β	ϕ	σ	β	ϕ	σ	β	ϕ	σ
0.016	0.900	0.675	0.016	0.950	0.483	0.016	0.980	0.308
0.025	0.900	0.363	0.025	0.950	0.260	0.025	0.980	0.166
0.029	0.900	0.135	0.029	0.950	0.096	0.029	0.980	0.061

Estimates

T=500	m=5							
0.023	0.891	0.666	0.025	0.940	0.484	0.028	0.969	0.315
(0.870)	(0.025)	(0.044)	(0.487)	(0.020)	(0.036)	(0.267)	(0.017)	(0.030)
0.029	0.888	0.370	0.030	0.938	0.269	0.030	0.969	0.174
(0.852)	(0.031)	(0.037)	(0.488)	(0.024)	(0.031)	(0.272)	(0.019)	(0.025)
0.030	0.868	0.150	0.030	0.927	0.107	0.030	0.952	0.072
(1.125)	(0.089)	(0.045)	(0.693)	(0.065)	(0.035)	(0.788)	(0.103)	(0.032)
T=500	m=20							
0.019	0.891	0.671	0.021	0.941	0.480	0.024	0.971	0.307
(0.901)	(0.024)	(0.029)	(0.496)	(0.019)	(0.024)	(0.261)	(0.015)	(0.020)
0.028	0.890	0.361	0.029	0.940	0.260	0.029	0.970	0.167
(0.833)	(0.026)	(0.022)	(0.470)	(0.020)	(0.019)	(0.256)	(0.016)	(0.016)
0.030	0.886	0.137	0.030	0.937	0.099	0.030	0.968	0.064
(0.864)	(0.036)	(0.019)	(0.498)	(0.027)	(0.016)	(0.283)	(0.022)	(0.012)
T=1000	m=5							
0.023	0.895	0.666	0.026	0.944	0.483	0.029	0.974	0.313
(0.827)	(0.017)	(0.031)	(0.439)	(0.013)	(0.025)	(0.214)	(0.010)	(0.021)
0.029	0.892	0.369	0.030	0.943	0.267	0.030	0.974	0.172
(0.805)	(0.021)	(0.026)	(0.439)	(0.015)	(0.022)	(0.218)	(0.011)	(0.018)
0.030	0.883	0.145	0.030	0.938	0.104	0.030	0.972	0.066
(0.898)	(0.043)	(0.030)	(0.492)	(0.027)	(0.022)	(0.248)	(0.016)	(0.015)
T=1000	m=20							
0.019	0.896	0.672	0.021	0.946	0.481	0.024	0.976	0.307
(0.850)	(0.016)	(0.021)	(0.446)	(0.012)	(0.018)	(0.211)	(0.009)	(0.014)
0.028	0.896	0.361	0.029	0.946	0.260	0.030	0.975	0.167
(0.782)	(0.017)	(0.016)	(0.422)	(0.013)	(0.014)	(0.209)	(0.009)	(0.011)
0.030	0.894	0.137	0.030	0.944	0.098	0.030	0.974	0.063
(0.792)	(0.023)	(0.013)	(0.437)	(0.016)	(0.010)	(0.221)	(0.011)	(0.008)

Simulation results are reported in Table (5.1) and (5.2) for both **parameterization 1** and **parameterization 2**. Each table has five portions. The first portion presents nine sets of parameters for data generation. The other four portions are corresponding to four combinations of T and m. The layout of each portion is similar to the ones in Jacquier et al. (1994). Each portion has nine groups, separated by horizontal and vertical lines. Each group is corresponding to a parameter setting in the first portion at the same group location. There are three columns in each group. Each column is corresponding to a parameter, which is denoted by the symbol at the top of the table. In each column, two numbers are the mean of estimates, and the square root of mean squared errors.

By comparing the first portion in Table (5.1) with Bayesian estimates in Table 9.5 in Jacquier et al. (1994), we can see that biases and root mean squared errors are generally smaller than those from a univariate model. The purpose of comparisons is not about different methodologies, but to see most estimates are improved by increasing either T or m, which means that more data become available. We re-run these simulations, but feed the alternative parameterized model to the optimizer. Summary results are the same. The results are more clear through Figure (5.1), (5.2), and (5.3).

In these plots, solid lines indicate the true values, dash-dotted lines indicate Bayesian estimates from Jacquier et al. (1994), dashed lines indicate increasing m from 5 to 20 while $T = 500$, and dotted lines indicate increasing m from 5 to 20 while $T = 1000$. From these plots, we can see, except for rare cases, dashed and dotted lines are closer to corresponding solid lines than dash-dotted lines; dotted lines are closer to solid lines than dashed lines; the right ends of dashed and dotted lines are closer to solid lines than their left ends. This presents the estimation improvement by increasing either m or T.

All the estimates for α in Table (5.1) are computed from the estimates in Table (5.2), by using the relationship that is described by Equation (4.7). We basically get the same information from both tables.

5.3.2 Simulation Study - 2

The first simulation study is to illustrate how estimates can be improved if multiple time series can be observed. All time series are mutually independent, and generated upon the same observation transition from the same state trajectory.

This simulation study will introduce correlations among observed time series. To do that, we assume a non-diagonal covariance matrix Σ for $\vec{Z}_t$ in Equation (2.26). More specifically, we are interested in two structures that have been discussed. We focus on the AR1 structure in this simulation study.

Due to lack of real data, we only create a set of arbitrary parameters. The multivariate stochastic model that we simulate data from is

$$\begin{aligned} \vec{Y}_t &= \exp(X_t/2)\vec{Z}_t, \\ X_t &= \phi X_{t-1} + \eta_t, \end{aligned}$$

where $\vec{Z}_t \sim N_m(\vec{0}, \Sigma)$ and $\eta_t \sim N_1(0, \sigma_\eta^2)$ are mutually independent Gaussian white noises in vector and scaler forms. Moreover, $[\Sigma]_{i,j} = \sigma_z^2 \rho^{|i-j|}$ is an AR1 structure covariance matrix with equal variance. The parameter values that are used in the simulation are $(\sigma_z = 0.6, \rho = 0.9, \phi = 0.95, \sigma_\eta = 0.16)$. For this set of parameters, we still simulate data with length $T = 500$ and $T = 1000$. For each series length, $m = 5$ and $m = 20$ are two scenario. By such, we can see how estimates are improved. Table (5.3) collects simulation results. Similarly, we generate Figure (5.3) to present the table graphically. Recall the SVM can be reparameterized to have a non-zero intercept state transition. In this simulation, the connection between two parameterizations is

$$\sigma_z = \exp\left\{\frac{\alpha}{2(1-\phi)}\right\},$$

in which α is the intercept in the state transition of the equivalent parameterization:

$$\begin{aligned} \vec{Y}_t &= \exp(X_t/2)\vec{Z}_t, \\ X_t &= \alpha + \phi X_{t-1} + \eta_t, \end{aligned}$$

where $\vec{Z}_t \sim N_m(\vec{0}, \Sigma)$ and $\eta_t \sim N_1(0, \sigma_\eta^2)$ are mutually independent Gaussian white noises in vector and scaler forms. Moreover, $[\Sigma]_{i,j} = \rho^{|i-j|}$ is an AR1 structure covariance matrix with variance 1 in the equivalent parameterization.

In Table (5.3), the true parameter values are in the first row, and α is computed due to the equivalence of reparameterization. Below the first row, every two rows present results of a simulation of size 1000, with a particular combination of time series length T and number of observed time series m. The first row of two presents the mean of parameter estimates. The second row presents the square root of mean squared errors. Except σ_z, the pattern in the results is similar to the one in the first simulation. Parameter estimates are less biased, if the time series length or the number of time series increases. In Figure (5.4), the pattern is more clear. The solid lines indicate true parameter values. Dashed lines indicate $T = 500$. Dotted lines indicate $T = 1000$. Left ends of lines are corresponding to $m = 5$. Right ends of lines are corresponding to $m = 20$. It is easy to see that dotted lines are closer to solid lines than dashed ones. Right ends of both dotted and dashed lines are more close to the solid lines than their left ends. At the present, we are not clear about why σ_z is different, although the equivalent parameter α follows the pattern. However, this might be an indicator that the alternative parameterization might be preferred in this situation, as well as in a purely Bayesian inference situation.

Table 5.3: Summary of Parameter Estimates in the Second Simulation. Model is a single multiplicative factor model. Observation innovations are multivariate Gaussian random variables with an AR1 correlation structure.

T	m	ϕ	σ	ρ	σ_z	α
		0.9500	0.1600	0.9000	0.6000	-0.0511
500	5	0.9358	0.1695	0.8997	0.6370	-0.0592
		(0.0293)	(0.0301)	(0.0062)	(0.0621)	(0.0293)
500	20	0.9393	0.1615	0.8999	0.6361	-0.0558
		(0.0224)	(0.0168)	(0.0039)	(0.0595)	(0.0224)
1000	5	0.9416	0.1675	0.8998	0.6383	-0.0531
		(0.0185)	(0.0217)	(0.0044)	0.(0530)	(0.0180)
1000	20	0.9449	0.1611	0.9000	0.6370	-0.0502
		(0.0139	(0.0116)	(0.0027)	(0.0502)	(0.0144)

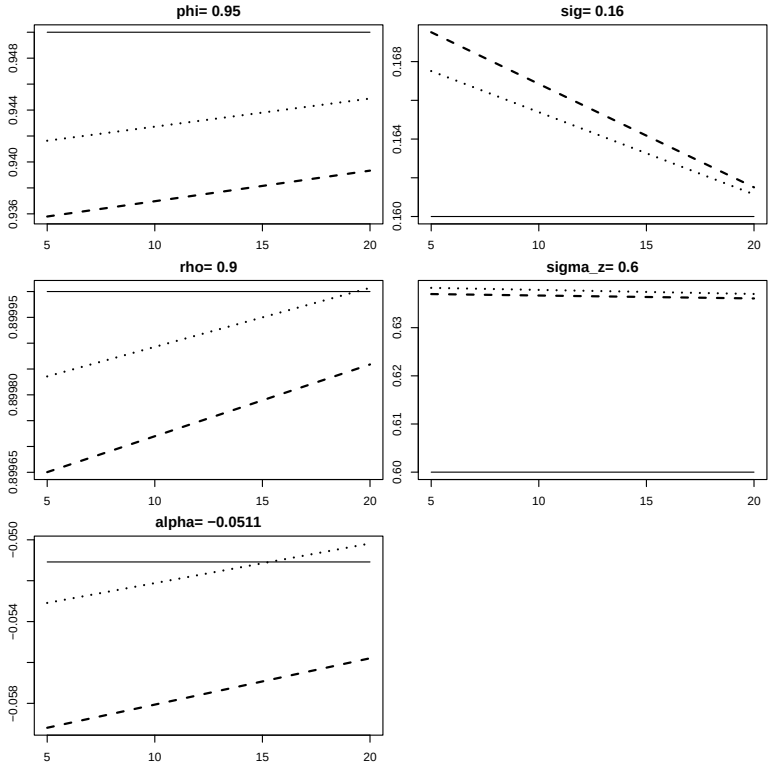

Figure 5.4: Summary of Parameter Estimates in the Second Simulation. Solid horizontal lines indicate true values. Dashed lines indicate $T = 500$. Dotted lines indicate $T = 1000$. The trend of dashed and dotted lines indicates the trend by increasing m from 5 to 20.

5.3.3 Simulation Study - 3

The third simulation study has same settings and scenario as those in the second simulation, except that the covariance matrix $[\Sigma]_{i,j} = \sigma_z^2 \rho^{|i-j|^2}$ has a Gaussian structure. The results are presented in the same way as those in the second simulation. In the simulation, we compute the inverse and determinant analytically. However, as m increases, the complexity increases, due to the Gaussian polynomial. This is the most time consuming simulation in our study, although a single fitting is tolerable. Table (5.4) and Figure (5.5) summarize the results. Results have a similar pattern to previous ones.

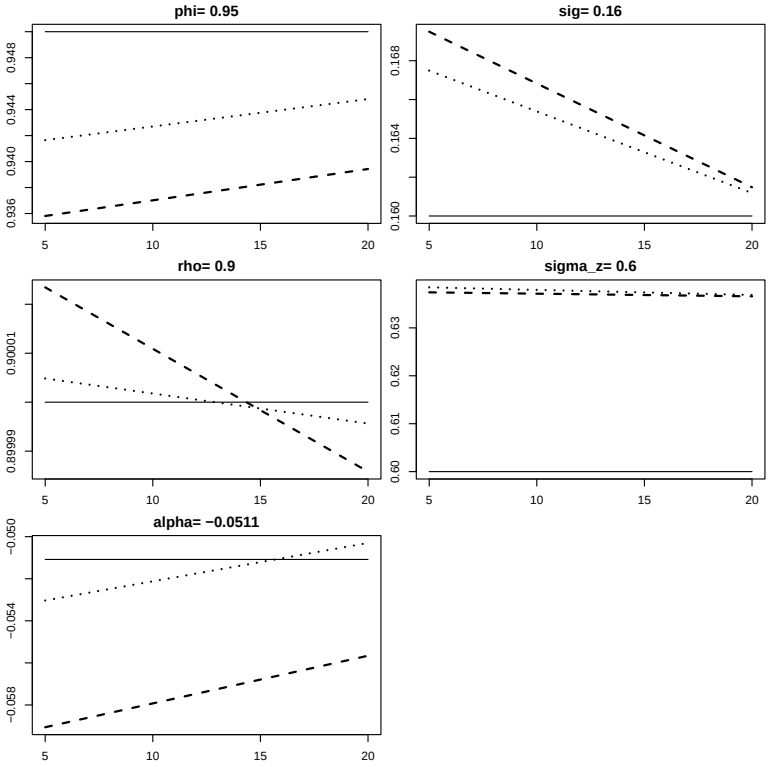

Figure 5.5: Summary of Parameter Estimates in the Third Simulation. Solid horizontal line indicates the true value. Dashed line indicates $T = 500$. Dotted line indicates $T = 1000$. The trend of dashed and dotted lines indicates the trend by increasing m from 5 to 20.

Table 5.4: Summary of Parameter Estimates in the Third Simulation. Model is a single multiplicative factor model. Observation innovations are multivariate Gaussian random variables with an Gaussian correlation structure.

T	m	ϕ	σ	ρ	σ_z	α
		0.9500	0.1600	0.9000	0.6000	−0.0511
500	5	0.9358	0.1695	0.9000	0.6374	−0.0591
		(0.0293)	(0.0301)	(0.0016)	(0.0608)	(0.0292)
500	20	0.9394	0.1615	0.9000	0.6366	−0.0557
		(0.0226)	(0.0169)	(0.0003)	(0.0598)	(0.0224)
1000	5	0.9416	0.1675	0.9000	0.6385	−0.0530
		(0.0185)	(0.0217)	(0.0012)	(0.0524)	(0.0179)
1000	20	0.9448	0.1612	0.9000	0.6369	−0.0503
		(0.0140)	(0.0117)	(0.0002)	(0.0503)	(0.0144)

5.3.4 Simulation Study for Small Sample Sizes

We have presented results regarding the asymptotic property. It is worth to study small sample size situations as well. Besides the fact that estimates will have large deviations, we find the existence of multimodal likelihood, if true values for ϕ, σ, T, and m all are relatively small. β is not relevant. To illustrate that, we simulate an example with $\phi = 0.2, \sigma = 0.3, \beta = 0.1, T = 500$, and $m = 1$. We search the log-likelihood surface, and identify a multimodal region what contains the maximum likelihood estimate. Figure (5.6) is the contour that presents two modes. The log-likelihood is apparently flat over a wide range of ϕ. We generate 1000 simulated samples. Estimates are presented in a scatter matrix plot in Figure (5.7). It reveals that estimates of ϕ and those of σ have a strange pattern, which implies that an optimizer is very likely to be trapped near the boundary of σ.

5.3.5 Filtering Examples

Previous simulations present results related to parameter inference. In this section, we present corresponding filtering results. The parameter settings are arbitrary. The states are generated from the following univariate AR1 process:

$$X_t = 0.98X_{t-1} + 0.16\eta_t,$$

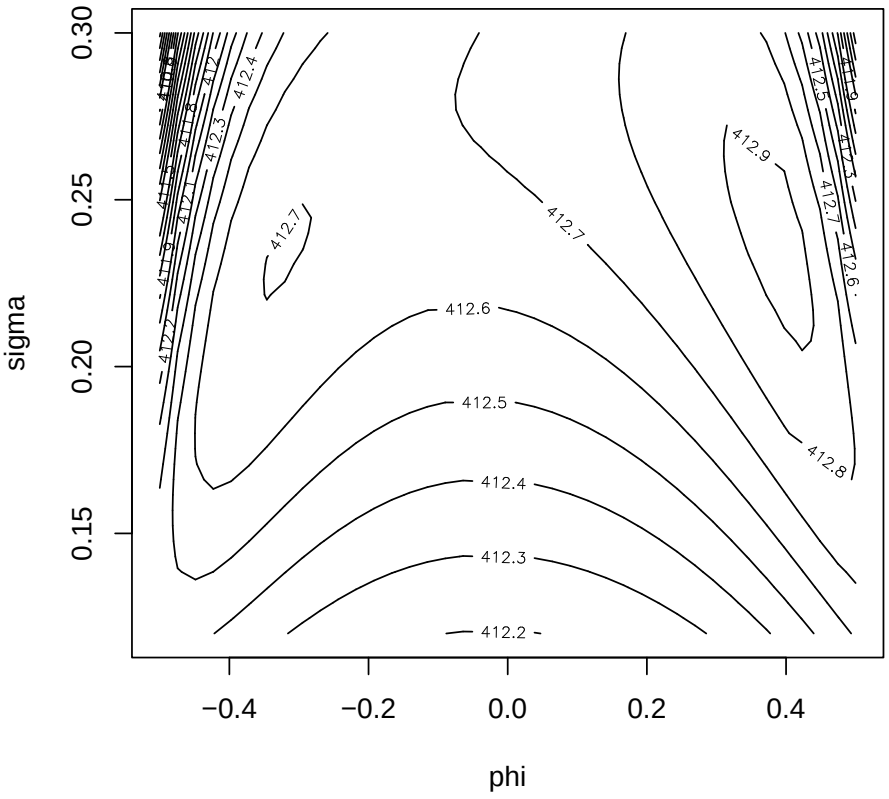

Figure 5.6: Multimodal Log-likelihood

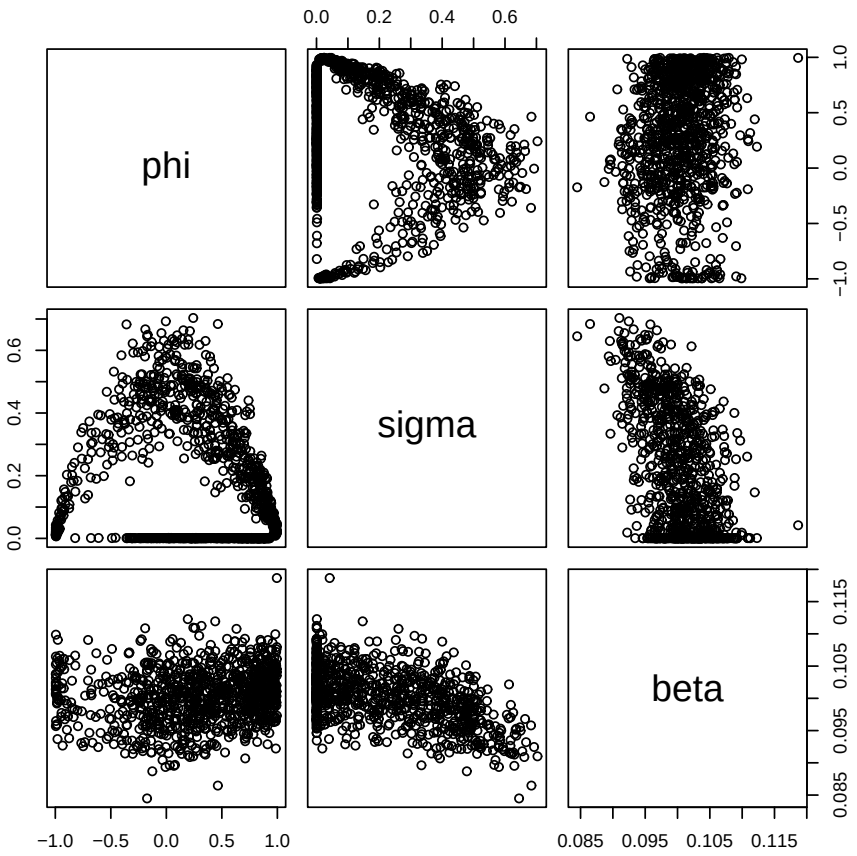

Figure 5.7: Estimates when Multi-modality Presents.

where $\eta_t \sim N(0,1)$ is a white noise. Observation transitions are

$$\vec{Y_t} = \exp(X_t/2)\vec{Z_t},$$

where $\vec{Z_t} \sim N(0, 0.36R))$, and R is a correlation matrix. We simulate three scenarios with three structures for R, which are identity, AR1 correlation matrix with $\rho = 0.9$, and Gaussian correlation matrix with $\rho = 0.9$. The dimension of R is either 5 or 20. In each scenario, we filter out the states, by using the true parameter values. Results are presented in Figure (5.8), (5.9), and (5.10).

Instead of plotting filtered states X_t, we plot the transformed states in terms of volatilities $\exp(X_t/2)$, which is of interest in financial literature. Figure (5.8) clearly presents that the filtered volatilities are close the the true volatilities. Although it is hard to tell which filtered volatility trajectory is closer to the true one, the one from $m = 20$ has narrower confidence intervals, which is what we expect. Meanwhile, Figure (5.9) and (5.10) present similar results, but the confidence intervals are wider. This is understood, because observed time series in these two scenarios are positively correlated. In addition, instead of assuming we know the true parameter values, Figure (5.11), (5.12), and (5.13) present respective filtering results with the estimated parameter values. Results are similar.

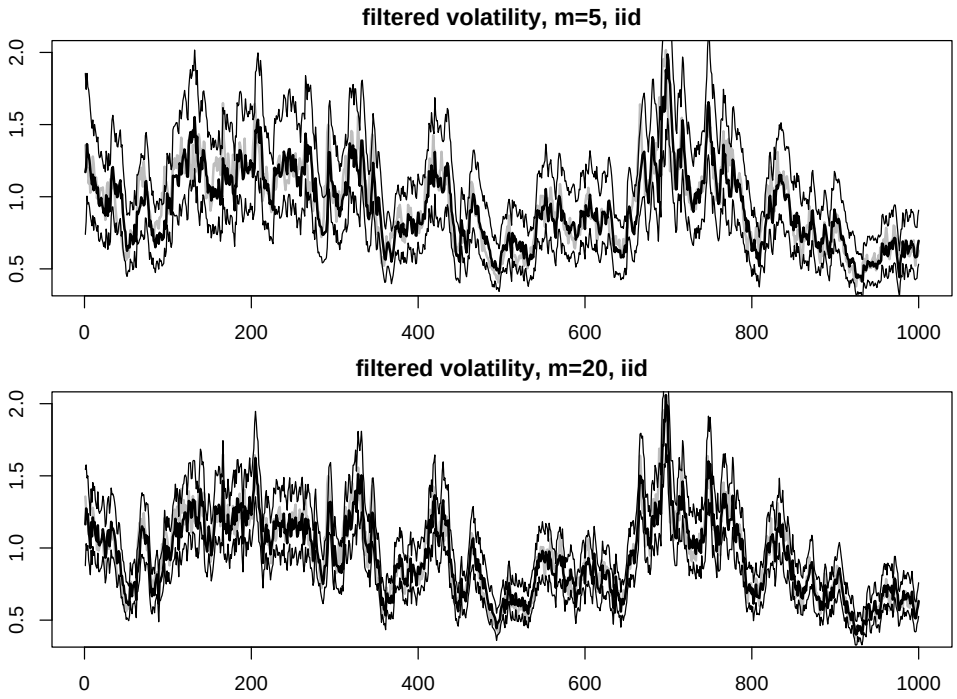

Figure 5.8: Filtered Volatilities from Different Dimensional Time Series, with an Identity Correlation Matrix. The upper portion presents filtered volatilities from a 5-dimensional time series. The true volatilities are in gray color. Dark solid line denotes filtered volatilities. Upper and lower thin lines denote 95% confidence interval for individual filtered volatilities. The bottom portion presents those from the 20-dimensional time series.

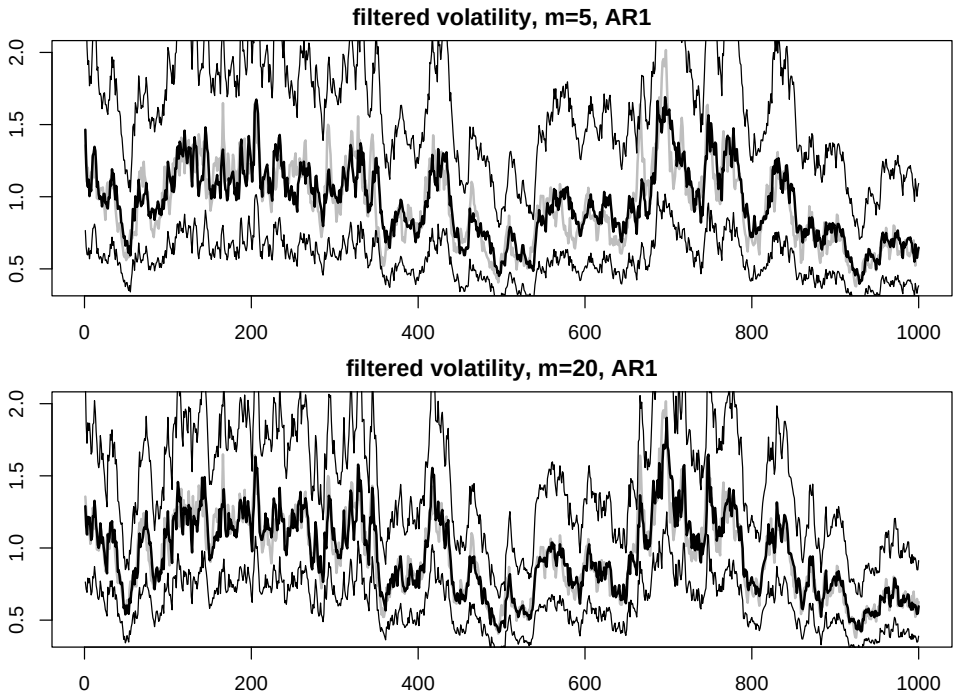

Figure 5.9: Filtered Volatilities from Different Dimensional Time Series, with an AR1 Correlation Matrix. The upper portion presents filtered volatilities from a 5-dimensional time series. The true volatilities are in gray color. Dark solid line denotes filtered volatilities. Upper and lower thin lines denote 95% confidence interval for individual filtered volatilities. The bottom portion presents those from the 20-dimensional time series.

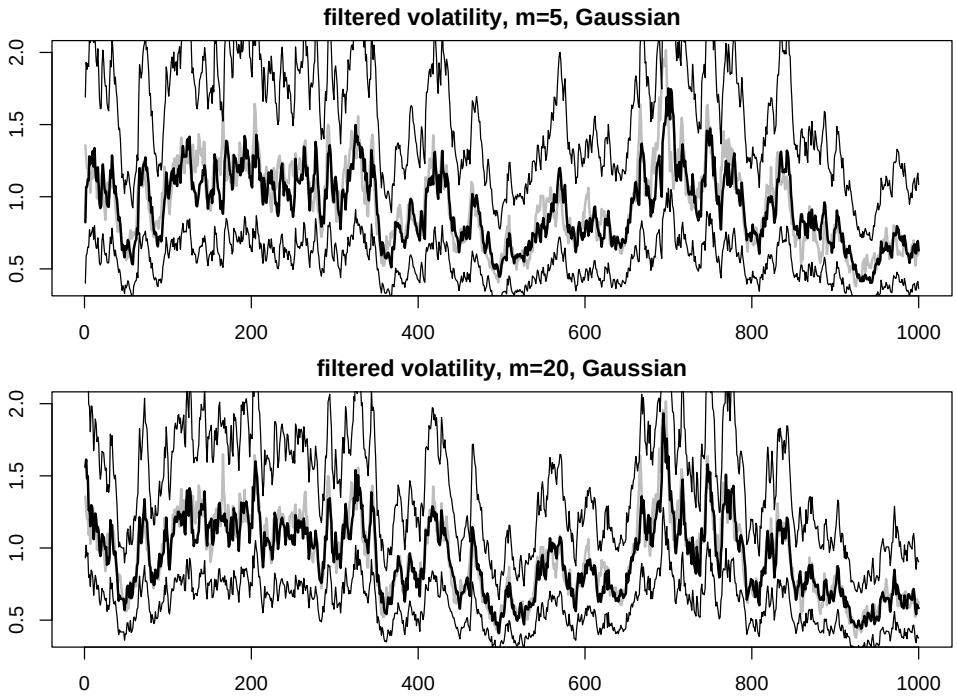

Figure 5.10: Filtered Volatilities from Different Dimensional Time Series, with a Gaussian Correlation Matrix. The upper portion presents filtered volatilities from a 5-dimensional time series. The true volatilities are in gray color. Dark solid line denotes filtered volatilities. Upper and lower thin lines denote 95% confidence interval for individual filtered volatilities. The bottom portion presents those from the 20-dimensional time series.

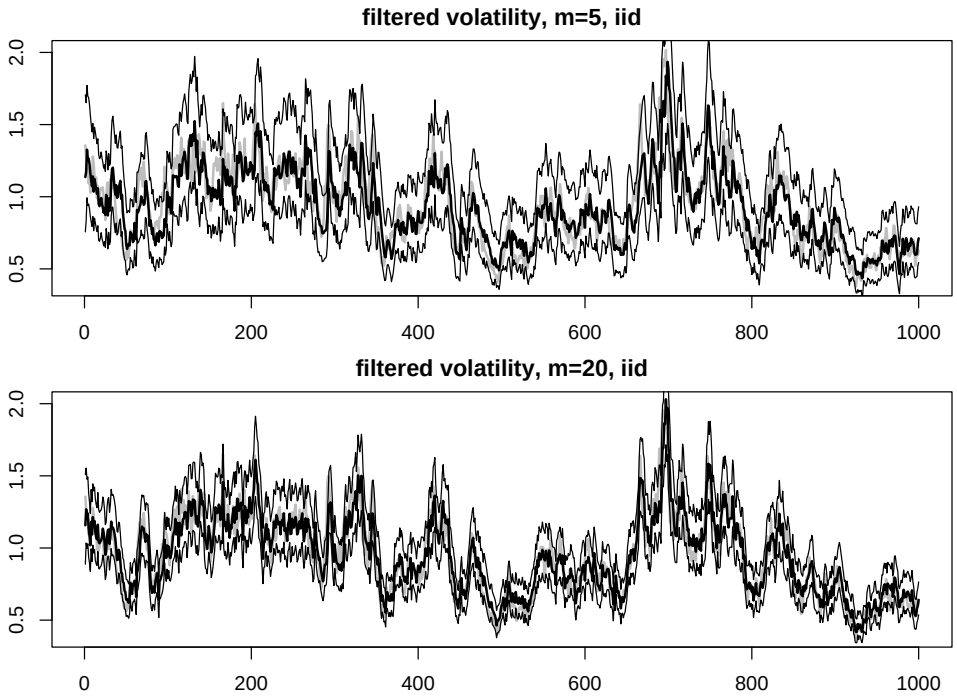

Figure 5.11: Filtered Volatilities (using estimates) from Different Dimensional Time Series, with an Identity Correlation Matrix. The upper portion presents filtered volatilities from a 5-dimensional time series. The true volatilities are in gray color. Dark solid line denotes filtered volatilities. Upper and lower thin lines denote 95% confidence interval for individual filtered volatilities. The bottom portion presents those from the 20-dimensional time series.

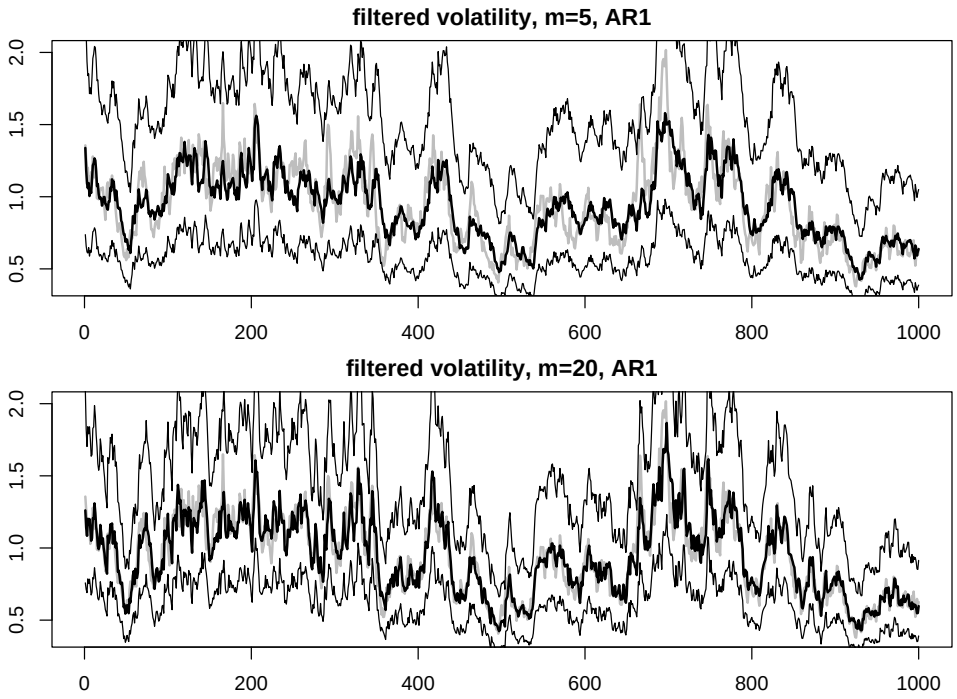

Figure 5.12: Filtered Volatilities (using estimates) from Different Dimensional Time Series, with an AR1 Correlation Matrix. The upper portion presents filtered volatilities from a 5-dimensional time series. The true volatilities are in gray color. Dark solid line denotes filtered volatilities. Upper and lower thin lines denote 95% confidence interval for individual filtered volatilities. The bottom portion presents those from the 20-dimensional time series.

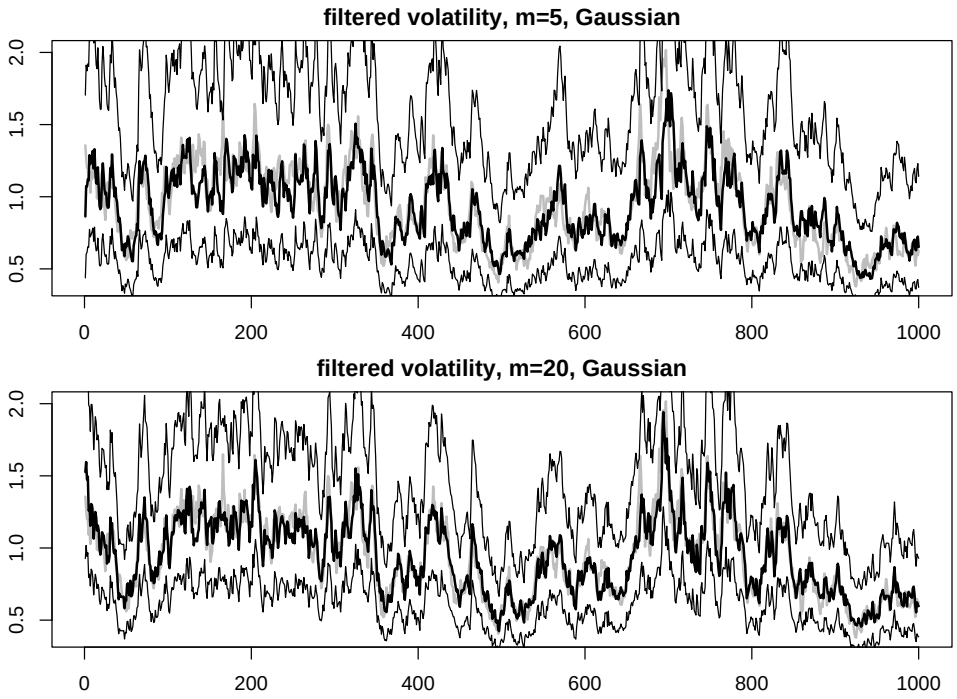

Figure 5.13: Filtered Volatilities (using estimates) from Different Dimensional Time Series, with a Gaussian Correlation Matrix. The upper portion presents filtered volatilities from a 5-dimensional time series. The true volatilities are in gray color. Dark solid line denotes filtered volatilities. Upper and lower thin lines denote 95% confidence interval for individual filtered volatilities. The bottom portion presents those from the 20-dimensional time series.

Chapter 6

Conclusion and Discussion

6.1 Conclusion

Motivated by proposing a suitable multivariate SVM for the term structure of interest rates, we are in favor of a low state dimensional multivariate SVM. We point out that multivariate Gaussian innovations must be more carefully studied, instead of being prescribed just for the sake of mathematical convenience. After reviewing the current difficulties in SVM inference literature, we first propose a systematic procedure, in order to guide the simulated maximum likelihood inference. We also realize that all the difficulties are from limited knowledge about individual states, which has no hope to be improved if no additional information is obtainable. In situations that state and parameter inference are equally important, existing approaches can hardly be satisfactory. On the contrary, the multivariate SVM that we are interested in can provide more information about individual states when the dimension of the observed time series increases. In addition to that, such property leads to a computationally inexpensive inference approach, which is the saddlepoint-approximated-likelihood inference. Simulation results confirm that the analytic approximation improves both state inference and parameter inference in most of the cases, when the dimension or the time series length increase.

6.2 Discussion

This research opens a door to several possible future directions. First, we have only considered a one factor model, it would be desirable if multiple factors can be modeled. Or at least additional noise can be attached. For example, consider the following example:

$$\begin{aligned} \vec{Y}_t &= \exp(X_t/2)\vec{Z}_t + \vec{e}_t, \\ X_t &= \phi X_{t-1} + \eta_t, \end{aligned}$$

which has an extra multivariate Gaussian innovation $\vec{e}_t$, in comparing with Equation (2.26). Though this model still has one stochastic volatility factor, the instantaneous correlation matrix of $\vec{Y}_t$ is not constant, which is very desirable in some financial applications. To see that, suppose the covariance matrix for $\vec{Z}_t$ is Σ, that for $\vec{e}_t$ is Λ, and the correlation matrix for $\vec{Y}_t$ is R, then the $(i,j)^{th}$ entry of R is

$$R_{ij} = \frac{\exp(X_t)\Sigma_{ij} + \Lambda_{ij}}{\sqrt{\exp(X_t)\Sigma_{ii} + \Lambda_{ii}}\sqrt{\exp(X_t)\Sigma_{jj} + \Lambda_{jj}}},$$

which is not a constant. A multiple factor model will have dynamic correlation as well, however more latent processes must be introduced, and identifiability issues will arise. Besides multiplicative factor models, additive factor models seem more familiar in the literature. The saddlepoint approach may also apply to the additive class. Besides the above extensions, in order to get a more theoretical justification, we would like to start the study on the convergence rate of the saddlepoint approximated log-likelihood to the true log-likelihood.

Bibliography

Albeverio, S., Lytvynov, E., and Mahnig, A. (2004). A Model of the Term Structure of Interest Rates Based on Levy Fields. *Stochastic Processes and Their Applications*, **114**: 251–263.

Baaquie, B. E. (2001). Quantum Field Theory of Treasury Bonds. *Physical Review E*, **64**: 016121.

Baaquie, B. E. (2002). Quantum Field Theory of Forward Rates with Stochastic Volatility. *Physical Review E*, **65**: 056122.

Baaquie, B. E. and Srikant, M. (2004). Finite Hedging in Field Theory Models of Interest Rates. *Physical Review E*, **69**: 036130.

Bester, C. A. (2004). Random Field and Affine Models for Interest Rates: An Empirical Comparison. Working paper.

Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy*, **81**: 637–654.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**: 307–327.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control.* Prentice Hall, 3rd. edition.

Brace, A., Gatarek, D., and Musiela, M. (1997). The Market Model of Interest Rate Dynamics. *Mathematical Finance*, **7**(2): 127–147.

Brigo, D. and Mercurio, F. (2001). *Interest Rate Models: Theory and Practice*. Springer.

Daniels, H. E. (1954). Saddlepoint Approximations in Statistics. *Annals of Mathematical Statistics*, **25**: 631–650.

Doucet, A., de Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Durham, G. B., Gallant, A. R., Ait-Sahalia, Y., and Brandt, M. W. (2002). Numerical Techniques for Maximum Likelihood Estimation of Continuous-time Diffusion Processes. *Journal of Business & Economic Statistics*, **20**: 297–316.

Efron, B. (1979). Bootstrap Methods: Another Look At The Jackknife. *Annals of Statistics*, **7**: 1–26.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation. *Econometrica*, **50**: 987–1008.

Fabozzi, F. J., editor (1997). *The Handbook of Fixed Income Securities*. Irwin.

Fouque, J.-P., Papanicolaou, G., and Sircar, K. R. (2000). *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press.

Gall, J., pap, G., and van Zuijlen, M. C. A. (2004). Maximum Likelihood Estimator of The Volatility of Forward Rates Driven by Geometric Spatial AR Sheet. *Journal of Applied Mathematics*, **4**: 293–309.

Gall, J., pap, G., and van Zuijlen, M. C. A. (2006). Forward Interest Rate Curves In Discrete Time Settings Driven By Random Fields. *Computers & Mathematics With Applications*, **3-4**: 387–396.

Gallant, A. R., Hsieh, D., and Tauchen, G. (1997). Estimation of Stochastic Volatility Models with Diagnostics. *Journal of Econometrics*, **81**: 159–192.

Gallant, A. R. and Tauchen, G. (1996). Which Moments to Match? *Econometric Theory*, **12**: 657–681.

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

Goldstein, R. S. (2000). The Term Structure of Interest Rates as a Random Field. *The Review of Financial Studies*, **13**(2): 365–384.

Gordon, N. (1993). Novel Approach to Nonlinear and Non-Gaussian Bayesian State Estimation. *Proceedings IEE-F*, **140**: 107–113.

Goutis, C. and Casella, G. (1999). Explaining the Saddlepoint Approximation. *The American Statistician*, **53**(3): 216–224.

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate Stochastic Variance Models. *Review of Economic Studies*, **61**(2): 247–264.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Heath, D., Jarrow, R., and Morton, A. (1992). Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation. *Econometrica*, **60**: 77–105.

Hull, J. (2002). *Options, Futures and Other Derivatives*. Prentice Hall.

Hürzeler, M. and Künsch, H. R. (2001). *Sequential Monte Carlo Methods in Practice*, chapter Approximating and Maximising the Likelihood for a General State-Space Model, pages 159–175. Springer.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian Analysis of Stochastic Volatility Models. *Journal of Business and Economic Statistics*, **12**(4): 371–417.

James, J. and Webber, N. (2000). *Interest Rate Modelling*. Wiley.

Jarrow, R. A. (1996). *Modelling Fixed Income Securities and Interest Rate Options*. McGraw-Hill.

Jones, R. H. (1980). Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. *Technometrics*, **22**(3): 389–395.

Kalman, R. E. (1960). Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. *Transactions of the ASME - Journal of Basic Engineering, Series D*, **82**: 35–45.

Kelley, C. T. (1995). *Iterative Methods for Linear and Nonlinear Equations*. SIAM.

Kelley, C. T. (1999). *Iterative Methods for Optimization*. SIAM.

Kennedy, D. P. (1994). The Term Structure of Interest Rates as a Gaussian Random Field. *Mathematical finance*, **4**: 247–258.

Kennedy, D. P. (1997). Characterizing Gaussian Models of the Term Structure of Interest Rates. *Mathematical Finance*, **7**(2): 107–116.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies*, **65**(3): 361–393.

Kimmel, R. L. (2004). Modeling the Term Structure of Interest Rates: A New Approach. *Journal of Financial Economics*, **72**: 143–183.

Loh, W.-L. and Lam, T.-K. (2000). Estimating Structured Correlation Matrices in Smooth Gaussian Random Field Models. *The Annals of Statistics*, **28**: 880–904.

Longstaff, F. and Schwartz, E. (2001). Valuing American Options By Simulation: A Simple Least-Squares Approach. *Review of Financial Studies*, **14**: 113–147.

Longstaff, F. A., Santa-Clara, P., and Schwartz, E. S. (2001). The Relative Valuation of Caps and Swaptions: Theory and Empirical Evidence. *The Journal of Finance*, **LVI**(6): 2067–2109.

Masreliez, C. J. (1975). Approximate Non-Gaussian Filtering With Linear State And Observation Relations. *IEEE transactions on automatic control*, **20**(1): 107–110.

Merton, R. C. (1990). *Continuous-Time Finance.* Basil Blackwell.

Musiela, M. and Rutkowski, M. (1997). *Martingale Methods in Financial Modelling.* Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer.

Nelder, J. A. and Mead, R. (1965). A Simplex Algorithm for Function Minimization. *Computer Journal*, **7**: 308–313.

Oksendal, B. (1995). *Stochastic Differential Equations: An Introduction with Applications.* Springer.

Pang, K. (1999). Calibration of Gaussian Heath, Jarrow and Morton and Random Field Interest Rate Term Structure Models. *Review of Derivatices Research*, **2**: 15–345.

Pitt, M. K. and Shephard, N. (1999a). Analytic Convergence Rates and Parameterization Issues for the Gibbs Sampler Applied to State Space Models. *Journal of Time Series Analysis*, **20**(1): 63–85.

Pitt, M. K. and Shephard, N. (1999b). Time-Varying Covariances: A Factor Stochastic Volatility Approach. *Bayesian Statistics*, **6**: 547–570.

Pourahmadi, M. (1999). Joint Mean-covariance Models with Applications to Longitudinal Data. *Biometrika*, **86**: 677–690.

Rebonato, R. (2002). *Modern Pricing of Interest-Rate Derivatives.* Princeton university Press.

Ripley, B. D. (1987). *Stochastic Simulation.* John Wiley & Sons.

Rutkowski, M. (1998). Dynamics of Spot, Forward, and Futures Libor Rates. *International Journal of Theoretical and Applied Finance*, **1**(3): 425–445.

Santa-Clara, P. (2001). The Dynamics of the Forward Interest Rate Curve with Stochastic String Shocks. *The Review of Financial Studies*, **14**(1): 149–185.

Schervish, M. J. (1995). *Theory of Statistics*. Springer.

Shephard, N., editor (2005). *Stochastic Volatility: Selected Readings*. Oxford.

Shimada, J. and Tsukuda, Y. (2005). Estimation of Stochastic Volatility Models: An Approximation to the Nonlinear State Space Representation. *Communications in Statistics – Simulation and Computation*, **34**: 429–450.

Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian Statistics without Tears: A Sampling-Resampling Perspective. *The American Statistician*, **46**(2): 84–88.

Stein, M. L. (1999). *Interpolation of Spatial Data*. Springer.

Sundaresan, S. M. (2000). Continuous-Time Methods in Finance: A Review and an Assessment. *The Journal of Finance*, **55**(4): 1569–1622.

Taylor, S. J. (1980). Conjectured Models for Trends in Financial Prices, Tests and Forecasts. *Journal of the Royal Statistical Society, Series A*, **143**: 338–362.

Taylor, S. J. (1982). *Time Series Analysis: Theory and Practice*, chapter Financial Returns Modelled by the Product of Two Stochastic Processes - A Study of Daily Sugar Prices 1961-79, pages 203–226. North-Holland.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions. *Journal of the American Statistical Association*, **84**(407): 710–716.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Vasicek, O. (1977). An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics*, **5**: 177–188.

von Neumann, J. (1951). Various Techniques in Connection with Random Digits. *NBS Appl. Math. Ser.*, **12**: 36–38.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions I.* Springer.

Appendix

Appendix A

Derivations and Proofs

A.1 Derivations for Equality 4.3 and 4.4

The following requires the knowledge of stochastic differential equations (SDE). An accessible reference is Oksendal (1995). The definition of Ornstein-Uhlenbeck processes in terms of stochastic differential equation and Ito's integral is as follows.

$$dX_t = -\theta(X_t - \mu)dt + \sigma dW_t,\ \theta > 0, \tag{A.1}$$

which is the solution to

$$X_t = X_0 e^{-\theta t} + \mu(1 - e^{-\theta t}) + \int_0^t \sigma e^{\theta(s-t)} dW_s,\ \theta > 0. \tag{A.2}$$

Step-by-step, the solution is obtained as follows. First, multiply both sides of Equation (A.1) by $e^{\theta t}$, and get:

$$e^{\theta t} dX_t = -\theta e^{\theta t}(X_t - \mu)dt + \sigma e^{\theta t} dW_t. \tag{A.3}$$

Compute $d(e^{\theta t} X_t)$ by using Ito formula, and replace $e^{\theta t} dX_t$ by the right hand side of Equation (A.3). We get:

$$d(e^{\theta t} X_t) = \theta e^{\theta t} X_t dt + e^{\theta t} dX_t \tag{A.4}$$

$$= \theta e^{\theta t} X_t dt - \theta e^{\theta t}(X_t - \mu)dt + \sigma e^{\theta t} dW_t \tag{A.5}$$

$$= \mu\theta e^{\theta t} dt + \sigma e^{\theta t} dW_t. \tag{A.6}$$

Therefore

$$e^{\theta t}X_t = e^{\theta \cdot 0}X_0 + \int_0^t \mu\theta e^{\theta s}ds + \int_0^t \sigma e^{\theta s}dW_s \quad (A.7)$$

$$= X_0 + \mu(e^{\theta t} - 1) + \int_0^t \sigma e^{\theta s}dW_s \quad (A.8)$$

Dividing both sides by $e^{\theta t}$, we get Equation (A.2).

In particular, we are interested in

$$X_t = \mu + \int_{-\infty}^t \sigma e^{\theta(s-t)}dW_s,\ \theta > 0,$$

for t is sufficiently large, which is equivalent to say the initial value has no long term effect on the process. Without loss of generality, we can assume $\mu = 0$, and for $t_1 \leq t_2$,

$$\begin{aligned}
Cov(X_{t_1}, X_{t_2}) &= E\left[\int_{-\infty}^{t_1} \sigma e^{\theta(s-t_1)}dW_s \int_{-\infty}^{t_2} \sigma e^{\theta(u-t_2)}dW_u\right] \\
&= E\left[\int_{-\infty}^{t_1} \sigma e^{\theta(s-t_1)}dW_s \int_{-\infty}^{t_1} \sigma e^{\theta(u-t_2)}dW_u\right. \\
&\quad \left. + \int_{-\infty}^{t_1} \sigma e^{\theta(s-t_1)}dW_s \int_{t_1}^{t_2} \sigma e^{\theta(u-t_2)}dW_u\right] \\
&= E\left[\int_{-\infty}^{t_1} \sigma e^{\theta(s-t_1)}dW_s \int_{-\infty}^{t_1} \sigma e^{\theta(u-t_2)}dW_u\right] \\
&= \int_{-\infty}^{t_1} \sigma^2 e^{\theta(2s-t_1-t_2)}ds \\
&= \frac{\sigma^2}{2\theta}e^{-\theta|t_2-t_1|}
\end{aligned}$$

The correlation is then $e^{-\theta|t_2-t_1|}$. And variance is $\frac{\sigma^2}{2\theta}$. Assigning $|t_2 - t_1| = 1$, we get Equation (4.3). By equality

$$\frac{1}{1-\phi^2}\sigma_\epsilon^2 = \frac{1}{2\theta}\sigma_\omega^2,$$

we get the equality (4.4).

A.2 Key Steps in the Saddlepoint Approximation

This is a self-contained reference due to Goutis and Casella (1999). The task of a saddlepoint approximation is to compute

$$\int_A f(x)dx,$$

where $f(x)$ is positive. Compute the first order approximation, by expanding $h(x) = \log f(x)$ in Taylor's series about x_0,

$$f(x) \approx \exp\left\{h(x_0) + (x - x_0)h'(x_0) + \frac{(x - x_0)^2}{2}h''(x_0)\right\}.$$

By choosing $x_0 = \hat{x}$, where $h(x)$ is maximized, we can eliminate the linear term. And

$$\int_A f(x)dx \approx \exp\{h(\hat{x})\}\left(-\frac{2\pi}{h''(\hat{x})}\right)^{1/2}\int_A \phi\left(x, \hat{x}, -\frac{1}{h''(\hat{x})}\right)dx,$$

where the integrand of the last integral is a Gaussian density function. For $A = \mathbb{R}$, the last integral is 1. The quality of the approximation depends on the quadratic approximation of $h(x)$ around $\hat{x}$, which is usually satisfactory if $f(x)$ is a likelihood function. Thus $\hat{x}$ is the maximum likelihood estimate. Well behaved loglikelihood functions $h(x)$ are known to be approximately quadratic at its local maximum.

A.3 Derivation of Saddlepoint Approximation

Assume $x \sim N(\mu, \sigma^2)$ and $p(y|x) \sim N(0, \beta^2 \exp(x))$. The task is to get $p(x|y)$, but eventually $E(x|y)$, and $\mathrm{Var}(x|y)$, using saddlepoint approximation. We also need $p(y)$ for the likelihood. The notations used in the next subsections follow Goutis and Casella (1999). At the present, we only conduct the first order approximation to achieve efficiency.

A.3.1 $p(y)$

$$\begin{aligned}
p(y) &= \int p(y|x)p(x)dx, \\
p(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\
p(y|x) &= \frac{1}{\sqrt{2\pi}\beta \exp\{x/2\}} \exp\left\{-\frac{y^2}{2\beta^2 \exp\{x\}}\right\} \\
&= \frac{1}{\sqrt{2\pi}\beta} \exp\left\{-\frac{y^2}{2\beta^2 \exp\{x\}} - \frac{x}{2}\right\}
\end{aligned}$$

Let

$$\begin{aligned}
m_1(y,x) &= p(y|x)p(x) \\
&= \frac{1}{2\pi\beta\sigma} \exp\left\{-\frac{y^2}{2\beta^2 \exp\{x\}} - \frac{x}{2} - \frac{(x-\mu)^2}{2\sigma^2}\right\}, \\
k_1(y,x) &= \log m_1(y,x) \\
&= -\log(2\pi\beta\sigma) - \frac{y^2 \exp\{-x\}}{2\beta^2} - \frac{x}{2} - \frac{(x-\mu)^2}{2\sigma^2}.
\end{aligned}$$

Then

$$\begin{aligned}
\frac{\partial k_1(y,x)}{\partial x} &= \frac{y^2 \exp(-x)}{2\beta^2} - \frac{1}{2} - \frac{x-\mu}{\sigma^2} \\
\frac{\partial^2 k_1(y,x)}{\partial x^2} &= -\frac{y^2 \exp(-x)}{2\beta^2} - \frac{1}{\sigma^2},
\end{aligned}$$

and

$$\begin{aligned} p(y) &\approx \int \exp\left\{ k_1(y,\hat{x}(y) + \frac{(x-\hat{x}(y))^2}{2} \frac{\partial^2 k_1(y,x)}{\partial x^2}\Big|_{\hat{x}(y)} \right\} dx \\ &= \exp\{k_1(y,\hat{x}(y))\} \left(-\frac{2\pi}{\frac{\partial^2 k_1(y,x)}{\partial x^2}\Big|_{\hat{x}(y)}} \right)^{1/2}, \end{aligned}$$

where $\hat{x}(y)$ solves $\frac{\partial k_1(y,x)}{\partial x} = 0$ for x in y.

A.3.2 $M_{x|y}(t) = E(e^{xt}|y)$

$$E(e^{xt}|y) = \int e^{xt} p(x|y)dx = \int e^{xt} \frac{p(y|x)p(x)}{p(y)} dx = \frac{1}{p(y)} \int e^{xt} p(y|x)p(x)dx$$

Let

$$\begin{aligned} m_M(y,x,t) &= e^{xt} p(y|x)p(x) \\ &= \frac{1}{2\pi\beta\sigma} \exp\left\{ xt - \frac{y^2}{2\beta^2 \exp\{x\}} - \frac{x}{2} - \frac{(x-\mu)^2}{2\sigma^2} \right\}, \\ k_M(y,x,t) &= \log m_2(y,x) \\ &= -\log(2\pi\beta\sigma) + xt - \frac{y^2 \exp\{-x\}}{2\beta^2} - \frac{x}{2} - \frac{(x-\mu)^2}{2\sigma^2}. \end{aligned}$$

And

$$\begin{aligned} \frac{\partial k_M(y,x,t)}{\partial x} &= t + \frac{y^2 \exp(-x)}{2\beta^2} - \frac{1}{2} - \frac{x-\mu}{\sigma^2} \\ \frac{\partial^2 k_M(y,x,t)}{\partial x^2} &= -\frac{y^2 \exp(-x)}{2\beta^2} - \frac{1}{\sigma^2}, \end{aligned}$$

and

$$\begin{aligned} \int e^{xt} p(y|x)p(x)dx &\approx \int \exp\left\{ k_M(y,\hat{x}(y,t),t) + \frac{(x-\hat{x}(y,t))^2}{2} \frac{\partial^2 k_M(y,x,t)}{\partial x^2}\Big|_{\hat{x}(y,t)} \right\} dx \\ &= \exp\{k_M(y,\hat{x}(y,t),t)\} \left(-\frac{2\pi}{\frac{\partial^2 k_M(y,x,t)}{\partial x^2}\Big|_{\hat{x}(y,t)}} \right)^{1/2} = u(y,\hat{x}(y,t),t), \end{aligned}$$

where $\hat{x}(y,t)$ solves $\frac{\partial k_M(y,x,t)}{\partial x} = 0$ for x in y. $p(y)$ is obtained from the previous section. In addition, we need the following evaluated at $x = \hat{x}(y,t)$, and $t = 0$, for $m = 1, 2$.

$$\frac{\partial^m}{\partial t^m} u(y,\hat{x}(y,t),t),$$

while is quite involving. Finite difference methods are usually used for computing those quantities, see Tierney et al. (1989). Here, with some aid from some symbolic computing system, we will be able to get explicit expressions, which will be given next, in a form of guidance. The lengthy formulae also follow, but at the present, we just give outputs from

the ***Maxima*** system, without further verifications. In practice, implementations do not use analytic formulae, but the finite difference method. It is understood that analytic solution is not feasible in general.

To see that, we show the complexity and provide the following guidance for computing the quantities in symbolic computing systems. First, redefine

$$\begin{aligned} k_M(y,x,t) &= xt + k_M^*(y,x(t)) \\ u(y,x,t) &= \exp(xt)\exp(k_M^*(y,x(t)) \left(-\frac{2\pi}{\left.\frac{\partial^2 k_M(y,x,t)}{\partial x^2}\right|_{\hat{x}(y,t)}} \right)^{1/2} \\ &= m(x,t)h(x(t)) = u(x,t), \end{aligned}$$

where $m(x,t) = \exp(xt)$ is a simple form that explicitly involves both $x(t)$ and t, and $h(x(t))$ does not involve t explicitly. In addition, y is a constant. First, we can show, by differentiating both sides of $\frac{k_M(y,x,t)}{\partial x} = 0$ w.r.t t,

$$\frac{\partial}{\partial t}x(t) = \frac{2\beta^2\sigma^2}{\sigma^2 y^2 \exp(-x) + 2\beta^2},$$

which is not a function of t, explicitly, either. Then

$$\begin{aligned} \frac{\partial}{\partial t}u(x,t) &= \left[\frac{\partial}{\partial t}m(x,t)\right] h(x(t)) + m(x,t)\left[\frac{\partial}{\partial t}h(x(t))\right] \\ \frac{\partial^2}{\partial t^2}u(x,t) &= \left[\frac{\partial^2}{\partial t^2}m(x,t)\right] h(x(t)) + \left[\frac{\partial}{\partial t}m(x,t)\right]\left[\frac{\partial}{\partial t}h(x(t))\right] + \\ &\quad \left[\frac{\partial}{\partial t}m(x,t)\right]\left[\frac{\partial}{\partial t}h(x(t))\right] + m(x,t)\left[\frac{\partial^2}{\partial t^2}h(x(t))\right], \end{aligned}$$

where

$$\begin{aligned} \frac{\partial}{\partial t}h(x(t)) &= \left[\frac{\partial}{\partial x}h(x(t))\right]\left[\frac{\partial}{\partial t}x(t)\right], \text{ is not an explicit function of } t, \\ \frac{\partial^2}{\partial t^2}h(x(t)) &= \frac{\partial}{\partial x}\left[\frac{\partial}{\partial t}h(x(t))\right]\left[\frac{\partial}{\partial t}x(t)\right], \text{ is not an explicit function of } t, \\ \frac{\partial}{\partial t}m(x,t) &= \exp(xt)\left[\frac{\partial}{\partial t}x(t) + 1\right] \\ \frac{\partial^2}{\partial t^2}m(x,t) &= \left[\frac{\partial}{\partial t}m(x,t)\right]\left[\frac{\partial}{\partial t}x(t) + 1\right] + \exp(xt)\frac{\partial}{\partial t}\left[\frac{\partial}{\partial t}x(t)\right] \\ &= \left[\frac{\partial}{\partial t}m(x,t)\right]\left[\frac{\partial}{\partial t}x(t) + 1\right] + \exp(xt)\left(\frac{\partial}{\partial x}\left[\frac{\partial}{\partial t}x(t)\right]\right)\left[\frac{\partial}{\partial t}x(t)\right]. \end{aligned}$$

All the complex parts that do not involve t explicitly with the exception of x, can be computed by a symbolic system easily. Parts that involve t explicitly among those above are not hard to compute. Plug in $\hat{x}(0)$ and $t = 0$, in place of x and t, to get

$$\frac{\partial}{\partial t} u(y, \hat{x}(y,t), t)\Big|_{t=0} \text{ and } \frac{\partial^2}{\partial t^2} u(y, \hat{x}(y,t), t)\Big|_{t=0}.$$

The first derivative can be finally simplified to a ratio, whose numerator is

$$\beta\,\sigma\,\left(\sigma^4\,x\,y^4+\sigma^4\,y^4-6\,\beta^2\,\sigma^2\,x\,e^x\,y^2+2\,\beta^2\,\sigma^4\,t\,e^x\,y^2+2\,\beta^2\,\mu\,\sigma^2\,e^x\,y^2-2\,\beta^2\,\sigma^2\,e^x\,y^2+8\right.$$
$$\left.\beta^4\,x\,e^{2\,x}-4\,\beta^4\,\sigma^2\,t\,e^{2\,x}+2\,\beta^4\,\sigma^2\,e^{2\,x}-4\,\beta^4\,\mu\,e^{2\,x}\right)\,e^{-\frac{e^{-x}\,y^2}{2\,\beta^2}-\frac{x^2}{2\,\sigma^2}+t\,x+\frac{\mu\,x}{\sigma^2}-\frac{\mu^2}{2\,\sigma^2}},$$

and denominator is

$$\beta\sigma\,\left(\sigma^2\,y^2-2\,\beta^2\,e^x\right)^3\,\sqrt{\frac{\pi}{\sigma^2\,y^2-2\,\beta^2\,e^x}}.$$

The second derivative is also a ratio, whose numerator is

$$\left(\left(\beta\,\sigma^5\,x+\beta\,\sigma^5\right)\,e^{t\,x+\frac{\mu\,x}{\sigma^2}}\,y^4+\left(-6\,\beta^3\,\sigma^3\,x+2\,\beta^3\,\sigma^5\,t+\left(2\,\beta^3\,\mu-2\,\beta^3\right)\,\sigma^3\right)\right.$$
$$\left.e^{t\,x+\frac{\mu\,x}{\sigma^2}+x}\,y^2+\left(8\,\beta^5\,\sigma\,x-4\,\beta^5\,\sigma^3\,t+2\,\beta^5\,\sigma^3-4\,\beta^5\,\mu\,\sigma\right)\,e^{t\,x+\frac{\mu\,x}{\sigma^2}+2\,x}\right)\,e^{-\frac{e^{-x}\,y^2}{2\,\beta^2}},$$

and denominator is

$$\left(\beta\,\sigma^6\,\sigma\,e^{\frac{x^2}{2\,\sigma^2}+\frac{\mu^2}{2\,\sigma^2}}\,y^6-6\,\beta^3\,\sigma^5\,e^{\frac{x^2}{2\,\sigma^2}+x+\frac{\mu^2}{2\,\sigma^2}}\,y^4+12\,\beta^5\,\sigma^3\,e^{\frac{x^2}{2\,\sigma^2}+2\,x+\frac{\mu^2}{2\,\sigma^2}}\,y^2-\right.$$
$$\left.8\,\beta^7\,\sigma\,e^{\frac{x^2}{2\,\sigma^2}+3\,x+\frac{\mu^2}{2\,\sigma^2}}\right)\,\sqrt{\frac{\pi}{\sigma^2\,y^2-2\,\beta^2\,e^x}}.$$

A.4 Integrating out a Class of State Variables

A.4.1 Statement of the Problem

Consider specification by Equation (2.26) and (2.27), which is as follows:

$$\begin{aligned} \vec{Y}_t &= \exp(X_t/2)\vec{Z}_t \\ X_t &= \phi X_{t-1} + \eta_t, \end{aligned}$$

where

$$\begin{aligned} X_t|\vec{Y}_{t-1} &\sim N(\mu, \sigma^2) \\ \vec{Z}_t &\sim N_m(\vec{0}, \Sigma) \\ \eta_t &\sim N_1(0, \sigma_\eta^2). \end{aligned}$$

Thus, we have the following:

$$p(\vec{Y}_t|X_t, \vec{Y}_{t-1}) \sim N_m(\vec{0}, \exp(X_t)\Sigma)$$

Our tasks, by using the plug-in version of the saddlepoint approximation, are

(a) Compute $p(\vec{Y}_t|\vec{Y}_{t-1}) = \int p(\vec{Y}_t|X_t, \vec{Y}_{t-1})p(X_t|\vec{Y}_{t-1})dX_t$,

(b) Approximate $p(X_{t+1}|\vec{Y}_t)$ by Gaussian.

The function that we are interested in is $k(\vec{Y}_t, X_t) = \log\left[p(\vec{Y}_t|X_t, \vec{Y}_{t-1})p(X_t)\right]$, which equals

$$\mathrm{C} - \frac{mX_t}{2} - \frac{\exp(-X_t)}{2}Y_t'\Sigma^{-1}Y_t - \frac{1}{2\sigma^2}(X_t - \mu)^2,$$

where the quantity C is independent of X_t and equal to,

$$-\frac{1+m}{2}\log(2\pi) - \frac{1}{2}\log\|\Sigma\| - \frac{1}{2}\log\sigma^2.$$

We need to compute the gradient and hessian of $k(\vec{Y}_t, X_t)$ with respect to X_t, which are:

$$\frac{\partial k(\vec{Y}_t, X_t)}{\partial X_t} = -\frac{m}{2} + \frac{\exp(-X_t)}{2}Y_t'\Sigma^{-1}Y_t - \frac{1}{\sigma^2}(X_t - \mu) \quad \text{(A.9)}$$

$$\frac{\partial^2 k(\vec{Y}_t, X_t)}{\partial X_t^2} = -\frac{\exp(-X_t)}{2}Y_t'\Sigma^{-1}Y_t - \frac{1}{\sigma^2}. \quad \text{(A.10)}$$

A.4.2 Maximization

Function $k(\vec{Y}_t, X_t)$ is concave, according to Equation (A.10). For the same reason, the gradient is monotonely decreasing in X_t. In addition, because $\frac{\partial k(\vec{Y}_t, X_t)}{\partial X_t} > 0$, as $X_t \to -\infty$, and $\frac{\partial k(\vec{Y}_t, X_t)}{\partial X_t} < 0$, as $X_t \to \infty$, there is a unique solution to $\frac{\partial k(\vec{Y}_t, X_t)}{\partial X_t} = 0$. Therefore, there is a global maximum for $k(\vec{Y}_t, X_t)$.

Regarding the fact that $k(\vec{Y}_t, X_t)$ has analytic first and second derivatives, and hessian is always negative, this maximization problem can be solved by several standard methods; see Kelley (1995; 1999).

A.4.3 Computation of $p(\vec{Y}_t|\vec{Y}_{t-1})$

Suppose $k(\vec{Y}_t, X_t)$ is maximized at $\hat{X}_t$. The approximation is

$$p(\vec{Y}_t|\vec{Y}_{t-1})) \approx \exp\{k(\vec{Y}_t, \hat{X}_t)\} \left(-\frac{2\pi}{\frac{\partial^2 k(\vec{Y}_t, X_t)}{\partial X_t^2}\Big|_{\hat{X}_t}} \right)^{1/2},$$

due to the saddlepoint method directly.

A.4.4 Approximation of $p(X_{t+1}|\vec{Y}_t)$ by Gaussian

One of suitable approximations is to approximate the mean by

$$\mathrm{E}(X_{t+1}|\vec{Y}_t) \approx \phi \hat{X}_t,$$

and approximate variance by

$$\mathrm{Var}(X_{t+1}|\vec{Y}_t) \approx \phi^2 \left(-\frac{\partial^2 k(\vec{Y}_t, X_t)}{\partial X_t^2}\Big|_{\hat{X}_t} \right)^{-1} + \sigma_\epsilon^2.$$

This approximation does not compute posterior mean and variance through the saddlepoint method described in Tierney et al. (1989). Fortunately, such approximations still have a satisfactory convergence rate to true values; see Schervish (1995).

A.4.5 Special Cases for Σ

Computation may have some problems due to computing matrix inversion and determinant. For two special cases that we have discussed, inversion and determinant have explicit expressions.

The general purpose of interest is a covariance matrix Σ, which can be factorized as a correlation matrix R, pre-multiplied and post-multiplied by the same diagonal matrix Δ, i.e. $\Sigma = \Delta R \Delta$. The diagonal entries of Δ are δ_j's.

Then the inversion and determinant are expressed in the following:

$$\begin{aligned} \Sigma^{-1} &= \Delta^{-1} R^{-1} \Delta^{-1} \\ \|\Sigma\| &= \|\Delta\| \|R\| \|\Delta\|. \end{aligned}$$

Exponential Correlation

An exponential correlation matrix R of dimension m is a matrix, whose entry is defined by $R_{i,j} = \rho^{|i-j|}$. Following quantities are either directly or indirectly used in the

likelihood construction.

$$
\begin{aligned}
R &= LDL' \\
L_{i,j} &= \begin{cases} \rho^{|i-j|} & i \geq j \\ 0, & i < j \end{cases} \\
D_{i,j} &= \begin{cases} 1 & i = j = 1 \\ 1-\rho^2, & i = j > 1 \\ 0 & i \neq j \end{cases} \\
\|R\| &= (1-\rho^2)^{m-1} \\
\|\Sigma\| &= (1-\rho^2)^{m-1} \prod_{j=1}^{m} \delta_j^2 \\
L^{-1} &= T \\
&= \begin{cases} 1, & i = j \\ -\rho, & i-j = 1 \\ 0, & \text{otherwise} \end{cases} \\
(L')^{-1} &= T' \\
Y'\Sigma^{-1}Y &= Y'\Delta^{-1}R^{-1}\Delta^{-1}Y \\
&= Y'\Delta^{-1}(LDL')^{-1}\Delta^{-1}Y \\
&= Y'\Delta^{-1}(L')^{-1}D^{-1}L^{-1}\Delta^{-1}Y \\
&= \left(\frac{y_1}{\delta_1}\right)^2 + \frac{1}{(1-\rho^2)}\sum_{j=1}^{m}\left(\frac{y_{j+1}}{\delta_{j+1}} - \rho\frac{y_j}{\delta_j}\right)^2
\end{aligned}
$$

Gaussian Correlation

An Gaussian correlation matrix R of dimension m is a matrix, whose entry is defined by $R_{i,j} = \rho^{|i-j|^2}$. Following quantities are either directly or indirectly used in the

likelihood construction.

$$
\begin{aligned}
R &= LL' \\
L &= \begin{cases} \rho^{|i-j|^2} \frac{\prod_{l=i-j+1}^{i-1}(1-\rho^{2l})}{\sqrt{\prod_{m=1}^{j-1}(1-\rho^{2m})}} & i \geq j \\ 0, & i < j \end{cases} \\
\|R\| &= \prod_{i=1}^{m} L_{i,i}^2 \\
\|\Sigma\| &= \|R\| \prod_{j=1}^{m} \delta_j^2 \\
L^{-1} &= \begin{cases} (-\rho)^{i-j} \frac{G(j-1,i-1;\rho^2)}{\sqrt{\prod_{m=1}^{i-1}(1-\rho^{2m})}} & i \geq j \\ 0, & i < j \end{cases} \\
G(m,n;q) &= \frac{\prod_{j=1}^{m}(1-q^{n+1-j})}{\prod_{j=1}^{m}(1-q^j)} \\
(L')^{-1} &= (L^{-1})' \\
Y'\Sigma^{-1}Y &= Y'\Delta^{-1}R^{-1}\Delta^{-1}Y \\
&= Y'\Delta^{-1}(LL')^{-1}\Delta^{-1}Y \\
&= Y'\Delta^{-1}(L')^{-1}L^{-1}\Delta^{-1}Y
\end{aligned}
$$