Copyright © 2000 by Fang Liu
All rights reserved

BAYESIAN TIME SERIES: ANALYSIS METHODS USING SIMULATION-BASED COMPUTATION

by

Fang Liu

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Professor Mike West, Supervisor

_____
Brani Vidakovic

_____
Merlise Clyde

_____
Gary Rosner

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2000

ABSTRACT

(Statistics)

BAYESIAN TIME SERIES: ANALYSIS METHODS USING SIMULATION-BASED COMPUTATION

by

Fang Liu

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Professor Mike West, Supervisor

_____
Brani Vidakovic

_____
Merlise Clyde

_____
Gary Rosner

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2000

Abstract

This dissertation introduces new simulation-based analysis approaches, including both sequential and off-line learning algorithms, for various Bayesian time series models. We provide a Markov Chain Monte Carlo (MCMC) method for an autoregressive (AR) model with innovations following exponential power distributions using the fact that an exponential power distribution is a scale mixture of normals. This model has application in signal processing, specifically image processing, with orthogonal wavelet transformations put forth in [58]. We discuss our experience in applying the proposed algorithm to data from a grass image. As an alternative to MCMC methods, a generic sequential algorithm is proposed for a wide class of Bayesian dynamic models, including those with fixed parameters. This combines old ideas of kernel smoothing via shrinkage ([92], [93]) for modeling fixed parameter with newer ideas of auxiliary particle filtering ([67]) for the time varying states. It is shown here that our specific smoothing approaches can interpret and suggest modifications to techniques that add "artificial evolution noise" to fixed model parameters at each time point to address problems of sample attrition and prior:data conflict. Our new approach permits smoothing and regeneration of sample points of model parameters without the "loss of historical information" inherent in the "artificial evolution noise" method; the computational load required by our method is meaningfully reduced from the earlier kernel smoothing algorithms for Bayesian posterior simulation. Simple examples are used to demonstrate the efficacy of our algorithm. Following general discussions of the generic sequential algorithm, we turn our attention to the modeling of multivariate stochastic variance matrices in financial time series, an important application area of Bayesian dynamic models and their analysis approaches. The emphasis is on some research problems in applying Monte Carlo

methods to multivariate, potentially singular, Dynamic Linear Models (DLMs) with variance discounting. Some theoretical results about singular Wishart/matrix-Beta distribution are derived based on the work of [86]. A simplified example is used to demonstrate the feasibility of the proposed sequential algorithm in such models using the theoretical results developed. A Forward Filtering Backward Sampling algorithm is simultaneously proposed for the DLM of interest. We also analyze with the proposed sequential algorithm a volatility model closely related to variance discounting models - a multivariate stochastic volatility model in dynamic factor representations studied with MCMC in [1] and by other researchers. With 30 fixed parameters, the factor model is used to demonstrate the performance of our algorithm in models of modest dimensions. We conclude with some discussion of general issues of practical relevance, and suggestions for further development.

Acknowledgements

I would like to express my deep gratitude to my advisor, Professor Mike West. I am grateful for the precious experience of working with him. His constant input of insights has made this thesis possible; his continual encouragement has inspired me to progress; and his devotion to research has set me a life time example.

I would like to thank Professor Brani Vidakovic, Merlise Clyde and Gary Rosner for taking interest in my work and making suggestions for improvement. Dr. Jose Quintana at CDC North America first exposed me to the open questions in financial time series modeling and provided patient research guidance. The final revision of the thesis was made while I was a member of the Global Dynamic Asset Allocation Group at CDC North America. I appreciate my supervisors and colleagues, especially Dr. Jose Quintana, for their support. I thank the faculty and my fellow graduate students at ISDS for providing such a cooperative environment of research and study. I am also indebted to Dr. Michael Pitt and Omar Aguilar who generously shared with me their data and MCMC simulation programs which are used in the applied examples for the sequential algorithm.

My interest in scientific research first developed in my early childhood under the influence of my parents, who have never failed their support in my pursuit of knowledge. My sister has remained a powerful "challenger" as well as the most intimate friend throughout my student life. To them, and my numerous friends, I owe special thanks.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation studies Monte Carlo (MC) simulation methods for a variety of Bayesian time series models. In Bayesian methodology, unknown model parameters are considered as random variables; one's knowledge of the unknown parameters is expressed in terms of the posterior distribution. With the exception of a few standard conjugate models, closed-form posterior analysis is not available for the majority of realistic Bayesian models. These models are most conveniently analyzed via random draws from the posterior distribution of the model parameters. Currently, importance sampling and Markov Chain Monte Carlo (MCMC) are the most popular posterior simulation approaches in Bayesian inference. Importance sampling is relatively old and dates back at least to the 60's while MCMC became popular in mainstream statistics only in the last decade.

In applied statistics, time series models are employed to describe dynamic systems – systems that evolve over time. Analysis of such systems can proceed in two fashions – on-line and off-line. Off-line analysis is based on the data available at a certain time point, isolating the consecutive inferences made on the ever growing data set. On-Line, or the so called sequential analysis, is a recursive process to utilize the analysis result on a previously available data set and the recent data points to obtain

new inference. MCMC methods, when applied to time series models, are mostly off-line, while most sequential methods are variations of importance sampling. Off-line MCMC methods are in general more efficient, because each iteration is drawn using the complete data set. In sequential analysis, early imputations are based on less data, which makes it usually not as efficient as MCMC. But sequential methods provide more prompt response to new data and, on most occasions, independent samples. Hence it is of great interest in the fields of automatic control, signal processing, finance and applied statistics. We introduce new sequential and off-line simulation schemes for times series models with application in various engineering fields and finance.

The first part of the thesis includes chapter 2, which develops the theory and a MCMC method to analyze an exponential power model that arises in image signal processing with orthogonal wavelet decompositions. Wavelet decomposition is a powerful tool in image processing ([58]). Random wavelet coefficients are modeled with exponential power distributions, for which there is no closed-form Bayesian analysis. We use an autoregressive (AR) process to model the dependency between the coefficients and use the theory that an exponential power distribution is a scale mixture of normals ([90]) to propose a Gibbs sampler. The Gibbs sampler is then applied to a fraction of grass image and some simulation issues are discussed.

In the second part of the dissertation, we propose a generic on-line algorithm for general time series models. Chapter 3 gives an overview of sequential MC methods in the framework of sequential importance sampling. Most of these methods deal with the time-varying parameters, and sequential simulation of fixed model parameters is a difficult question. In chapter 4, we propose a general algorithm that combines the new method of auxiliary particle filters ([67]) with the old idea of kernel smoothing with shrinkage ([92], [93]). Our new approach permits smoothing and regeneration of

sample points of model parameters without the "loss of historical information" inherent in some earlier methods and the heavy computational load in kernel smoothing. It is shown that our specific smoothing approaches can interpret and suggest modifications to techniques that add "artificial evolution noise" to fixed model parameters at each time point to address problems of sample attrition and prior:data conflict. This algorithm is applied to an analytically tractable AR model, compared with its analytical posterior, and a stochastic volatility model, compared with MCMC results.

The third part of the thesis addresses simulation issues in dynamic volatility models for financial time series. Chapter 5 studies variance discounting in multivariate Dynamic Linear Models (DLMs), an important class of Bayesian time series models. For these models, variance matrices are often modeled via discounting. When both the variance and the regression coefficients are unknown, there are no analytical updating equations and simulation-based estimation methods have not been developed for such systems. In particular, multivariate DLMs with singular observation variance matrices have not been fully discussed in academic literature, although these models turn out to be quite useful in practice. We present some problems encountered in discounting and updating through MC simulation, develop some necessary theoretical results, and use them to apply the sequential algorithm in Chapter 4. Discussion of the discount model is extended to MCMC techniques, which are often used to monitor and facilitate sequential simulation. Chapter 6 develops an alternative variance-covariance model – a multivariate stochastic volatility model using dynamic factor representations. We implement this model with the general algorithm proposed in Chapter 4 on exchange return data. The feasibility of our algorithm in moderately high dimensional models is demonstrated via this example of 30 unknown fixed parameters. Chapter 7 concludes with summarization and discussion of future directions.

Chapter 2

Modeling Wavelet Coefficients with Exponential Power Distributions

2.1 Image Processing with Wavelet Decompositions

In computer image processing, it is well known that it is difficult to analyze an image from the gray level of the image pixels. The gray levels change with the lighting condition, and more importantly, pictures of various sizes with the same content should be given the same interpretation as far as the content itself is concerned, i.e. image signals need to be analyzed with a scale invariant representation. Multiresolution representations of image information can be partially scale invariant and provide a hierarchical framework for interpreting the image information (see [51]). The details of the image captured at coarse resolutions give the large-scale structure of the scene, while the finer level details correspond to the subtle structure. Bertley ([15]) and Crowley ([23]) studied the Laplacian representation, a pyramidal implementation for computing the signal details at various resolution levels. Among other things, the Laplacian data structure has the drawback that the coefficients at different resolution levels are correlated by the design of the decomposition approach itself, thus one can

not tell the correlation between transformation coefficients caused by the inherent redundancy of the representation from that reflecting the character of the data. This disadvantage can be overcome with orthogonal wavelet representations.

Wavelets are classes of functions $\psi(x)$ whose translations and dilations $(\sqrt{s}\psi(sx-t))_{(s,t)\in R^+\times R}$ can be used for expansions of Lesbegue $L^2(R)$ functions. (It is the conventional assumption that 2 dimensional images are $L^2(R^2)$ functions.) Multiresolution representations of $L^2(R)$ can always be found with a set of orthogonal wavelet basis functions. See [87] for a general treatment of orthogonal wavelet transformations and their applications, and [63] for wavelet based Bayesian models. [58] explored the multiresolution analysis of 2 dimensional signals with orthogonal wavelet decompositions and the application to data compression, texture discrimination and fractal analysis. In these applications, the statistical properties of the wavelet coefficients turn out to be useful information. In compact coding of wavelet representations for example, the knowledge of the distribution of the wavelet coefficients provides a basis for efficient coding methods. Straightforwardly, imagine a situation in which the data of an image in the wavelet domain are to be transferred with binary code, and the discrete wavelet coefficients range from 0 to 255, which occupy from 1 to 8 bits. It seems natural to represent the most frequent coefficient with a 1-bit number, and the least frequent coefficient with an 8-bit number. With orthogonal wavelet transformations, since the wavelet basis functions are orthogonal within and between levels, the coefficients are not correlated by the construction of the vector subspaces, which makes it easier to study the distribution of the coefficients. Although theoretically the wavelet coefficients could follow any distribution, Mallat ([58]) noticed that in practice, for all resolutions and orientations, the histogram of the coefficients are "symmetrical peaks centered around zero". He further proposed that the wavelet coefficients could be modeled with exponential power distributions and gave the estimation of the dis-

tribution parameters using method of moments. Assuming independence between the coefficients, Mallat's method ignores the coefficient correlations brought about by the intrinsic continuity of the images.

An exponential power density is a complicated function of its two parameters (we restrict our discussion to centered exponential power distributions), which forbids conjugate Bayesian analysis. West ([90]) showed that an exponential power distribution is a scale mixture of normals and identified the scale mixing density. This finding opened the possibility for Bayesian inference of exponential power models. This chapter attempts to analyze, in the Bayesian framework, the wavelet coefficients of some image signals using West's normal mixture theorem for exponential power distributions. We give a brief overview of orthogonal wavelet decompositions in the framework of multiresolution analysis in this section. In section 2, we examine some grass image data and propose a model that generalizes Mallat's ([58]) by taking into account the correlations between coefficients within resolution levels. Section 3 introduces the theory that exponential power distribution is a scale mixture of normals. Section 4 presents a Gibbs sampler for the proposed model and the MC simulation results.

2.1.1 Multiresolution Analysis

We intend to give a brief intuitive introduction to multiresolution analysis and orthogonal wavelet transformation in these two subsections, rather than a strict mathematical treatment. Interested readers are referred to [58] for details and [57] for the mathematical foundation. Although image data have to be discretized when stored or processed in the computer, we base our discussion on continuous signals with finite energy - $L^2(R)$ functions, with the understanding that the properties discussed apply to discrete signals and the corresponding discrete wavelet transformation. One

dimensional signals are considered here, extension to two dimension will be made later.

Multiresolution approximations of L^2 functions are based on a sequence of "pyramidal structured" closed subspaces V_j such that

$$V_j \subset V_{j+1} \tag{2.1}$$

for $\forall j \in Z$. The orthogonal projection of a function $f(x) \in L^2$ in the subspace V_j, $A_j f(x)$, is the approximation of $f(x)$ at level j. The hierarchical structure of the V_j subspaces in (2.1) guarantees that the approximation of a signal at resolution $j+1$ contains all the information to compute the approximation at the lower resolution j. The approximation operation is similar at all levels and $f(x) \in V_j$ iff $f(2x) \in V_{j+1}$. As the resolution level increases to infinity, the approximation converges to the original signal, and when the resolution level decreases to zero, the approximated signal converges to zero, i.e. $lim_{j\to\infty} V_j$ is dense in L^2 and $lim_{j\to-\infty} V_j = \{0\}$. For any such multiresolution approximation of L^2, $V_j, j \in Z$, there exists a unique function $\phi \in V_0$ whose integer translations $\{\phi(x-k), k \in Z\}$ is the orthonormal basis of V_0, and whose dilations by 2^j, $\{2^{j/2}\phi(2^j x - k), k \in Z\}$ is the orthonormal basis of V_j.

2.1.2 Orthogonal Wavelet Transformation

[58] showed that a multiresolution representation can be obtained using orthogonal wavelet basis functions as the basis of $V_j, j \in Z$. The first wavelet basis was found in 1910 by Harr. In analogy with the well known Fourier transformations, which transform information from the time domain to the frequency domain, wavelet transformations express information in wavelet domain. Unlike Fourier transformation whose basis functions are restricted to sinusoidal functions, the functional form of wavelets can be chosen to suit specific problems. And wavelet transformations are better localized in both time and frequency domains, which means that wavelet

representations can capture rapid local changes with relatively few coefficients. As mentioned above, orthogonal wavelet transformations do not introduce correlations between coefficients; on the contrary, they tend to simplify the dependence structure in the original data. It also is inclined to centralize the energy in the transformed data, and the total energy concentrates in only a few coefficients. This has led to the use of wavelet transformations in data denoising and data compression by shrinkage. The shrinkage of wavelet coefficients with Bayesian methods is explored in [20] and [22].

In the multiresolution analysis framework, the wavelet $\phi(x)$ whose translations and dilations are used for the basis of the V_j spaces is called the father wavelet. Since for any j, $V_j \subset V_{j+1}$, one can always find the orthogonal complement of V_j in V_{j+1}, O_j. For any father wavelet $\phi(x)$, there is always a mother wavelet function $\psi(x)$, such that $\{2^{j/2}\psi(2^j x-k), k \in Z\}$ is an orthonormal basis of O_j and $\{2^{j/2}\psi(2^j x-k), (j,k) \in Z^2\}$ is an orthonormal basis of $L^2(R)$.

Suppose the signal f is at the beginning described with an arbitrary resolution $j+1$, then the orthogonal projection of the signal on V_j, $A_j f$, would be the approximation of the signal at that coarser level. The difference between the information in V_j from that in V_{j+1} however, can be completely captured by O_j. $A_j f$ can again be represented with the next lower level subspaces V_{j-1} and O_{j-1}. Let the orthogonal projection of the signal in O_j be represented by $D_j f$, then the original signal is represented by

$$(A_J f, (D_i f)_{J \leq i \leq j})$$

for any $J < j, J \in Z$.

This set of discrete signals consists of the father wavelet signal at the coarse level J and all the mother wavelet signals at the levels J to j, it is called an orthogonal wavelet representation of f.

2.1.3 Orthogonal Wavelet Transformation in 2 Dimension

Orthogonal wavelet transformations can be extended to any positive dimension. For image processing, 2 dimensions are required. Suppose the signal is a function $f(x,y) \in L^2(R^2)$. One can still find a sequence of subspaces $V_j, j \in Z$ in $L^2(R^2)$ with properties discussed in the previous subsection. There still exists a unique father wavelet function $\Phi(x,y)$ whose translations and dilations $2^j\Phi(2^jx-n, 2^jy-m)_{(m,n)\in Z^2}$ give an orthogonal basis of each V_j. Each subspace V_j could be decomposed into the tensor product of two identical subspaces V_j^1 in $L^2(R)$. The scaling function can be written as $\Phi(x,y) = \phi(x)\phi(y)$ where $\phi(x)$ is the scaling function in V_j^1. Let O_j be the orthogonal complement of V_j in V_{j+1}, and $\psi(x)$ be the one dimensional mother wavelet for $\phi(x)$. The wavelets

$$\begin{aligned} \Psi^1(x,y) &= \phi(x)\psi(y) \\ \Psi^2(x,y) &= \psi(x)\phi(y) \\ \Psi^3(x,y) &= \psi(x)\psi(y) \end{aligned} \tag{2.2}$$

are the mother wavelets for $\Phi(x,y)$, i.e

$$\begin{aligned} &2^j\Psi^1(2^jx-n, 2^jy-m), \\ &2^j\Psi^2(2^jx-n, 2^jy-m), \\ &2^j\Psi^3(2^jx-n, 2^jy-m), \qquad (n,m) \in Z^2 \end{aligned} \tag{2.3}$$

is an orthogonal basis of O_j and

$$\begin{aligned} &2^j\Psi^1(2^jx-n, 2^jy-m), \\ &2^j\Psi^2(2^jx-n, 2^jy-m), \\ &2^j\Psi^3(2^jx-n, 2^jy-m), \qquad (n,m,j) \in Z^3 \end{aligned} \tag{2.4}$$

is an orthogonal basis of $L^2(R^2)$. As in one dimension, the approximation of a 2 dimensional signal in a certain subspace is given by the inner product of the signal

with the basis of that subspace.

2.2 Modeling Random Wavelet Coefficients with AR Processes and Exponential Power Distributions

We examine the orthogonal wavelet coefficients of the grass image in Figure 2.1. The image is represented with a 187×188 matrix of the grey levels of the discrete pixels; a 3 level 2 dimensional discrete wavelet transformation is applied to the matrix and the coefficients are plotted in Figure 2.2. (The wavelet coefficients are labeled following the convention of the wavelets package in S+.)

Figure 2.1: Grass image

The histogram plots of the wavelet coefficients have high peaks and fat tails (Figure 2.3). The fatter-than-normal tails are more obvious from the normal Q-Q plot (Figure 2.4).

Observations of this kind prompted Mallat to use centered exponential power

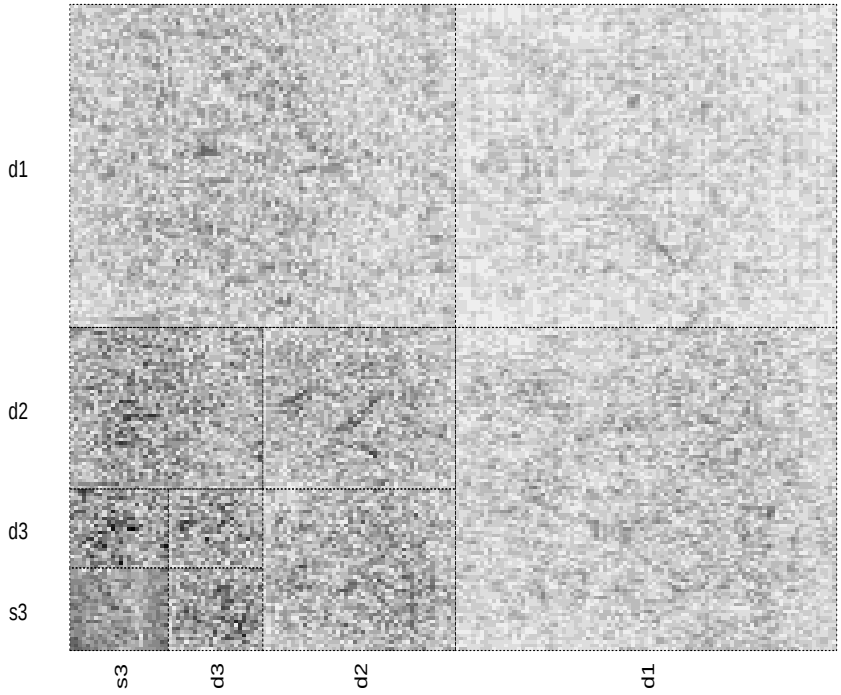

Figure 2.2: Wavelet coefficients of grass image

distributions to model wavelet coefficients. A centered exponential power distribution has the density form

$$f(d) = Ce^{-\frac{1}{2}|\frac{d}{\alpha}|^{\beta}}, 0 < \alpha, \beta.$$

This distribution was studied by [13] in Bayesian robust analysis. Two special cases in the exponential power family are normal (when $\beta = 2$) and double exponential (when $\beta = 1$). β can be regarded as a measure of kurtosis, and α a measure of scale. The mean of the above distribution is 0 and the variance is given by $2^{2/\beta}\frac{\Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})}\alpha^2$.

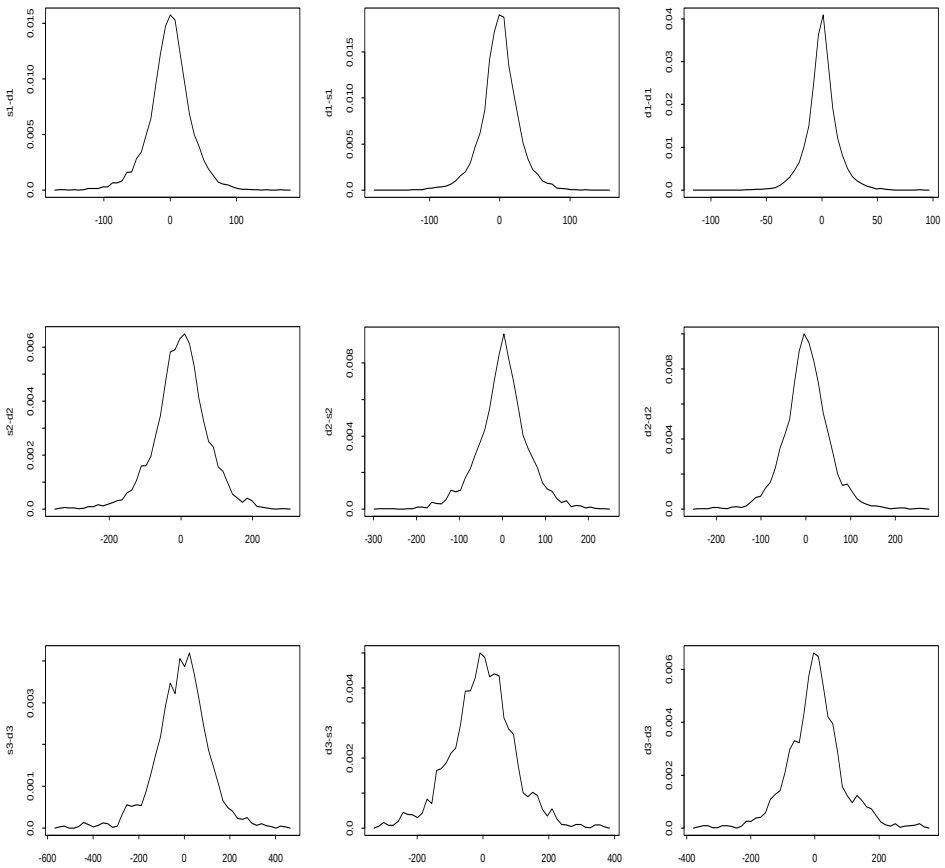

Figure 2.3: Empirical densities of wavelet coefficients

We denote the distribution by $EPD(\alpha, \beta)$.

On the two-dimensional autocorrelation plot (see Figure 2.5), the first or second order autocorrelations are significant while the higher order autocorrelations appear negligible.

We start from modeling data in one dimension. We denote a row of the wavelet coefficients with $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ and model $\mathbf{X}$ with an AR(1) process with exponential power noise, i.e.

$$x_t = \theta x_{t-1} + \epsilon_t \tag{2.5}$$

where $\epsilon_t \sim EPD(\alpha, \beta), t = 2, 3, \ldots, T$ and $-1 < \theta < 1$.

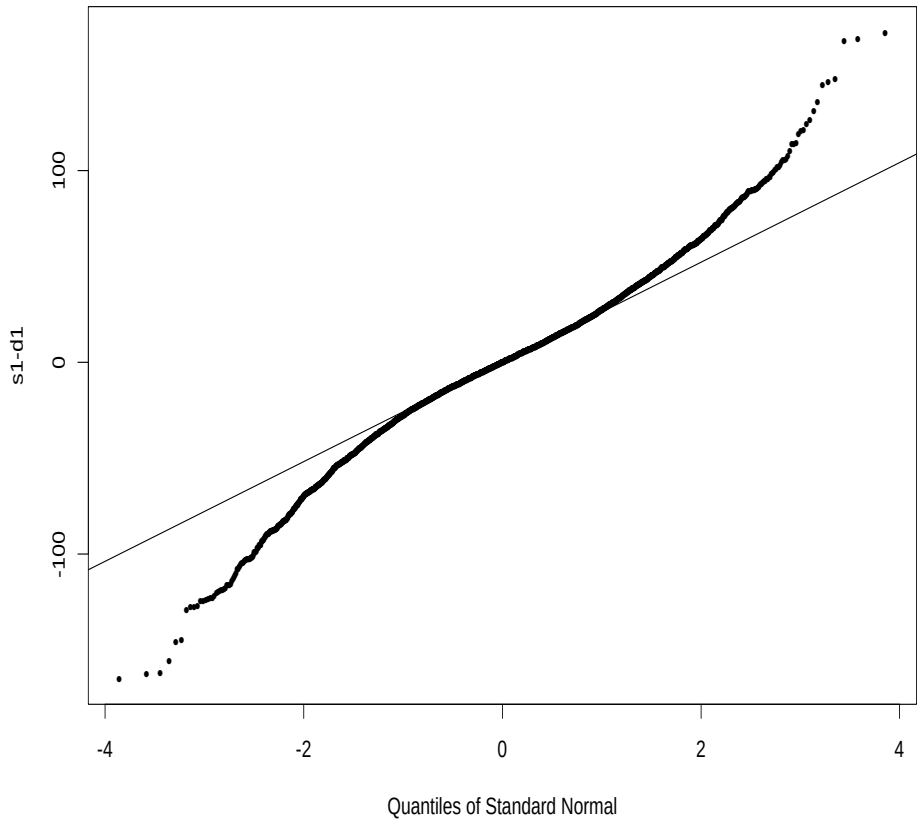

Figure 2.4: Normal Q-Q plot of wavelet coefficients S1-D1

Note that with nonzero autocorrelation, the marginal distribution of x_t is no longer exponential power distributed, but it does have high peak and fat tails, especially when θ is close to zero.

2.3 Exponential Power Distributions As Scale Mixtures of Normals

West ([90]) proved that an exponential power family distribution with $1 \leq \beta \leq 2$ is a mixture of normals and identified the mixing distribution. The idea of a normal

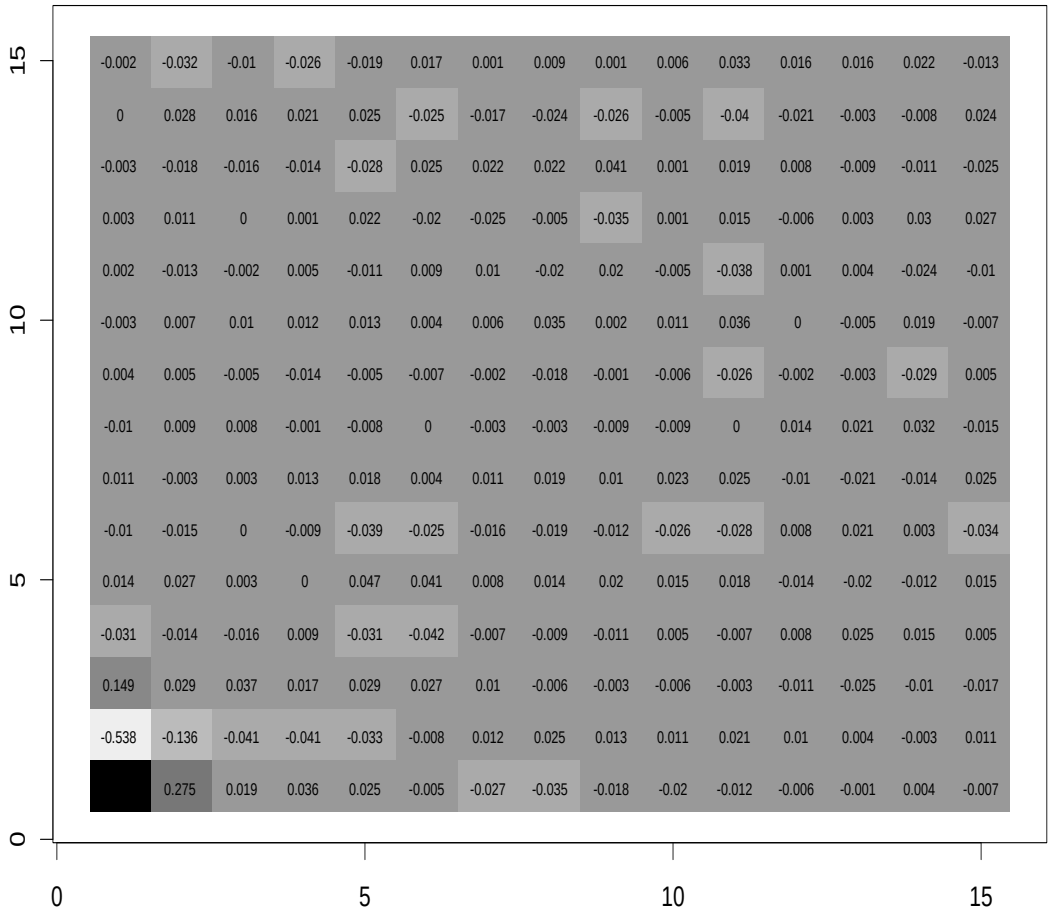

Figure 2.5: Correlations of wavelet coefficients S1-D1

scale mixture is that if Y follows the standard normal distribution, and σ has some distribution on $(0, \infty)$ with density $h(\sigma)$, then $X = Y\sigma$ is said to be a scale mixture of normals, with the scale mixing density $h(\sigma)$. Many continuous, unimodal and symmetric distributions on the real line are scale mixture of normals. In addition to exponential power family, examples include Student t and the stable family. The mixture-of-normal structure of these distributions is useful in deriving some theoretical properties of the distributions and in practice are often used in robustness studies. Here we exploit this structure of the exponential power distribution in the MC simulation of the proposed AR(1) model.

West's result can be summarized as follows. If ϵ follows a centered exponential power distribution with kurtosis parameter β and scale parameter α, then its density, as given on page 157 of Box & Tiao ([13]), is

$$p(\epsilon) = [\Gamma(1 + \frac{1}{\beta})2^{1+\frac{1}{\beta}}\alpha]^{-1} exp(-\frac{1}{2}|\frac{\epsilon}{\alpha}|^{\beta}) \tag{2.6}$$

where $\beta \geq 1$ and $\alpha > 0$.

This density can be written as

$$\int f(\epsilon|\phi)h(\phi)d\phi \tag{2.7}$$

where $\epsilon|\phi \sim N(0, \phi^{-1}\alpha^2 2^{\frac{2}{\beta}})$ and $h(\phi) \propto \phi^{-\frac{1}{2}} p_{\frac{\beta}{2}}(\frac{\phi}{2})$. Here $p_{\frac{\beta}{2}}(.)$ is the density of positive stable distribution.

Stable distributions were first developed and thoroughly studied by Levy in the 1920's. By the central limit theorem, the sum of random variables with finite variances tends to normal in distribution. When the variances are not finite, the sum tends to stable in distribution. Stable distributions have fatter tails than normal and have been used in modeling economic data (see [59]). Except for a few special cases, the density function is not available for the stable family. Despite that, samples can

be drawn, and we will use the positive stable sampler given by Devroy ([24]) in our MC simulation. The sampler, along with some relevant properties of stable distributions, is given in the appendix. Notice that there are various parameterizations for exponential power and stable distributions. [90] follows [32]'s notation while [24]'s is different. To avoid confusion, we follow [24] and show in the appendix how the scale mixing density (2.7) can be reached, rephrasing the theorem in [90] with [24]'s notation.

2.4 Monte Carlo Simulation

2.4.1 Metropolis-Hastings Algorithm and Gibbs Sampler

It is not possible to identify the posterior distribution of the parameter β with any standard distribution, so that we have to resort to MC simulation in order to make posterior inference. Metropolis-Hastings algorithms are widely used in MC simulation. The idea is that the integration of a function of a certain random variable over its density function can be approximated by drawing samples in a Markov Chain whose stationary distribution is the distribution of the random variable. Metropolis-Hastings algorithm provides a general framework to construct such Markov chains. The algorithm was first proposed by Metropolis et al ([61]), later generalized by Hastings ([44]). Suppose the distribution to be sampled from is $\pi(\theta)$, at each sampling stage j, the sample at the next stage $\theta^{(j+1)}$ is chosen by sampling a candidate θ^* from a proposal distribution $q(.|\theta^{(j)})$. θ^* is accepted with the probability $min(1, \frac{\pi(\theta^*)q(\theta^{(j)}|\theta^*)}{\pi(\theta^{(j)})q(\theta^*|\theta^{(j)})})$. If θ^* is accepted, then set $\theta^{(j+1)} = \theta^*$, otherwise set $\theta^{(j+1)} = \theta^{(j)}$.

The Gibbs sampler is a special case of Metropolis-Hastings: for each parameter, when the proposal distribution is the full conditional distribution (the distribution of that parameter given all the other parameters and data), the acceptance rate is 1.

It is one of the most popular MCMC algorithms in statistical applications.

2.4.2 A Gibbs Sampler for AR(1) with EPD Noise

In model (2.5), x_1 should be considered a fixed value in the likelihood since it does not involve any random innovation; and $\epsilon_t, t = 2, \ldots, T$, are identically distributed independent exponential power variables. For each ϵ_t, introduce a scale parameter ϕ_t as in (2.7), the full conditionals in model (2.5) can be identified. With reference priors: constant for θ and β, and α^{-1} for α, the full conditionals are:

$$
\begin{aligned}
p(\theta|.) &\propto exp[-\frac{1}{2^{2/\beta+1}\alpha^2}\sum_{t=2}^{n}\phi_t(x_t-\theta x_{t-1})^2] \qquad (2.8)\\
&\sim N(\frac{\sum_2^n \phi_t x_t x_{t-1}}{\sum_2^n \phi_t x_{t-1}^2}, \frac{\alpha^2 2^{2/\beta}}{\sum_2^n \phi_t x_{t-1}^2});\\
p(\alpha^2|.) &\propto \alpha^{-1}\alpha^{-(n-1)}exp[-\frac{1}{2^{2/\beta+1}\alpha^2}\sum_{t=2}^{n}\phi_t(x_t-\theta x_{t-1})^2]\\
&\propto \alpha^{-n}exp[-\frac{1}{2^{2/\beta+1}\alpha^2}\sum_{t=2}^{n}\phi_t(x_t-\theta x_{t-1})^2]\\
&\sim Inverse-Gamma(\frac{n}{2}, \frac{1}{2^{2/\beta+1}}\sum_{t=2}^{n}\phi_t(x_t-\theta x_{t-1})^2);\\
p(\phi_t|.) &\propto \phi_t^{-\frac{1}{2}}p_{\frac{\beta}{2}}(\frac{\phi_t}{2})\phi_t^{\frac{1}{2}}exp[-\frac{\phi_t}{2^{2/\beta+1}\alpha^2}(x_t-\theta x_{t-1})^2]\\
&\propto p_{\frac{\beta}{2}}(\frac{\phi_t}{2})exp[-\frac{\phi_t}{2^{2/\beta+1}\alpha^2}(x_t-\theta x_{t-1})^2];\\
p(\beta|.) &\propto C(\beta)^{n-1}exp(-\frac{1}{2}\sum_{2}^{n}|\frac{\epsilon_t}{\alpha}|^\beta),
\end{aligned}
$$

where $C(\beta) = [\Gamma(1+\frac{1}{\beta})2^{1+\frac{1}{\beta}}]^{-1}$, and $t = 2, \ldots, T$.

The full conditionals of θ and α^2 are normal and inverse-gamma respectively. The

ϕ_t and β do not follow any standard distributions, but samples can be drawn from their full conditionals with rejection steps. For each ϕ_t, use $p_{\frac{\beta}{2}}(\frac{\phi_t}{2})$ as proposal. At iteration $j+1$, the rejection step for ϕ_t is

- Draw $\phi_t^* \sim p_{\frac{\beta}{2}}(\frac{\phi_t}{2})$
- Calculate

$$\begin{aligned} r &= min\{1, \frac{exp[-\frac{\phi_t^*}{2^{2/\beta+1}\alpha^2}(x_t - \theta x_{t-1})^2]}{exp[-\frac{\phi_t^{(j)}}{2^{2/\beta+1}\alpha^2}(x_t - \theta x_{t-1})^2]}\} \qquad (2.9) \\ &= min\{1, exp[\frac{1}{2^{2/\beta+1}\alpha^2}(x_t - \theta x_{t-1})^2(\phi_t^{(j)} - \phi_t^*)]\} \end{aligned}$$

$$(2.10)$$

- Accept ϕ_t^* with probability r. If ϕ_t^* is accepted, set $\phi_t^{(j+1)} = \phi_t^*$, otherwise set $\phi_t^{(j+1)} = \phi_t^{(j)}$.

For β, use $\beta - 1 \sim Beta(b_1, b_2)$ as proposal, then at iteration j, the rejection step is

- Draw $\tilde{\beta} \sim Beta(b_1, b_2)$, and let $\beta^* = 1 + \tilde{\beta}$; represent the density of β^* by $B(b^*/b_1, b_2)$
- Calculate

$$r = min\{1, \frac{C(\beta^*)^{n-1} exp(-\frac{1}{2}\sum_2^n |\frac{\epsilon_t}{\alpha}|^{\beta^*})/B(\beta^*/b_1, b_2)}{C(\beta^{(j)})^{n-1} exp(-\frac{1}{2}\sum_2^n |\frac{\epsilon_t}{\alpha}|^{\beta^{(j)}})/B(\beta^{(j)}/b_1, b_2)}\} \qquad (2.11)$$

- Accept β^* with probability r. If β^* is accepted, set $\beta^{(j+1)} = \beta^*$, otherwise set $\beta^{(j+1)} = \beta^{(j)}$

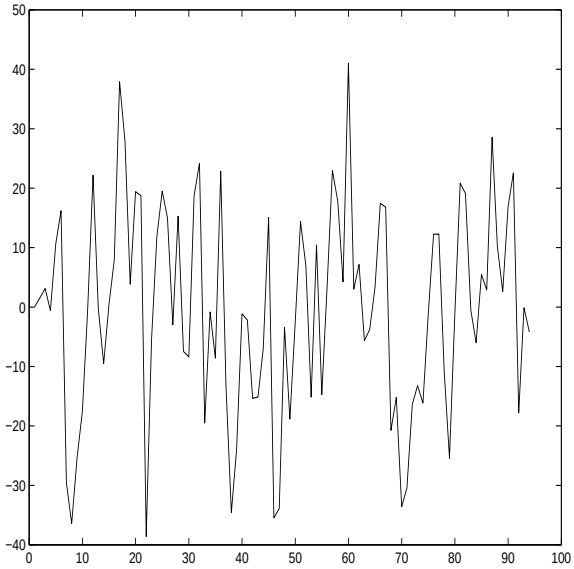

Figure 2.6: Simulated wavelet coefficients

2.4.3 Implementing the Gibbs Sampler

To test our Gibbs sampler, we first generate a series of 94 data points from model (2.5), with parameters $\theta = 0.5, \alpha^2 = 150, \beta = 1.5$ and $x_1 = 0$. The innovations ϵ_t are generated with a Metropolis-Hastings algorithm using the scale-mixture-of-normal structure of ϵ_t. The proposal distribution we use for ϕ_t in generating ϵ_t is $p_{\frac{\beta}{2}}(\frac{\phi}{2})$. To remove the dependency between neighboring ϵ_t in the Markov Chain, we take one sample of ϵ_t out of every 10 samples generated. This artificial series is plotted in Figure 2.6.

Experiments show that the samples of β and θ mix quickly: even when the starting samples are far from the posterior center, they quickly find their way; α^2 may converge less rapidly given an extreme starting sample. The proposal distribution for β used is $1+Beta(14, 14)$. We use the true value as the starting point, run the Gibbs sampler for 2000 iterations and take one sample out of every 2, hence end up with 1000 effective samples. The posterior $0.05, 0.25, 0.50, 0.75$ and 0.95 quantiles well contain the true

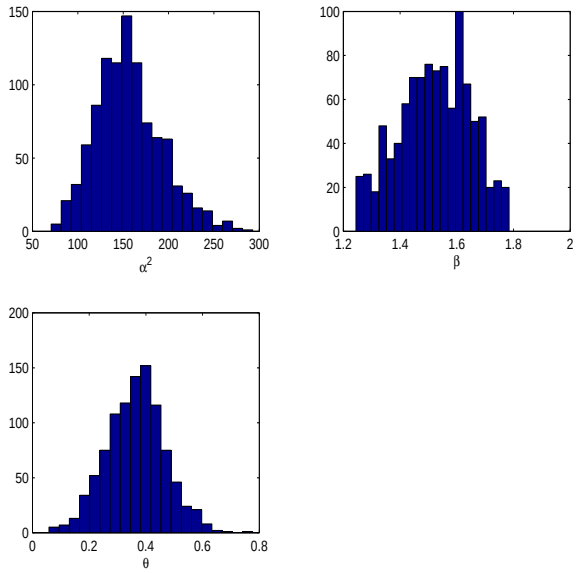

Figure 2.7: Histograms of the posterior samples using simulated data

values of the parameters as shown in Table 2.1. Figure 2.7 plots the histogram and Figure 2.8 plots the trace of the posterior samples. Note that there are a few peaks in the histogram of β suggesting the chain is stuck at a few places, which could also be seen in the trace plot. The $0.05, 0.25, 0.50$ and 0.75 quantiles of the acceptance rate for β are 0.0031, 0.1392, 0.6564 and 1.

	0.025	0.25	0.50	0.75	0.975
α^2 :	102.4437	130.9729	152.6982	177.4681	223.9640
β :	1.2969	1.4339	1.5291	1.6171	1.7104
θ :	0.1896	0.2952	0.3677	0.4300	0.5342

Table 2.1: Posterior quantiles of the unknown parameters

We then tried this algorithm with the first row of the wavelet coefficients from S1-D1. The data are a vector of 94 points as shown in Figure 2.9. Figures 2.10 and 2.11 give the histogram and trace plots based on posterior samples of 3000. A set of 1000 spaced 3 apart was subsampled. For this set of data, we used 1+Beta(12,12) as the proposal for β. We have tried different parameters for the Beta distribution in the

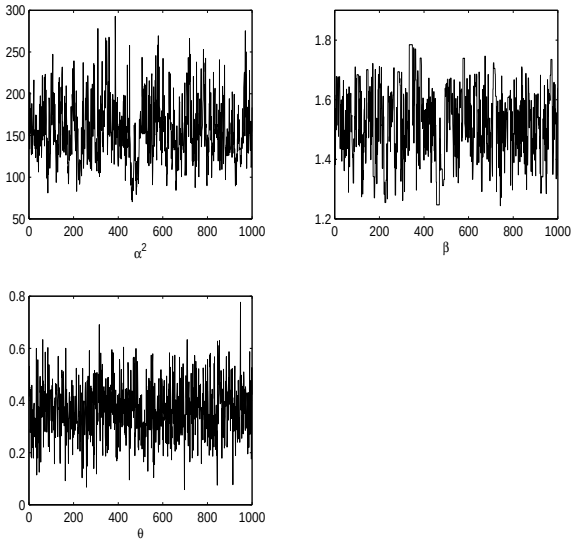

Figure 2.8: Trace plots of the posterior samples using simulated data

proposal for β; the problem of slow mixing did not go away, even when subsamples were drawn 20 apart from the initial Markov Chain, it was still apparent that the chain for β sticks occasionally. As obvious from the trace plot, α^2 and β are highly correlated. This seems to be the main reason for slow mixing. The slow mixing problem happens with the artificial data set too, but there, changing the proposal for β seems to make improvement. To solve this problem, some quick mixing recipes are to be explored.

2.4.4 Discussion

To process image signals, extensions of this model need to be made to two dimensional data. The two dimensional Gaussian Markov random field models ([5]) have been used for spatially distributed data. A future research direction is to explore two dimensional exponential power Markov random field model for two dimensional wavelet coefficients.

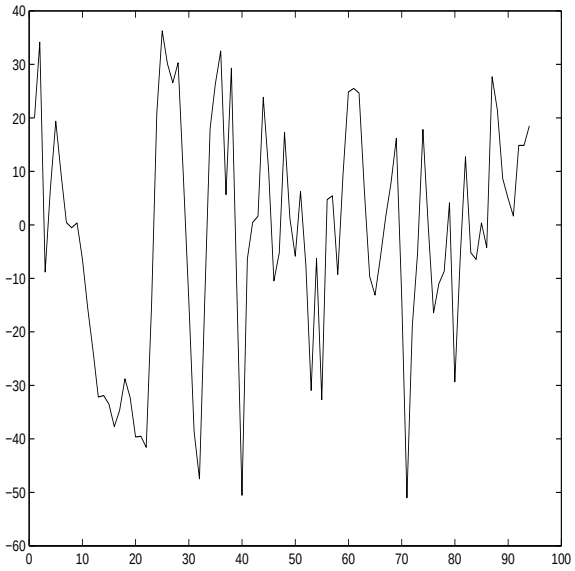

Figure 2.9: Wavelet coefficients: first row from S1-D1

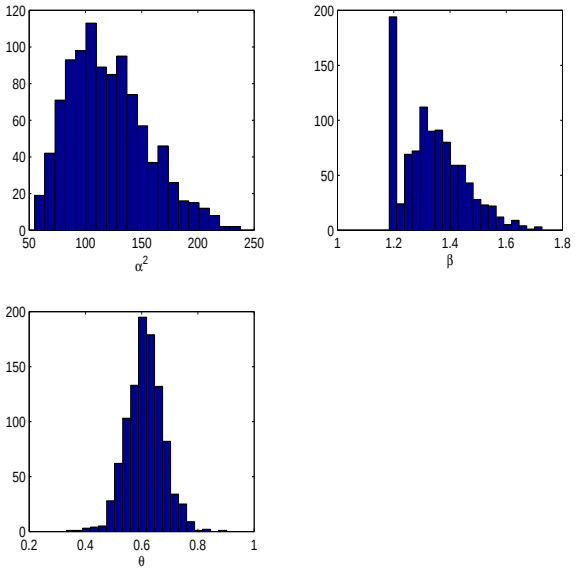

Figure 2.10: Histograms of the posterior samples using real data

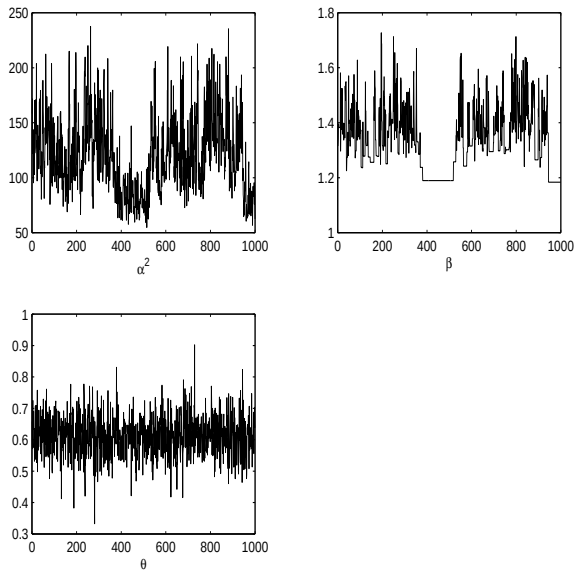

Figure 2.11: Trace plots of the posterior samples using real data

Chapter 3

Sequential Importance Sampling

3.1 State Space Models and Sequential Inference

State space models are a class of dynamic models for sequentially observed data $\boldsymbol{y}_t, t = 1, 2, \ldots$. In these models, some unknown underlying time-varying vectors $\boldsymbol{x}_t$, the so called states, along with some fixed parameters, are assumed to drive the observations. The state is typically assumed to evolve over time following a Markovian process, which possibly involves some fixed parameters. We use $\boldsymbol{\theta}$ to represent the vector of all the fixed parameters in a dynamic model. The time-varying parameters $\boldsymbol{x}_t$ reflect changes between time points while the assumption of fixed $\boldsymbol{\theta}$ meets the necessity of fully preserving information over time.

State space models are specified at each time t by the observation equation, defining the observation-state relationship

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}) \tag{3.1}$$

and the evolution equation, defining the Markovian transition density for the states

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}). \tag{3.2}$$

Each $\boldsymbol{y}_t$ is conditionally independent of the other states and observations given the

current state $\boldsymbol{x}_t$ and the parameter $\boldsymbol{\theta}$, and $\boldsymbol{x}_t$ is conditionally independent of earlier states and observations given $\boldsymbol{x}_{t-1}$ and $\boldsymbol{\theta}$. This covers a very broad class of interesting models in practice. Closed form recursive estimation and forecasting equations are available only when both of the equations are linear with Gaussian noise. For nonlinear or non-Gaussian problems, there are generally no analytic solutions. The sequential analysis of such systems has to resort to deterministic approximation or Monte Carlo simulation. Research on sequential learning algorithms is not restricted to state space models, but given that this class of models is one of the most important types of dynamic models, we put our emphasis on state space models, covering others when relevant.

For several decades researchers involved in sequential analysis of dynamic models, in both statistics and various engineering fields, have been using discrete numerical approximations to sequentially update posterior distributions in various "mixture modeling" frameworks. This literature has involved methods for both time-evolving states and fixed parameters, and is exemplified by the important class of adaptive multi-process models used in Bayesian forecasting since the early 1970s [42, 85, 94]. During the 1980s, this naturally led to larger-scale analyses using discrete grids of parameter values, though the combinatorial "explosion" of grid sizes with increasing parameter dimension limited this line of development. Novel methods using efficient quadrature-based, adaptive numerical integration ideas were introduced in the late 1980s ([68, 70, 69]). This involved useful methods in which discrete grids – of both fixed model parameters and state variables – themselves change over time as data are processed, sequentially adapting the discrete posterior approximations by generating new "samples" as well as associated "weights". This work recognized the utility of the Markov evolution equations of dynamic models in connection with the generation of new grids of values for time-evolving state variables. It similarly recognized

and addressed the practically critical issues of "diminishing weights" on unchanging grids of parameter values, and the associated need for some method of interpolation and smoothing to "regenerate" grids of values for fixed model parameters. In these respects, this adaptive deterministic approach naturally anticipated future developments of simulation-based approaches. Again, however, parameter dimension limited the broader applicability of such approaches.

The end of the 1980s saw a developing interest in simulation-based methods (see review papers [26] and [56]). Parallel developments in the 1990s have led to publication of different approaches, of which almost all can be considered as various forms of sequential importance sampling. Most of these methods were for the recursive sampling of the state variables. Recognizing the need for some kind of interpolation/smoothing of "old" parameter samples, West ([92],[93]) used weighted mixtures of kernel densities as the proposal for sequential importance sampling at each time point. West's approach was later extended to include variable shapes of multivariate kernels to reflect changing patterns of dependencies among parameters in different regions of parameter space [38], as explicitly anticipated in West ([92],[93]). In these kernel density estimation based methods, the calculation of the importance weights for each sample requires the evaluation of all the kernels for that state, thus without techniques to reduce the number of kernels, the computation load is in the order of the square of the sample size.

Gordon's bootstrap filter ([39]) and Kitagawa's filter and smoother ([50]) drew samples of a new state from the samples of the previous state according to the system equation. In their case, the priors of the states serve as the proposal in sequential importance sampling. This straightforward method involves less computation than the kernel smoothing approach: for each sample, it only requires sampling from the evolution equation and evaluating the likelihood. However, the drawback is also

obvious: since the samples are drawn without utilizing the information in the likelihood, in the case of prior and likelihood conflict the simulation becomes rather inefficient. To overcome this problem, [50] exploited fixed-lag smoothing, processing the data in a batch when possible to avoid information loss between the sampling of consecutive states. [39] suggested "roughening" and "prior editing". Roughening adds an independent jitter to each sample drawn in the updating procedure. This smooths the sample from the prior but introduces extra noise. Prior editing suggests throwing away the initial samples that have low likelihood probability to avoid carrying samples with low weights. Roughening and prior editing are ad hoc methods to resolve prior-likelihood conflict; although lacking sound theoretical justification, they bear great similarity to theoretically more rigorous methods developed later. [2] implemented bootstrap filter in a multitarget tracking model.

Kong ([52]) and Liu et al ([56]) studied sequential imputation methods for missing data in dynamic models. [52] designed sequential imputation approach for the class of models in which at each stage the missing data can be sampled from its posterior, and the predictive probability can be evaluated. Unlike in state space models, the missing data in these models do not have to form a Markov chain, and the observed data do not have to be conditionally independent. The posterior of the missing data at each imputation stage is used as the proposal distribution. For this class of models, they discussed the measurement of sampling efficiency and explained why the variance of the posterior weights tend to increase over time but not strictly so. [56] applied [52]'s method to blind deconvolution and proposed a rejuvenation procedure to resolve the dilemma between resampling and keeping the weights in simulation. Coinciding with some other authors, they made the point that early imputations are primarily based on priors and are very likely to lose predictability for the future observations. They proposed the use of a Gibbs sampler to process the first batch of signals and

provide the initial samples. Along the line of rejuvenation, [54] proposed rejection controlled sequential importance sampling for state space models, which is designed to simultaneously reduce Monte Carlo variation and retain independent samples.

Berzuini et al ([4]) proposed a sampling-resampling approach for general state space models and showed that each resampling step contributes an additional variance component to the Monte Carlo integration result. They further put forth a sampler that combines sequential importance sampling with the Metropolis-Hastings algorithm. Since in this method, iterations are involved at each sampling stage, the convergence rate could be slow. Pitt and Shephard ([67]) introduced an auxiliary variable in sequential importance sampling. Their idea is in spirit analogous to Gordon et al's prior editing; the improvement is that after sampling, adjustment is made to compensate for "prior editing". We will elaborate on this efficient algorithm in later sections.

Most of these works highlighted the utility of the convolution structure of Markov evolution equations for state variables in generating Monte Carlo samples of time-evolving states. The treatment of fixed model parameters was different, however. West's kernel density idea can be naturally applied to the fixed parameters. At each time step, the mixture of the fixed parameters around the current samples provides a continuous approximation of the posterior density; with high dimensional parameters, however, this method suffers heavy computational burden. Gordon et al's roughening idea, extended to fixed model parameters, leads to a synthetic method of generating new sample points for parameters via the convolution implied by the "artificial evolution". As used in its original form for dynamic states, this idea suffers loss of information. Those methods whose sample randomness exclusively comes from the evolution equation, when applied to model fixed parameters, would not be able to produce new samples for the state variable: the samples for a new state would

only be samples for the old state reweighted.

This chapter gives an overview of the techniques for time-varying parameters. The dependency on $\boldsymbol{\theta}$ is suppressed from our notation with the implicit assumption that $\boldsymbol{\theta}$ is known. We introduce Bayesian importance sampling in the general setting in section 2, and sequential importance sampling in section 3. Section 4 introduces a variation of Bayesian sequential importance sampling, auxiliary particle filters, which will be incorporated in our general algorithm. The last section discusses efficiency issues and resampling, an ad hoc technique to deal with the problem of sample degeneracy.

3.2 Bayesian Importance Sampling

3.2.1 The Importance Sampling Algorithm

Importance sampling appeared in the literature as early as the 60's ([41])). [78] proposed using a sampling/importance resampling algorithm for drawing imputations from the posterior distribution in Bayesian missing data problem. [84] discussed rejection and weighted bootstrapping methods to obtain posterior samples via random draws from the prior distribution. This section offers an introduction to Bayesian importance sampling; detailed illustration of the application of importance sampling methods in Bayesian inference can be found in [35].

In Bayesian models, the purpose of analysis is often to evaluate the expected value of a certain integrable function $f(\boldsymbol{x})$ of the unknown parameters $\boldsymbol{x}$ with respect to their joint posterior density $p(\boldsymbol{x}|D)$, where D denotes the data, i.e. to calculate

$$I(f(\boldsymbol{x})) = E_{p(\boldsymbol{x}|D)}(f(\boldsymbol{x})) = \int f(\boldsymbol{x})p(\boldsymbol{x}|D)d\boldsymbol{x} \tag{3.3}$$

If we can draw N i.i.d. samples from the posterior density $p(\boldsymbol{x}|D)$, then the above

integration can be approximated by $\frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{x}^{(j)})$. In most cases however, sampling directly from the posterior distribution $p(\boldsymbol{x}|D)$ is not possible. Importance sampling is a systematic estimation approach that theoretically can be applied to any situation in which $E_t(f_t)$ exists.

The basic idea of Bayesian importance sampling is that since direct sampling from the posterior distribution is not possible, it is desirable to find a proposal probability density function $g(\boldsymbol{x}|D)$ that is close to the posterior distribution but can be sampled from. This density function is called the importance function. Samples from the importance function, along with some "correction for the target posterior distribution", can be used to approximate the integration $I(f(\boldsymbol{x})) = E_{p(\boldsymbol{x}|D)}(f(\boldsymbol{x}))$. This in mathematical language is

$$\begin{aligned} I(f(\boldsymbol{x})) &= \int f(\boldsymbol{x})p(\boldsymbol{x}|D)d\boldsymbol{x} \\ &= \int f(\boldsymbol{x})g(\boldsymbol{x}|D)\frac{p(\boldsymbol{x}|D)}{g(\boldsymbol{x}|D)}d\boldsymbol{x} \end{aligned} \tag{3.4}$$

which can be approximated by

$$\widehat{I}(f(\boldsymbol{x})) = \sum_{j=1}^{N} f(\boldsymbol{x}^{(j)})\omega^{(j)} \tag{3.5}$$

where $\boldsymbol{x}^{(j)}, j = 1, \ldots, N$, are i.i.d. samples from the proposal $g(\boldsymbol{x}|D)$, and

$$\omega^{(j)} = \frac{1}{N}\frac{p(\boldsymbol{x}^{(j)}|D)}{g(\boldsymbol{x}^{(j)}|D)} \tag{3.6}$$

$g(\boldsymbol{x}|D)$ is often chosen to be a standard distribution that resembles the true posterior $p(\boldsymbol{x}|D)$, thus evaluating $g(\boldsymbol{x}^{(j)}|D)$ in (3.6) is generally straightforward; it is

usually not as straightforward to calculate $p(\boldsymbol{x}|D)$ in order to get $\omega^{(j)}$. From Bayes theorem,

$$p(\boldsymbol{x}|D) = \frac{p(\boldsymbol{x})f(D|\boldsymbol{x})}{f(D)}. \tag{3.7}$$

While $p(\boldsymbol{x})$ and $f(D|\boldsymbol{x})$ are specified as the prior and likelihood distributions in the model, the marginal data density $f(D)$ has to be evaluated through an integration:

$$f(D) = \int f(D|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \tag{3.8}$$

which may not be obtained analytically. The importance sampling idea in (3.5) can be used again to approximate the integration:

$$\begin{aligned} f(D) &= \int g(\boldsymbol{x}|D)\frac{f(D|\boldsymbol{x})p(\boldsymbol{x})}{g(\boldsymbol{x}|D)}d\boldsymbol{x} \\ &\approx \frac{1}{N}\sum_{j=1}^{N}\omega'^{(j)} \end{aligned} \tag{3.9}$$

where $\boldsymbol{x}^{(j)}, j = 1, \ldots, N$, are i.i.d. samples from $g(\boldsymbol{x}|D)$, and

$$\omega'^{(j)} = \frac{f(D|\boldsymbol{x}^{(j)})p(\boldsymbol{x}^{(j)})}{g(\boldsymbol{x}^{(j)}|D)}.$$

It is not accidental that the same importance function is selected in (3.5) and (3.9): by "recycling" the importance function $g(\boldsymbol{x}|D)$, one avoids the trouble of drawing two sets of samples of $\boldsymbol{x}$.

Plug (3.7) and (3.9) into (3.6), we get

$$\omega^{(j)} = \frac{p(\boldsymbol{x}^{(j)})f(D|\boldsymbol{x}^{(j)})/g(\boldsymbol{x}^{(j)}|D)}{\sum_{k=1}^{N}p(\boldsymbol{x}^{(k)})f(D|\boldsymbol{x}^{(k)})/g(\boldsymbol{x}^{(k)}|D)}$$

$$= \frac{\omega'^{(j)}}{\sum_{k=1}^{N} \omega'^{(k)}}. \tag{3.10}$$

$\omega'^{(j)}$ is called the unnormalized weight, which is obtained when the unnormalized posterior density $p(\boldsymbol{x})f(D|\boldsymbol{x})$ is used in place of the posterior density in (3.6); $\omega^{(j)}$ is the normalized weight, calculated by dividing the unnormalized weights with their summation, so that $\omega^{(j)}$ sum to one.

These samples with importance weights are proper samples from the posterior. One may choose to keep the weights and use (3.5) to approximate the integration of interest, or to get evenly weighted samples through resampling. The resampling step simply involves sampling from $\mathbf{x}^{(j)}, j = 1, \ldots, N$, each with probability $\omega^{(j)}$.Although the treatment of the importance weights may be of no importance in some applications, in the sequential context, about which we are concerned, it would affect the sampling efficiency as the MC errors tend to accumulate over time.

This is the widely used Bayesian importance sampling. In using it, one has to choose the importance function, which is the key to sampling efficiency. In the next subsection, we are going to show that, under certain conditions, $\widehat{I}(f(\boldsymbol{x}))$ converges to its population counterpart almost surely and address the problem of MC efficiency along with the selection of importance functions.

3.2.2 Convergence and Optimal Importance Functions

[36] studied the conditions under which the numerical approximations of the posterior moments converge almost surely to the true values as the number of MC replications increases. These results alone may be of little practical value because there is no way to know the convergence rate. Their practical value is in providing some guidelines to construct the importance functions $g(\boldsymbol{x}|D)$ and to evaluate the numerical efficiency. In summary, Geweke's convergence theorems are as the following:

Assumption 1: The product of the prior density and the likelihood function is proportional to a proper probability density function.

Assumption 2: $\boldsymbol{x}^{(j)}, j = 1, \ldots, N$, are i.i.d. samples from $g(\boldsymbol{x}|D)$.

Assumption 3: the support of $g(\boldsymbol{x}|D)$ includes the support of $p(\boldsymbol{x}|D)$.

Assumption 4: $I(f(\boldsymbol{x}))$ exists and is finite.

Theorem 3.1 *Under assumptions 1-4, $\widehat{I}(f(\boldsymbol{x}))$ converges to $I(f(\boldsymbol{x}))$ almost surely.*

Theorem 3.2 *Under assumptions 1-4, suppose $E[\omega(\boldsymbol{x})] < \infty$ and $E_{p(\boldsymbol{x}|D)}[f^2(\boldsymbol{x})\omega(\boldsymbol{x})] < \infty$, then $\sqrt{N}[\widehat{I}(f(\boldsymbol{x})) - I(f(\boldsymbol{x}))]$ converges to $N(0, \sigma^2)$ in distribution, where*

$$\sigma^2 = E_{p(\boldsymbol{x}|D)}\{[f(\boldsymbol{x}) - I(f(\boldsymbol{x})]^2\omega(\boldsymbol{x})\}$$

and

$$\widehat{\sigma}^2 = \sum_{j=1}^{N}[f(\boldsymbol{x}_i) - I(f(\boldsymbol{x}))]^2\omega_i^2$$

converges to σ^2 almost surely.

The optimal importance function which minimizes the numerical error is given by:

Theorem 3.3 *Under assumptions 1-4, if the mean deviation of $f(\boldsymbol{x})$, $md[f(\boldsymbol{x})] \equiv f(\boldsymbol{x}) - I(f(\boldsymbol{x})) < \infty$, then the importance sampling density that minimizes σ^2 has density $(f(\boldsymbol{x}) - I(f(\boldsymbol{x})))p(\boldsymbol{x})$ and for this choice $\sigma^2 = \{md[f(\boldsymbol{x})]\}^2$.*

It is impractical to select a different importance function for each function $f(\boldsymbol{x})$ whose expectation is to be estimated and devise schemes to sample from each importance function thus selected. The strength of Theorem 3.3 is that it suggests that

importance sampling densities with tails thicker than the posterior density, which is a character of $(f(\boldsymbol{x})-I(f(\boldsymbol{x})))p(\boldsymbol{x})$, might be more efficient. Theorem 3.2 also confirms this point: compare two candidate importance functions with thick and thin tails respectively; the one with thick tails is likely to give small importance weights when the value of $[f(\boldsymbol{x})-I(f(\boldsymbol{x})]^2$ is large while the one with thinner tails gives relatively large importance weights when the value of $[f(\boldsymbol{x})-I(f(\boldsymbol{x})]^2$ is large. Other things being equal, the thick-tailed importance function should have smaller numerical error σ^2. It is important that tails of the importance function should not decay more rapidly than the posterior distribution. Important function with rapidly decaying tails tend to result in samples in the tail area with extremely large weights; this is appreciated in the empirical work of [99], [34] and [3].

Not incorporating $f(\boldsymbol{x})$, the most ideal importance sampling function is the posterior distribution itself, in which case the importance weights are uniformly $\frac{1}{N}$ and the samples are i.i.d.; in contrast, when the weights are very uneven, a few samples carry most of the weights, the samples are effectively highly dependent and the numerical integration value might oscillate abruptly even when the sample size is large. The variance of the importance weights is often (but not uniquely) used as a measure of the sampling degeneracy; see [97], [17].

Unfortunately, it is rarely the case that the optimal importance function can be easily sampled from and the importance weights for such importance function can be easily calculated, hence there is a need to find approximations of the optimal importance function. The approximation could be done through MC simulation or analytical work. The MC simulation methods are criticized for the lack of convergence results and the great computation expense. Analytical approximation can sometimes be found by utilizing the special structure of the model (see [26] and [80] for examples), generally, the Taylor's expansion applies as follows.

Assume $l(\boldsymbol{x}) = lnp(\boldsymbol{x}|D)$ is twice differentiable with respect to $\boldsymbol{x}$. Then, with

$$l'(\boldsymbol{x}) = \frac{\partial l(\boldsymbol{x})}{\partial \boldsymbol{x}}|\boldsymbol{x}=\boldsymbol{x}_0$$

and

$$l''(\boldsymbol{x}) = \frac{\partial l^2(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^t}|\boldsymbol{x}=\boldsymbol{x}_0,$$

the second order Taylor's expansion of $l(\boldsymbol{x})$ at $\boldsymbol{x}_0$ is

$$l(\boldsymbol{x}) \approx l(\boldsymbol{x}_0) + (l'(\boldsymbol{x}_0))^t(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^t l''(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0).$$

If $l''(\boldsymbol{x}_0)$ is negative definite, then letting

$$\boldsymbol{\Sigma}(\boldsymbol{x}_0) = -l''(\boldsymbol{x}_0)^{-1}$$

and

$$\mathbf{m}(\boldsymbol{x}_0) = \Sigma(\boldsymbol{x}_0)l'(\boldsymbol{x}_0)$$

yields

$$l(\boldsymbol{x}) \approx C - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0 - \mathbf{m}(\boldsymbol{x}_0))^t \boldsymbol{\Sigma}(\boldsymbol{x}_0)^{-1}(\boldsymbol{x} - \boldsymbol{x}_0 - \mathbf{m}(\boldsymbol{x}_0)). \tag{3.11}$$

This suggests a normal importance function for $\boldsymbol{x}$,

$$g(\boldsymbol{x}|D) \sim N(\mathbf{m}(\boldsymbol{x}_0) + \boldsymbol{x}_0, \boldsymbol{\Sigma}(\boldsymbol{x}_0)).$$

The variance matrix is often calculated at the mode of the posterior distribution; the importance weight is calculated as in (3.15). In practice, this normal importance function may perform poorly if the tails of the posterior decay slower or the posterior is substantially asymmetric. [36] suggested using split-normal or split-t to adjust the importance sampling density in directions along each axis.

3.3 Sequential Importance Sampling

The aim of sequential importance sampling is to approximate

$$I(f_t) = \int f_t(\boldsymbol{x}_t) p(\boldsymbol{x}_t|D_t) d\boldsymbol{x}_t$$

at each time through recursive simulation. This can be achieved via calculation similar to that in the previous section, utilizing the fact that at time t, the prior distribution of $\boldsymbol{x}_t$ is its distribution given the information up to the previous time, $p(\boldsymbol{x}_t|D_{t-1})$.

¿From Bayes theorem, at any time t,

$$p(\boldsymbol{x}_t|D_t) = \frac{p(\boldsymbol{x}_t|D_{t-1}) f(\boldsymbol{y}_t|\boldsymbol{x}_t, D_{t-1})}{f(\boldsymbol{y}_t|D_{t-1})} \tag{3.12}$$

Noticing that all the information in D_{t-1} related to $\boldsymbol{x}_t$ lies in the posterior MC samples $\{\boldsymbol{x}_{t-1}^{(j)}, j = 1, \ldots, N\}$, we have

$$\begin{aligned} p(\boldsymbol{x}_t|D_{t-1}) &= \int p(\boldsymbol{x}_{t-1}|D_{t-1}) p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) d\boldsymbol{x}_{t-1} \\ &\approx \sum_{j=1}^{N} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)}) \omega_{t-1}^{(j)} \end{aligned} \tag{3.13}$$

where $\boldsymbol{x}_{t-1}^{(j)}$ is a sample from $p(\boldsymbol{x}_{t-1}|D_{t-1})$ with weight $\omega_{t-1}^{(j)}$.

Thus,

$$\begin{aligned} I(f_t) &= \int f_t(\boldsymbol{x}_t) p(\boldsymbol{x}_t|D_t) d\boldsymbol{x}_t \\ &= \int f_t(\boldsymbol{x}_t) \frac{p(\boldsymbol{x}_t|D_{t-1}) f(\boldsymbol{y}_t|D_{t-1}, \boldsymbol{x}_t)}{f(\boldsymbol{y}_t|D_{t-1})} d\boldsymbol{x}_t \\ &\approx \frac{1}{f(\boldsymbol{y}_t|D_{t-1})} \int f_t(\boldsymbol{x}_t) \sum_{j=1}^{N} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)}) \omega_{t-1}^{(j)} f(\boldsymbol{y}_t|\boldsymbol{x}_t) d\boldsymbol{x}_t \end{aligned}$$

$$= \frac{1}{f(\boldsymbol{y}_t|D_{t-1})}\sum_{j=1}^{N}\int f_t(\boldsymbol{x}_t)g(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)},\boldsymbol{y}_t)\frac{p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)})f(\boldsymbol{y}_t|\boldsymbol{x}_t)\omega_{t-1}^{(j)}}{g(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)},\boldsymbol{y}_t)}d\boldsymbol{x}_t$$

$$\approx \frac{1}{f(\boldsymbol{y}_t|D_{t-1})}\sum_{j=1}^{N} f_t(\boldsymbol{x}_t^{(j)})\omega_t^{\prime(j)} \tag{3.14}$$

where $\boldsymbol{x}_t^{(j)}$ is a sample from $g(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)},\boldsymbol{y}_t)$, and

$$\omega_t^{\prime(j)} = \frac{p(\boldsymbol{x}_t^{(j)}|\boldsymbol{x}_{t-1}^{(j)})f(\boldsymbol{y}_t|\boldsymbol{x}_t^{(j)})}{g(\boldsymbol{x}_t^{(j)}|\boldsymbol{x}_{t-1}^{(j)},\boldsymbol{y}_t)}\omega_{t-1}^{(j)}. \tag{3.15}$$

Notice that

$$\begin{aligned}
f(\boldsymbol{y}_t|D_{t-1}) &= \int p(\boldsymbol{x}_{t-1}|D_{t-1})p(\boldsymbol{y}_t|\boldsymbol{x}_{t-1})d\boldsymbol{x}_{t-1} \\
&\approx \sum_{j=1}^{N} p(\boldsymbol{y}_t|\boldsymbol{x}_{t-1}^{(j)})\omega_{t-1}^{(j)} \\
&= \sum_{j=1}^{N}\omega_{t-1}^{(j)}\int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)})f(\boldsymbol{y}_t|(D_{t-1}),\boldsymbol{x}_t)d\boldsymbol{x}_t \\
&= \sum_{j=1}^{N}\omega_{t-1}^{(j)}\int g(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)},\boldsymbol{y}_t)\frac{p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)})f(\boldsymbol{y}_t|\boldsymbol{x}_t)}{g(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)},\boldsymbol{y}_t)}d\boldsymbol{x}_t \\
&\approx \sum_{j=1}^{N}\omega_{t-1}^{(j)}\frac{p(\boldsymbol{x}_t^{(j)}|\boldsymbol{x}_{t-1}^{(j)})f(\boldsymbol{y}_t|\boldsymbol{x}_t^{(j)})}{g(\boldsymbol{x}_t^{(j)}|\boldsymbol{x}_{t-1}^{(j)},\boldsymbol{y}_t)} \\
&= \sum_{j=1}^{N}\omega_t^{\prime(j)}
\end{aligned} \tag{3.16}$$

Thus,

$$I(f_t) \approx \sum_{j=1}^{N} f_t(\boldsymbol{x}_t^{(j)})\omega_t^{(j)} \tag{3.17}$$

where

$$\omega_t^{(j)} = \frac{\omega_t^{\prime(j)}}{\sum_{k=1}^{N}\omega_t^{\prime(k)}}$$

The simulation goes as follows:

(1) After observing the first datum $\boldsymbol{y}_1$, sample $\boldsymbol{x}_1^{(j)}, j = 1, \ldots, N$, from $g(\boldsymbol{x}_1|\boldsymbol{y}_1)$. For each sample $\boldsymbol{x}_1^{(j)}$, evaluate the unnormalized weights

$$\omega_1'^{(j)} = \frac{p(\boldsymbol{x}_1^{(j)}|D_0) f(\boldsymbol{y}_1|\boldsymbol{x}_1^{(j)})}{g(\boldsymbol{x}_1^{(j)}|\boldsymbol{y}_1)}$$

and normalize them to get $\omega_1^{(j)}$;

(2) For $t = 2$, sample $\boldsymbol{x}_t^{(j)}, j = 1, \ldots, N$, from $g(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^{(j)}, \boldsymbol{y}_t)$, and calculate the weights

$$\omega_t'^{(j)} = \frac{p(\boldsymbol{x}_t^{(j)}|\boldsymbol{x}_{t-1}^{(j)}) f(\boldsymbol{y}_t|\boldsymbol{x}_t^{(j)})}{g(\boldsymbol{x}_t^{(j)}|\boldsymbol{x}_{t-1}^{(j)}, \boldsymbol{y}_t)} \omega_{t-1}^{(j)}$$

and normalize to get $\omega_t^{(j)}$;

(3) Repeat step (2) for $t = 3, 4, \ldots$

At any time t, one obtains

$$\widehat{I}(f_t) = \sum_{j=1}^{N} f_t(\boldsymbol{x}_t^{(j)}) \omega_t^{(j)}.$$

In light of the discussion in the previous section, the proposal density should have a fatter tail than the posterior to be sampled from. A natural such choice is the prior distribution at each time step. In the case of dramatic deviation between prior and posterior, this method becomes inefficient. In sequential importance sampling, the repeated approximation of continuous posterior distributions with discrete samples of limited size naturally leads to accumulation of MC errors, hence the efficiency issue is more of a concern. We discuss this further in the next section and introduce a more efficient algorithm.

3.4 Filtering with Auxiliary Variables

When the optimal importance function can not be directly sampled from, apart from looking for an approximation of it, another way of improving efficiency is through an auxiliary variable; examples include [4], [67], [55].

The idea is to introduce a random variable j which takes value in $1, \ldots, N$. Standing at time t, suppose we have a sample of current states $\{\boldsymbol{x}_t^{(1)}, ..., \boldsymbol{x}_t^{(N)}\}$ and associated weights $\{\omega_t^{(1)}, ..., \omega_t^{(N)}\}$ that together represent a Monte Carlo importance sample approximation to the posterior $p(\boldsymbol{x}_t|D_t)$. This includes, of course, the special case of equal weights in which we have a direct posterior sample. Time evolves to $t+1$, we observe $\boldsymbol{y}_{t+1}$, and want to generate a sample from the posterior $p(\boldsymbol{x}_{t+1}|\mathbf{D}_{t+1})$. Theoretically,

$$p(\boldsymbol{x}_{t+1}|\mathbf{D}_{t+1}) \propto p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})p(\boldsymbol{x}_{t+1}|\mathbf{D}_t) \tag{3.18}$$

where $p(\boldsymbol{x}_{t+1}|\mathbf{D}_t)$ is the prior density of $\boldsymbol{x}_{t+1}$ and $p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$ is the likelihood function. The second term here – the prior for the state at time $t+1$ – is implied by the state equation as $\int p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\mathbf{D}_t)d\boldsymbol{x}_t$. Under the Monte Carlo approximation to $p(\boldsymbol{x}_t|\mathbf{D}_t)$, this integral is replaced by a weighted summation over the sample points $\boldsymbol{x}_t^{(k)}$, so that the required update in equation (3.18) becomes

$$p(\boldsymbol{x}_{t+1}|\mathbf{D}_{t+1}) \propto p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}) \sum_{k=1}^{N} \omega_t^{(k)} p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(k)}). \tag{3.19}$$

As mentioned before, to generate Monte Carlo approximations to this density, an old and natural idea is to sample from $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(k)})$ for $k = 1, ..., N$, evaluate the corresponding values of the weighted likelihood function $\omega_t^{(k)} p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$ at each draw, and then use the normalized weights as the new weights of the samples (see ([39], [50])). This basic "particle filter" is an importance sampling method closely related to those of West ([92],[93]). A key problem is that the sampled points come

from the current prior of $\boldsymbol{x}_{t+1}$ and the resulting weights may be very small on many points in cases of meaningful separation of the prior and the likelihood function based on $\boldsymbol{y}_{t+1}$. West ([92],[93]) developed an effective method of adaptive importance sampling to address this. The idea of auxiliary particle filtering of Pitt and Shephard is similar in spirit but has real computational advantages; this works as follows. Incorporate the likelihood function $p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$ under the summation in equation (3.19) to give

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{D}_{t+1}) \propto \sum_{k=1}^{N} \omega_t^{(k)} p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(k)}) p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$$

and generate samples of the current state as follows. For each $k = 1, \ldots, N$, select an estimate $\boldsymbol{\mu}_{t+1}^{(k)}$ of $\boldsymbol{x}_{t+1}$, such as the mean or mode of $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(k)})$. Evaluate the weights $g_{t+1}^{(k)} \propto \omega_t^{(k)} p(\boldsymbol{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(k)})$. A large value of $g_{t+1}^{(k)}$ indicates that $\boldsymbol{x}_t^{(k)}$, when "evolving" to time $t+1$, is likely to be more consistent with the datum $\boldsymbol{y}_{t+1}$ than otherwise. Then indicators j are sampled with probabilities proportional to $g_{t+1}^{(j)}$, and values $\boldsymbol{x}_{t+1}^{(j)}$ of the current state are drawn from $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(j)})$ based on these "auxiliary" indicators. These sampled states are essentially importance samples from the proposal distribution $\omega_t^{(j)} p(\boldsymbol{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(j)}) p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(j)})$ and have associated weights

$$\omega_{t+1}^{(j)} = \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}^{(j)})}{p(\boldsymbol{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(j)})}. \tag{3.20}$$

Posterior inferences at time $t+1$ can be based directly on these sampled values and weights, or we may resample according to the importance weights $\omega_{t+1}^{(j)}$ to obtain an equally weighted set of states representing a direct Monte Carlo approximation to the required posterior in equation (3.18).

In essence, the auxiliary particle filter approach selects the index variable J in light of the likelihood function, and given the index, sample $\boldsymbol{x}_{t+1}$ from the evolution equation. The auxiliary variable J serves the function of Gordon et al ([39])'s "prior

editing" - picking out the samples of the old state that favors the new likelihood; the importance weights correct the distortion of the posterior introduced by "prior editing". Some other variations of using the auxiliary variable include marginalization of $\boldsymbol{x}_{t+1}$, and Hastings Independence Chain approach in [55].

3.5 Efficiency and Importance Resampling

We have mentioned that the variance of the importance weights indicates the sampling efficiency; the larger the variance, the less efficient the importance sampling scheme. As time goes on, it could happen that the weight becomes concentrated on a few trajectories while a large number of trajectories carry very small weight. When this happens, resampling is often used to eliminate the trajectories with weak importance weight and duplicate trajectories with strong importance weight. Resampling, however, introduces Monte Carlo errors, thus too frequent resampling decreases efficiency. In practice, people judge the efficiency against a threshold using effective sample size and resample only when the effective sample size is sufficiently small. We follow [52] and [55] for the definition of effective sample size.

Effective sample size and the variance of the importance weights both reflect sample efficiency; as a matter of fact, the former is defined based on the latter. Suppose $\omega_t^{(j)}(j = 1, \ldots, N)$ are the normalized importance weights, let $W(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|D)}{g(\boldsymbol{x}|D)}$. From (3.6), it is obvious that $E_g(W(\boldsymbol{x})) = 1$, so that $C^2 = \frac{1}{N}\sum_{j=1}^{N}(N\omega_t^{(j)} - 1)^2$ approximates the variance of the importance weight as N is large. Effective sample size (ESS) is defined as

$$ESS = \frac{N}{1 + C^2}.$$

The intuition is as follows. Suppose N samples have been drawn, each with weight

either $\frac{1}{M}$ or 0, where M is the number of weights that are not 0. Then

$$\begin{aligned} C^2 &= \frac{(N-M)^2}{M} + N - M \\ &= N/M - 1 \end{aligned}$$

and

$$M = N/(1+C^2).$$

The effective sample size is the number of samples with non-zero weight, that is

$$ESS = M = \frac{N}{1+C^2},$$

and can be estimated by $\frac{1}{\sum_{j=1}^{N}(\omega_t^{(j)})^2}$.

With resampling, the simulation scheme in Section 2.3 goes with an additional step. At each time, ESS is calculated and compared with a threshold N_0; if $ESS > N_0$, nothing needs to be done; otherwise, $\boldsymbol{x}_t^{(j)}, j = 1, \ldots, N$, are resampled with replacement, each with probability $\omega_t^{(j)}$, and the new samples have uniform weight $\frac{1}{N}$.

Resampling is a practical technique to deal with the sample degeneracy problem. It does not eliminate the dependency in the sample caused by large sample weights; the simple convergence result in Section 3.2.2 is lost.

Chapter 4

A Generic Algorithm For Sequential Learning of States and Parameters

In sequential importance sampling approaches, the system evolution equation plays a central role in both the sampling step and the reweighting step. It describes the tendency of movement of a state from its previous state; given the distribution of the previous state, the system equation provides the prior distribution for the current state. For parameters that are assumed to remain constant over time, the posterior distribution at one time is simply its prior distribution for the next time. Imagine using the sequential importance sampling for the model fixed parameter vector $\boldsymbol{\theta}$. Given the posterior samples at time t, $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$, the prior distribution at time $t+1$ would only have probabilities at these discrete samples, and 0 probability elsewhere, so drawing new samples of $\boldsymbol{\theta}$ is not possible. Naively applying the sampling importance sampling (SIS) in the previous chapter leads to reweighting the fixed set of initial samples according to the likelihood function at each time. As one can imagine, sample attrition could be a serious problem.

We are going to introduce two approaches to get around the sample attrition problem: in section 1, adding artificial evolution noise for the fixed parameter, in section 2, smoothing with kernel density estimation via shrinkage. The close rela-

tionship between these two approaches will be revealed in section 3. Section 4 and 5 provides a general algorithm to sample the states and model fixed parameters simultaneously. Section 6 demonstrates the implementation of the algorithm with two examples: AR(1) and univariate stochastic volatility models.

4.1 Artificial Evolutions for Parameters

In dealing with time varying states, Gordon et al ([39]) suggested an approach to reducing the sample degeneracy/attrition problem by adding small random disturbances to state particles between time steps, in addition to any existing evolution noise contributions. In the literature since then, this idea has been extrapolated to fixed model parameters. One version of the idea adds small random perturbations to all the parameter particles under the posterior at each time point before evolving to the next. That is, consider a different model in which $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta}_t$ at time t, and simply include $\boldsymbol{\theta}_t$ in an augmented state vector. The approach usually adds an independent, zero-mean normal increment to the parameter at each time. That is,

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \boldsymbol{\omega}_{t+1} \\ \boldsymbol{\omega}_{t+1} &\sim N(\mathbf{0}, \mathbf{W}_{t+1}) \end{aligned} \tag{4.1}$$

for some specified variance matrix $\boldsymbol{W}_{t+1}$ and where $\boldsymbol{\theta}_t$ and $\boldsymbol{\omega}_{t+1}$ are conditionally independent given D_t. With the model recast with the corresponding augmented state vector, the standard filtering methods for states alone, such as the auxiliary particle filter, now apply. The key motivating idea is that the artificial evolution provides the mechanism for generating new parameter values at each time step in the simulation analysis, so helping to address the sample "attrition" issue in reweighting methods that stay with the same sets of parameter points between time steps.

Among the various issues and drawbacks of this approach, the key one is simply

that fixed model parameters are, well, fixed! Pretending that they are in fact time-varying implies an artificial "loss of information" between time points, resulting in posteriors that are, eventually, far too diffuse relative to the theoretical posteriors for the actual fixed parameters. To date there has been no resolution of this issue: if one adopts a model in which all parameters are subject to independent random shocks at each time point, the precision of resulting inferences is inevitably limited.

However, we show below that a reconciliation with kernel smoothing methods leads to a modification of this artificial evolution method in which the problem of information loss is avoided. First we discuss the basic form of our kernel smoothing method.

4.2 Kernel Smoothing with Shrinkage

Understanding the imperative to develop some method of smoothing for approximation of the required density $p(\boldsymbol{\theta}|D_t)$, West ([93]) developed kernel smoothing methods that provided the basis for rather effective adaptive importance sampling techniques. This represented extension to sequential analysis of basic mixture modeling ideas in West ([92]).

Standing at time t, suppose we have current posterior parameter samples $\boldsymbol{\theta}_t^{(j)}$ and weights $\omega_t^{(j)}$, $(j = 1, ..., N)$, providing a discrete Monte Carlo approximation to $p(\boldsymbol{\theta}|D_t)$. Here the t suffix on $\boldsymbol{\theta}$ indicates that the samples are from the time t posterior; $\boldsymbol{\theta}$ is not time-varying. Write $\overline{\boldsymbol{\theta}}_t$ and $\mathbf{V}_t$ for the Monte Carlo posterior mean and variance matrix of $p(\boldsymbol{\theta}|D_t)$, computed from the Monte Carlo samples $\{\boldsymbol{\theta}_t^{(j)}\}$ with weights $\{\omega_t^{(j)}\}$. The smooth kernel density form of West ([92], [93]) is given by

$$p(\boldsymbol{\theta}|D_t) \approx \sum_{j=1}^{N} \omega_t^{(j)} N(\boldsymbol{\theta}|\mathbf{m}_t^{(j)}, h^2\mathbf{V}_t) \tag{4.2}$$

with the following defining components. First, $N(\cdot|\mathbf{m}, \boldsymbol{S})$ is a multivariate normal

density mean $\mathbf{m}$ and variance matrix $\boldsymbol{S}$, so that the above density is a mixture of $N(\boldsymbol{\theta}|\mathbf{m}_t^{(j)}, h^2\mathbf{V}_t)$ distributions weighted by the sample weights $\omega_t^{(j)}$. Kernel rotation and scaling uses $\mathbf{V}_t$, the Monte Carlo posterior variance, and the overall scale of kernels is a function of the smoothing parameter $h > 0$, the specification of which follows conventional approaches (Silverman 1986) in which h is a slowly decreasing function of N. Hence kernel components are naturally more concentrated about their locations $\mathbf{m}_t^{(j)}$ for larger N. West ([92], [93]) suggests taking slightly smaller values than the conventional kernel methods as a general rule. As we discuss below, our new work has led to a quite different perspective on this issue.

The kernel locations $\mathbf{m}_t^{(j)}$ are specified using a shrinkage rule introduced by West. Standard kernel methods would suggest $\mathbf{m}_t^{(j)} = \boldsymbol{\theta}_t^{(j)}$ so that kernels are located about existing sample values. However, this results in a kernel density function that is *over-dispersed* relative to the posterior sample, in the sense that the variance of the resulting mixture of normals is $(1 + h^2)\mathbf{V}_t$, always larger than $\mathbf{V}_t$. This is a most significant flaw in the sequential simulation; an over-dispersed approximation to $p(\boldsymbol{\theta}|D_t)$ will lead to an over-dispersed approximation to $p(\boldsymbol{\theta}|D_{t+1})$, and the consequent "loss of information" will build up as the operation is repeated at future times. To correct this, West introduced the novel idea of shrinkage of kernel locations. Take

$$\mathbf{m}_t^{(j)} = a\boldsymbol{\theta}_t^{(j)} + (1-a)\overline{\boldsymbol{\theta}}_t \tag{4.3}$$

where $a = \sqrt{1-h^2}$. With these kernel locations, the resulting normal mixture retains the mean $\overline{\boldsymbol{\theta}}_t$ and now has the correct variance $\mathbf{V}_t$, hence the over-dispersion is trivially corrected.

4.3 Reinterpreting Artificial Parameter Evolutions

The undesirable "loss of information" in the method of Gordon et al is implicit in equation (4.1). The Monte Carlo approximation $\{\boldsymbol{\theta}_t^{(j)}, \omega_t^{(j)}\}$ to $p(\boldsymbol{\theta}|D_t)$ has mean $\overline{\boldsymbol{\theta}}_t$

and variance matrix $\boldsymbol{V}_t$. Hence, in the evolution in equation (4.1) with the innovation $\boldsymbol{\omega}_{t+1}$ independent of $\boldsymbol{\theta}_t$ as proposed, the implied prior $p(\boldsymbol{\theta}_{t+1}|D_t)$ has the correct mean $\overline{\boldsymbol{\theta}}_t$ but variance matrix $\boldsymbol{V}_t + \boldsymbol{W}_{t+1}$. The loss of information is explicitly represented by the component $\boldsymbol{W}_{t+1}$. Now, there is a close tie-in between this method and the kernel smoothing approach. To see this clearly, note that the Monte Carlo approximation to $p(\boldsymbol{\theta}_{t+1}|D_t)$ implied by equation (4.1) is also a kernel form, namely

$$p(\boldsymbol{\theta}_{t+1}|D_t) \approx \sum_{j=1}^{N} \omega_t^{(j)} N(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t^{(j)}, \boldsymbol{W}_{t+1}) \tag{4.4}$$

and this, as we have seen, is over-dispersed relative to the required or "target" variance $\boldsymbol{V}_t$.

It turns out that we can correct for this over-dispersion by modifying the Gordon et al method as follows. The key is to note that our kernel method is effective due to the use of location shrinkage. This shrinkage pushes samples $\boldsymbol{\theta}_t^{(j)}$ values towards their mean $\overline{\boldsymbol{\theta}}_t$ before adding a small degree of "noise" implied by the normal kernel. This suggests that the artificial evolution method should be modified by introducing correlations between $\boldsymbol{\theta}_t$ and the random shock $\boldsymbol{\omega}_{t+1}$. Assuming a non-zero covariance matrix, note that the artificial evolution equation (4.1) implies

$$V(\boldsymbol{\theta}_{t+1}|D_t) = V(\boldsymbol{\theta}_t|D_t) + \boldsymbol{W}_{t+1} + 2C(\boldsymbol{\theta}_t, \boldsymbol{\omega}_{t+1}|D_t).$$

To correct to "no information lost" implies that we set

$$V(\boldsymbol{\theta}_{t+1}|D_t) = V(\boldsymbol{\theta}_t|D_t) = \boldsymbol{V}_t,$$

which then implies

$$C(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t|D_t) = -\boldsymbol{W}_{t+1}/2.$$

Hence there must be a structure of negative correlations to remove the unwanted information loss effect. In the case of approximate joint normality of $(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t|D_t)$, this

would then imply the conditional normal evolution in which

$$p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) = N(\boldsymbol{\theta}_{t+1}|\boldsymbol{A}_{t+1}\boldsymbol{\theta}_t + (\boldsymbol{I} - \boldsymbol{A}_{t+1})\overline{\boldsymbol{\theta}}_t, (\boldsymbol{I} - \boldsymbol{A}_{t+1}^2)\boldsymbol{V}_t) \tag{4.5}$$

where $\boldsymbol{A}_{t+1} = \boldsymbol{I} - \boldsymbol{W}_{t+1}\boldsymbol{V}_t^{-1}/2$.

The resulting Monte Carlo approximation to $p(\boldsymbol{\theta}_{t+1}|D_t)$ is then a generalized kernel form with complicated shrinkage patterns induced by the shrinkage matrix $\boldsymbol{A}_{t+1}$. We restrict here to the very special case in which the evolution variance matrix $\boldsymbol{W}_{t+1}$ is specified using a standard discount factor technique. Specifically, take

$$\boldsymbol{W}_{t+1} = \boldsymbol{V}_t(\frac{1}{\delta} - 1)$$

where δ is a discount factor in $(0, 1]$, typically around $0.95 - 0.99$. In this case, $\boldsymbol{A}_{t+1} = a\boldsymbol{I}$ with $a = (3\delta - 1)/2\delta$ and the conditional evolution density above reduces

$$p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) \sim N(\boldsymbol{\theta}_{t+1}|a\boldsymbol{\theta}_t + (1 - a)\overline{\boldsymbol{\theta}}_t, h^2\boldsymbol{V}_t) \tag{4.6}$$

where $h^2 = 1 - a^2$, so that $h^2 = 1 - ((3\delta - 1)/2\delta)^2$, and we note that $a = \sqrt{1 - h^2}$. The resulting Monte Carlo approximation to $p(\boldsymbol{\theta}_{t+1}|D_t)$ is then precisely of the kernel form of equation (4.2), but now with a controlling smoothing parameter h specified directly via the discount factor. Thus we have the following theorem:

Theorem 4.1 *Assume* $\boldsymbol{\theta}_t|D_t \sim N(\overline{\boldsymbol{\theta}}_t, \boldsymbol{V}_t)$ *and* $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\omega}_{t+1}$. *If*

$$\boldsymbol{\omega}_{t+1} \sim N(\boldsymbol{0}, \boldsymbol{W}_{t+1})$$

and

$$C(\boldsymbol{\theta}_t, \boldsymbol{\omega}_{t+1}) = -\frac{\boldsymbol{W}_{t+1}}{2},$$

then

$$\boldsymbol{\theta}_{t+1}|\boldsymbol{D}_t \sim N(\overline{\boldsymbol{\theta}}_t, \boldsymbol{V}_t).$$

In addition, if

$$\boldsymbol{W}_{t+1} = \boldsymbol{V}_t(\frac{1}{\delta} - 1),$$

$$h^2 = 1 - (\frac{3\delta - 1}{2\delta})^2$$

and

$$a = \sqrt{1 - h^2},$$

then

$$\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t \sim N((1-a)\overline{\boldsymbol{\theta}}_t + a\boldsymbol{\theta}_t, h^2\boldsymbol{V}_t).$$

We therefore have a modification of the method of Gordon et al that connects directly with kernel smoothing with shrinkage, and justifies the basic idea of an artificial evolution for fixed model parameters in a modification that removes the problem of information loss over time. The modified "artificial" evolution model of equation (4.6) may be adopted directly, without reference to the motivating discussion involving normal posteriors. This is clear from the following general result. Suppose $p(\boldsymbol{\theta}_t|D_t)$ has a finite mean $\overline{\boldsymbol{\theta}}_t$ and variance matrix $\boldsymbol{V}_t$, *whatever the global form of the distribution may be.* Suppose in addition that $\boldsymbol{\theta}_{t+1}$ is generated by the evolution model specified by equation (4.6). It is then easily seen that the mean and variance matrix of the implied marginal distribution $p(\boldsymbol{\theta}_{t+1}|D_t)$ are also $\overline{\boldsymbol{\theta}}_t$ and $\boldsymbol{V}_t$. Hence the connection with kernel smoothing with shrinkage, and the adjustment to fix the problem of information loss over time in artificial evolution approaches, is quite general.

One of the practical relevances of the above theorem is that it provides a direct method of specifying the scale of kernels. A traditional specification of the smoothing parameter h in normal kernel density estimation is

$$h = c/N^{1/(1+4m)} \tag{4.7}$$

with

$$c = \frac{4}{1+2m}^{\frac{1}{1+4m}}$$

where p is the number of unknown fixed parameters. [91] and [83] suggested that, though the form of sample size dependence present in this formula can be justified and appears sensible, rather smaller values of the constant c are often desirable. [91] discussed the selection of the smoothing parameter with model-based approaches, but left more general, model independent guidelines for choosing this parameter an open research topic. Theorem 4.1 explicitly establishes a connection between the smoothing parameter in kernel density estimation via shrinkage with variance discount factor. Based on this relationship, the smoothing parameter h (and a) can be determined via the single discount factor δ. The practical applicability of this will be demonstrated in the applied examples later.

4.4 Filtering for States and Parameters

In the general model with fixed parameters $\boldsymbol{\theta}$, extend the sample-based framework as follows. Standing at time t, we now have a combined sample

$$\{\mathbf{x}_t^{(j)}, \boldsymbol{\theta}_t^{(j)} : \ j = 1, \ldots, N\}$$

and associated weights

$$\{\omega_t^{(j)} : \ j = 1, \ldots, N\}$$

representing an importance sample approximation to the time t posterior $p(\mathbf{x}_t, \boldsymbol{\theta}|D_t)$ for both parameter and state. Note that the t suffix on the $\boldsymbol{\theta}$ samples here indicate that they are from the time t posterior, *not* that $\boldsymbol{\theta}$ is time-varying. Time evolves to $t+1$, we observe $\mathbf{y}_{t+1}$, and now want to generate a sample from $p(\mathbf{x}_{t+1}, \boldsymbol{\theta}|\boldsymbol{D}_{t+1})$. Bayes theorem gives this as

$$p(\mathbf{x}_{t+1}, \boldsymbol{\theta}|\boldsymbol{D}_{t+1}) \ \propto \ p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \boldsymbol{\theta}) p(\mathbf{x}_{t+1}, \boldsymbol{\theta}|\boldsymbol{D}_t)$$

$$\propto \quad p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \boldsymbol{\theta})p(\mathbf{x}_{t+1}|\boldsymbol{\theta}, \boldsymbol{D}_t)p(\boldsymbol{\theta}|\boldsymbol{D}_t), \tag{4.8}$$

where the form chosen in the last equation makes explicit the notion that the theoretical density function $p(\boldsymbol{\theta}|D_t)$ is an important ingredient in the update.

If $\boldsymbol{\theta}$ were known, equation (4.8) simplifies: $p(\boldsymbol{\theta}|D_t)$ is degenerate and we drop the known parameter from the conditioning statements. This leads to equation (3.18) and the auxiliary particle method applies for filtering on the state vector. We deal with the fixed parameter part $p(\boldsymbol{\theta}|D_t)$ with the kernel smoothing with mixtures presented in (4.2). We have

$$\begin{aligned} & p(\mathbf{x}_{t+1}|\boldsymbol{\theta}, \boldsymbol{D}_t)p(\boldsymbol{\theta}|\boldsymbol{D}_t) \\ \approx \quad & p(\mathbf{x}_{t+1}|\boldsymbol{\theta}, \boldsymbol{D}_t) \sum_{j=1}^{N} \omega_t^{(j)} N(\boldsymbol{\theta}|\mathbf{m}_t^{(j)}, h^2\mathbf{V}_t) \\ \propto \quad & \sum_{j=1}^{N} \omega_t^{(j)} p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(j)}, \boldsymbol{\theta}) N(\boldsymbol{\theta}|\mathbf{m}_t^{(j)}, h^2\mathbf{V}_t) \end{aligned} \tag{4.9}$$

which, in (4.8) above, leads to

$$p(\mathbf{x}_{t+1}, \boldsymbol{\theta}|\boldsymbol{D}_{t+1}) \propto \sum_{j=1}^{N} \omega_t^{(j)} p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \boldsymbol{\theta}) p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(j)}, \boldsymbol{\theta}) N(\boldsymbol{\theta}|\mathbf{m}_t^{(j)}, h^2\mathbf{V}_t).$$

4.5 A General Algorithm

Return now to the general filtering problem, that of sampling the posterior in equation (4.8). We have available the Monte Carlo sample $(\mathbf{x}_t^{(j)}, \boldsymbol{\theta}_t^{(j)})$ and weights $\omega_t^{(j)}$ $(j = 1, \ldots, N)$, representing the joint posterior $p(\mathbf{x}_t, \boldsymbol{\theta}|\boldsymbol{D}_t)$. Again, the suffix t on

the parameter samples indicates the time t posterior, not time-variation. We adopt the kernel form of equation (4.2) as the marginal density for the parameter, following the earlier discussion. With the equivalent interpretation of this as arising from an artificial evolution with correlation structure, as just discussed, we can now apply an extended version of the auxiliary particle filter algorithm, incorporating the parameter with the state. The resulting general algorithm runs as follows.

1. For each $j = 1, ..., N$, identify the prior point estimates of $(\mathbf{x}_t, \boldsymbol{\theta})$ given by $(\mu_{t+1}^{(j)}, \mathbf{m}_t^{(j)})$ where

$$\mu_{t+1}^{(j)} = E(\mathbf{x}_{t+1}|\mathbf{x}_t^{(j)}, \boldsymbol{\theta}_t^{(j)}).$$

   may be computed from the state evolution density and $\mathbf{m}_t^{(j)} = a\boldsymbol{\theta}_t^{(j)} + (1-a)\overline{\boldsymbol{\theta}}_t$ is the j^{th} kernel location from equation (4.3).

2. Sample an auxiliary integer variable from the set $\{1, ..., N\}$ with probabilities proportional to

$$g_{t+1}^{(j)} \propto \omega_t^{(j)} p(\mathbf{y}_{t+1}|\mu_{t+1}^{(j)}, \mathbf{m}_t^{(j)});$$

   call the sampled index k.

3. Sample a new parameter vector $\boldsymbol{\theta}_{t+1}^{(k)}$ from the k^{th} normal component of the kernel density, namely

$$\boldsymbol{\theta}_{t+1}^{(k)} \sim N(\cdot|\mathbf{m}_t^{(k)}, h^2\boldsymbol{V}_t).$$

4. Sample a value of the current state vector $\mathbf{x}_{t+1}^{(k)}$ from the system equation

$$p(\cdot|\mathbf{x}_t^{(k)}, \boldsymbol{\theta}_{t+1}^{(k)}).$$

5. Evaluate the corresponding weight

$$\omega_{t+1}^{(k)} \propto \frac{p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}^{(k)}, \boldsymbol{\theta}_{t+1}^{(k)})}{p(\mathbf{y}_{t+1}|\mu_{t+1}^{(k)}, \mathbf{m}_t^{(k)})}.$$

6. Repeat step (2)-(5) a large number of times to produce a final posterior approximation $(\mathbf{x}_{t+1}^{(k)}, \boldsymbol{\theta}_{t+1}^{(k)})$ with weights $\omega_{t+1}^{(k)}$, as required.

Note that the Monte Carlo sample size N can be different at each time point, if required. Also, we might be interested in over-sampling a rather larger set of values and then resampling according to the weights above in order to produce an equally weighted final sample. Resampling can be applied before step 2 using a certain threshold for effective sample size. As with the auxiliary particle filters, this algorithm applies as far as the prior for the states can be sampled from and the likelihood function can be evaluated. The computational load of our algorithm at each time step is in the order of the sample size, which is about the square root of the amount of computation required by earlier kernel smoothing methods.

4.6 Examples

4.6.1 AR(1)

As a simple example in which the sequential updating is available in closed form for comparison, consider the AR(1) model in which $\mathbf{y}_t = x_t$, a scalar, with $x_{t+1} \sim N(x_t\phi, 1)$. Here $\boldsymbol{\theta} = \phi$, a single parameter, and there is no unobserved state variable. Hence the focus is exclusively on the efficacy of learning the parameter (steps 1 and 4 of the general algorithm above are vacuous).

A realization of length 897 was generated from this AR(1) model at $\phi = 0.8$ and $x_1 = 0$. The sequential analysis was then performed over times $t = 1, \ldots, 897$, and the posterior approximation at $t = 897$ compared to the exact (normal) posterior for ϕ. The prior used for ϕ is $N(0.6, 0.25)$. The simulation-based analysis used $N = 5000$ sample points throughout. Figure 4.1 graphs the time trajectories of the posterior sample quantiles (2.5%, 25%, 50%, 75%, 97.5%) together with the exact posterior

quantiles. Agreement is remarkable across the entire time period, with very little evidence of "build up" of approximation error. Table 4.1 provides a numerical comparison of exact and approximate quantiles for $p(\phi|D_{897})$ which further illustrates the accuracy. By comparison, when step 3 above is applied without shrinkage, the sample mean oscillates dramatically, indicating over-reaction to likelihood information. Without shrinkage, the sample mean is nowhere close to the true posterior mean even for this simple model.

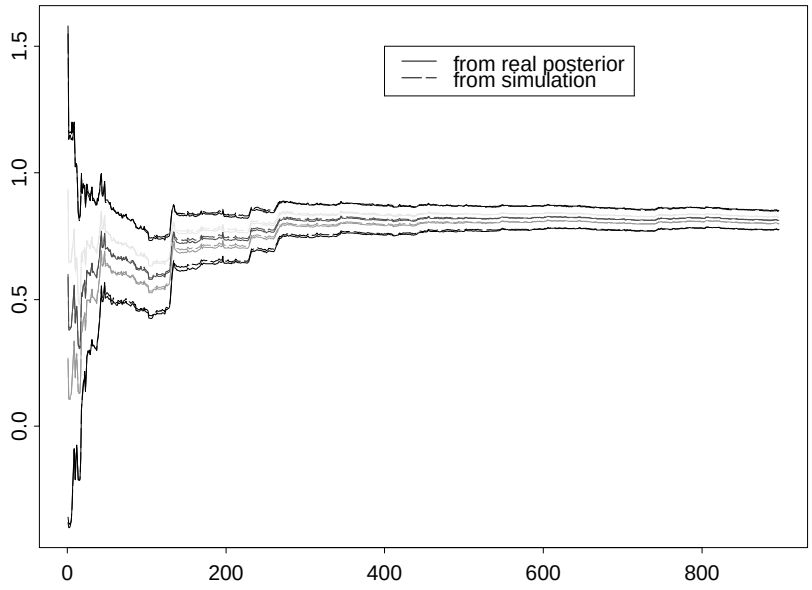

Figure 4.1: Time trajectories of posterior quantiles (2.5%, 25%, 50%, 75%, 97.5%) of the posteriors for the AR(1) parameter ϕ.

	0.025	0.25	0.50	0.75	0.975
exact:	0.7755	0.8005	0.8136	0.8267	0.8517
approx:	0.7758	0.7993	0.8119	0.8242	0.8482

Table 4.1: Posterior quantiles from the posterior for the AR(1) parameter ϕ at $t = 897$.

In fact, when the real prior is normal, it can be shown theoretically that kernel smoothing with shrinkage samples from the proper MC approximation of the posterior. To see this, suppose at a certain time the prior distribution of the fixed parameter $\boldsymbol{\theta}$ is normal with mean $\boldsymbol{\theta}_0$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, and $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$ are prior samples. Our simulation strategy gives samples from

$$\begin{aligned}
& \sum_{j=1}^{N} N(\boldsymbol{\theta}|a\boldsymbol{\theta}^{(j)} + (1-a)\overline{\boldsymbol{\theta}}, \boldsymbol{V}(1-a^2))L(\boldsymbol{y}_{t+1}|\boldsymbol{\theta}, \boldsymbol{x}_{t+1}) \\
\overset{app}{\propto} & \int N(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(0)}, \boldsymbol{\Sigma}_\theta) N(\boldsymbol{\theta}|a\boldsymbol{\theta}^* + (1-a)\boldsymbol{\theta}_0, \boldsymbol{V}(1-a^2)) d\boldsymbol{\theta}^* L(\boldsymbol{y}_{t+1}|\boldsymbol{\theta}, \boldsymbol{x}_{t+1}) \\
\propto & N(\boldsymbol{\theta}|\boldsymbol{\theta}_0, a^2\boldsymbol{\Sigma}_\theta + \boldsymbol{V}(1-a^2))L(\boldsymbol{y}_{t+1}|\boldsymbol{\theta}, \boldsymbol{x}_{t+1}) \\
\overset{app}{\propto} & N(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_\theta)L(\boldsymbol{y}_{t+1}|\boldsymbol{\theta}, \boldsymbol{x}_{t+1}).
\end{aligned}$$

The theory of kernel density estimation assures that any density function can be arbitrarily well approximated by a mixture of normals; with this special case, we have shown that if the "real" prior of $\boldsymbol{\theta}$ is normal at each time, *smoothing via shrinkage* gives samples from approximation of the proper posterior, regardless of the likelihood.

4.6.2 Stochastic Volatility

In the second example, we implement a univariate stochastic volatility model:

$$y_{t+1} = \epsilon_t exp(h_t/2), \epsilon_t \sim N(0,1),$$

$$h_{t+1} - \mu = \phi(h_t - \mu) + \eta_{t+1}, \eta_{t+1} \sim N(0, \sigma_\eta^2),$$

$$\beta = exp(\mu/2),$$

where $\boldsymbol{\epsilon}_t$ and η_t are mutually and internally independent normal processes with variance 1 and σ^2 respectively. Here h_t is the log of the volatility of the series, $\boldsymbol{\phi}$ describes

the persistence in the volatility shocks and σ^2 is the volatility of the log-volatility innovation. This model and its variations constitute an important class of volatility models that have recently attracted much attention in econometric literatures. More details of stochastic volatility models will be presented in the next chapter, here we fit this simple model with the proposed sequential updating algorithm to the weekday closes (difference of the log of the series) on the Pound Sterling/US dollar exchange rate from Oct/1/1981 on for 900 weekdays. This dataset has been previously analyzed using quasi-maximum likelihood in [43], and using MCMC in a Bayesian framework in [48]. The closest to our work is [67], which uses auxiliary particle filters on h_t and stratification on the fixed parameters.

We tried generating samples from some dispersed priors for the parameters: $\phi \sim 2Beta(20, 1.5) - 1, \sigma^2_{y_{it}} \sim 0.05/\chi^2_5$ and a diffused prior for $log\beta$, and applied particle filters using these samples as prior samples. This did not yield results close to that from MCMC even for updating of 50 steps with sample size 5000. When the priors are really dispersed, particle filters lose predictability because a great number of samples at early stages are essentially discarded at later stages.

To get samples from a relatively concentrated prior, we use MCMC method to process data at times 1-300 and proceed via sequential sampling using the MCMC samples as prior samples. Table 4.2 shows the posterior mean from the particle filters compared with that from the MCMC methods: the first two rows are posterior means at time 350 and the last two rows show the posterior means at time 900. In both cases, the sample size is 5000. Not only do MCMC and particle filters give acceptably close results in both cases, but the deviation between the two results is not systematically widened from time 350 to time 900.

	ϕ	β	σ_η
smoothing(.$\|\boldsymbol{Y}_{350}$)	0.969	0.724	0.115
MCMC (.$\|\boldsymbol{Y}_{350}$)	0.973	0.737	0.114
smoothing(.$\|\boldsymbol{Y}_{900}$)	0.975	0.634	0.123
MCMC (.$\|\boldsymbol{Y}_{900}$)	0.979	0.630	0.107

Table 4.2: Posterior means of model fixed parameters at time 350 and time 900

Chapter 5

Sequential Variance Learning through Discounting

5.1 Stochastic Variance Models

Volatility plays a central role in much of modern finance theory. In options pricing, volatility of the future spot price of the underlying asset is an important input in the pricing function. In asset pricing theory, the risk premium is determined by the covariance between the return of the asset and the return of the market portfolio. In modern portfolio theory, the optimal portfolio is constructed through a compromise of risk and return, where the risk of a portfolio is determined by the variances of and the covariances between the assets in the portfolio.

The assumption of constant volatility is well known to be unrealistic. The standard technique to model changing volatility in Bayesian Dynamic Linear Models (DLMs) is variance discounting, which discounts the degrees of freedom in the posterior inverse Wishart distribution of the covariance matrix to allow for more uncertainty about the covariance in the next step prior. It is basically a random walk type of model and does not have real forecasting capability. However, a closed-form sequential learning equation is available for a broad class of dynamic linear models with

variance discounting; in its limiting form, those updating equations agree with the practically widely used exponentially weighted moving average. The simplicity and analytical tractability remain a great appeal for the variance discounting approach.

In recent years, research in finance and econometrics has devoted significant efforts to come up with more sophisticated stochastic variance models; two major types of volatility models - ARCH/GARCH models and stochastic volatility (SV) models have been developed. ARCH([27])/GARCH([8]) types of models appeared in the early 1980's; they model the conditional variance as a moving average of the squared past observation errors and an autoregression of the past conditional variances. Numerous extensions with univariate structure were rapidly developed, among which Exponential GARCH (EGARCH) ([65]) assumes the logarithm of the conditional variance follow an ARMA process, guaranteeing that all the conditional variances are positive. Integrated GARCH(1,1) (IGARCH) ([29]) has coefficients of the GARCH process summing to one, it is non-stationary with the variance process analogous to ARIMA(0,1,1), thus captures the persistence in the conditional variance. ARCH-in-Mean ([30]) models assume the conditional mean to be an explicit function (usually linear or logarithmic) of the conditional variance, reflecting the trade-off between risk and return in many theories in finance. Maximum likelihood estimation of GARCH models is usually straightforward even for multivariate models. Empirical studies of foreign exchange and equity markets have shown that GARCH models can provide somewhat better estimates than the simple moving averages. Unfortunately, generalization of GARCH models to the multivariate case has mostly been difficult due to the large number of unknown parameters; restrictions have to be imposed to simplify the model. [10] implemented a multivariate GARCH(1,1)-in-Mean model with constant diagonal coefficient matrix in the process for the conditional volatility. [9] proposed a multivariate GARCH representation in which the conditional correlation

is constant over time. Reviews of literatures on GARCH models are given in [21] and [12].

The standard stochastic volatility model assumes an AR(1) process for the log of the volatility; it is the discrete time version of the continuous time model on which modern derivative pricing theory is based, and its properties are easy to derive. Estimation is very difficult because the likelihood is not tractable; implementation of stochastic volatility models are commonly through quasi-maximum likelihood ([64] and [43]) and method of moments ([60]). Alternatively, people have used MCMC techniques: [46] and [79] implemented a single move Gibbs sampler; [82] improved it by sampling all the log volatilities in a group using a Metropolis Hastings algorithm; [48] used an offset mixture of normals approximation to log-Chi-square distribution coupled with an importance reweighting procedure to improve the efficiency. [43] generalized SV to multivariate setting, with constraints similar to the constant conditional correlation GARCH model.

To remove the additional restrictions in multivariate volatility models, factor representation has been used to reach a parsimonious yet flexible parameterization. [28] proposed a K-factor GARCH model, in which the conditional covariance matrix depends on the conditional variances of K orthogonal linear combinations of $\boldsymbol{y}_t$; [31] suggested a simple two-stage estimation procedure for this model; [11] gave conditions for covariance stationarity of the K-factor GARCH models and showed how multivariate IGARCH models allow for the possibility of co-persistence in variance. [25] proposed a latent factor model in which the common movements in volatility are represented by a single unobserved latent factor subject to ARCH effect; later, [49] replaced the ARCH process with an SV process. [47] discussed using MCMC methods on the factor-volatility model in which both the idiosyncratic observation errors and the volatility processes are independent; [48] suggested using the offset

mixture approach in simulation. Subsequently, [81] and [1] independently implemented factor-SV models of moderate dimensions; recently, [19] provided a scalable MCMC sampling algorithm: a reduced blocking scheme for sampling the factor loading matrix and factors, the use of [18] method for sampling the parameters of the univariate SV model, and the sampling of the loading matrix via a tuned M-H step based on a local approximation to the current density of the sample, marginalized over the factors.

The rest of the thesis addresses sequential learning in multivariate dynamic volatility models. This chapter aims at providing a sequential learning scheme for the traditional variance discounting DLM in multivariate framework. We set out the basis for variance discounting, especially from the view point of a random walk process for the conditional variance which is required by sequential simulation. Some theoretical developments are made to address this process and other issues that have arisen in the practical application of variance discounting to financial time series. A random walk process is provided for the variance matrix in singular DLMs and a particle filtering scheme proposed for general, potentially singular, vector DLMs. In the mean time, an MCMC simulation scheme is put forth to facilitate the implementation of the sequential simulation. In chapter 6, following the line of [1], we apply the proposed particle filtering scheme to the multivariate SV model in factor representation. This model not only poses new challenges for sequential simulation in multivariate context, but also possesses interesting connections with the traditional variance discounting model. In chapter 7, we conclude with some discussions and thoughts about future directions.

5.2 Dynamic Linear Model and Variance Discounting in Financial Time Series Modeling

Bayesian dynamic linear models, a class of dynamic linear regression models, have experienced growth in real life application in recent years. Among those dedicated to apply Bayesian models to financial time series, [73] proposed using multivariate Bayesian DLMs in the framework of [94], [72] and [76] to forecast currency returns. In contrast to the widely used static regression models, Bayesian DLMs allows the regression coefficients to change over time, thus are better at capturing market trends. In addition, Bayesian DLMs provide the mechanism to naturally incorporate external information as input apart from historical data. This advantage of Bayesian methodology is fully appreciated and utilized in the well known work of [7]. These models were later developed into a relative seemingly unrelated regression model with relative dynamic shrinkage ([71]), which is able to model simultaneously asset returns across asset classes and locations. This model uses relative return in place of the "absolute" return so that it is invariant to the choice of base currency. It also exploits information pooling, or the so called shrinkage technique (pioneered in [98]), to improve estimation in cross sectional regression by "borrowing strength" from other sections when estimation is made in the regression for one section. These techniques materially improved the forecasting performance while posing new challenges to the research of Bayesian DLMs: with the linear transformation that transforms the absolute return to the relative return, the observation noise follows a singular distribution; "shrinkage" makes the updating and forecasting more difficult by adding an extra equation to the model.

These models assume unknown variances and covariances and model them via variance discounting. Closed form updating and forecasting is not available for this category of practically interesting multivariate DLMs. Currently ad hoc deterministic

approximation approaches are used for parameter estimation and forecasting. The wide application of MC techniques in applied statistics has naturally led people to consider simulation based learning methods for multivariate DLMs. As is obvious in the previous chapters, in MC simulation of state space models, the system equation that links the states at consecutive time steps plays a central role. In multivariate DLMs, the states are the time-varying regression coefficients and the variances of the noise terms. The evolution equation for the regression coefficients is defined by the modeler as part of the model setup. With variance discounting, however, the modeler does not explicitly specify a Markovian process for the variances. For univariate conjugate DLMs, in which the precision, which is the inverse of the variance, has a Gamma form in prior and posterior, an evolution equation that involves an independent beta variable conveniently "explains" the discounting process. For multivariate DLMs, [86] and [74], [75] introduced a similar stochastic equation for the precision process using matrix-variate Beta distributions. Constrained by the fact that matrix Beta distributions are only defined for parameters in a certain integer range, [86] used a fixed degree of freedom for the Wishart in the posterior of the precision matrix at all time steps. Since using the "growth" of the degrees of freedom in the posterior distribution to reflect the accumulation of information over time is a key concept in variance discounting, we find this treatment undesirable. [53] made an attempt to generalize [86]'s variance discounting process to incorporate the singular case. However, an important detail was neglected in [53]'s work, which essentially made the discounting processes permitted by his theories unrealistic.

Along the line of [86] and [53], in this chapter we attempt to provide the theoretical basis for variance discounting in singular models and a sensible discount rule that compromises the limited domain of the parameters in the matrix-variate Beta distribution with the real need that the degrees of freedom of the posterior

increase over time. Based on these results, we apply our particle filter to potentially singular discount DLMs and demonstrate the feasibility of our algorithm with a simplified example. As noted from our experience, sequential MC techniques should always be monitored and used with standard non-sequential techniques, typically MCMC. Efficient MCMC approaches for multivariate DLMs are to be developed, presumably due to the afore mentioned theoretical challenges and the difficulty in simulating matrix variate random variables. We propose a forward filtering backward sampling algorithm with Metropolis-Hasting rejection sampling imbedded in the backward sampling step for the DLMs of interest.

5.3 General Forms of Bayesian DLMs and the Concept of Variance Discounting

In this section, we introduce the general forms of DLM following [94] with emphasis on the concept of variance discounting.

A basic univariate DLM takes the form

$$\begin{aligned} Y_t &= \mathbf{F}_t'\boldsymbol{\theta}_t + \nu_t, \nu_t \sim N(0, V_t) \\ \boldsymbol{\theta}_t &= \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \boldsymbol{\omega}_t \sim N(0, \mathbf{W}_t) \end{aligned} \tag{5.1}$$

in which

- Y_t is a scalar observation, the response variable;
- $\mathbf{F}_t$ is a known n-vector of regression covariates;
- $\boldsymbol{\theta}_t$ is a vector of unknown time-varying regression coefficients, which follows a Markov process given by the system equation;

- $\mathbf{G}_t$ is $n \times n$ matrix describing how the regression coefficients evolve linearly over time;

- ν_t and $\boldsymbol{\omega}_t$ are internally and mutually independent series of observation noise and system evolution noise. Their variances V_t and $\mathbf{W}_t$ are sometimes assumed to be known.

Despite their simple forms, DLMs incorporate a variety of traditional time series models such as Holt's linear growth model ([45]), exponentially weighted regression ([14]), ARIMA ([13]). These constant models may be locally correct, but the model parameters could change markedly over time. Bayesian DLMs provide means of responding and adapting to changes in underlying conditions and related variables; the freedom that the modelers have to choose $\mathbf{G}_t$, V_t and $\mathbf{W}_t$ and to adjust the priors on-line allows monitoring and intervention. Given V_t and $\mathbf{W}_t$, closed-form normal updating, forecasting and smoothing equations are available as described in [94].

In applying this model, one has to specify V_t and $\mathbf{W}_t$, which can be a crucial step. $\mathbf{W}_t$ is usually treated with the standard variance discounting approach; this chapter focuses the discussion on V_t. The simplest way to deal with V_t is to assume it to remain constant over time. If an inverse-gamma prior is used for V, then the model is still analytically tractable: the conjugate inverse-gamma distribution for V is kept at each updating step; conditional on V, the smoothing, filtering and forecasting equations have the usual normal form; with V integrated out, the marginal smoothing, filtering and forecasting distributions follow t distributions.

The disadvantage of the fixed V assumption is that it eventually leads to convergence in the posterior of V. Following the notation of [94], the posterior of the precision $\phi = V^{-1}$ at time t is given by $(\phi|D_t) \sim G(n_t/2, d_t/2)$, where the d.f. n_t is increased by 1 for each observation. Thus, as $t \to \infty$, $n_t \to \infty$ and the posterior converges about the mode. With the usual point estimate $S_t = d_t/n_t$ of V, the posterior

asymptotically degenerates with $|\phi - S_t| \to 0$ with probability one. Although this is an asymptotic result, the posterior can become quite precise rapidly as n_t increases, leading to under-adaptation to new data so far as the learning on the variance is concerned. The problem lies with the assumption of constancy of V, conflicting with the belief in change over time.

The simplest way of modeling changes in V is via some form of random walk for V. At time $t-1$, suppose the precision ϕ_{t-1} has the usual posterior

$$(\phi_{t-1}|D_{t-1}) \sim G(n_{t-1}/2, d_{t-1}/2) \tag{5.2}$$

Proceeding to time t, it is desirable to retain the gamma form of distribution for the resulting distribution $(\phi_t|D_{t-1})$, as it is conjugate to the likelihood function for updating based on the next observation Y_t. This led to the development of "variance discounting" to model a decay of information about the precision between time points, while retaining the gamma form of posterior and prior distributions. The degree of the decay is determined by a constant variance discount factor δ with $0 < \delta \leq 1$. $\delta = 1$ means a constant precision, thus no decay of information, the smaller the δ, the larger the decay. Based on the time $t-1$ posterior (5.2), suppose that ϕ_t is derived from ϕ_{t-1} by some "random walk" model resulting in the time t prior distribution

$$(\phi_t|D_{t-1}) \sim G(\delta n_{t-1}/2, \delta d_{t-1}/2). \tag{5.3}$$

This has the same location as (5.2), i.e. $E(\phi_t|D_{t-1}) = E(\phi_{t-1}|D_{t-1})$, but increased dispersion through the discounting of the degrees of freedom parameter.

A formal model underlies this use of variance discounting, derived from a special case of the result in [86]. Suppose that, given D_{t-1}, $\gamma_t \sim Beta(\delta n_{t-1}/2, (1-\delta)n_{t-1}/2)$ independently of ϕ_{t-1}, such that $E(\gamma_t|D_{t-1}) = \delta$. Set

$$\phi_t = \gamma_t \phi_{t-1}/\delta \tag{5.4}$$

It can be deduced that under (5.2) and (5.4), the marginal distribution of ϕ_t given D_{t-1} is (5.3). Hence equation (5.4) can be introduced as a system equation to formally model the stochastic process of the precision ϕ_t. The maintenance of the conjugate gamma prior and posterior enables continued closed-form sequential updating.

Our discussion so far has been restricted to univariate DLM where the response variable Y_t is a scalar. Generalization can be made to multivariate or matrix-variate DLM.

A framework for matrix-variate DLMs with unknown cross series structure is presented in [72], [76] and [77]. This class of conjugate Matrix DLMs has observation and system equations generalizing those for the univariate DLM:

$$\begin{aligned} \boldsymbol{y}_t' &= \mathbf{F}_t'\boldsymbol{\Theta}_t + \boldsymbol{\nu}_t', \boldsymbol{\nu}_t \sim N(\mathbf{0}, V_t\boldsymbol{\Sigma}) \\ \boldsymbol{\Theta}_t &= \mathbf{G}_t\boldsymbol{\Theta}_{t-1} + \boldsymbol{\Omega}_t, \boldsymbol{\Omega}_t \sim N(\mathbf{0}, \mathbf{W}_t, \boldsymbol{\Sigma}) \end{aligned} \tag{5.5}$$

where

- $\boldsymbol{y}_t = (Y_{t1}, \ldots, Y_{tq})'$ is a q-vector of observation;
- $\mathbf{F}_t$ is an n-vector;
- $\boldsymbol{\nu}_t$ is a q-vector of observation noise terms;
- $\boldsymbol{\Theta}_t$ is an $n \times q$ matrix whose columns are the state vectors of the individual univariate DLMs for the scalar series y_{tj};
- $\boldsymbol{\Omega}_t = (\boldsymbol{\omega}_{t1}, \ldots, \boldsymbol{\omega}_{tq})$, is an $n \times q$ matrix whose columns are the evolution errors of the individual DLMs.

If V_t and $\mathbf{W}_t$ are known, and the initial prior for $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Sigma}$ is normal/inverse Wishart

$$(\boldsymbol{\Theta}_0, \Sigma|D_0) \sim NW_{n_0}^{-1}(\mathbf{m}_0, \mathbf{C}_0, \mathbf{S}_0),$$

then closed-form updating and forecasting equations are available. The key features of this model are: the series share the same defining elements $\mathbf{F}_t$, $\mathbf{G}_t$ and $\mathbf{W}_t$; the observation noise terms have a constant correlation structure but a common time-varying scale in their variances; the correlation structure in the observation noise terms also defines the correlations between the state vectors in the individual DLMs. These features put strict restrictions on the systems that can be modeled with matrix DLMs. Some applications, for example the dynamic seemingly unrelated regression used to model correlated time series with series-specific factors in [71], require a more general and flexible setup.

A general multivariate DLM has the following form

$$\begin{aligned} \boldsymbol{y}_t &= \mathbf{F}_t\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \boldsymbol{\nu}_t \sim N(\mathbf{0}, \mathbf{V}_t) \\ \boldsymbol{\theta}_t &= \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \boldsymbol{\omega}_t \sim N(\mathbf{0}, \mathbf{W}_t) \end{aligned} \tag{5.6}$$

where

- $\boldsymbol{y}_t$ is an m-vector of observation;
- $\mathbf{F}_t$ is a known $m \times n$ matrix;
- $\boldsymbol{\nu}_t$ is a m-vector of observation noise.

The other assumptions are the same as in (5.1). When $\mathbf{V}_t$ and $\mathbf{W}_t$ are unknown, there are not closed-form updating and forecasting equations for this model. In search of the model that underlies variance discounting in multivariate framework, [86] proposed a stochastic model for the observation precision $\boldsymbol{\Phi}_t = \mathbf{V}_t^{-1}$:

$$\begin{aligned} \boldsymbol{\Phi}_t &= U(\boldsymbol{\Phi}_{t-1})'\mathbf{B}_t U(\boldsymbol{\Phi}_{t-1})/\delta \\ \mathbf{B}_t &\sim B(\delta n_{t-1}/2, (1-\delta)n_{t-1}/2) \\ \boldsymbol{\Phi}_{t-1} &\sim W(\mathbf{S}_{t-1}, n_{t-1}) \end{aligned} \tag{5.7}$$

where $U(\mathbf{\Phi}_{t-1})$ denotes the upper Cholesky decomposition of $\mathbf{\Phi}_{t-1}$; $W(\mathbf{S}_{t-1}, n_{t-1})$ the Wishart distribution with shape matrix $\mathbf{S}_{t-1}$ and d.f. n_{t-1}; $B(\delta n_{t-1}/2, (1-\delta)n_t/2)$ is the matrix Beta distribution whose definition will be given in the next section; and given D_{t-1}, $\mathbf{\Phi}_{t-1}$ and $\mathbf{B}_t$ are conditionally independent. Like in the univariate case, (5.7) leads to the marginal of $\mathbf{\Phi}_t$ following $W(\delta\mathbf{S}_{t-1}, \delta n_{t-1})$. Matrix Beta distributions are defined through two Wishart variables, and the convolution between Beta and Wishart is used here to get a relatively dispersed Wishart from a Wishart with higher precision.

The readers are referred to [62] for the general theory of the Wishart and matrix Beta distribution; [94] and [13] also provide comprehensive background of the (inverse) Wishart distribution. It is worth pointing out that two sets of notations are used in this literature – [62] and [86]'s d.f. for Wishart is actually the d.f. in [94] and [13] plus the dimension m minus 1. So in variance discounting, [94] advises using a very small positive number as the d.f. in the prior for variance, representing a weak prior, whereas in the notation of [86], the equivalent prior would have d.f. a bit greater than $m-1$. We shall follow the convention of [62] and [86] which is more convenient as far as the matrix Beta distribution is concerned.

In the above literature, discussions on Wishart distributions are restricted to the cases where the shape matrix is positive definite, and the d.f. is no less than the dimension m. The Wishart and matrix Beta defined in these literatures yield δ falling outside the normally adopted range for discount factors, 0.90 to 0.99. To make the discounting practically feasible, [86] extended Wishart distribution to the pseudo case, and extended multivariate Beta distribution to rank-1. The usual convolution was extended to this case so that the discounting process of (5.7) with proper discount factor is justified.

In using (5.7) to forecast global asset returns, [74] used a singular linear trans-

formation to make the model forecast invariant to the choice of the base currency. Simply speaking, if $\boldsymbol{y}_t$ denotes the unobserved absolute returns of assets, the transformation to the observed relative returns $\boldsymbol{y}_t^{(o)}$ involves left multiplying $\boldsymbol{y}_t$ by a matrix $\mathbf{D}$. In the case of four asset sections, for example, $\mathbf{D}$ is a block sparse matrix with the structure

$$\mathbf{D} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{I} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ -\mathbf{I} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

If the unobserved absolute returns can be modeled by the dynamic regression

$$\boldsymbol{y}_t = \mathbf{F}_t\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \boldsymbol{\nu}_t \sim N(\mathbf{0}, \mathbf{V}_t)$$

and some process for $\boldsymbol{\theta}_t$, then the equivalent observation equation for $\boldsymbol{y}_t^{(o)}$ is

$$\boldsymbol{y}_t^{(o)} = \mathbf{F}_t^{(o)}\boldsymbol{\theta}_t + \boldsymbol{\nu}_t^{(o)}, \boldsymbol{\nu}_t^{(o)} \sim N(\mathbf{0}, \mathbf{V}_t^{(o)})$$

where $\boldsymbol{y}_t^{(o)} = \mathbf{D}\boldsymbol{y}_t$, $\mathbf{F}_t^{(o)} = \mathbf{D}\mathbf{F}_t$, and $\mathbf{V}_t^{(o)} = \mathbf{D}\mathbf{V}_t\mathbf{D}'$. Obviously, $\mathbf{F}_t^{(o)}$ and $\mathbf{V}_t^{(o)}$ are singular.

Singular DLMs are also potential models for series that split, merge or emerge. Suppose a trader uses (5.6) to forecast the returns of 10 equities. One day, a new stock draws his attention; he immediately wants to incorporate it in the forecasting model. The new stock however only has 3 months of historical data, while the stocks in the current model have over 3 years of data. Now the question arises: how to treat the series of unequal lengths? One can use the last 3 months' data and throw away some data for the series currently in the model. Alternatively, the trader may want to make good use of all the data available and set up this model as eleven dimensional from the beginning. In this model, the variance matrices for the returns in the first 33 months would be singular, with the variance for the new stock always zero. Conceivably, at the time of adding the new stock's data to the analysis, there

is a conflict between the prior for the variance matrix in a 10 dimensional subspace and the likelihood which requires the variance matrix in 11 dimensional space. With the flexibility provided by Bayesian methodology, this can be trivially overcome.

Even when singularity is not caused by model setup or the data series itself as in the above examples, it could still appear in numerical computation. These questions address the need to generalize the results of [86] to singular Wisharts. [53] attempted to make such generalization, but like [13] and [62], his results only permit the range of discount factor far below the values used in practice. In the next two sections, we derive the singular Wishart density, the singular Wishart - Beta convolution and propose a sequential simulation and a MCMC scheme for multivariate DLMs based on our results and [86].

5.4 Some Theoretical Problems in Singular Variance Discounting

The properties of Wishart distributions when the shape matrix $\mathbf{S}$ is positive definite, and the d.f. n is greater than the dimension $m-1$ are well known.

Definition 5.1 *A positive definite matrix* $\boldsymbol{A}$ *is* $W_m(n, \boldsymbol{S})$ *(*$n > m-1$, $\boldsymbol{S} > 0$*) distributed iff it has the density function*

$$\frac{1}{2^{mn/2}\Gamma_m(n/2)(det\boldsymbol{S})^{n/2}} etr(-\boldsymbol{S}^{-1}\boldsymbol{A}/2)(det\boldsymbol{A})^{(n-m-1)/2} \tag{5.8}$$

where $\dot{\Gamma_m}()$ *is the multivariate gamma function.*

When integer $n \leq m-1$, and $\mathbf{S}$ is positive definite, $W_m(n, \mathbf{S})$ is sometimes called pseudo Wishart; when $\mathbf{S}$ is singular, and n is no less than the rank of $\mathbf{S}$, $W_m(n, \mathbf{S})$ is called singular Wishart; when $\mathbf{S}$ is singular, and n is less than the rank of $\mathbf{S}$, $W_m(n, \mathbf{S})$

is called pseudo-singular Wishart; see [6] for further discussion of pseudo and singular Wisharts.

The evolution for multivariate variance discounting in (5.7) uses pseudo Wishart, for which [86] deduced the following density:

Theorem 5.1 *If* $\boldsymbol{A}$ *is* $W_m(n, \boldsymbol{S})$ *with* $0 < n \leq m-1$, n *is integer,* $\boldsymbol{S} > 0$, *then the density function of* $\boldsymbol{A}$ *is*

$$\frac{\pi^{(-mn+n^2)/2}}{2^{mn}\Gamma_n(n/2)(det\boldsymbol{S})^{n/2}} etr(-\boldsymbol{S}^{-1}\boldsymbol{A}/2)(det\boldsymbol{L})^{(n-m-1)/2} \tag{5.9}$$

where $\boldsymbol{A}$ *represented as* $\boldsymbol{A} = \boldsymbol{H}_1\boldsymbol{L}\boldsymbol{H}_1'$, $\boldsymbol{H}_1 \in V_{n,m}$ *and* $\boldsymbol{L} = diag(l_1, ..., l_n), l_1 > l_2 > \ldots > 0$

This only provided the density for pseudo Wishart when the d.f. is integer; some researchers have explored the possibility to generalize it to real-valued d.f. [66] proved that the characteristic function for pseudo Wishart with non-integer d.f. does not exist, which concluded that pseudo Wishart does not exist unless the d.f. is integer.

The evolution for variance discounting in (5.7) is a direct application of the following Wishart-Beta convolution from [86]:

Theorem 5.2 *Let* $m > 1$ *be an integer and let integer* $p > m-1$. *Let* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be independent, where* $\boldsymbol{A} \sim W_m(1, \boldsymbol{S})$ *and* $\boldsymbol{B} \sim W_m(p, \boldsymbol{S})$. *Put* $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{T}'\boldsymbol{T}$ *where* T *is an upper-triangular* $m \times m$ *matrix with positive diagonal elements. Let* $\boldsymbol{U}$ *be the* $m \times m$ *symmetric matrix defined by* $\boldsymbol{A} = \boldsymbol{T}'\boldsymbol{U}\boldsymbol{T}$. *Then* $\boldsymbol{A}+\boldsymbol{B}$ *and* $\boldsymbol{U}$ *are independent;* $\boldsymbol{A} + \boldsymbol{B}$ *is* $W_m(p+1, \boldsymbol{\Sigma})$ *and the density function of* $\boldsymbol{U}$ *on the space* $S_{m,1}^+$ *is*

$$\pi^{(-m+1)/2}\frac{\Gamma_m((p+1)/2)}{\Gamma(1/2)\Gamma_m(p/2)} L^{-m/2} det(\mathbf{I}_m - \mathbf{U})^{(p-m-1)/2} \tag{5.10}$$

where $\mathbf{U} = \mathbf{H}_1 L \mathbf{H}_1', \mathbf{H}_1 \in V_{1,m}, L \in R$. A matrix $\mathbf{U}$ with the above density is said to have matrix Beta distribution $B_m(1/2, p/2)$, and the matrix $\mathbf{V} = \mathbf{I}_m - \mathbf{U}$ is said to have matrix Beta distribution $B_m(p/2, 1/2)$. This relationship also holds when $\mathbf{A} \sim W_m(q, \mathbf{S})$ where $q > m-1$. Since it is not of direct use in the MC schemes we are proposing here, we omit it in our presentation and refer interested readers to [62].

Theorem 5.2 can be used to generalize (5.2) - (5.4) from univariate to multivariate. Notice that (5.2) - (5.4) go through for any positive d.f., thus discounting is unconstrained in univariate models; Theorem 5.2 requires the d.f. to be integer, hence restricting the way discounting is conducted.

To generalize the above results for positive definite shape matrix $\mathbf{S}$ to singular $\mathbf{S}$, we need to define the volume element (d$\mathbf{Z}$) for an $m \times n$ matrix $\mathbf{Z}$ of rank $r, (m > n > r)$, which is its differential form of the maximum degree, as in the following theorem:

Theorem 5.3 *Let $\boldsymbol{Z}$ be $m \times n$ matrix of rank $r(m > n > r)$ and $\boldsymbol{Z} = \boldsymbol{H}_r \boldsymbol{D}_r \boldsymbol{P}_r'$ where $\boldsymbol{H}_r \in V_{r,m}$, $\boldsymbol{D}_r$ is $r \times r$ diagonal matrix with elements $D_{11} > D_{22} ... > D_{rr} > 0$ and $\boldsymbol{P}_r \in V_{r,n}$, then*

$$(d\boldsymbol{Z}) = 2^{-r} det \boldsymbol{D}_r^{m-n+2} \prod_{i<j}^{r} (D_{ii}^2 - D_{jj}^2)(\boldsymbol{H}_r' d\boldsymbol{H}_r) \Lambda(d\boldsymbol{D}_r) \Lambda(\boldsymbol{P}_r' d\boldsymbol{P}_r).$$

Proof: Decompose $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{H}_1 \mathbf{D} \mathbf{P}'$ where $\mathbf{H}_1 \in V_{n,m}$, $\mathbf{D}$ is $n \times n$ diagonal matrix with elements $D_{11} > D_{22} ... > D_{rr} > 0$, $D_{r+1,r+1} = ... = D_{n,n} = 0$ and $\mathbf{P} \in \bigcirc(n)$. Then $\mathbf{H}_2 \in V_{n-m,m}$ can be found such that $\mathbf{H} = (\mathbf{H}_1 \mathbf{H}_2) \in \bigcirc(m)$. Since

$$d\mathbf{Z} = d\mathbf{H}_1 \mathbf{D} \mathbf{P}' + \mathbf{H}_1 d\mathbf{D} \mathbf{P}' + \mathbf{H}_1 \mathbf{D} d\mathbf{P}' \tag{5.11}$$

it follows that

$$\mathbf{H}' d\mathbf{Z} \mathbf{P} = \begin{pmatrix} \mathbf{H}_1' \\ \mathbf{H}_2' \end{pmatrix} d\mathbf{H}_1 \mathbf{D} + \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} d\mathbf{D} + \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} \mathbf{D} d\mathbf{P}' \mathbf{P}$$

$$= \begin{pmatrix} \mathbf{H}_1' d\mathbf{H}_1 \mathbf{D} + d\mathbf{D} + \mathbf{D} d\mathbf{P}' P \\ \mathbf{H}_2' d\mathbf{H}_1 \mathbf{D} \end{pmatrix}. \tag{5.12}$$

Since $(d\mathbf{Z}) = (\mathbf{H}' d\mathbf{Z}\mathbf{P})$, calculating the exterior product of the right hand side of (5.12) gives the solution. The exterior product of the elements of $\mathbf{H}_2' d\mathbf{H}_1 \mathbf{D}$ is

$$\mathbf{H}_2' d\mathbf{H}_1 \mathbf{D} = (\prod_{i=1}^{r} D_{ii})^{m-n} \Lambda_{i=1}^{r} \Lambda_{j=n+1}^{m} \mathbf{h}_j' d\mathbf{h}_i. \tag{5.13}$$

Let $\mathbf{T} = \mathbf{H}_1' d\mathbf{H}_1 \mathbf{D} + d\mathbf{D} + \mathbf{D} d\mathbf{P}' \mathbf{P}$. Due to the skew symmetry of $\mathbf{H}_1' d\mathbf{H}_1$ and $\mathbf{P}' d\mathbf{P}$,

$$(\mathbf{T}) = \Lambda_{i=1}^{r} \mathbf{T}_{ii} \Lambda_{i<j\leq r} \mathbf{T}_{ij} \Lambda_{i<j\leq r} \mathbf{T}_{ji} \Lambda_{j=r+1}^{n} \Lambda_{i=1}^{r} \mathbf{T}_{ij} \Lambda_{i=r+1}^{n} \Lambda_{j=1}^{r} \mathbf{T}_{ij} \tag{5.14}$$

where

$$\mathbf{T}_{ij} = \begin{Bmatrix} \mathbf{h}_i' d\mathbf{h}_j \mathbf{D}_{jj} & i > r, j \leq r \\ \mathbf{D}_{ii} d\mathbf{P}_i' \mathbf{P}_j & i \leq r, j > r \end{Bmatrix}. \tag{5.15}$$

By following the same treatment of the first 3 factors as that in the proof of Theorem 5 in [86], we can get

$$\begin{aligned} & (\mathbf{H}_2' d\mathbf{H}_1 \mathbf{D}) \Lambda(\mathbf{T}) \qquad (5.16) \\ = \; & (\textstyle\prod_{i=1}^{r} D_{ii})^{m+n-2r} \Lambda_{i=1}^{r} \Lambda_{j=n+1}^{m} \mathbf{h}_j' d\mathbf{h}_i \prod_{i<j}^{r} (D_{ii}^2 - D_{jj}^2) \Lambda_{i=1}^{r} \mathbf{D}_{ii} \\ & \Lambda_{i=1}^{r} \Lambda_{j=i+1}^{r} \mathbf{h}_j' d\mathbf{h}_i \Lambda_{j=i+1}^{r} \Lambda_{i=1}^{r} \mathbf{P}_j' d\mathbf{P}_i \Lambda_{i=r+1}^{n} \Lambda_{j=1}^{r} \mathbf{h}_i' d\mathbf{h}_j \Lambda_{j=r+1}^{n} \Lambda_{i=1}^{r} \mathbf{P}_j' d\mathbf{P}_i \\ = \; & det \mathbf{D}_r^{m+n-2r} \prod_{i<j}^{r} (D_{ii}^2 - D_{jj}^2) (\mathbf{H}_r' d\mathbf{H}_r) \Lambda(d\mathbf{D}) \Lambda(\mathbf{P}_r' d\mathbf{P}_r). \end{aligned}$$

Allowing for arbitrary assignment of signs to the columns of $\mathbf{P}$ makes $\mathbf{Z}$ the image of 2^r decompositions $\mathbf{Z} = \mathbf{H}_r \mathbf{D}_r \mathbf{P}_r'$ such that the density has to be divided by 2^r.

Based on the above volume element, now we give the density function of singular Wishart:

Theorem 5.4 *Let $m > n > 0$ be integers. The density for $\boldsymbol{S} \sim W_m(n, \boldsymbol{\Sigma}), (rank(\boldsymbol{\Sigma}) = r < n)$ distribution on the space $S^+_{m,r}$ of rank $-r$ $m \times m$ matrix with distinct positive eigenvalues with respect to $d\boldsymbol{S}$ is given by*

$$\frac{2^{-rn}}{\Gamma_r(n/2)}(det\mathbf{D}_\Sigma)^{-\frac{n}{2}}etr(-\boldsymbol{\Sigma}^-\mathbf{S}/2)(det\mathbf{L})^{\frac{n-m-1}{2}}, \tag{5.17}$$

where $\mathbf{S}$ represented as $\mathbf{S} = \mathbf{H}_1\mathbf{L}\mathbf{H}_1'$, $\mathbf{H}_1 \in V_{r,m}$ and $\mathbf{L} = diag(l_1, ..., l_r), l_1 > l_2 > \ldots > 0$, $\boldsymbol{\Sigma} = \mathbf{H}_\Sigma\mathbf{D}_\Sigma\mathbf{H}_\Sigma'$ where $\mathbf{H}_\Sigma \in V_{r,m}$ and $\mathbf{D}_\Sigma$ is $r \times r$ diagonal matrix with $d_1 > d_2 \ldots$, if $(\mathbf{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^-)\mathbf{S} = 0$, 0 otherwise.

Proof: Let $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$ where $\boldsymbol{y}_i \sim N(0, \boldsymbol{\Sigma})$, and let $\mathbf{S} = \boldsymbol{y}\boldsymbol{y}'$. Then $\mathbf{S}$ has r distinct eigenvalues almost surely. Let $\boldsymbol{\Sigma} = \mathbf{H}_\Sigma\mathbf{D}_\Sigma\mathbf{H}_\Sigma'$ where $\mathbf{H}_\Sigma \in V_{r,m}$ and $\mathbf{D}_\Sigma$ is $r \times r$ diagonal matrix with $d_1 > d_2 \ldots$. Then the density function of $\boldsymbol{y}$, analogous to Theorem 3.1.1 in [62] , is given by

$$\begin{aligned} & (2\pi)^{-rn/2}det\mathbf{D}_\Sigma^{-n/2}etr(-\boldsymbol{y}'\mathbf{H}_\Sigma\mathbf{D}_\Sigma^{-1}\mathbf{H}_\Sigma'\boldsymbol{y}/2)d\boldsymbol{y} \qquad (5.18) \\ = \quad & (2\pi)^{-rn/2}det\mathbf{D}_\Sigma^{-n/2}etr(-\frac{1}{2}\boldsymbol{\Sigma}^-\mathbf{S})d\boldsymbol{y}, \end{aligned}$$

if $(\mathbf{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^-)\boldsymbol{y} = 0$, 0 elsewhere.

Decompose $\boldsymbol{y}$ as $\boldsymbol{y} = \mathbf{H}_1\mathbf{D}\mathbf{P}'$ as in the above theorem. Then $\mathbf{S} = \mathbf{H}_1\mathbf{L}\mathbf{H}_1'$ where $\mathbf{L} = \mathbf{D}^2$. Noticing that $(d\mathbf{L}) = 2^r|\mathbf{D}|(d\mathbf{D})$, plug (d$\boldsymbol{y}$) into (5.18) to get

$$(2\pi)^{-rn/2}det\mathbf{D}_\Sigma^{-n/2}etr(-\frac{1}{2}\boldsymbol{\Sigma}^-\mathbf{S})2^{-2r}det\mathbf{L}^{m+n-2r-1}\prod_{i<j<r}(l_{ii}-l_{jj})(\mathbf{H}_1'd\mathbf{H}_1)\Lambda(d\mathbf{L})\Lambda(\mathbf{P}'d\mathbf{P}). \tag{5.19}$$

By Theorem 2.1.15 in [62], integrate out $(\mathbf{P}'d\mathbf{P})$ and compare with (d$\mathbf{S}$) defined by Theorem 2 in [86] to get the density function for $\mathbf{S}$. Since $(\mathbf{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^-)\boldsymbol{y}_i = 0$, the Wishart density is restricted to the hyper surface $(\mathbf{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^-)\mathbf{S} = 0$.

[6] derived this density function in another way, which failed to give the constant part; the part they gave that involved $\mathbf{S}$ is in agreement with ours. [6]'s approach is not able to give the density of pseudo-singular Wishart, which can be calculated very similarly to Theorem 5.4 with our approach. Since it does not appear in any literature we are aware of, for the sake of completeness, we also include it here:

Theorem 5.5 *Let $m > n > 0$ be integers, the density for $\boldsymbol{S} \sim W_m(n, \boldsymbol{\Sigma}), (m > rank(\boldsymbol{\Sigma}) = r > n)$ distribution on the space $S^+_{m,n}$ of rank $-$ n $m \times m$ matrix with distinct positive eigenvalues with respect to $d\boldsymbol{S}$ is given by*

$$\frac{2^{-rn}\pi^{(n-r)n/2}}{\Gamma_n(n/2)}(det\mathbf{D}_\Sigma)^{-\frac{n}{2}}etr(-\boldsymbol{\Sigma}^-\mathbf{S}/2)(det\mathbf{L})^{\frac{n-m-1}{2}} \tag{5.20}$$

where $\mathbf{S}$ represented as $\mathbf{S} = \mathbf{H}_1\mathbf{L}\mathbf{H}_1'$, $\mathbf{H}_1 \in V_{n,m}$ and $\mathbf{L} = diag(l_1, ..., l_n), l_1 > l_2 > \ldots > 0$, if $(\mathbf{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^-)\mathbf{S} = 0$, 0 otherwise.

The above densities set the basis for closed form updating equation for multivariate DLMs in which $\boldsymbol{\theta}_t = 0$. This equation, which is given in the next section, is used in our Gibbs sampler for general multivariate DLMs.

The following theorem generalizes Theorem 5.2 to singular Wishart, providing the random walk system equation for the singular conditional variance matrix. The properties of a generalized inverse of a singular matrix used here are quoted in the appendix.

Theorem 5.6 *If $\boldsymbol{H} \sim W_m(n+1, \boldsymbol{E})$ where $\boldsymbol{E}$ is of rank r, $0 < r < m$, $n > m-1$, let $\boldsymbol{H} = \boldsymbol{S}'\boldsymbol{S}$, where $\boldsymbol{S}$ is $r \times m$, $\boldsymbol{Q} \sim B_r(\frac{n}{2}, \frac{1}{2})$; then $\boldsymbol{S}'\boldsymbol{Q}\boldsymbol{S} \sim W_m(n, \boldsymbol{E})$.*

Proof: Write $\mathbf{E} = \mathbf{R}'\mathbf{R}_{r\times m}$ and $\mathbf{H} = \mathbf{S}'\mathbf{S}_{r\times m}$. Since $\mathbf{H} \sim W_m(n+1, \mathbf{E})$, we know that $(\mathbf{I} - \mathbf{E}\mathbf{E}^-)\mathbf{H} = 0$, which is $(\mathbf{I} - \mathbf{R}'\mathbf{R}\mathbf{R}^-\mathbf{R}'^-)\mathbf{H} = 0$, yielding $(\mathbf{I} - \mathbf{R}'\mathbf{R}'^-)\mathbf{H} = 0$.

Plug in $\mathbf{H} = \mathbf{S}'\mathbf{S}$, we get $(\mathbf{I} - \mathbf{R}'\mathbf{R}'^{-})\mathbf{S}'\mathbf{S} = 0$, which gives $(\mathbf{I} - \mathbf{R}'\mathbf{R}'^{-})\mathbf{S}' = 0$; transposing both sides and moving the second item to the right, we get $\mathbf{S} = \mathbf{S}\mathbf{R}^{-}\mathbf{R}$.

On the other hand, from $\mathbf{H} \sim W_m(n+1, \mathbf{E})$ we have

$$\mathbf{R}'^{-}\mathbf{H}\mathbf{R}^{-} = \mathbf{R}'^{-}\mathbf{S}'\mathbf{S}\mathbf{R}^{-} \sim W_r(n+1, \mathbf{I}).$$

For $\mathbf{Q} \sim B_r(\frac{n}{2}, \frac{1}{2})$,

$$\mathbf{R}'^{-}\mathbf{S}'\mathbf{Q}\mathbf{S}\mathbf{R}^{-} \sim W_r(n, \mathbf{I}).$$

Left multiply both sides by $\mathbf{R}'$ and right multiply both side by $\mathbf{R}$, using the fact that $\mathbf{S} = \mathbf{S}\mathbf{R}^{-}\mathbf{R}$, we get

$$\mathbf{S}'\mathbf{Q}\mathbf{S} \sim W_m(n, \mathbf{E}).$$

5.5 Discounting and Updating through Sequential Simulation

Now with the theorems to deal with singular Wishart, we generalize variance discounting and updating equations to a simplified multivariate DLM in which the observation covariance matrix is singular. For the general form of multivariate DLM with unknown variances, closed-form updating for the states is not available. Our discussion will be on multivariate DLMs, in which the regression coefficients $\boldsymbol{\theta}_t$ are $\mathbf{0}$:

$$\boldsymbol{y}_t = \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Phi}_t^{-}) \tag{5.21}$$

where $\boldsymbol{y}_t$ and $\boldsymbol{\epsilon}_t$ are m-vectors, $\boldsymbol{\Phi}_t$ is a singular precision matrix. This model interests us for two reasons: first, closed-form updating for variances alone is required by the Gibbs sampler for general DLMs that we will propose in the next section; second, in sequential simulation, the main difficulty is the treatment of the variances. This model focuses on the variance and provides a case to preliminarily test the feasibility

of MC methods for more complicated models. We apply auxiliary particle filters to the general form of multivariate DLM, implement it with the simplified model (5.21) and compare our MC result with the true posterior distribution.

5.5.1 Sequential Discounting and Updating Equations

For model (5.21), assume conjugate prior $\mathbf{\Phi}_0 \sim W(d_0, \mathbf{S}_0)$, where d_0 is normally chosen to be a little bit greater than m, $\mathbf{S}_0$ is of rank $r < m$. Assume at any time $t-1$, the posterior has the form

$$\mathbf{\Phi}_{t-1} \sim W(d_{t-1}, \mathbf{S}_{t-1}). \tag{5.22}$$

Then the prior distribution of $\mathbf{\Phi}_t$ can be obtained through random walk system equation as in Theorem (5.6). Let

$$\mathbf{\Phi}_{t-1} = \mathbf{T}'\mathbf{T} \tag{5.23}$$

where $\mathbf{T}$ is $r \times m$,

$$\mathbf{Q} \sim B_r(\frac{d_{t-1}-1}{2}, \frac{1}{2}) \tag{5.24}$$

independently of $\mathbf{\Phi}_{t-1}$, and

$$\mathbf{\Phi}_t = \frac{d_{t-1}}{d_{t-1}-1}\mathbf{T}'\mathbf{Q}\mathbf{T}. \tag{5.25}$$

Then

$$\mathbf{\Phi}_t | D_{t-1} \sim W_m(d_{t-1}-1, \frac{d_{t-1}}{d_{t-1}-1}\mathbf{S}_{t-1}) \tag{5.26}$$

is the prior for $\mathbf{\Phi}_t$, which has the same mean as $W(d_{t-1}, \mathbf{S}_{t-1})$, but is more dispersed.

Given the Wishart prior for $\mathbf{\Phi}_t$ and a discount factor $0 < \delta \leq 1$, the updating using the data point $\mathbf{y}_t$ could proceed analytically as

$$
\begin{aligned}
& f(\mathbf{\Phi}_t|\boldsymbol{y}_t) \qquad (5.27)\\
\propto\ & p(\mathbf{\Phi}_t|D_{t-1})f(\boldsymbol{y}_t|\mathbf{\Phi}_t)\\
\propto\ & etr(-\tfrac{1}{2}(\delta^{-1}\mathbf{S}_{t-1})^{-}\mathbf{\Phi}_t)det\mathbf{L}_t^{\frac{\delta d_{t-1}-m-1}{2}}det\mathbf{L}_t^{\frac{1}{2}}exp(-\boldsymbol{y}_t'\mathbf{\Phi}_t\boldsymbol{y}_t)/2\\
\propto\ & det\mathbf{L}^{\frac{\delta d_{t-1}+1-m-1}{2}}etr(-\tfrac{1}{2}((\delta^{-1}\mathbf{S}_{t-1})^{-}+\boldsymbol{y}_t\boldsymbol{y}_t')\mathbf{\Phi}_t)\\
\sim\ & W(\delta d_{t-1}+1,[(\delta^{-1}\mathbf{S}_{t-1})^{-}+\boldsymbol{y}_t\boldsymbol{y}_t']^{-})
\end{aligned}
$$

if $(\mathbf{I}-\mathbf{\Phi}_t^{-}\mathbf{\Phi}_t)\boldsymbol{y}_t=0$, 0 otherwise.

For this model, sequential variance discounting and parameter updating analytically and through computation should yield the same result. Despite this, differences exist between the two methods. (5.27) goes through for any choice of discount factor $0<\delta<1$; in simulation, however, δ is restricted to be $\frac{d_{t-1}-1}{d_{t-1}}$. This restriction is due to the fact that the multivariate Beta distribution is defined through the Wishart distribution, and $W_m(n,\mathbf{\Sigma})$, when $n<m-1$ is defined only for integer n. For more complicated models of interest, (5.6) for example, there is no closed form updating equation and MC methods have to be employed.

5.5.2 On the Discount Rule in Simulation

When variance discounting is applied in MC simulation, the fashion of discounting is restricted by the parameter domain in which the Wishart distribution is defined. Unlike discounting and updating analytically, where any discount factor between 0 and 1 can be used, in simulation, the discount rule has to guarantee that the d.f. in the posterior distribution is integer. Fixing the d.f. of the posterior of $\mathbf{\Phi}_t$ at an integer n and the discount factor δ at $\frac{n-1}{n}$ used in [86] is subject to dispute. To illustrate this, under the fixed-d.f. discount rule, consider a series containing only 10 data points, the posterior of the observation variance from the small data set would have d.f. n,

whereas when it comes to the point that 10,000 data points have accumulated, the posterior from this relative large data set would also have d.f. n. The information in the 10,000 data points would be discounted to the same degree as the information in the 10 data points is discounted. If a shock happens at time 10001, the model is supposed to "balance" it with the long history of data and react "calmly", but under the fixed-degree-of-freedom discount rule, the model would respond as if it has only seen 10 data points; or it may happen in the opposite way: the model would respond to a shock at time 11 as confidently as if it has information from 10000 data. It is obvious that the d.f. of the posterior of $\mathbf{\Phi}_t$ should increase over time to reflect the accumulation of information, especially at the beginning of the series.

A compromise idea is to specify an ideal discount factor δ, and calculate the d.f. of the posterior of $\mathbf{\Phi}_t$ at each t, use the integer part of them as the real d.f. of the posteriors. The d.f. thus obtained often remains unchanged across time; on these occasions, discounting needs to be done between updating steps. Occasionally, the d.f. increases by 1; the discount step should be skipped when this happens. The idea is to keep the posterior d.f. increasing reasonably in the integer domain. In practice, it may be favorable to combine this guideline of choosing the d.f. with external information to decide when to discount.

5.5.3 Auxiliary Particle Filtering for Multivariate DLMs

We apply auxiliary particle filtering to the general multivariate DLM (5.6) with singular variance. For ease of presentation, the variance in the system evolution $\mathbf{W}_t$ is assumed to be known; for unknown $\mathbf{W}_t$, the treatment for $\mathbf{\Phi}_t$ here can be used. Suppose at time t, we have available the Monte Carlo samples $(\mathbf{\Phi}_t^{(j)}, \boldsymbol{\theta}_t^{(j)})$ and weights $\omega_t^{(j)}$ $(j = 1, \ldots, N)$, representing the joint posterior $p(\mathbf{\Phi}_t, \boldsymbol{\theta}_t | \boldsymbol{D}_t)$.

1. For each $j = 1, ..., N$, identify the prior point estimates of $(\mathbf{\Phi}_t, \boldsymbol{\theta}_t)$

$$\begin{aligned}(\widehat{\mathbf{\Phi}}_{t+1}, \widehat{\boldsymbol{\theta}}_{t+1}) &= E(\mathbf{\Phi}_{t+1}, \boldsymbol{\theta}_{t+1} | \mathbf{\Phi}_t^{(j)}, \boldsymbol{\theta}_t^{(j)}) \\ &= (\mathbf{\Phi}_t^{(j)}, \mathbf{G}_t \boldsymbol{\theta}_t^{(j)}).\end{aligned} \tag{5.28}$$

2. Sample an auxiliary integer variable from the set $\{1, ..., N\}$ with probabilities proportional to

$$g_{t+1}^{(j)} \propto \omega_t^{(j)} p(\mathbf{y}_{t+1} | \widehat{\mathbf{\Phi}}_{t+1}, \widehat{\boldsymbol{\theta}}_{t+1});$$

call the sampled index k.

3. Sample a value of the current state $\mathbf{\Phi}_{t+1}^{(k)}$ from equation (5.25) and $\boldsymbol{\theta}_{t+1}^{(k)}$ from the second equation in (5.6) based on samples $\mathbf{\Phi}_t^{(k)}$ and $\boldsymbol{\theta}_t^{(k)}$.

4. Evaluate the corresponding weight

$$\omega_{t+1}^{(k)} \propto \frac{N(\mathbf{y}_{t+1} | \mathbf{F} \boldsymbol{\theta}_{t+1}^{(k)}, \mathbf{\Phi}_{t+1}^{(k)\ -})}{N(\mathbf{y}_{t+1} | \mathbf{F} \widehat{\boldsymbol{\theta}_{t+1}}, \widehat{\mathbf{\Phi}_{t+1}}^{\ -})}.$$

5. Repeat step (2)-(5) a large number of times to produce a final posterior approximation $(\mathbf{\Phi}_{t+1}^{(k)}, \boldsymbol{\theta}_{t+1}^{(k)})$ with weights $\omega_{t+1}^{(k)}$, as required.

Step 3 requires sampling from $B_r(\frac{d_{t-1}-1}{2}, \frac{1}{2})$, which can be done using Theorem 5.2. Step 4 involves the calculation of the singular normal density in the likelihood; this density is given in the appendix. Notice that this algorithm applies to multivariate DLMs with non-singular $\mathbf{\Phi}_t$ as well, in which case (5.7) should be used in place of (5.25) in step 3.

As an artificial example, we choose a rank-3, 4-dimensional matrix as the shape of the prior for the covariance matrix:

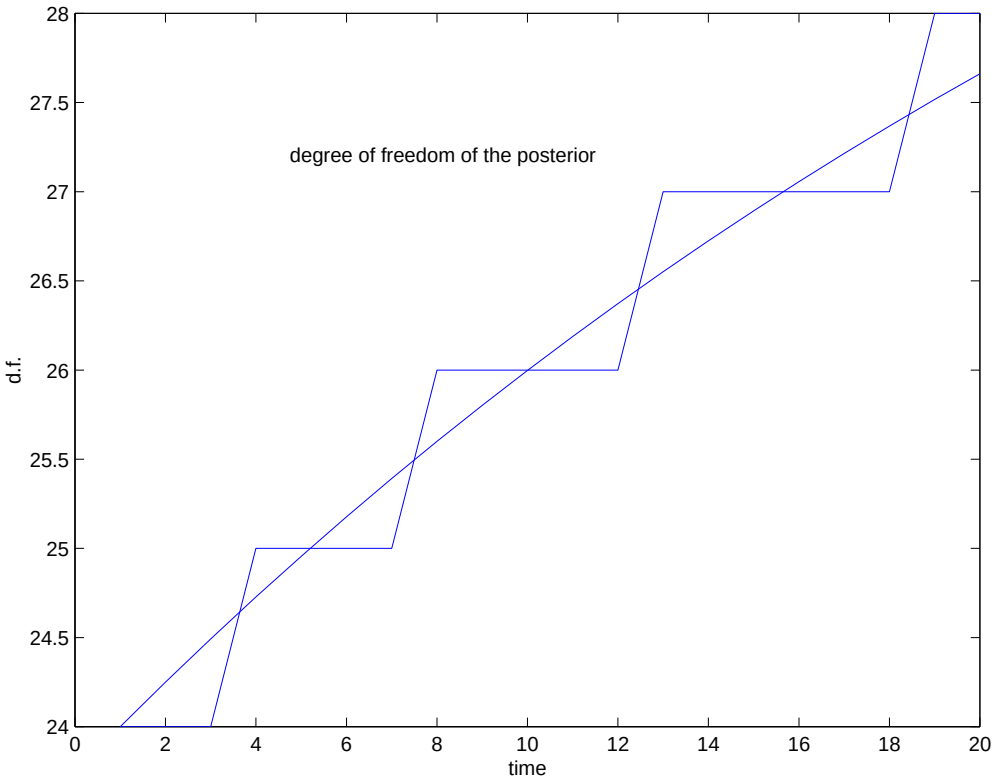

Figure 5.1: d.f. in the posterior distribution

$$\mathbf{S}_0 = \begin{pmatrix} 1.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 \end{pmatrix},$$

whose eigenvalues are $\{3.7763, 1, 0.4237, 0\}$. As noticed from our experience, particle filters can be inefficient when very dispersed priors are used, thus we use a moderate $n_0 = 24$. Let $\delta = 0.97$. According to the discount rule discussed in the last section, we calculate the posterior d.f. for step 1 to $T = 20$, and plot them in Figure 5.1. The smooth curve (which is actually an exponential curve although it looks like a straight line) is the d.f. obtained with normal discounting, and the step line shows the d.f. actually used in simulation.

We generate a series of $\{\Phi_t, \boldsymbol{y}_t, t = 1, \ldots, T\}$ according to the system and observation equations and the discount rule that yields the posterior d.f. as the step line in Figure 5.1. We draw 5000 samples from $W_m(n_0, \mathbf{S}_0)$ as the initial sample, and perform particle filtering for 20 steps, keeping the sample size 5000 at each step. To verify that our samples are actually from the posterior $W_4(n_T, \mathbf{S}_T)$, we use the Bartlett decomposition of Wishart variable:

Theorem 5.7 *Let* $\boldsymbol{A}$ *be* $W_m(n, \boldsymbol{I}_m)$*, where* $n \geq m$ *is an integer, and put* $\boldsymbol{A} = \boldsymbol{U}'\boldsymbol{U}$*, where* $\boldsymbol{U}$ *is an upper-triangular* $m \times m$ *matrix with positive diagonal elements. Then the elements* $u_{ij}(1 \leq i \leq j \leq m)$ *of* $\boldsymbol{U}$ *are independent, and each* u_{ii}^2 *is* $\chi^2_{n-i+1}(i = 1, ..., m)$ *while each* u_{ij} *is* $N(0,1)(1 \leq i \leq j \leq m)$.

The derivation of this theorem can go backward stepwise, so the decomposition is necessary and sufficient condition for a matrix to be Wishart distributed. To verify that our posterior samples are from the real posterior distribution $W(d_T, \mathbf{S}_T)$, decompose $\mathbf{S}_T$ into $\mathbf{R}'\mathbf{R}$, where $\mathbf{R}$ is $r \times m$. If a sample $\mathbf{\Phi}_T^{(j)}$ is from $W(d_T, \mathbf{S}_T)$, then $\mathbf{P}_T = \mathbf{R}'^{-}\mathbf{\Phi}_T^{(j)}\mathbf{R}^{-} \sim W(d_T, \mathbf{I})$. Suppose $\mathbf{P}_T = \mathbf{T}_T'\mathbf{T}_T$, if $\mathbf{\Phi}_t^{(j)} \sim W(d_T, \mathbf{S}_T)$, then $t_{11}^2 \sim \chi^2_{28}, t_{22}^2 \sim \chi^2_{27}, t_{33}^2 \sim \chi^2_{26}$ and $t_{ij} \sim N(0,1), 1 \leq i < j \leq 3$. We generate 5000 random samples from these distributions and in Figure 5.2 plot 0.01, 0.05, 0.25, 0.50, 0.75, 0.95 and 0.99 quantiles of the 5000 χ^2 samples vs. the same quantiles of t_{ii}^2; Figure 5.3 compares the quantiles of the 5000 $N(0,1)$ samples with the off-diagonal elements of $\mathbf{T}_T$. From these pictures, it is evident that the posterior samples are in agreement with their theoretical distribution.

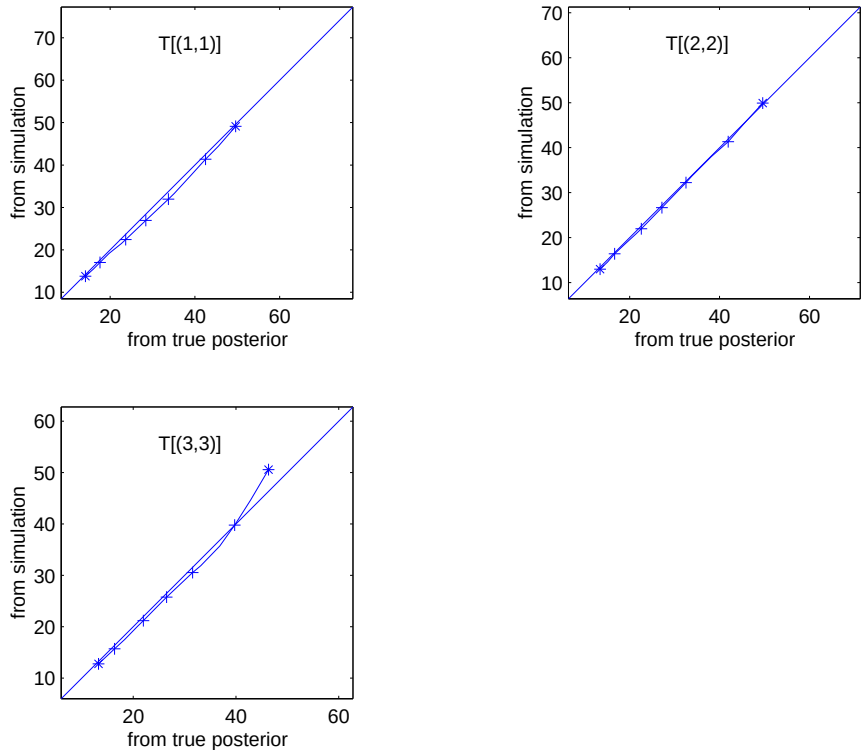

Figure 5.2: Samples from χ^2 vs. samples from t_{ii}^2

5.6 An Evolution Equation for the Precision Matrices Using the Bartlett Decomposition

Having introduced the Bartlett decomposition in the last section, we would like to point out that discounting of a Wishart variable can be performed via manipulating the diagonal elements of its Bartlett decomposition. This can be used to provide an evolution equation for the precision matrices in variance discounting. The following theorem parallels Theorem 5.6.

Theorem 5.8 *Suppose* $0 < \delta \leq 1$ *and* $m \times m$ *matrix* $\boldsymbol{B} \sim W(d, \boldsymbol{S}), d > 0$. *Let* $\boldsymbol{U}$ *be the Cholesky decomposition of* $\boldsymbol{S}$, *and* $\boldsymbol{A} = \boldsymbol{U}'^{-1}\boldsymbol{B}\boldsymbol{U}^{-1}$. *Let* $\boldsymbol{P}$ *be the Cholesky decomposition of* $\boldsymbol{A}$ *with elements denoted by* $p_{ij}, 1 \leq i \leq j \leq m$. *Let* $\boldsymbol{P}^*$ *be an upper triangular matrix with elements* $p_{ii}^* = \sqrt{r}_i p_{ii}, i = 1, 2, \ldots, m$, *and* $p_{ij}^* = p_{ij}, 1 \leq i <$

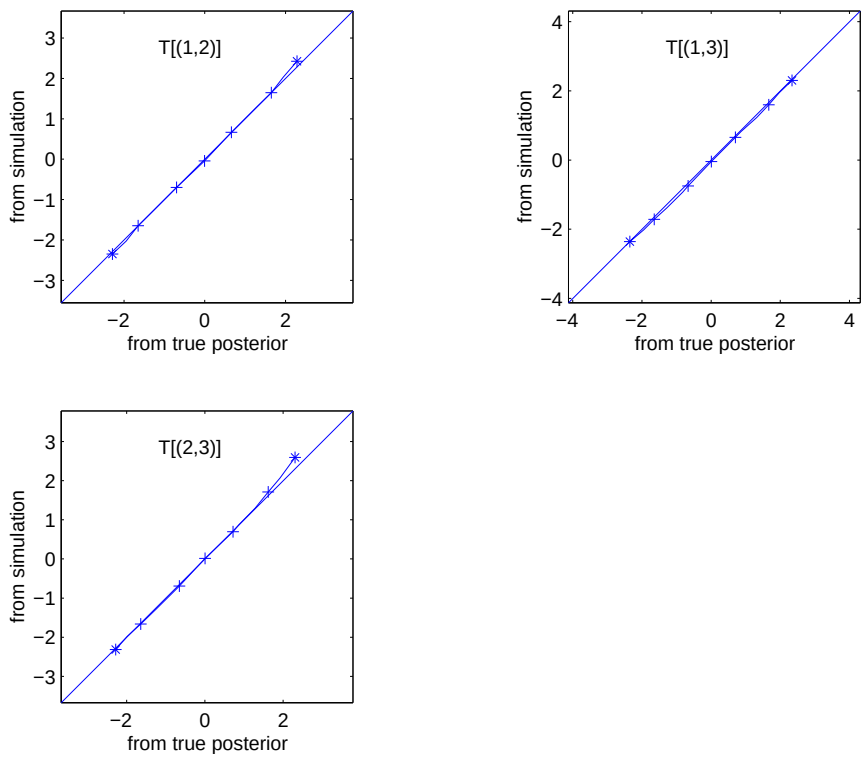

Figure 5.3: Samples from $N(0,1)$ vs. samples from u_{ij}

$j \leq m$, *where* $r_i \sim Beta(\delta_i(d+m-i)/2, (1-\delta_i)(d+m-i)/2)$ *and* $\delta_i = \frac{\delta d+m-i}{d+m-i}$, *then* $\boldsymbol{U}^{*\prime}(\boldsymbol{P}^*)'\boldsymbol{P}^*\boldsymbol{U}^* \sim W(\delta d, \boldsymbol{S})$.

Proof: Since $\mathbf{B} \sim W(d, \mathbf{S})$ and $\mathbf{A} = \mathbf{U}'^{-1}\mathbf{B}\mathbf{U}^{-1}$, $\mathbf{A} \sim W_m(d, \mathbf{I})$. Given that $\mathbf{P}$ is the Cholesky decomposition of $\mathbf{A}$, according to Theorem 5.7, we know $p_{ii}^2 \sim \chi^2(d+m-i), i = 1, ..., m$, and $p_{ij} \sim N(0,1), 1 \leq i < j \leq m$. Because $\chi^2(d+m-i)$ is equivalent to $Gamma((d+m-i)/2, 1/2)$, and $r_i \sim beta(\delta_i(d+m-i)/2, (1-\delta_i)(d+m-i)/2)$, we have $p_{ii}^{*2} = r_i p_{ii}^2 \sim Gamma(\delta_i(d+m-i)/2, 1/2)$, i.e. $p_{ii}^{*2} \sim \chi^2(\delta_i(d+m-i))$ which is $\chi^2(\delta d+m-i)$. So $(\mathbf{P}^*)'\mathbf{P}^* \sim W(\delta d, \mathbf{I})$ and $\mathbf{U}^{*\prime}(\mathbf{P}^*)'\mathbf{P}^*\mathbf{U}^* \sim W(\delta d, \mathbf{S})$.

Generalization of Bartlett decomposition can be made to singular Wishart matrices, whose Bartlett decomposition have 0's in the lower part of the upper triangular matrix $\mathbf{U}$ in Theorem 5.7. And the above theorem can correspondingly be generalized

to "evolve" singular Wishart variables.

While Theorem 5.6 uses a matrix-Beta variable to introduce random noise, Theorem 5.8 uses some independent Beta variables, hence avoiding matrix-Beta distribution and the its difficulty afore mentioned. However, Theorem 5.8 involves $\mathbf{S}$, the shape matrix of the Wishart distribution; estimators of $\mathbf{S}$ have to be found in order to use Theorem 5.8 in MC simulation.

5.7 MCMC for Multivariate DLMs

MCMC is often used as a standard to calibrate and monitor sequential MC simulation. As detailed in [94], when the variances in both the observation equation and the system equation are known, closed-form updating is available, but in general, there is no neat, conjugate analysis to enable sequential learning about the unknown variance matrices. In cases when $\mathbf{V}_t = V$, constant for all time t, the unknown parameters in $\mathbf{V}$ introduce complications that can, in principle, be handled using some form of approximate analysis. Analytic approximations are developed in [88], and numerical approximation techniques are given in [94]. In this section we propose a MCMC simulation scheme to learn about the state variable $\mathbf{\Theta}$ and precision matrix $\mathbf{\Phi}_t = \mathbf{V}_t^-$ in the model defined by (5.6) and (5.23) to (5.25), based on Forward Filtering Backward Sampling (FFBS) and Gibbs sampler.

5.7.1 Forward Filtering Backward Sampling

[16] and [33] provided the basic and original development of an efficient method to sample the set of state vectors at all time from the posterior distribution $p(\mathbf{\Theta}|D_t)$ (where $\mathbf{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T\}$) in normal DLMs. The idea goes as follows. Exploiting the Markov structure of the evolution equation of the DLM, we may write

$$p(\boldsymbol{\Theta}|D_T) = p(\boldsymbol{\theta}_T|D_T)p(\boldsymbol{\theta}_{T-1}|\boldsymbol{\theta}_T, D_{T-1}) \ldots p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, D_1). \tag{5.29}$$

We may sample the entire $\boldsymbol{\Theta}$ by sequentially simulating the individual states:

(1) Sample $\boldsymbol{\theta}_T$ from $p(\boldsymbol{\theta}_T|D_T) \sim N(\mathbf{m}_T, \mathbf{C}_T)$

(2) For $t = T-1, \ldots, 1$, sample $\boldsymbol{\theta}_t$ from $(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, D_t)$

The required distribution $(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, D_t)$ is obtained in the filtering recurrences:

$$(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, D_t) \sim N(\mathbf{h}_t, \mathbf{H}_t)$$

where

$$\mathbf{h}_t = \mathbf{m}_t + \mathbf{B}_t(\mathbf{h}_{t+1} - \mathbf{a}_{t+1})$$

and

$$\mathbf{H}_t = \mathbf{C}_t - \mathbf{B}_t\mathbf{R}_{t+1}\mathbf{B}_t'$$

with

$$\mathbf{B}_t = \mathbf{C}_t\mathbf{G}_{t+1}\mathbf{R}_{t+1}'$$

for each t.

The process of sampling $\boldsymbol{\Theta}$ starts from running the standard DLM updating for $t = 1$ to $t = T$ and obtain the quantities $\mathbf{m}_t, \mathbf{C}_t, \mathbf{a}_t, \mathbf{R}_t, \mathbf{B}_t$ at each stage. At $t = T$, sample a vector $\boldsymbol{\theta}_T$ from $p(\boldsymbol{\theta}_T|D_T)$, then for $t = T-1, \ldots, 1$, go backwards through time computing $\mathbf{B}_t, \mathbf{h}_t, \mathbf{H}_t$ and sampling $\boldsymbol{\theta}_t$ at each step.

5.7.2 A Gibbs Sampler

Now consider the model (5.6), (5.23) - (5.25). If the $\mathbf{\Phi} = \{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_T\}$ were known, then the above FFBS algorithm could be applied to get samples of $\mathbf{\Theta}$. Now that $\mathbf{\Phi}$ needs to be sampled at the same time, we may think of a Gibbs sampler, starting from some initial sample, iterating between sampling $\mathbf{\Theta}$ given the current sample of $\mathbf{\Phi}$ and sampling $\mathbf{\Phi}$ given the samples of $\mathbf{\Theta}$.

Assume that we know how to sample from the full conditional of $\mathbf{\Phi} = \{\mathbf{\Phi}_T, \ldots, \mathbf{\Phi}_1\}$, $p(\mathbf{\Phi}|D_T, \mathbf{\Theta})$, the Gibbs sampler runs as follows:

(1) Draw initial samples $\mathbf{\Phi}^{(0)}$; set the index of iteration $j = 1$.

(2) Sample $\mathbf{\Theta}^{(j)}$ from $p(\mathbf{\Theta}|D_T, \mathbf{\Phi}^{(j-1)})$ according to the FFBS algorithm above.

(3) Sample $\mathbf{\Phi}^{(j)}$ from its full conditional distribution $p(\mathbf{\Phi}|D_t, \mathbf{\Theta}^{(j)})$.

(4) Increase j by 1 and repeat (2),(3) until the number of iteration is reached.

The question remains how to sample from $p(\mathbf{\Phi}|D_T, \mathbf{\Theta}^{(j)})$. We would like to have a FFBS type of algorithm for $\mathbf{\Phi}$, but the distribution $p(\mathbf{\Phi}_t|\mathbf{\Phi}_{t+1}, D_T)$ is not tractable. We have to include a Metropolis-Hasting step in FFBS, using the so called "Metropolis-Hastings-within-Gibbs"([61]).

5.7.3 Sampling from the Full Conditional Distribution of $\mathbf{\Phi}$

Similar to (5.29), we may write the full conditional distribution of $\mathbf{\Phi}$ as

$$p(\mathbf{\Phi}|D_T, \mathbf{\Theta}^{(j)}) = p(\mathbf{\Phi}_T|D_T, \mathbf{\Theta}^{(j)})p(\mathbf{\Phi}_{T-1}|\mathbf{\Phi}_T, D_{T-1}, \mathbf{\Theta}^{(j)}) \ldots p(\mathbf{\Phi}_1|\mathbf{\Phi}_2, D_1, \mathbf{\Theta}^{(j)}) \tag{5.30}$$

For each $t = T-1, \ldots, 1$,

$$p(\mathbf{\Phi}_t|\mathbf{\Phi}_{t+1}, D_t, \mathbf{\Theta}^{(j)}) = \frac{p(\mathbf{\Phi}_{t+1}|\mathbf{\Phi}_t)p(\mathbf{\Phi}_t|D_t, \mathbf{\Theta}^{(j)})}{p(\mathbf{\Phi}_{t+1}|D_t, \mathbf{\Theta}^{(j)})}$$

Let us take a closer look at each component of the right hand side:

- $p(\mathbf{\Phi}_t|D_t, \mathbf{\Theta}^{(j)})$, the posterior distribution of $\mathbf{\Phi}_t$ given a known set of state variables $\mathbf{\Theta}$. Given $\mathbf{\Theta}$, the model is equivalent to (5.21), thus the posterior distribution is $W(d_t, \mathbf{S}_t)$; the sequential updating equation is given by (5.27).

- $p(\mathbf{\Phi}_{t+1}|D_t, \mathbf{\Theta}^{(j)})$, the prior distribution of $\mathbf{\Phi}_{t+1}$ given a known set of state variable $\mathbf{\Theta}$, which is given in (5.26)

- $p(\mathbf{\Phi}_{t+1}|\mathbf{\Phi}_t)$, the system evolution density for $\mathbf{\Phi}_t$. [53] (Appendix A to Chapter 6, (k)) attempts to provide such a density, but the restraint that $n_2 > m-1$ makes it not helpful here. It is straight forward by utilizing the singular matrix Beta density in [86] and Theorem 2.1.6 in [62].

Theorem 5.9 *If* $\boldsymbol{Q} \sim B_m(n_1/2, 1/2)$, *and* $\boldsymbol{\Phi} = \boldsymbol{T}'\boldsymbol{Q}\boldsymbol{T}$ *where* $\boldsymbol{T}$ *is non-singular, then* $\boldsymbol{\Phi}$ *has density*

$$p(\boldsymbol{\Phi}) \propto det\boldsymbol{T}^{-n_1} det\boldsymbol{\Phi}^{\frac{n_1-m-1}{2}} \boldsymbol{L}^{-\frac{m}{2}},$$

where $\boldsymbol{I} - \boldsymbol{T}'^{-1}\boldsymbol{\Phi}\boldsymbol{T}^{-1} = \boldsymbol{H}_1\boldsymbol{L}\boldsymbol{H}_1'$, $\boldsymbol{L}$ *is diagonal, and* $\boldsymbol{H}_1 \in O(m)$.

Step (3) to sample $\mathbf{\Phi}^{(j)}$ from its full conditional distribution $p(\mathbf{\Phi}|D_t, \mathbf{\Theta}^{(j)})$ in the afore mentioned Gibbs sampler goes as follows:

(1) For $t = 1, 2, \ldots, T$, sequentially compute the posterior density of $\mathbf{\Phi}_t$ using (5.27).

(2) Sample $\mathbf{\Phi}_T^{(j)}$ from $W(d_T, \mathbf{S}_T)$.

(3) For $t = T-1$, sample $\mathbf{\Phi}_t$ from $p(\mathbf{\Phi}_t|\mathbf{\Phi}_{t+1}, D_t, \mathbf{\Theta}^{(j)})$ by following a metropolis step: sample $\mathbf{\Phi}_t^{(n)}$ from $p(\mathbf{\Phi}_t|D_t, \mathbf{\Theta}^{(j)})$, calculate

$$r = \frac{p(\mathbf{\Phi}_{t+1}^{(j)}|\mathbf{\Phi}_t^{(n)})/p(\mathbf{\Phi}_{t+1}^{(n)}|D_t, \mathbf{\Theta}^{(j)})}{p(\mathbf{\Phi}_{t+1}^{(j)}|\mathbf{\Phi}_t^{(j-1)})/p(\mathbf{\Phi}_{t+1}^{(j-1)}|D_t, \mathbf{\Theta}^{(j)})}$$

and let $\mathbf{\Phi}_t^{(j)} = \mathbf{\Phi}_t^{(n)}$ with probability $max\{1, r\}$. If $\mathbf{\Phi}_t^{(n)}$ is not accepted, set $\mathbf{\Phi}_t^{(j)} = \mathbf{\Phi}_t^{(j-1)}$.

(4)For each $t = T-2, \ldots, 1$, repeat (3) to get samples of $\mathbf{\Phi}_t^{(j)}$.

Chapter 6

Sequential Variance Learning through Factor Stochastic Volatility Models

Variance discounting is one step ahead from the constant-variance models in that it allows the conditional variance to change over time. However, structured patterns of change, often available in real problems, are missing in this type of model. The forecast uncertainty keeps increasing as the forecast horizon gets longer, finally approaching infinity. This conflicts with the mean reverting phenomena observed in many financial time series: the volatility tends to return to some fixed level instead of "randomly walking" all over an infinite domain. SV models explicitly describe mean reversion in the model setup.

Factor representations have long been used as a powerful tool in multivariate analysis. In finance, Arbitrage Pricing Theory ascribes the movements in financial prices to some common unknown underlying factors. In using discount models, principle components of the posterior estimates of the variances are frequently looked at to understand the pattern of change in the variance structure and the mechanism driving the changes. When SV is generalized to multivariate models, factor representations are also preferred; instead of modeling the SV process on each of the series, representing the variance-covariance structure with a small number of unknown un-

derlying factors and studying the factor processes with SV models greatly reduces the parameter dimension.

In studies of dynamic latent factor models with multivariate stochastic volatility components, recent Bayesian work has developed both MCMC methods and aspects of sequential analysis using versions of auxiliary particle filtering for states [1, 81]. In these models, the state variables are latent volatilities of both common factor processes and of residual/idiosyncratic random terms specific to observed time series. In the application in exchange rate modeling, forecasting and portfolio analysis, Aguilar and West ([1]) use MCMC methods to fit these complicated dynamic models to historical data and then perform sequential particle filtering over a long stretch of further data that provides the context for sequential forecasting and portfolio construction. In that example, these authors fix a full set of constant model parameters at estimated values taken as the means of posterior distributions based on the MCMC analysis of the initial (and very long) data stretch. The results are very positive from the financial time series modeling viewpoint. For many practical purposes, an extension of that approach that involves periodic reanalysis of some recent historical data using full MCMC methods, followed by sequential analysis using auxiliary particle filtering on just the time-varying states with model parameters fixed at most recently estimated values, is quite satisfactory. However, from the viewpoint of the use of sequential simulation technology in more interesting and complicated models, this setting provides a very nice and somewhat challenging test-bed, especially when considering multivariate time series of moderate dimensions. Hence our interest in exploring the general sequential algorithm of the previous section in this context.

6.1 A Factor Model

We adopt the context and notation of [1], noting the closely similar developments in Shephard and Pitt ([67]). Begin with a $q-$variate time series of observations, in $\mathbf{y}_t$, $(t = 1, 2, \ldots)$. In our example, this is a vector of observed daily exchange rates of a set of $q = 6$ national currencies relative to the US dollar. The dynamic latent factor model is structured as follows.

At each time t, we have

$$\mathbf{y}_t = \boldsymbol{\alpha}_t + \mathbf{x}\mathbf{f}_t + \boldsymbol{\epsilon}_t \tag{6.1}$$

with the following ingredients.

- $\mathbf{y}_t$ is the q-vector of observation and $\boldsymbol{\alpha}_t$ is a q-vector representing a local level of the series.
- $\mathbf{x}$ is a $q \times k$ matrix called the *factor loadings matrix*.
- $\mathbf{f}_t$ is a $k-$vector which represents the vector of *latent factors* at time t; the $\mathbf{f}_t$ are assumed to be conditionally independent over time and distributed as $N(\mathbf{f}_t|\mathbf{0}, \mathbf{H}_t)$ where $\mathbf{H}_t$ =diag$(h_{t1}, \ldots, h_{tk})$ is the diagonal matrix of instantaneous factor variances.
- $\boldsymbol{\epsilon}_t \sim N(\boldsymbol{\epsilon}_t|\mathbf{0}, \Psi)$ are idiosyncratic noise terms, assumed to be conditionally independent over time and with a diagonal variance matrix Ψ =diag$(\psi_1, \ldots, \psi_k)$. The elements of Ψ are called the *idiosyncratic noise variances* of the series. We note that Aguilar and West ([1]) use an extension of this model that, as in Shephard and Pitt ([67]), has time-varying idiosyncratic noise variances, but we do not consider that here.
- $\boldsymbol{\epsilon}_t$ and $\mathbf{f}_s$ are mutually independent for all t, s.

6.2 Factor Model Constraints

Identification is a common concern in factor models. First is the number of factors allowed; the observation variance is $\mathbf{x}\mathbf{H}_t\mathbf{x}' + \Psi$, by counting the number of free elements in the observation variance matrix and the number of free parameters in $\mathbf{x}\mathbf{H}_t\mathbf{x}' + \Psi$, one can get an upper limit of the number of factors allowed. Since the purpose of using factor representation is to use a small number of factors to account for the variance structure in a large number of series, this is not normally an issue here. Secondly, for any pair of $\mathbf{x}, \mathbf{f}_t$, the pair $\mathbf{x}\mathbf{A}, \mathbf{A}^{-1}\mathbf{f}_t$, where $\mathbf{A}$ is a nonsingular square matrix, would equally suit the model. Like other authors, including Geweke and Zhou ([37]), Aguilar and West ([1]) adopts a factor loading matrix of the form

$$\mathbf{x} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ x_{2,1} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ x_{k,1} & x_{k,2} & x_{k,3} & \cdots & 1 \\ x_{k+1,1} & x_{k+1,2} & x_{k+1,3} & \cdots & x_{k+1,k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{q,1} & x_{q,2} & x_{q,3} & \cdots & x_{q,k} \end{pmatrix}. \tag{6.2}$$

The reduced number of parameters in $\mathbf{x}$ ensures mathematical identification of the model and the lower-triangular form provides a nominal identification of the factors: the first series is driven by the first factor alone, the second series is driven by the first two factors, and so forth.

6.3 Stochastic Volatility for Factors

Stochastic volatility structures are defined for the sequences of conditional variances of the factors. For each $i = 1, \ldots, k$, define $\lambda_{ti} = \log(h_{ti})$, and write $\boldsymbol{\lambda}_t = (\lambda_{t1}, \ldots, \lambda_{tk})$. The set of log factor variances $\{\boldsymbol{\lambda}_t\}$ is modeled as a vector autoregression of order one, VAR(1), to capture correlations in fluctuations in volatility levels.

Specifically,

$$\boldsymbol{\lambda}_t = \boldsymbol{\mu} + \Phi(\boldsymbol{\lambda}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\omega}_t \tag{6.3}$$

with the following ingredients: $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ is the underlying stationary volatility level, Φ =diag$(\phi_1, \ldots, \phi_k)$ is a diagonal matrix with individual AR parameters ϕ_i for factor volatility process λ_{ti}, and the innovations vectors $\boldsymbol{\omega}_t$ are conditionally independent and normal,

$$\boldsymbol{\omega}_t \sim N(\boldsymbol{\omega}_t|\mathbf{0}, \mathbf{U}) \tag{6.4}$$

for some innovations variance matrix $\mathbf{U}$. This model differs from that of Shephard and Pitt ([81]) in several respects, an important one being that we allow non-zero off-diagonal entries in $\mathbf{U}$ to estimate dependencies in changes in volatility patterns across the factors. This turns out to be empirically supported and practically relevant in short-term exchange rate modeling. We note that Aguilar and West ([1]) also develop stochastic volatility model components for the variances Ψ of the idiosyncratic errors, but we do not explore that here.

6.4 Sequential Analysis

We analyze the one-day-ahead returns on exchange rates over a period of several years in the 1990s, as in [1]. Taking s_{ti} as the spot rate in US dollars for currency i on day t, the returns are simply $y_{ti} = s_{ti}/s_{t-1,i} - 1$ for currency $i = 1, \ldots, q = 6$. The currencies are, in order, the Deutsch mark/Mark (DEM), Japanese Yen (JPY), Canadian Dollar (CAD), French Franc (FRF), British Pound (GBP) and Spanish Peseta (ESP). Here we explore analysis of the returns over the period 12/1/92 to 8/9/96, a total of 964 observations. We adopt the model as structured above, and take an assumedly fixed return level $\boldsymbol{\alpha} = \boldsymbol{\alpha}$.

We first performed intensive Bayesian analysis of the first 914 observations using the MCMC simulation approach of [1]. At $t = 914$, we then have a full sample

from the actual posterior, based on data up to that time point, for all past latent factors, their volatilities, and all fixed model parameters. In terms of proceeding ahead sequentially, we identify the relevant state variables and parameters as follows. First note that we can reduce the model equation (6.1) by integrating out the latent factors to give the conditional observation distribution

$$\mathbf{y}_t \sim N(\mathbf{y}_t|\mathbf{0}, \mathbf{x}\mathbf{H}_t\mathbf{x}' + \Psi). \tag{6.5}$$

Now introduce the definitions of the state variable

$$\mathbf{x}_t \equiv \mathbf{H}_t$$

at time t, and the fixed model parameters

$$\boldsymbol{\theta} = \{\mathbf{x}, \Psi, \boldsymbol{\mu}, \Phi, \mathbf{U}\}.$$

In our example, the state variable $\mathbf{x}_t$ is $3-$dimensional, and the parameter $\boldsymbol{\theta}$ is $36-$dimensional. As we discuss below, the sequential component of the study reported here treats $\mathbf{U}$ as fixed at a value based on the MCMC analysis of the first 914 observations, so that $\boldsymbol{\theta}$ reduces to 30 free model parameters, and the posterior at each time point is in 33 dimensions. For reference, the estimate of $\mathbf{U}$ is

$$E(\mathbf{U}|D_{914}) = \begin{pmatrix} 0.0171 & 0.0027 & 0.0009 \\ 0.0027 & 0.0194 & 0.0013 \\ 0.0009 & 0.0013 & 0.0174 \end{pmatrix}$$

based on the initial MCMC analysis over $t = 1, \ldots, 914$. The posterior standard deviations of elements of $\mathbf{U}$ at $t = 914$ are all of the order of 0.001-0.003, so there is a fair degree of uncertainty about $\mathbf{U}$ that is being ignored in the sequential analysis and comparison.

To connect with the general dynamic model framework, note that we now have the observation equation (3.1) defined by the model equation (6.5), and the evolution

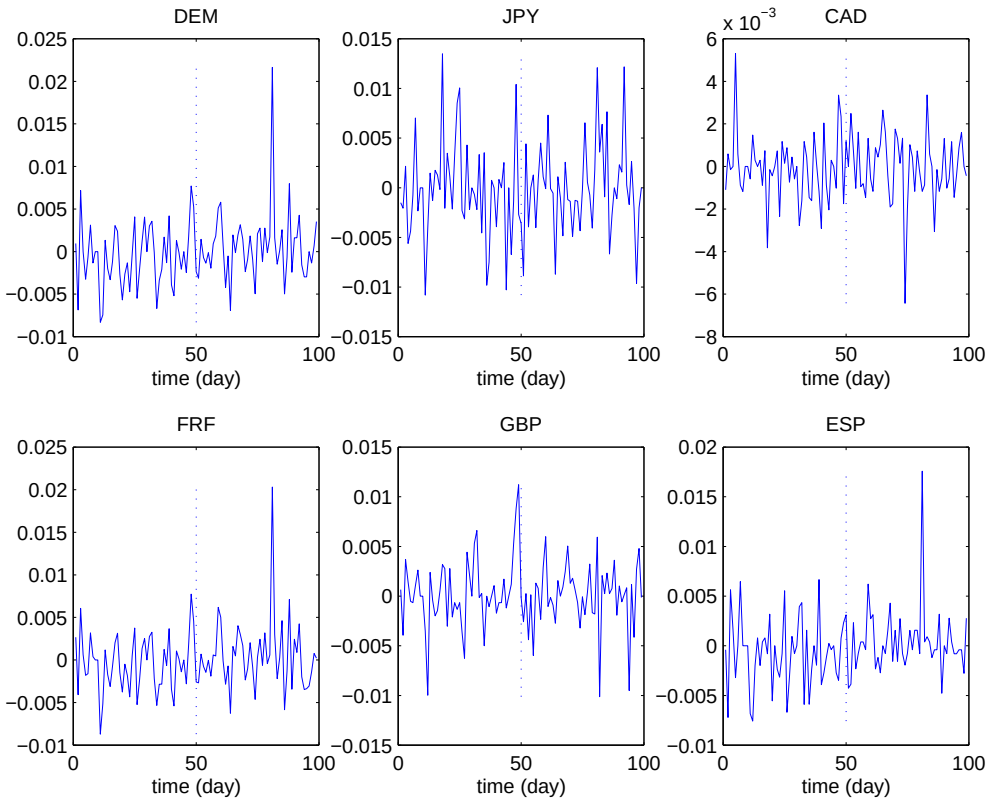

Figure 6.1: Exchange rate time series

equation (3.2) given implicitly by the stochastic volatility model equations (6.3) and (6.4). We can now apply the general sequential filtering algorithm, and do so starting at time $t = 914$ with the full posterior $p(\mathbf{x}_{914}, \boldsymbol{\theta}|D_{914})$ available as a large Monte Carlo sample based on the MCMC analysis of all the data up to that time. It is relevant to note that the context here – with an informed prior $p(\mathbf{x}_{914}, \boldsymbol{\theta}|D_{914})$ based on past data, is precisely that facing practical analysts in many fields in which further analysis, at least over short stretches of data, is required to be sequential. As noted above, we make one change: for reasons discussed below we fix the VAR(1) innovations variance matrix $\mathbf{U}$ at the estimate $E(\mathbf{U}|D_{914})$ and so the parameter $\boldsymbol{\theta}$ is reduced by removal of $\mathbf{U}$. Hence the particle filtering applies to the 3 state variables at each time point and the 30 model parameters. We then proceed to analyze further data, sequentially, over $t = 915, 916, \ldots$. Figure 6.1 displays a stretch of the data running from $t = 864$ to $t = 964$ with $t = 914$ marked. Our sequential filtering methods produce Monte Carlo approximations to each $p(\mathbf{x}_t, \boldsymbol{\theta}|D_t)$ over $t = 915, 916, \ldots$. Throughout the analysis, the Monte Carlo sample size is fixed at $N = 9000$ at each step. The kernel shrinkage and shapes are defined via the discount factor $\delta = 0.99$ which implies $a = 0.995$ and $h = 0.1$. A final technical point to note is that we operate with the kernel method on parameters transformed so that normal kernels are appropriate; thus each of the μ_j and ϕ_j parameters is transformed to the logit scale, and the variance parameters ψ_j are logged (this follows [92, 93]).

Our experiences in this study mirror those of using the straight auxiliary particle filtering method when the parameters are assumed fixed [1]. That is, filtering on the volatilities is a more-or-less standard problem, and the state variable is in only 3 dimensions so performance is expected to be excellent. The questions of accuracy and performance in the extended context with a larger number of parameters are now much more interesting, however, due to the difficulties inherent in deal-

ing with discrete samples in higher dimensional parameter spaces. Inevitably, the accuracy of approximation is degraded relative to simple filtering on two or three time-evolving states. One way to define "performance" here is via comparison of the sequentially computed Monte Carlo approximations to posteriors with those based on a full MCMC analysis that refits the entire data set up to specified time points of interest. Our discussion here focuses specifically on this aspect in connection with inferences on the fixed model parameters. For a chosen set of times during the period of 50 observations between $t = 914$ and 964, we re-ran the full MCMC analysis of the factor model based on all the data up to that time point and explored comparisons with the sequential filtering-based approximations in which we begin filtering at $t = 914$. At any time t, the posterior from the MCMC analysis plays the role of the "true" posterior, or at least the "gold standard" by which to assess the performance of the filtering algorithm. Some relevant summaries appear in Figures 6.2 to 6.11 inclusive. The first set, Figures 6.2–6.6, display summaries of the univariate marginal posterior distributions at $t = 924$. We refer to this as the 10-step analysis, as the sequential filter is run for just 10 steps from the starting position at $t = 914$. For each of the 30 fixed parameters, we display quantile plots comparing quantiles of the approximate posteriors from the MCMC and sequential analyses. The graphs indicate (with crosses) the posterior quantiles at 1, 5, 25, 50, 75, 95 and 99% of the posteriors, graphing the filtering-based quantiles versus those from the MCMC. The $y = x$ line is also drawn. From these graphs, it is evident that posterior margins are excellent agreement (we could have added approximate intervals to these plots, based on methods of Bayesian density estimation, to represent uncertainty about the estimated quantile functions; for the large sample sizes of 9000 here, such intervals are extremely narrow except in the extreme tails, and just obscure the plots.) The APF and MCMC-based posterior margins are the same for all practical purposes. Only in

the very extreme upper tail of two of the VAR model parameters – μ_1 and the logit of ϕ_1 – are there any deviations at all, and here the APF posterior is very slightly heavier tailed than that from the MCMC, but the differences are hardly worth a mention.

The remaining graphs, Figures 6.7–6.11, display similar quantile plots comparing the APF and MCMC posteriors $t = 964$. This provides a similar comparison but now for a 50-step analysis, the sequential filter running for 50 time points from the starting position at $t = 914$. Again we see a high degree of concordance in general, although the longer filtering period has introduced some discrepancies between the marginal posteriors, especially in the extreme tails of several of the margins. Some of the bigger apparent differences appear in the parameters Φ and $\boldsymbol{\mu}$ of the VAR volatility model component, indicated in Figures 6.8 and 6.10. Also noteworthy is the fact that this period of 50 observations includes a point at around $t = 940$ where the series exhibits a real outlier, peaking markedly in the DEM, FRF, ESP and CAN series. Such events challenge sequential methods of any kind, and may play a role here in inducing small additional inaccuracies in the APF approximations by skewing the distribution of posterior weights at that time point. We do have ranges of relevant methods for model monitoring and adaptation to handle such events ([89, 95, 96] and [94]), though such methods are not applied in this study.

One additional aspect of the analysis worth noting is that the distributions of the sets of sequentially updated weights $\omega_t^{(j)}$ remain very well-behaved across the 50 update points. The shape is smooth and unimodal near the norm of $1/N = 1/9000$, with few weights deviating really far at all. Even at the outlier point the maximum weight is only is 0.004, fewer than 200 of the 9000 weights are less than $0.1/9000$, and only 16 exceed $10/9000$. All in all, we can view the analysis as indicating the utility of the filtering approach even over rather longer time intervals.

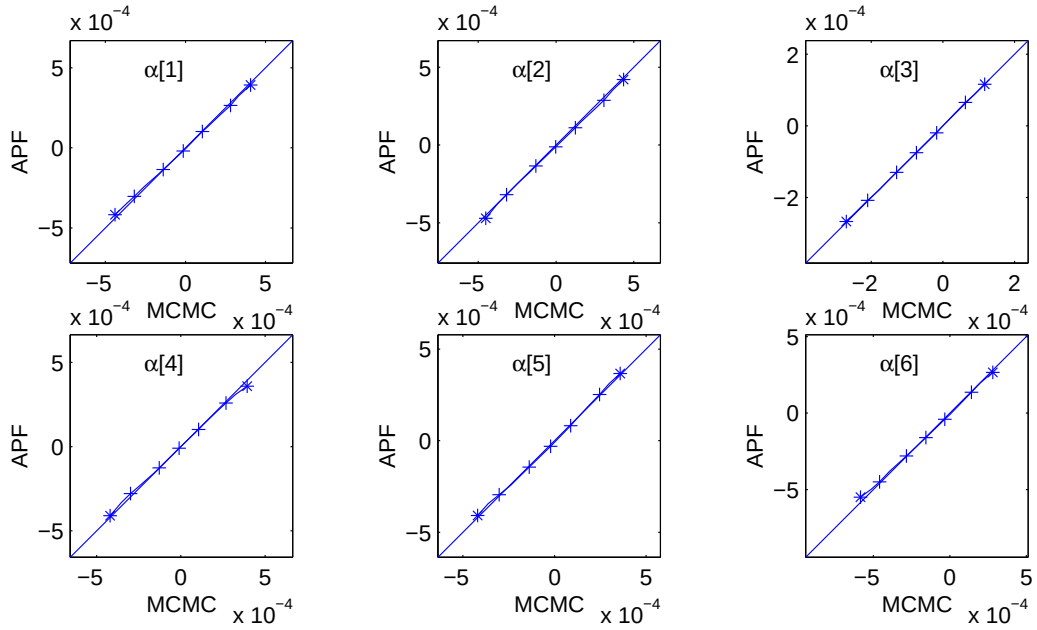

Figure 6.2: Q-Q plots of posterior samples of the α_j parameters in the 10-step analysis

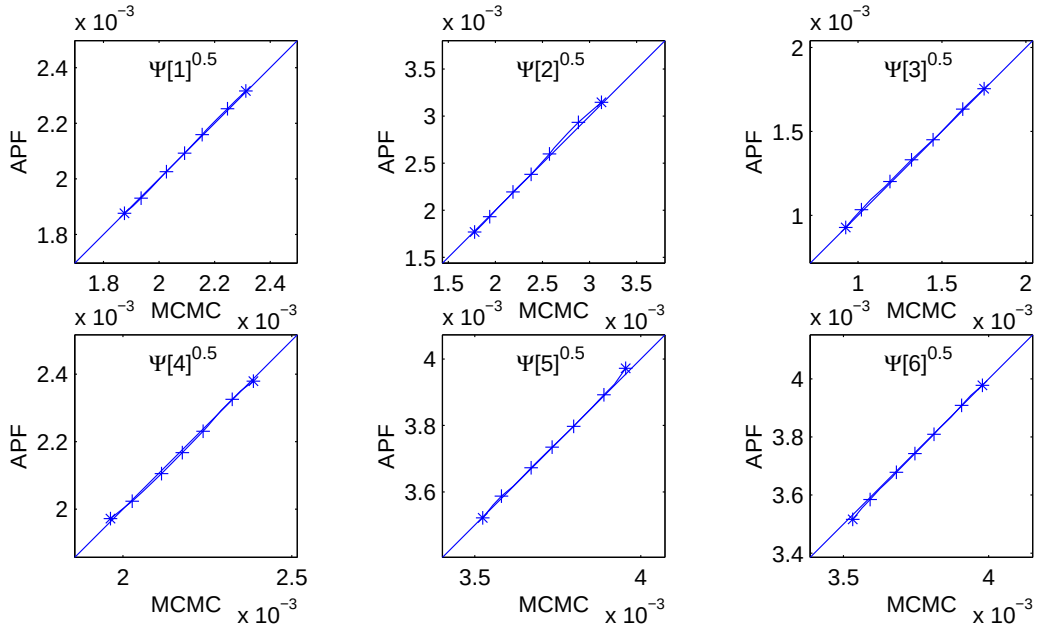

Figure 6.3: Q-Q plots of posterior samples of the $\sqrt{\psi_j}$ parameters in the 10-step analysis

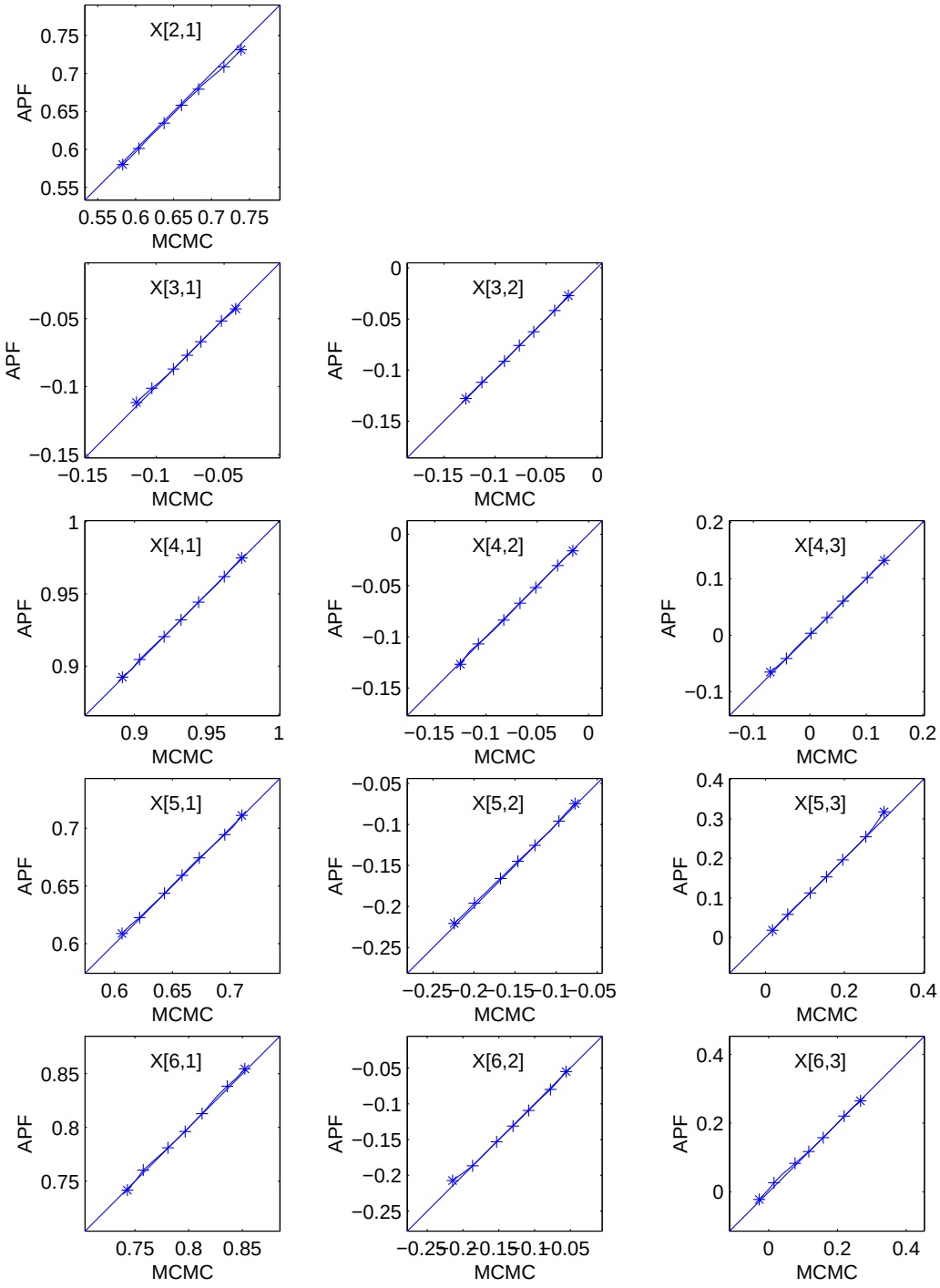

Figure 6.4: Q-Q plots of posterior samples of the X_{ij} parameters in the 10-step analysis

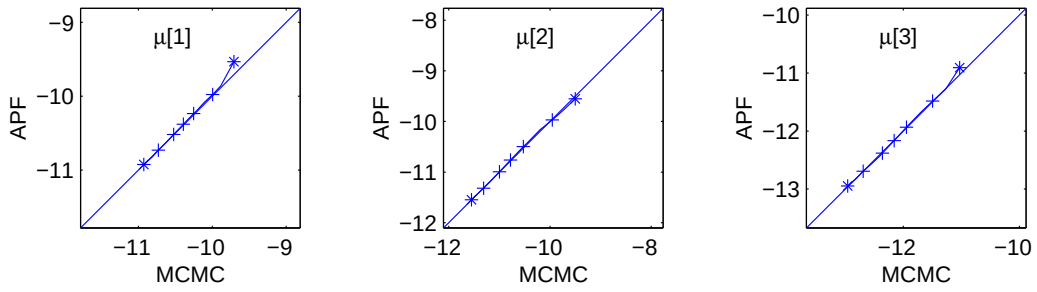

Figure 6.5: Q-Q plots of posterior samples of the μ_j parameters in the 10-step analysis

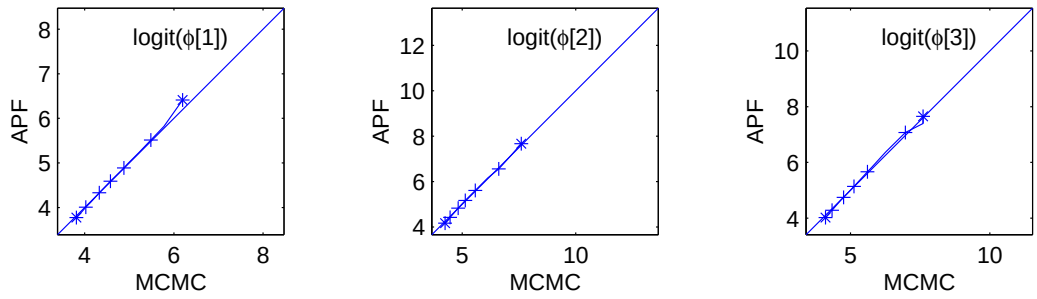

Figure 6.6: Q-Q plots of posterior samples of the logits of the ϕ_j parameters in the 10-step analysis

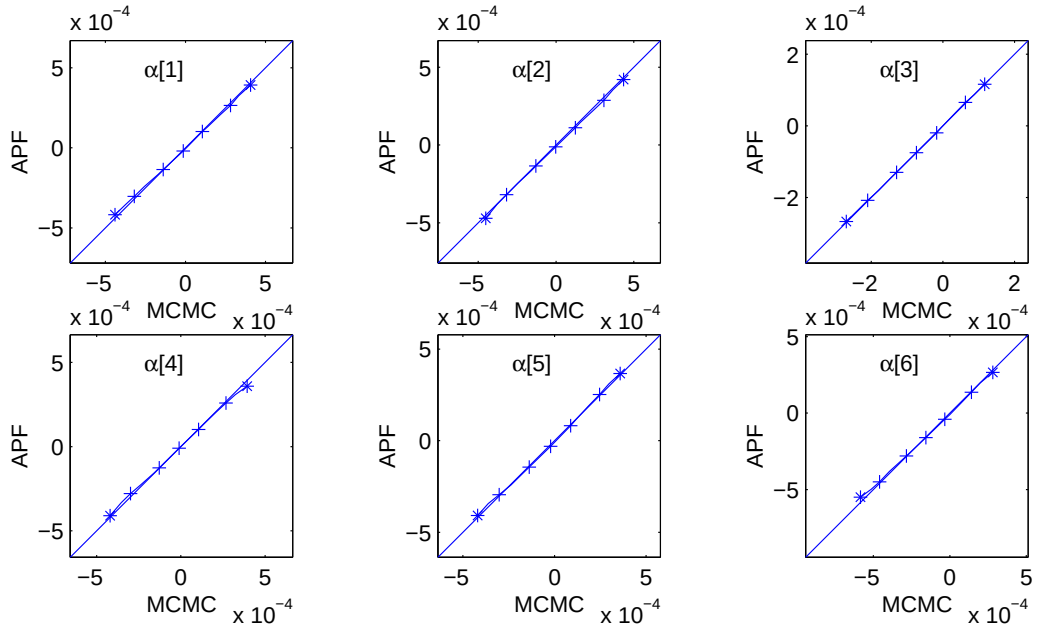

Figure 6.7: Q-Q plots of posterior samples of the α_j parameters in the 50-step analysis

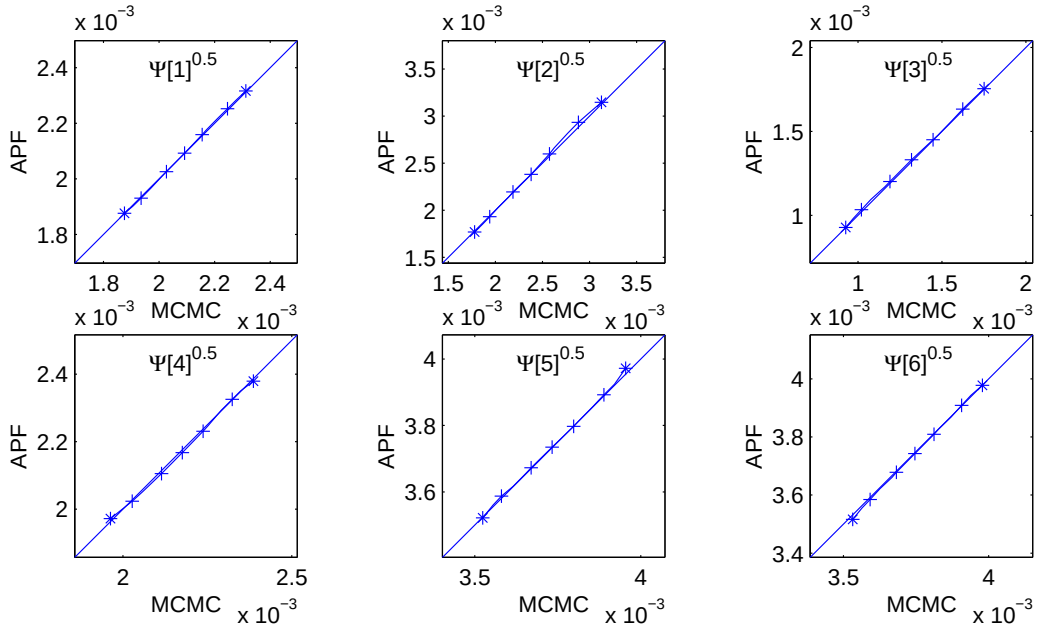

Figure 6.8: Q-Q plots of posterior samples of the $\sqrt{\psi_j}$ parameters in the 50-step analysis

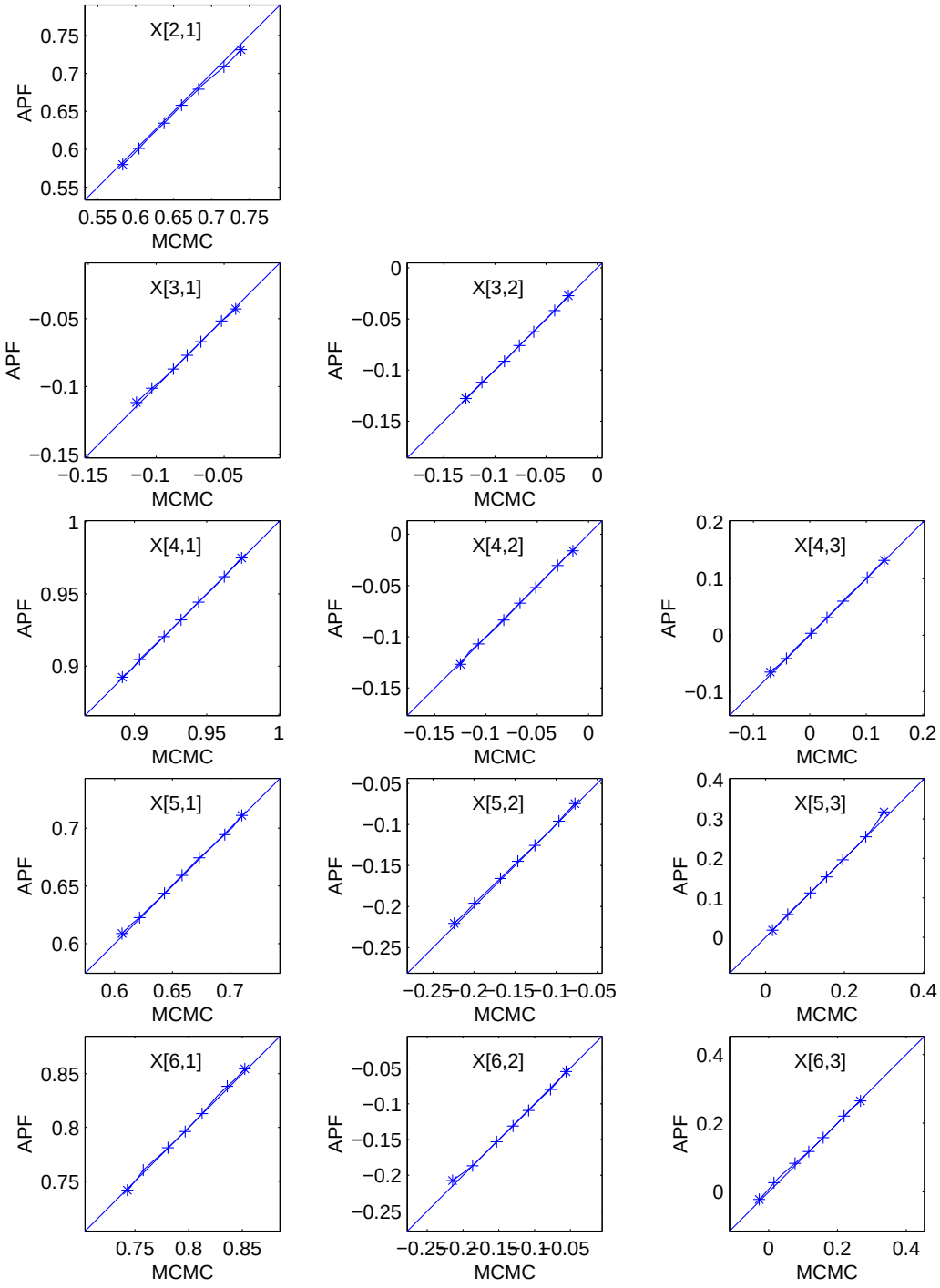

Figure 6.9: Q-Q plots of posterior samples of the X_{ij} parameters in the 50-step analysis

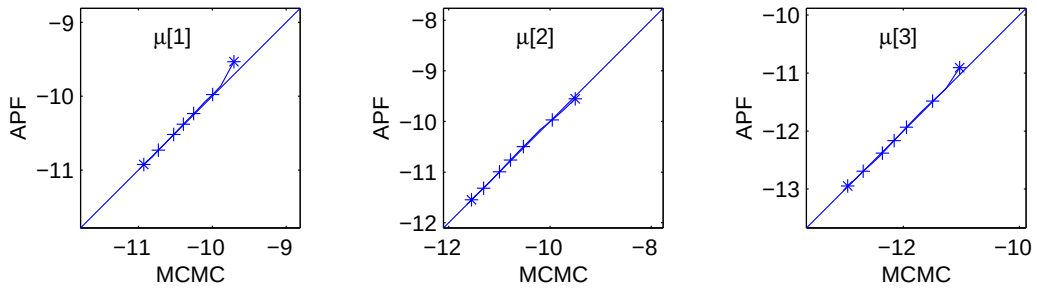

Figure 6.10: Q-Q plots of posterior samples of the μ_j parameters in the 50-step analysis

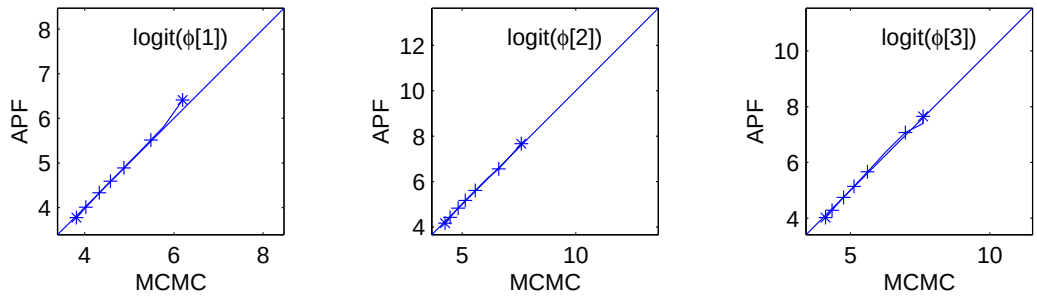

Figure 6.11: Q-Q plots of posterior samples of the logits of the ϕ_j parameters in the 50-step analysis

As earlier mentioned, the sequential simulation analysis fixes the volatility innovations variance matrix $\mathbf{U}$ at its prior mean $E(\mathbf{U}|D_{914})$. This is because we have no easy way of incorporating a structured set of parameters such as $\mathbf{U}$ in the kernel framework – normal distributions do not apply to symmetric positive definite matrices of parameters. It may be that fixing this set of parameters induces some inaccuracies in the filtering analysis compared to the MCMC analysis. With this in mind, we should expect to see some differences between the APF posterior and the MCMC, and these differences can be expected to be most marked in the margins for the volatility model parameters $\boldsymbol{\mu}$ and, most particularly, Φ. The largest differences in the 50-step analysis do indeed relate to Φ, suggesting that some of the differences generally may indeed be due to the lack of proper accounting for the uncertainties about $\mathbf{U}$. Looking ahead, it is of interest to anticipate developments of kernel methods that allow for such structuring – perhaps using normal kernels with elaborate reparametrisations, or perhaps with a combination of normal and non-normal kernels – though for the moment we have no way of doing this.

Chapter 7

Discussion and Future Directions

7.1 On Exponential Power Models

We have attempted to analyze some wavelet coefficients from image data using an AR(1) model with exponential power noise. From a modeling point of view, this model is intended to analyze image data. By nature, the data are not directional and the grey level of one pixel is affected not only by the "previous" pixel, but also all the other neighboring pixels. Modeling correlations from multiple directions would of course introduce more complexity. It is interesting to see if the one directional model is able to represent the correlations that are not directional. In practice, there is a great number of one-dimensional directional signals, for example, physical waves and sound waves conducted along a solid. Our model might be extended to analyze those data.

In terms of MC simulation, our algorithm works relatively well for the simulated "well behaved" series. The implementation on the real wavelet coefficients discloses problems to be improved. First, the second order autocorrelation of the series is above 0.1 so that an AR(2) model might be more appropriate for this data. Second, the samples of β get stuck. Other proposal distribution for β, uniform on $(1, 2)$, or

truncated $2Beta(.,.)$ may help. But on the trace plot, samples of β and α^2 are highly correlated: when β gets stuck, α^2 does not move much either. This seems to be the main reason for slow mixing. Reparameterization, more directional proposal for β, and other novel techniques to improve mixing in MCMC should be considered here.

This AR(1) model with exponential power noise is the first step towards more complicated models that can analyze two-dimensional texture images. For such purposes, wavelet decompositions are powerful tools. The empirical densities of random wavelet coefficients are high-peaked and fat-tailed, which characterize an exponential power distribution. The intrinsic continuity of an image yields the dependence between its wavelet coefficients. These facts motivate the exploration of using autoregressive models with exponential power innovations to model wavelet coefficients. One can expect future developments of this basic model that can handle correlations of higher orders, wavelet coefficients of two dimensions, and coefficients from different resolution levels. This one dimensional model itself, when expanded to AR(p), is a potentially useful model in communicational signal processing and other engineering fields. In those applications, the relevant order of the AR process is in the range of 10 to 30; it is very important that samples could be easily drawn from the full conditional distribution of the AR coefficients, which is a multivariate normal in our Gibbs sampler.

7.2 On Multivariate Stochastic Variance Models

We have explored sequential MC methods along with multivariate volatility models. From the modeling point of view, the two classes of volatility models we have studied, discount models in Chapter 6 and factor models in Chapter 7, present some similarities in format. With the series-specific idiosyncratic variance $\mathbf{\Psi}$ aside, the marginal variance $\mathbf{x}\mathbf{H}_t\mathbf{x}'$ in the factor SV model seems very close to the decomposition

$\mathbf{S} = \mathbf{H}_t \mathbf{L} \mathbf{H}_1'$ used in the singular Wishart density given in Theorem 5.4. The structure of $\mathbf{S}$ itself reminds us of a principle component decomposition, in which $\mathbf{H}_{m\times r}$ is a point on the Stiefel manifold, satisfying $\mathbf{H}'\mathbf{H} = I_r$, i.e. with orthonormal columns. The lower triangular constraint for $\mathbf{x}$ is a convenient assumption to get around the identification problem; theoretically it can well be replaced by orthonormal-column constraint, though MC simulation of such matrices will not be as easy. This resemblance in format discloses the close link between the mathematical tools used to study matrix variate distributions, analytical manifolds and the exterior differential forms defined on them, and the parsimonious empirical models designed to learn about variance matrices. Similarities also exist between the evolution equation (5.26) and $\mathbf{x}\mathbf{H}_t\mathbf{x}'$. In (5.26), $\mathbf{Q}$ is a matrix-beta variable used to introduce some randomness and to "discount" information, while $\mathbf{H}_t$ promises mean-reversion in addition to randomness. The former has mathematical tractability and the latter describes the real process more accurately. One potentially interesting future research direction would be to create some mixture of the two models that combines more reasonable assumptions with analytical tractability.

7.3 On Particle Filters

In our experience, the role of Theorem 4.1 in determining the value of smoothing parameter h used in kernel smoothing via shrinkage is worth noting. Ours coincides with [91] and [83]'s observation that the conventional choice of h given by (4.7) is problematic, especially for models of relatively high dimension. Among the examples that we give for our general algorithm, the AR(1) model has one unknown fixed parameter while the univariate stochastic volatility model has 3 and the multivariate factor model has 30. In our experiment, using (4.7) to determine h does not cause apparent "deformity" in the posterior samples in the AR and univariate volatility

model although it can result in seriously "distorted" samples in the factor model. Let us label the univariate stochastic volatility model as model A and the dynamic factor model model B. According to Theorem 4.1, the equivalent discount factor δ implied by h is $\frac{1}{3-2\sqrt{1-h^2}}$, which, for the h of conventional choice, is 0.7905 in model A, and 0.4635 in model B. 0.4635 is far below the normal range of δ, which may explain why the conventional choice for h does not work for model B. Note that the conventional choice of h is also affected by the sample size, for which we adopted 5000 in model A and 9000 in model B. The equivalent δ is an increasing function of the sample size N and a decreasing function of the dimension of the unknowns m and it is more sensitive to m than to N for the range of m and N in our models. To see the degree of sensitivity, for model A, the equivalent δ falls to 0.6291 only when N reduces to 50, while for model B, δ increases to 0.8711 only when N is as large as 10^{50}. We may conclude that, in kernel smoothing via shrinkage, with the conventional choice for the smoothing parameter, unrealistically large MC sample sizes are required in problems with as many as 30 fixed parameters.

In the moderate dimensional factor model, the analysis certainly indicates the feasibility of sequential simulation-based filtering using the extended auxiliary particle filtering algorithm that incorporates several parameters in addition to state variables. Performance relative to the (almost) equivalent MCMC analysis is excellent; for most practical purposes, the results are in good agreement with the MCMC results even in the 50-step analysis where some minor differences in tail behavior are noted. We have indicated some possible reasons for these differences that are not related to the specific algorithm nor the sequential context. If we ignore those issues and assume that all differences arise due to the inaccuracies inherent in sequential particle filtering, it is clear that there should be room for improvement.

Before discussing some ideas and suggestions for improvements, we want to stress

the relevance of context and goals. Sequential filtering inherently induces approximation errors that may tend to build up over time. In applied work, such as in using dynamic factor models in financial analysis, this must be accounted for and corrected. In existing application of factor models with collaborators in the banking industry, the sequential filtering methods are used over only short time scales – the 5 days of a working week with daily time series such as the exchange rate returns here. This is coupled with periodic updating based on a full MCMC analysis of a longer historical stretch of data (i.e., MCMC at the weekend based on the last several months of data). The horizon of 10 days in the example is therefore very relevant, whereas the 50 day horizon is very long and perhaps unrealistic. In this context, the differences between the filtering-based and MCMC-based posterior quantile functions at lower time steps are quite negligible relative to those at the longer time step. This experience and perspective is consistent with our long-held view that sequential simulation-based filtering methods must always be combined with some form of periodic re-calibration based on off-line analysis performed with much more computational time available than the filtering methods are designed to accommodate.

Some final comments relate to possible extensions of the filtering algorithm that may improve posterior approximations. Questions of accuracy and adequacy arise in connection with the approximation of (typical) posteriors that exhibit varying patterns of dependencies among parameters in different regions of the parameter space, and also varying patterns of tail-weight. Discrete, sample-based approximations inevitably suffer problems of generating points far enough into the tails of fatter-tailed posteriors, especially in higher dimensions. It is sometimes helpful to use fatter-tailed kernels, such as T kernels ([92, 93]) but this does not often help much and goes no way to addressing the real need for more sensitive analytic approximation of *local* structure in the posterior; the kernel mixture of equation (4.2) is *global* in

that the mixture components are based on the same "global" shrinkage center $\overline{\boldsymbol{\theta}}_t$ and each have the same scales and shapes as determined by $h^2\boldsymbol{V}_t$. Very large numbers of such kernels are needed to truly adequately approximate posteriors that may evidence tails of differing weight in different dimensions (and fatter than normal tails), highly non-linear relationships among the parameters and hence varying patterns of "local" scale and shape as we move around in $\boldsymbol{\theta}$ space. We need to complement this suggestion with modifications that allow "differential shrinkage centers." West ([92, 93]) discussed some of these issues, with suggestions about kernel methods with kernel-specific variance matrices in components in particular, and this idea was developed and implemented in certain non-sequential contexts in [38]. Development for implementation in sequential contexts remains an important research challenge.

A simple example helps to highlight these issues and underscores some suggestions for algorithmic extensions that follows. Consider a bimodal prior $p(\boldsymbol{\theta}|D_t)$ in which one mode has the shape of a unit normal distribution and the other that of a normal but with a much larger variance. In using a kernel approximation based on a prior sample, we would expect to do very much better using bimodal mixture in which sample points "near" one mode are shrunk towards that mode, and with kernel scalings that are higher for points "near" the second mode. The existing global kernel method uses global shrinkage to match the first two moments, but loses accuracy in less regular situations as this example indicates.

A specific research direction that reflects these considerations completes our discussion. To begin, consider the existing framework and recall that *any* density function $p(\boldsymbol{\theta}_t|D_t)$ may be arbitrarily well approximated by a mixture of normal distributions. Suppose therefore, for theoretical discussion, that the density has exactly such a form, namely

$$p(\boldsymbol{\theta}_t|D_t) = \sum_{r=1}^{R} q_r N(\boldsymbol{\theta}_t|\mathbf{B}_r, \mathbf{C}_r) \tag{7.1}$$

for some parameters R and $\{q_r, \mathbf{B}_r, \mathbf{C}_r : \ r = 1, \ldots, R\}$ (these will all depend on t though this is not made explicit in the notation, for clarity). In this case, the mean $\overline{\boldsymbol{\theta}}_t$ and variance matrix $\boldsymbol{V}_t$ are given by $\overline{\boldsymbol{\theta}}_t = \sum_{r=1}^{R} q_r \mathbf{B}_r$ and $\boldsymbol{V}_t = \sum_{r=1}^{R} q_r \{\mathbf{C}_r + (\mathbf{B}_r - \overline{\boldsymbol{\theta}}_t)(\mathbf{B}_r - \overline{\boldsymbol{\theta}}_t)'\}$. Suppose also, again, that $\boldsymbol{\theta}_{t+1}$ is generated by the evolution model specified by equation (4.6). It then easily follows that the implied marginal density $p(\boldsymbol{\theta}_{t+1}|D_t)$ is of the form

$$p(\boldsymbol{\theta}_{t+1}|D_t) = \sum_{r=1}^{R} q_k N(\boldsymbol{\theta}_{t+1}|a\mathbf{B}_r + (1-a)\overline{\boldsymbol{\theta}}_t, a^2\mathbf{C}_r + (1-a^2)\boldsymbol{V}_t). \qquad (7.2)$$

Now, in general, this is not the same as the density of $(\boldsymbol{\theta}_t|D_t)$, though the mean and variance matrix match as mentioned above. In practice, a will be quite close to 1 so that the two distributions will be close, but not precisely the same in general. The exception is the case of a normal $p(\boldsymbol{\theta}_t|D_t)$, i.e., the case $R = 1$, when both $p(\boldsymbol{\theta}_t|D_t)$ and $p(\boldsymbol{\theta}_{t+1}|D_t)$ are $N(\cdot|\overline{\boldsymbol{\theta}}_t, \boldsymbol{V}_t)$. Otherwise, in the case of quite non-normal priors, the component means $\mathbf{B}_r$ will be quite separated, the local variance matrices $\mathbf{C}_r$ quite different in scale and structure. Hence the location and scale/shape shrinkage effects in the components of the resulting mixture (7.2) tend to obscure the differences by the implied shrinking/averaging.

This discussion, linking back to the important use of normality of $\boldsymbol{\theta}_t$ in the theoretical tie-up between artificial evolution methods and kernel methods in Section 3.3, suggests the following development. Suppose that the distribution $p(\boldsymbol{\theta}_t|D_t)$ is indeed exactly of the form of equation (7.2). To focus on "local" structure in this distribution, introduce the component indicator variable r_t such that $r_t = r$ with probability q_r $(r = 1, \ldots, R)$. Then $(\boldsymbol{\theta}_t|r_t = r, D_t) \sim N(\cdot|\mathbf{B}_r, \mathbf{C}_r)$. At this point we can apply the same line of reasoning about an artificial evolution to smooth a set of $\boldsymbol{\theta}_t$ samples, but now explicitly including the indicator r_t provides a focus on the *local* structure. This suggests the modification of the key evolution equation (4.6) to the

local form

$$p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, r_t = r, D_t) \sim N(\boldsymbol{\theta}_{t+1}|a\boldsymbol{\theta}_t + (1-a)\mathbf{B}_r, h^2\mathbf{C}_r) \tag{7.3}$$

where a, h are as earlier defined. This conditional distribution is such that the implied marginal $p(\boldsymbol{\theta}_{t+1}|D_t)$ has precisely the same mixture form as $p(\boldsymbol{\theta}_t|D_t)$, so that the local structure is respected. This theoretical discussion therefore indicates and opens up a direction for development that, if implemented, can be expected to generate more accurate and efficient methods of smoothing posterior samples. To exploit this mixture theory will require, among other things, computationally and statistically efficient methods of identifying the parameters R and $\{q_r, \mathbf{B}_r, \mathbf{C}_r : \ r = 1, \ldots, R\}$ of the mixture in equation (7.1) based on an existing Monte Carlo sample (and weights) from that distribution. Some form of hierarchical clustering of sample points (and weights), such as utilized in [92, 93], will be needed, though a key emphasis lies on computational efficiency so new clustering methods will be needed. The connections with stratification methods in simulation are also evident and likely worth elaborating on. Such developments, while challenging, will directly contribute in this context to usefully extend and improve the existing algorithms for sequential filtering on both parameters and states in higher dimensional dynamic models.

Appendix A

Appendix to Chapter 2

A.1 Stable Distributions

The polar form of the log of the characteristic function of the $stable(b, \delta)$ distribution is

$$log(\phi(t)) = -|t|^b e^{-i\frac{\pi}{2} min(b, 2-b)\delta sgn(t)}.$$

The parameter δ is the skewness parameter. When $\delta = 0$ we obtain the symmetric stable distribution, whose characteristic function is $\phi(t) = e^{-|t|^b}$. Symmetric stables are unimodal with infinite tails. Special cases are Cauchy when $b = 1$ and normal when $b = 2$. The parameter b indicates how fast the tails decay; roughly speaking, the tails drop at the rate of $|x|^{-(1+b)}$ as $|x| \to \infty$. There are two infinite tails when $|\delta| \neq 1$ or $b \leq 1$. When $b < 1$ and $\delta = 1$, the distribution is restricted to a positive range.

Symmetric stables are scale mixtures of normals with positive stable mixing densities: if $q_b(x)$ is the $stable(b, 0)$ density, then

$$q_b(x) = \int_0^\infty N(x|0, 2y) p_{\frac{b}{2}}(y) dy$$

where $N(.|0, 2y)$ is normal density with mean 0 and variance $2y$, and $p_{\frac{b}{2}}(.)$ is the

$stable(\frac{b}{2}, 1)$ density.

Samples from $stable(b, 1)$ can be drawn according to the following theorem from Devroy ([24]). For $b < 1$, the density of a stable (b,1) can be written as

$$f(x) = \frac{bx^{\frac{1}{b-1}}}{(1-b)\pi} \int_0^\pi g(u) exp(-g(u)x^{\frac{b}{b-1}}) du$$

where $g(u) = (\frac{sin(bu)}{sin(u)})^{\frac{1}{1-b}} \frac{sin((1-b)u)}{sin(bu)}$. When U is uniformly distributed on $[0, 1]$ and E is independent of U and exponentially distributed, $(\frac{g(\pi U)}{E})^{\frac{1-b}{b}}$ is $stable(b, 1)$ distributed.

A.2 Exponential Power as a Mixture of Normals

We rephrase the theorem in Section 2 of [90] in Devroy's notation to clarify the confusions caused by various parameterizations.

If t' follows the exponential power distribution with location 0 and scale $\alpha(> 0)$, then its density, as given on page 157 of [13], is

$$p(t') = [\Gamma(1 + \frac{1}{b})2^{1+\frac{1}{b}}\alpha]^{-1} exp(-\frac{1}{2}|\frac{t'}{\alpha}|^b) \tag{A.1}$$

where $1 \leq b$.

Let $t = 2^{-\frac{1}{b}}\alpha^{-1}t'$. Then t has density

$$p(t) \propto exp(-|t|^b).$$

On the other hand, from page 455 of [24], the symmetric stable characteristic function is

$$\Psi(t) = exp(-|t|^b)$$

where $0 < b \leq 2$.

By definition,

$$\Psi(t) = \int_{-\infty}^{\infty} e^{itx} q_b(x) dx$$

where $q_b(x)$ is the stable(b, 0) density. As mentioned above,

$$q_b(x) = \int_{-\infty}^{\infty} N(x|0, 2y) p_{\frac{b}{2}}(y) dy$$

where $N(.|0, 2y)$ is normal density with mean 0 and variance $2y$, and $p_{\frac{b}{2}}(.)$ is the $stable(\frac{b}{2}, 1)$ density.

Thus for $1 \leq b \leq 2$

$$\begin{aligned} exp(-|t|^b) &= \int_{-\infty}^{\infty} e^{itx} \int_0^{\infty} N(x|0, 2y) p_{\frac{b}{2}}(y) dy dx \\ &= \int_0^{\infty} (\int_{-\infty}^{\infty} e^{itx} N(x|0, 2y) dx) p_{\frac{b}{2}}(y) dy \\ &= \int_0^{\infty} exp(-yt^2) p_{\frac{b}{2}}(y) dy \\ &= \int_0^{\infty} \sqrt{2y} exp(-\frac{2yt^2}{2}) \frac{1}{\sqrt{2y}} p_{\frac{b}{2}}(y) dy. \end{aligned} \tag{A.2}$$

Let $\phi = 2y$. Then the mixing density is $h(\phi) = \phi^{-\frac{1}{2}} p_{\frac{b}{2}}(\frac{\phi}{2})$, and given ϕ, $t \sim N(0, \phi^{-1})$. t' in (A.1) is also a scale mixture of normal: $t'|\phi \sim N(0, \phi^{-1}\alpha^2 2^{\frac{2}{b}})$ and ϕ has density $h(\phi)$.

Appendix B

Appendix to Chapter 5

B.1 Generalized Inverse of a Matrix

Definition of generalized inverse of a matrix. Let $\mathbf{A}$ be an $m \times n$ matrix. A generalized inverse or a g-inverse of $\mathbf{A}$ is an $n \times m$ matrix, denoted by $\mathbf{A}^-$, such that $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$.

A g-inverse $\mathbf{A}^-$ of $\mathbf{A}$ always exists but is not necessarily unique. We can make it unique by imposing additional conditions. See [40] for further reference.

An equivalent definition of generalized inverse of an $m \times n$ matrix $\mathbf{A}$ is an $n \times m$ matrix, denoted by $\mathbf{A}^-$, such that $\mathbf{X} = \mathbf{A}^-\mathbf{y}$ is a solution of $\mathbf{A}\mathbf{X} = \mathbf{y}$ for any $\mathbf{y}$ which makes the equation consistent.

Some properties of generalized inverse of a matrix. Let $\mathbf{A}^-$ be any g-inverse and $\mathbf{H} = \mathbf{A}^-\mathbf{A}$, then the following results hold:

1. $\mathbf{A}^-\mathbf{A}$ and $\mathbf{A}\mathbf{A}^-$ are idempotent.
2. $rank(\mathbf{A}) = rank(\mathbf{H}) = tr(\mathbf{H}) = rank(\mathbf{A}\mathbf{A}^-)$.
3. $rank(\mathbf{A}) \leq rank(\mathbf{A}^-)$.

4. A general solution of the consistent equation $\mathbf{AX} = \mathbf{y}$ is $\mathbf{A}^{-}\mathbf{y} + (\mathbf{I} - \mathbf{H})\mathbf{z}$ where $\mathbf{z}$ is arbitrary.

5. Let $\mathbf{A}$ be $m \times n$, then $\mathbf{A}^{-}\mathbf{A} = \mathbf{I}_n$ if and only if $rank(\mathbf{A}) = n$ and $\mathbf{A}\mathbf{A}^{-} = \mathbf{I}_m$ if and only if $rank(\mathbf{A}) = m$.

6. If $\mathbf{A}^{-}$ is any g-inverse of $\mathbf{A}$, then $(\mathbf{A}^{-})'$ is a g-inverse of $\mathbf{A}'$.

Note: Let $\mathbf{A}$ be an $m \times n$ matrix of rank $r(\leq min(m, n))$, then by a suitable interchange of rows and columns, $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{pmatrix} \tag{B.1}$$

where $\mathbf{A}_1$ is a nonsingular $r \times r$ sub matrix. Consider

$$\mathbf{A}^{-} = \begin{pmatrix} \mathbf{A}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{B.2}$$

It is easy to observe that $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$ so that $\mathbf{A}^{-}$ is a g-inverse with rank r. Alternatively, we can consider

$$\mathbf{A}^{-} = \begin{pmatrix} \mathbf{A}_1^{-1} & -\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{B.3}$$

where $\mathbf{B}$ is such that $\mathbf{A}_1\mathbf{B} = \mathbf{A}_2$ and $\mathbf{A}_3\mathbf{B} = \mathbf{A}_4$. Then $\mathbf{A}^{-}$ is a g-inverse with rank $min(m, n)$.

B.2 Definition and Properties of (Singular) Multivariate Normal Distribution

Covariance matrix $\mathbf{\Sigma} : p \times p$ is said to be singular or degenerate when the rank of $\mathbf{\Sigma}$ is less than p, say $r(< p)$. In this case, the total probability mass concentrates on a linear set of exactly r dimensions with probability one.

To have a natural definition of a singular multivariate normal distribution, consider the following factorization of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \mathbf{H} \begin{pmatrix} \mathbf{D}_\lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{H}' \tag{B.4}$$

$$\mathbf{H} = (\mathbf{H}_1, \mathbf{H}_2)$$

where $\mathbf{H} : p \times p$ is an orthogonal matrix, $\mathbf{H}_1 : p \times r$ being the matrix of the first r columns of $\mathbf{H}$, and $\mathbf{D}_\lambda = diag(\lambda_1, \ldots, \lambda_r)$ with $\lambda_i > 0$ for $i = 1, \ldots, r$. Putting $\mathbf{B} = \mathbf{H}_1 \mathbf{D}_{\sqrt{\lambda}}$, we have $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}'$. $\mathbf{B}$ is not unique, using one of such matrices, $\mathbf{B}$, and $\boldsymbol{\mu} \sim N_r(\mathbf{0}, \mathbf{I}_r)$, we consider $\mathbf{B}\boldsymbol{\mu} + \boldsymbol{\nu}$.

The characteristic function (c.f.) is

$$\begin{aligned} \mathbf{E}[exp\{i\mathbf{t}'(\mathbf{B}\boldsymbol{\mu} + \boldsymbol{\nu})\}] &= exp\{i\mathbf{t}'\boldsymbol{\nu} - \frac{1}{2}\mathbf{t}'(\mathbf{B}\mathbf{B}')\mathbf{t}\} \\ &= exp\{i\mathbf{t}'\boldsymbol{\nu} - \frac{1}{2}\mathbf{t}'\mathbf{\Sigma}\mathbf{t}\} \end{aligned}$$

Since the argument is obviously valid for $r = p$, the following definition includes both nonsingular and singular cases.

Definition of multivariate normal distribution. A random vector $\mathbf{x}$ in the p-dimensional Euclidean space R^p is said to have a p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$ if x has the same distribution as $\mathbf{B}\boldsymbol{\mu}+\boldsymbol{\nu}$, where $\mathbf{B} : p \times r$ is of rank r such that $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}'$ and $\boldsymbol{\mu} \sim N_r(\mathbf{0}, \mathbf{I}_r)$ where $r \leq p$. r is called the rank of the distribution. When $r = p$, the distribution is nonsingular and when $r < p$, it is singular. We write $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$

The singular density. The density of a p-variate singular normal distribution $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ may be defined in the following way. Let $\mathbf{y}' = \mathbf{H}'x = (\mathbf{y}_1', \mathbf{y}_2')$, $\mathbf{s}' = \mathbf{t}'\mathbf{H} =$

$(\mathbf{s}_1', \mathbf{s}_2')$, $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2') = \boldsymbol{\mu}'\mathbf{H}$, the c.f. of $\mathbf{y}$ can be written as

$$\begin{aligned} \phi_y(\mathbf{s}) &= \mathbf{E}[exp\{i\mathbf{s}'\mathbf{y}\}] = \mathbf{E}[exp\{i\mathbf{s}'\mathbf{H}'\mathbf{x}\}] \\ &= \mathbf{E}[exp\{i\mathbf{s}'(\mathbf{B}\boldsymbol{\mu} + \boldsymbol{\nu})\}] \\ &= exp(i\mathbf{s}_1'\boldsymbol{\theta}_1 - \frac{1}{2}\mathbf{s}_1'\mathbf{D}_\lambda\mathbf{s}_1 + i\mathbf{s}_2'\boldsymbol{\theta}_2). \end{aligned} \tag{B.5}$$

It follows that $\mathbf{y}_1 \sim N_r(\boldsymbol{\theta}_1, \mathbf{D}_\lambda), \mathbf{D}_\lambda > 0, \mathbf{y}_2 = \boldsymbol{\theta}_2$ with probability one, and they are independently distributed. Since $\mathbf{y}_2 - \boldsymbol{\theta}_2 = \mathbf{H}_2'(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0}$, the domain of the distribution is the r-dimensional linear subspace defined by $\mathbf{H}_2'(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0}$, where $\mathbf{H}_2 : p \times (p - r)$ satisfies $\mathbf{H}_2'\mathbf{H}_1 = 0$, and $\mathbf{H}_2'\mathbf{H}_2 = \mathbf{I}$. The p.d.f. of $\mathbf{y}_1$ is

$$\begin{aligned} f(\mathbf{y}_1) &= |2\pi\mathbf{D}_\lambda|^{-1/2} exp\{-\frac{1}{2}(\mathbf{y}_1 - \boldsymbol{\theta}_1)'\mathbf{D}_\lambda^{-1}(\mathbf{y}_1 - \boldsymbol{\theta}_1)\} \\ f(\mathbf{x}) &= |2\pi\mathbf{D}_\lambda|^{-1/2} exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{H}_1\mathbf{D}_\lambda^{-1}\mathbf{H}_1')(\mathbf{x} - \boldsymbol{\mu})\} \\ &= |2\pi\mathbf{D}_\lambda|^{-1/2} exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-}(\mathbf{x} - \boldsymbol{\mu})\} \end{aligned} \tag{B.6}$$

where x lies on the r-dimensional linear subspace defined by $\mathbf{H}_2'(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0}$, where $\mathbf{H}_2 : p \times (p - r)$ satisfies $\mathbf{H}_2'\mathbf{H}_1 = 0$, and $\mathbf{H}_2'\mathbf{H}_2 = \mathbf{I}$.

Some properties of multivariate normal distribution.

1. Characteristic function: The c.f. of $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \geq 0$ is $exp\{i\mathbf{t}'\boldsymbol{\nu} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\}$
2. Let $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \geq 0$, $\mathbf{A}$ be a $k \times p$ matrix, then $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \sim N_k(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$
3. Let $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')' \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \geq 0$, where $\mathbf{x}_1$ and $\mathbf{x}_2'$ are $r-$ and $(p - r)-$ vectors respectively.Let the corresponding partitions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

be $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{i,j}), i, j = 1, 2$, $\mathbf{x}_1$ and $\mathbf{x}_{2\mathrm{i}} = \mathbf{x}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-}\mathbf{x}_1$ are independently distributed and $\mathbf{x}_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{x}_{2\mathrm{i}} \sim N_r(\boldsymbol{\mu}_{2\mathrm{i}}, \boldsymbol{\Sigma}_{22\mathrm{i}})$ where $\boldsymbol{\mu}_{2\mathrm{i}} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-}\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_{22\mathrm{i}} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-}\boldsymbol{\Sigma}_{12}$.

Bibliography

[1] O Aguilar and M West. Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics*, June, 2000.

[2] D Avitzour. Stochastic simulation Bayesian approach to multitarget tracking. *IEE Proc.-Radar, Sonar Navig.*, 142:41–44, 1995.

[3] W Bauwens and F Richard. A 1-1 poly-t random variable generator with application to Monte Carlo integration. *Journal of Econometrics*, 29:19–46, 1985.

[4] C Berzuini, N G Best, W R Gilks, and C Larizza. Dynamic conditional independence models and Markov Chain Monte Carlo methods. *Journal of the American Statistical Association*, 92(440):1403–1412, 1997.

[5] J Besag and C L Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82:733–746, 1995.

[6] P Bhimasankaram and D Sengupta. Testing for the mean vector of a multivariate normal distribution with a possibly singular dispersion matrix and related results. *Statistics & Probability Letters*, 11:473–478, 1991.

[7] F Black and B Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43, 1992.

[8] T Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[9] T Bollerslev. Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH approach. *Review of Economics and Statistics*, 72:498–505, 1990.

[10] T Bollerslev and R F Engle. A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, 96(1):116–131, 1988.

[11] T Bollerslev and R F Engle. Common persistence in conditional variances. *Econometrica*, 61:167–186, 1993.

[12] T Bollerslev, R F Engle, and D B Nelson. ARCH models. *The Handbook of Econometrics*, 4:2959–3038, 1994.

[13] G E P Box and G C Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Massachussetts, 1973.

[14] R G Brown. *Smoothing, Forecasting and Prediction of Discrete Time Series*. Prentice-Hall, Englewood Cliffs, NJ, 1962.

[15] P J Burt and E H Adelson. The Laplacian pyramid as a compact image code. *IEEE Trans Commun*, 31:532–540, 1983.

[16] C K Carter and R Kohn. On Gibbs sampling for state space models. *Biometrika*, 81:541–553, 1994.

[17] R Chen and J S Liu. Predictive updating methods with application to Bayesian classification. *Journal of the Royal Statistical Society (Ser B)*, 58:397–415, 1996.

[18] S Chib, F Nardari, and N Shephard. Markov Chain Monte Carlo methods for generalized stochastic volatility models. Working paper, Nuffield College, Oxford University, 1998.

[19] S Chib, F Nardari, and N Shephard. Analysis of high dimensional multivariate stochastic volatility models. Working paper, Nuffield College, Oxford University, 1999.

[20] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch. Adaptive Bayesian wavelet shrinkage. *Journal of American Statististical Association*, 92(440):1413–1421, 1997.

[21] T Bollerslev R Chou and K Kroner. ARCH modeling in finance. *Journal of Econometrics*, 52:5–59, 1992.

[22] M. A. Clyde, G. Parmigiani, and B. Vidakovic. Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85:391–402, 1998.

[23] J Crowley. A representation for visual information. Technical report CMU-RI-TR-82-7, Carnegie Mellon University, 1987.

[24] L Devroy. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[25] F X Diebold and M Nerlove. The dynamics of exchange rate volatility: A multivariate latent ARCH model. *Journal of Applied Econometrics*, 4:1–21, 1989.

[26] A Doucet, S J Godsill, and C Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, forthcoming, 2000.

[27] R F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica*, 50:987–1008, 1982.

[28] R F Engle. Multivariate GARCH with factor structures-cointegration in variance. Unpublished paper, Department of Economics, UCSD, 1987.

[29] R F Engle and T Bollerslev. Modeling the persistence of conditional variances. *Econometric Reviews*, 5(1):1–50, 1986.

[30] R F Engle, D M Lilien, and R P Robins. Estimating time-varying risk premia in the term structure: The ARCH-M model. *Econometrica*, 55:391–407, 1987.

[31] R F Engle, V K Ng, and M Rothschild. Asset pricing with a factor ARCH covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics*, 45:213–238, 1990.

[32] W Feller. *An Introduction to Probability Theory and Its Applications.* Wiley, New York, 1966.

[33] S Fruhwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15:183–202, 1994.

[34] A R Gallant and J F Monahan. Explicitly infinite-dimensional Bayesian analysis of production technologies. *Journal of Econometrics*, 30:171–202, 1985.

[35] A Gelman, J B Carlin, H S Stern, and D B Rubin. *Bayesian Data Analysis.* Chapman and Hall, New York, 1995.

[36] J Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, Nov 1989.

[37] J F Geweke and G Zhou. Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies*, 9:557–587, 1996.

[38] G Givens and A E Raftery. Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association*, 91:132–141, 1996.

[39] N J Gordon, D J Salmond, and A F M Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, 1993.

[40] F A Graybill. *Introduction to Matrices with Application in Statistics*. Wadsworth, Belmont, CA, 1969.

[41] J M Hammersley and DC Handscomb. *Monte Carlo Methods*. New York: Wiley, 1964.

[42] P J Harrison and C F Stevens. Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society (Ser B)*, 38:205–247, 1976.

[43] A C Harvey, E Ruiz, and N Shephard. Multivariate stochastic variance models. *Review of Economic Studies*, 61:247–264, 1994.

[44] W K Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 1970.

[45] C C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. In *ONR Research Memo*, Carnegie Institute of Technology, 1957.

[46] E Jacquier, N G Polson, and P Rossi. Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, 12:371–388, 1994.

[47] E Jacquier, N G Polson, and P Rossi. Models and prior distributions for multivariate stochastic volatility. Unpublished paper, Graduate School of Business, University of Chicago, 1995.

[48] S Kim, N Shephard, and S Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65:361–393, 1998.

[49] M King, E Sentana, and S Wadhwani. Volatility and links between national stock markets. *Econometrica*, 62:901–933, 1994.

[50] G Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[51] J Koenderink. The structure of images. *Cybernetics*, 1984.

[52] A Kong, J S Liu, and W H Wong. Sequential imputation and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

[53] F Li. Time deformation modeling: Theory and application. Unpublished PhD thesis, Institute of Statistics and Decision Sciences, Duke University, 1997.

[54] J S Liu. Rejection control and sequential importance sampling. *Journal of American Statistical Association*, 93:1022–1031, 1998.

[55] J S Liu and R Chen. Blind deconvolution via sequential imputations. *Journal of the American Statistical Association*, 90(430):567–576, 1995.

[56] JS Liu and R Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 90:567–576, 1995.

[57] S Mallat. Multiresolution approximation and wavelet orthonormal bases of L^2. *Trans Amer Math Soc*, June 1989.

[58] S Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans on Pattan Analysis and Machine Intelligence*, 11(7), 1989.

[59] B Mandelbrot. The variation of certain speculative prices. *Journal of Business*, 36:394–419, 1963.

[60] A Meliino and S M Turnbull. Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, 45:239–265, 1990.

[61] N Metropolis, A W Rosenbluth M N Rosenbluth A H Teller, and E Teller. Equations of state calculations by fast computing machines. *Journal Chem Phys*, 21:1087–1091, 1953.

[62] R J Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, New York, 1982.

[63] P. Müller and B. Vidakovic. MCMC methods in wavelet shrinkage: Non-equally spaced regression, density and spectral density estimation. In P. Müller and B. Vidakovic, editors, *Bayesian Inference In Wavelet Based Models*, volume 141 of *Lecture Notes in Statistics*, pages 187 – 202. Springer Verlag, New York, 1999.

[64] D Nelson. The time series behavior of stock market volatility and returns. unpublished PhD thesis, Department of Economics, Massachusetts Institute of Technology, 1988.

[65] D Nelson. Conditional heteroscedasticity in asset pricing: A new approach. *Econometrica*, 59:347–370, 1991.

[66] S Peddada and D S P Richards. Proof of a conjecture of M L Eaton on the

characteristic function of the Wishart distribution. *The Annals of Probability*, 19(2):868–874, 1991.

[67] M Pitt and N Shephard. Filtering via simulation: Auxiliary particle filter. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

[68] A Pole. Transfer response models: A numerical approach. In J M Bernardo, M H DeGroot, D V Lindley, and A F M Smith, editors, *Bayesian Statistics 3*, Oxford University Press, 1988.

[69] A Pole and M West. Efficient Bayesian learning in non-linear dynamic models. *Journal of Forecasting*, 9:119–136, 1990.

[70] A Pole, M West, and P J Harrison. Non-normal and non-linear dynamic Bayesian modelling. In J C Spall, editor, *Bayesian Analysis of Time Series and Dynamic Models*. Marcel Dekker, New York, 1988.

[71] B H Putnam and J M Quintana. New Bayesian statistical approaches to estimating and evaluating models of exchange rates determination. In *American Statistical Association 1994 Proceedings of the Section on Bayesian Statistical Science*, 1994.

[72] J M Quintana. Multivariate Bayesian forecasting models. Unpublished PhD thesis, Department of Statistics, University of Warwick, 1987.

[73] J M Quintana. Optimal portfolios of forward currency contracts. In J O Berger, J M Bernardo, A P Dawid, and A F M Smith, editors, *Bayesian Statistics 4*, Oxford University Press, 1992.

[74] J M Quintana, V K Chopra, and B H Putnam. Global asset allocation: Stretching returns by shrinking forecasts. In *American Statistical Association 1995 Proceedings of the Section on Bayesian Statistical Science*, Bankers Trust Company, New York, 1995.

[75] J M Quintana and B H Putnam. Debating currency markets efficiency using dynamic multiple-factor models. In *American Statistical Association 1995 Proceedings of the Section on Bayesian Statistical Science*, Bankers Trust Company, New York, 1996.

[76] J M Quintana and M West. Multivariate time series analysis: New techniques applied to international exchange rate data. *The Statistician*, 36:275–281, 1987.

[77] J M Quintana and M West. Time series analysis of compositional data. In J M Bernardo, M H DeGroot, D V Lindley, and A F M Smith, editors, *Bayesian Statistics 3*, Oxford University Press, 1988.

[78] D B Rubin. Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of American Statistical Association*, 82:543–545, 1987.

[79] N Shephard. Fitting non-linear time series models with applications to stochastic variance models. *Journal of Applied Econometrics*, 8:135–152, 1993.

[80] N Shephard and M Pitt. Likelihood analysis of non-gaussian measurement time series. *Biometrika*, 84, 1997.

[81] N Shephard and M Pitt. Analysis of time varying covariances: A factor stochastic volatility approach (with discussion). In J O Berger, J M Bernardo, A P Dawid, and A F M Smith, editors, *Bayesian Statistics 6*, Oxford University Press, 1999.

[82] N Shephard and M K Pitt. Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84:653–667, 1997.

[83] B W Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[84] A F M Smith and A E Gelfand. Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

[85] A F M Smith and M West. Monitoring renal transplants: An application of the multi-process Kalman filter. *Biometrics*, 39:867–878, 1983.

[86] H Uhlig. On singular Wishart and singular multivariate beta distribution. *Annals of Statistics*, 22:395–405, 1994.

[87] G. G. Walter. *Wavelets and Other Orthogonal Systems with Applications*. CRC Press Inc., Boca Raton, FL, 1994.

[88] M West. Aspects of recursive Bayesian estimation. Unpublished PhD thesis, Department of Mathematics, University of Nottingham, 1982.

[89] M West. Bayesian model monitoring. *Journal of the Royal Statistical Society (Ser B)*, 48:70–78, 1986.

[90] M West. On scale mixture of normal distributions. *Biometrika*, 74(3), 1987.

[91] M West. Bayesian kernel density estimation. Discussion paper 90-A02, ISDS, Duke University, 1990.

[92] M West. Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society (Ser B)*, 55:409–422, 1993.

[93] M West. Mixture models, Monte Carlo, Bayesian updating and dynamic models. In J H Newton, editor, *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface*, pages 325–333, Interface Foundation of North America, Fairfax Station, Virginia, 1993.

[94] M West and J Harrison. *Bayesian Forecasting and Dynamic Models (2nd edition)*. Springer-Verlag, New York, 1997.

[95] M West and P J Harrison. Monitoring and adaptation in Bayesian forecasting models. *Journal of the American Statistical Association*, 81:741–750, 1986.

[96] M West and P J Harrison. Subjective intervention in formal models. *Journal of Forecasting*, 8:33–53, 1989.

[97] V S Zariskii, V B Svetnik, and L I Shimelevich. Monte Carlo techniques in problems of optimal data processing. *Automatic Remo Control*, 12:95–103, 1975.

[98] A Zellner, C Hong, and C Min. Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. *Journal of Econometrics*, 49:275–304, 1991.

[99] A Zellner and R Rossi. Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, 25:365–394, 1984.

Biography

Fang Liu was born on September, 20, 1972 in Tiantsin, China. She received her B.S. degree in computer and system sciences from Nankai University in June, 1995 and M.S. in statistics from Duke University in May, 1997.