DISSERTATION

ESTIMATION FOR STATE-SPACE MODELS AND BAYESIAN REGRESSION ANALYSIS WITH PARAMETER CONSTRAINTS

Submitted by
Gabriel A. Rodriguez-Yam
Department of Statistics

In partial fulfillment of the requirements
For The Degree of Doctor of Philosophy
Colorado State University
Fort Collins, Colorado
Fall 2003

COLORADO STATE UNIVERSITY

November 4, 2003

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY GABRIEL A. RODRIGUEZ-YAM ENTITLED: ESTIMATION FOR STATE-SPACE MODELS AND BAYESIAN REGRESSION ANALYSIS WITH PARAMETER CONSTRAINTS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Adviser

Co-Adviser

Department Chair

ABSTRACT OF DISSERTATION

ESTIMATION FOR STATE-SPACE MODELS AND BAYESIAN REGRESSION ANALYSIS WITH PARAMETER CONSTRAINTS

In the first part of this dissertation an estimation procedure for non-Gaussian state-space models is proposed. Typically, the likelihood function for non-Gaussian state-space models can not be computed explicitly and so simulation based procedures, such as importance sampling or MCMC, are commonly used to estimate model parameters. In this dissertation, we consider an alternative estimation procedure which is based on an approximation to the likelihood function. The approximation can be computed and maximized directly, resulting in a quick estimation procedure without resorting to simulation. Moreover, this approach is competitive with estimates produced using simulation-based procedures. The speed of this procedure makes it viable to fit a wide range of potential models to the data and allows for bootstrapping the parameter estimates.

In the second part of this dissertation an efficient Gibbs sampler for simulation of a multivariate normal random vector subject to inequality linear constraints is proposed. An application to a Bayesian linear model, where the regression parameters are subject to inequality linear constraints, is the primary motivation behind this research. Geweke (1991) and Robert (1995) have implemented the Gibbs sampler

to the multivariate normal distribution subject to inequality linear constraints while the multiple linear regression with inequality constraints are considered for example by Chen and Deeley (1996) and Geweke (1996). However, these implementations can often exhibit poor mixing and slow convergence. The Gibbs sampler developed in this dissertation overcomes these limitations. In addition, it allows for the number of constraints to exceed the vector size and is able to cope with equality linear constraints.

Gabriel A. Rodriguez-Yam
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Fall 2003

ACKNOWLEDGMENTS

This dissertation would not exist without the synchronized occurrence of a series of events. First of all, I am very much indebted to my adviser Richard Davis, for his invaluable help and support to this dissertation. I am indebted also to my co-adviser Louis Scharf, who brought the identification problem. Without their guidance it would not be possible to find the new solution to this problem. It is pleasing and rewarding to work with them.

I also would like to thank: Thomas Lee and Simon Tavener for serving on my committee, William Dunsmuir and Jay Breidt for their comments and helpful suggestions to the state-space modeling part of this research, Wolfgang Kober, from Data Fusion Corporation, who provided the identification problem and a grant during the year I spent in this part of my research, and Felix Gonzalez Cossio for serving on my master's committee.

I benefited from the learning atmosphere and courses from the Statistics Department. I would like especially to thank my professors Duane Boes, Peter Brockwell, Richard Davis, Hari Iyer and Mohammed Siddiqui.

I also wish to thank the staff in the Statistics Department for their support.

Finally, I gratefully acknowledge the scholarship from Consejo Nacional de Ciencia y Tecnología and the partial support of Universidad Autónoma Chapingo.

To María,
our son Gabriel,
and my parents

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

This dissertation studies parameter estimation in two nonstandard situations. The first deals with maximum likelihood estimation for non-Gaussian state-space models in which a closed form for the likelihood function does not exist. In the second problem, estimation for multiple linear regression with constraints on the regression parameters is considered. In the first case, maximum likelihood estimates are difficult to compute, while in the second case maximum likelihood estimates are computable but no longer enjoy the optimal properties associated with MLE. To address these problems, we will use an approximate likelihood approach for estimation in the first case and a Bayesian approach in the second problem.

We will consider non-Gaussian state-space models (SSM) for which the distribution of the t-th component of the time series of observations $Y_1, Y_2, \ldots,$ is assumed to be of the form,

$$p(y_t|\alpha_t, \alpha_{t-1}, \ldots, \alpha_1, y_{t-1}, \ldots, y_1; \boldsymbol{\theta}) = p(y_t|\alpha_t; \boldsymbol{\theta}).$$

Here, $\boldsymbol{\theta}$ is a vector of parameters and the "state process" $\{\alpha_t\}$ follows a stationary Gaussian autoregressive model of order p, i.e.,

$$\alpha_t = \gamma + \phi_1 \alpha_{t-1} + \ldots + \phi_p \alpha_{t-p} + \eta_t,$$

where p is an integer greater than zero and $\{\eta_t\} \sim$ iid $N(0, \sigma^2)$, $t = 0, \pm 1, \pm 2, \ldots$.

As an example of this setup, consider the time series shown in Figure 1.1 consisting of the daily counts of asthma presentations from January 1, 1990-December 31,1993 from a single hospital (Campbelltown) in a suburb of Sydney, Australia. Here, it might be plausible to model the counts Y_t by a Poisson distribution with rate $\lambda_t := e^{\alpha_t + \mathbf{x}_t^T \beta}$, where $\mathbf{x}_t$ is a vector of covariates observed at time t, β is the vector of regression coefficients, and $\{\alpha_t\}$ is the latent or "state" process. Models of this type have been used frequently for modeling counts of individuals infected by a rare disease, e.g., Harvey and Fernandes (1989); Chan and Ledolter (1995); Davis et al. (1998).

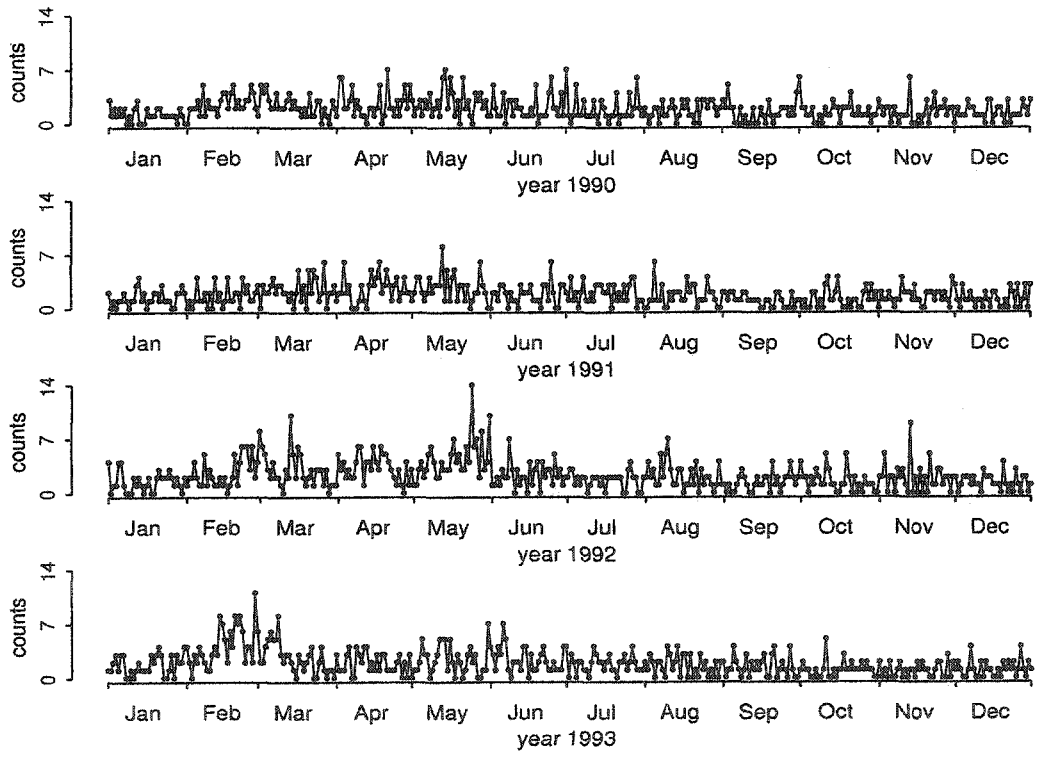

Figure 1.1: Asthma presentations at a Sydney hospital.

If $\psi := (\theta, \gamma, \phi_1, \ldots, \phi_p, \sigma^2)$ denotes the vector of parameters of the SSM, then the

likelihood based on the observations $y_1, \ldots, y_n$, becomes the n-fold integral

$$L(\psi; \mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda}) d\boldsymbol{\alpha}, \tag{1.1}$$

where $\mathbf{y} := (y_1, \ldots, y_n)$, $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n)$, and $\boldsymbol{\lambda} := (\gamma, \phi_1, \ldots, \phi_p, \sigma^2)$. Except for simple cases, this integral can not be computed explicitly and hence maximizing this function with respect to ψ is problematic.

Let $p_a(\boldsymbol{\alpha}|\mathbf{y}; \psi)$ be an approximation to the posterior distribution of the state vector $\boldsymbol{\alpha}$, then from (1.1)

$$L(\psi; \mathbf{y}) = \int \frac{p(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda})}{p_a(\boldsymbol{\alpha}|\mathbf{y}; \psi)} p_a(\boldsymbol{\alpha}|\mathbf{y}; \psi) d\boldsymbol{\alpha}. \tag{1.2}$$

Hence, if an iid sample $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(N)}$ can be "easily" drawn from $p_a(\boldsymbol{\alpha}|\mathbf{y}; \psi)$, by the method of *importance sampling* (Ripley, 1987, pages 122-123),

$$\hat{L}(\psi; \mathbf{y}) := \frac{1}{N} \sum_{i=1}^{N} \frac{p(\mathbf{y}|\boldsymbol{\alpha}^{(i)}; \boldsymbol{\theta}) p(\boldsymbol{\alpha}^{(i)}|\boldsymbol{\lambda})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y}; \psi)}, \tag{1.3}$$

is an unbiased estimator of $L(\psi; \mathbf{y})$. Also, by the strong law of large numbers, as $N \to \infty$,

$$\hat{L}(\psi; \mathbf{y}) \overset{a.s.}{\to} L(\psi; \mathbf{y}).$$

Once an importance density is available, an approximate MLE can be obtained by maximizing (1.3). However, because the objective function is random, so is the approximation to the MLE that results. The "accuracy" of the approximation depends on the "quality" of the importance function and on the number N of draws from this function on which the objective function is based. Here, the term accuracy refers to the closeness of the approximate MLE to the actual MLE, which can not be computed. For moderate n, large values of N can be used to obtain an accurate estimate. However, for n large, the accuracy of the estimate and the speed of the procedure are in conflict. For N large, the estimate is accurate, but at the expense

of increased computation time. A small value of N will speed up the process at the expense of decreased accuracy.

To construct an importance function, Durbin and Koopman (1997) assume that the observations come from a classical linear Gaussian SSM. This is the case when $p(y_t|\alpha_t;\boldsymbol{\theta})$ is assumed Gaussian. Let $g(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\theta})$ be the joint distribution of the vector of observations relative to this "working model". The parameters of the working model are found by making $g(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\theta})$ as close as possible to $p(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\theta})$, in a neighborhood of the smoothed state vector. The posterior distribution $g(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ of the state vector of the fitted working model is then considered an importance density by Durbin and Koopman (1997).

While the formulation of the importance density of the working model in which this importance function $g(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$ is based is easy when $p(y_t|\alpha_t;\boldsymbol{\theta})$ is a member of the "standard" exponential family of distributions, it can be tedious and difficult for other cases. One such "nonstandard" example is the stochastic volatility model, in which Sandmann and Koopman (1998) implement this method to find an approximate MLE of the parameters of this model. Their working model is based on the log of the squared observations. Since an observation with value zero can not be ruled out, this tranformation may in fact, cause problems.

Other related simulation-based procedures include the Monte Carlo Newton-Raphson and Monte Carlo EM algorithms. In the latter, the states are considered missing observations, so that in the E-step, the conditional expectation of the logarithm of the complete likelihood

$$\int \log L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}^{j-1})d\boldsymbol{\alpha},$$

which is again an n-fold integral, is required. Here, $L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) = p(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta})p(\boldsymbol{\alpha}|\boldsymbol{\lambda})$ and $\boldsymbol{\psi}^{j-1}$ is the value of the vector of parameters at the preceding iteration step.

As with the likelihood, this integral can not be computed explicitly except in simple cases. Chan and Ledolter (1995) use Monte Carlo approximation of the integral using a sample from the posterior distribution of the vector of states obtained via the Gibbs sampler. In the M-step, the approximation is optimized with respect to the vector of parameters. Thus, the iterative solution becomes random. This makes a stopping criterion difficult to implement. In addition, the need to draw a sample from the exact posterior at each iteration makes the procedure difficult and slow to implement and run.

In the Monte Carlo Newton-Raphson method, the first and second order derivatives required in the Newton-Rapshon iterations are n-fold integrals. Using a sample from the posterior distribution of the state vector, Kuk and Cheng (1997) approximate these integrals using Monte Carlo integration. Once again, the iterative solution ψ^j is random. Thus, the difficulties encountered in the Monte Carlo EM algorithm appear in this procedure also.

In this dissertation, an alternative to the Durbin and Koopman (1997) importance sampling procedure is proposed. Unlike their method, this new procedure is not based on any working model, which makes it easier to implement in the case when the distribution of the observations is not a member of the standard exponential family of distributions.

Also, a generalization of the estimation procedure given by Davis, et al. (1998) for modeling time series of counts, based on an analytical approximation to the likelihood function is considered. The approximation can be computed and maximized directly without resorting to simulation. For the values of N used in this dissertation, this approach is approximately 100 times faster than the importance sampling approach in (1.3), yet it provides competitive results. The speed of this procedure makes it

viable to fit a wide range of potential models to the data and allows for bootstrapping the parameter estimates.

The likelihood in (1.1) can be written as

$$L(\psi; \mathrm{y}) = p(y_1|\psi) \prod_{t=2}^{n} p(y_t|y_{1:t-1}; \psi), \tag{1.4}$$

where $y_{1:t-1} := (y_1, \ldots, y_{t-1})$. This representation has traditionally been used to estimate the likelihood of a non-Gaussian SSM, e.g., Kitagawa (1987); Hodges and Hale (1993); Hurzeler (1998); Pitt and Shepard (1999). In this approach, $p(y_1|\psi)$ and $p(y_t|y_{1:t-1}; \psi)$, $t \geq 2$ are estimated either by numerical integration or Monte Carlo simulation. The latter can be implemented via particle filtering procedures, suggested independently by various authors in the early 90's.

In this dissertation, the n factors of the likelihood in (1.4) are estimated using three particle filtering implementations. The estimates of the likelihood so obtained are compared with the importance sampling and analytical approximation to the likelihood proposed in this dissertation. Of special interest is a comparison of the speed of the procedures.

The second class of models considered in this dissertation is the multiple linear regression in which the regression parameters are subject to linear constraints of inequality and equality. The motivation behind this part of the dissertation was an identification problem in hyperspectral imaging, which consists in the analysis of a *mixing linear model.* In this model, the spectrum y of a mixed pixel is represented as a linear combination of component spectra, i.e.,

$$\mathrm{y} = \mathbf{X}\beta + \epsilon, \tag{1.5}$$

where the columns of the full rank matrix $\mathbf{X}$ contain the spectra of the k materials in a pixel, β is a vector consisting of the "abundances" of the materials in the pixel,

and $\epsilon \sim N(0, \sigma^2 \mathbf{I})$ is the noise of the model (see Manolakis and Shaw, 2002). Due to physical considerations, the abundance parameters are considered to be non-negative, i.e., $\boldsymbol{\beta} \geq 0$ and satisfy the sum-to-one constraint $\beta_1 + \ldots + \beta_k = 1$.

The mixing linear model fits into a more general framework, where the vector of regression coefficients $\boldsymbol{\beta}$ from the multiple linear regression in (1.5) is subject to a set of linear constraints given by

$$\mathbf{B}\boldsymbol{\beta} \leq \mathbf{b}, \quad \mathbf{C}\boldsymbol{\beta} = \mathbf{c},$$

where $\mathbf{B}$ and $\mathbf{C}$ are known matrices and $\mathbf{b}$ and $\mathbf{c}$ are known vectors.

In multiple regression analysis, a relationship of a response variable against a set of predictor variables is studied. In the classical formulation given in (1.5), the estimate of parameters is based on maximum likelihood estimation theory. Unlike the non-Gaussian SSM, the MLE estimates with constraints on the parameters, can be found in closed form. However, the optimal MLE properties that the unconstrained parameters possess are no longer valid for the constrained case.

Judge and Takayama (1966) and Liew (1976) give the inequality constrained least-squares (ICLS) estimate of $\boldsymbol{\beta}$ using the Dantzig-Cottle algorithm. The ICLS estimator reduces to the ordinary least squares estimator for a sufficiently large sample. Conditioning on knowledge of which constraints are binding and which are not, they compute an untruncated covariance matrix of the ICLS estimator. Geweke, (1986) points out that this variance matrix is incorrect, since in practice it is not known ahead of time which constraints will be binding. Thus, inferences based on this matrix can be seriously misleading (Lovell and Prescott, 1970).

The case when the vector of regression coefficients $\boldsymbol{\beta}$ from the multiple linear

regression in (1.5) is subject to a set of inequality linear constraints given by

$$\mathrm{B}\beta \leq \mathrm{b}, \tag{1.6}$$

has been analyzed from the Bayesian perspective. Geweke (1986) uses a prior that is the product of a conventional uninformative distribution and an indicator function representing the inequality constraints. The posterior distribution and expected values of functions of interest are then computed using importance sampling. In this case, an importance function is easy to find due to the simplicity of the prior. This method can be extremely slow especially when the truncation region has a small probability with respect to the unconstrained Gaussian distribution.

Gelfand et al. (1992) suggest a routine approach to analyze problems with constrained parameters using the Gibbs sampler. Geweke (1996) applies this procedure to the problem of multiple linear regression when the inequality linear constraints in (1.6) are linearly independent. However, this implementation may suffer from poor mixing (i.e., the chain does not move rapidly through the "entire" support of the posterior distribution). Due to the requirement of independent constraints, the number of constraints can not exceed the number of parameters. Also, equality linear constraints are not considered.

In Rodriguez-Yam et al. (2002), a Gibbs sampler implementation with good mixing is provided for the mixing linear model when only the non-negativity constraints on the abundance parameters are considered. For this case, the constraints are linearly independent and the number of inequality linear constraints coincides with the number of regression coefficients.

In this dissertation a new implementation of the Gibbs sampler for this constrained regression problem is proposed. In this implementation,

- the inequality linear constraints can be linearly dependent,
- more constraints than number of parameters can be handled, and
- equality linear constraints can be included.

Furthermore, this implementation has faster mixing, requiring substantially fewer iterations of the Markov chain than previously published Gibbs Sampler implementations.

The organization of the remaining chapters of this dissertation is as follows. In Chapter 2, a formulation of the state space model considered in this dissertation is given. An estimator based on an analytical approximation to the likelihood is provided and the performance of this estimate is compared with the importance sampling estimate of Durbin and Koopman (1997) via simulation studies based on Poisson and stochastic volatility models. Also, bootstrap bias corrections of these estimates are provided in the analysis of the polio and Pound-Dollar exchange rates datasets. The chapter concludes with two numerical examples that address the quality of the approximation to the posterior distribution of the state vector.

In Chapter 3, an introduction to particle filtering is provided and three particle filtering implementations are given. Estimation of the likelihood of the SSM based on these implementations are compared with the importance sampling and analytical approximation of the likelihood from Chapter 2, giving special attention to the computation time of the procedures.

In Chapter 4, a Bayesian framework is given for multiple linear regression when the regression parameters are subject to linear inequality and equality constraints. An efficient Gibbs sampler from the truncated multivariate normal distribution is provided. This implementation is then used to obtain an efficient Gibbs sampler for

the constrained multiple linear regression model. The procedure is implemented in two numerical examples, one containing only inequality linear constraints and the second containing inequality and equality linear constraints. In the latter, the Gibbs sampler is implemented in a model containing more inequality linear constraints than number of regression coefficients.

In the appendix, the innovations algorithm (Brockwell and Davis, 1991) is applied to compute the approximations to the likelihood and posterior distribution of the state vector given in Chapters 2 and 3. This method is applied to an example in which the observations are Poisson distributed.

CHAPTER 2

State Space Models

The class of state-space models (SSM) provides a flexible framework for modeling and describing a wide range of time series in a variety of disciplines. The books by Harvey (1989) and Durbin and Koopman (2001) contain extensive accounts of state-space models and their applications. One of the attractive features of state-space models is that many traditional models, such as ARMA and ARIMA, can be expressed in a linear state-space system. For linear and/or Gaussian state-space models, the Kalman filter can be used to compute predictors of the state-variables and one-step-ahead predictors of the observations. This allows for straightforward calculation of the likelihood in the Gaussian case. However, in many applications in which the Gaussian assumption is not realistic, the likelihood function is difficult to calculate, which makes maximum likelihood estimation problematic.

The state-space model that we consider in this dissertation has the following formulation: If $Y_1, Y_2, \ldots,$ represent the time series of observations and $\alpha_1, \alpha_2, \ldots$ the respective "state variables", then it is assumed that

$$p(y_t|\alpha_t, \alpha_{t-1}, \ldots, \alpha_1, y_{t-1}, \ldots, y_1) = p(y_t|\alpha_t) \tag{2.1}$$

belongs to a known parametric family of distributions. In addition, the state process

is assumed to follow a stationary Gaussian autoregressive model of order p, given by

$$\alpha_t = \gamma + \phi_1 \alpha_{t-1} + \ldots + \phi_p \alpha_{t-p} + \eta_t, \tag{2.2}$$

where p is an integer greater than zero and $\eta_t \sim$ iid $N(0,\ \sigma^2)$, $t = 0, \pm 1, \pm 2, \ldots$. Perhaps the most important special case is when the conditional distribution in (2.1) is a member of the exponential family, an extremely rich class of distributions. Durbin and Koopman (1997) and Kuk (1999), consider the following form for this family

$$p(y_t|\alpha_t) = e^{(\mathbf{x}_t^T \boldsymbol{\beta} + \alpha_t) y_t - b(\mathbf{x}_t^T \boldsymbol{\beta} + \alpha_t) + c(y_t)}, \tag{2.3}$$

where $\mathbf{x}_t$ is a vector of covariates observed at time t; $\boldsymbol{\beta}$ is a vector of parameters; and $b(.)$ and $c(.)$ are known real functions.

One special application that we will consider in more detail, is the case in which the time series $Y_1, \ldots, Y_n$ consist of counts. Here, it might be plausible to model Y_t by a Poisson distribution with rate $\lambda_t := e^{\alpha_t + \mathbf{x}_t^T \boldsymbol{\beta}}$ in which case, $p(y_t|\alpha_t)$ is a particular case of (2.3). Models of this type have been used for modeling counts of individuals infected by a rare disease, e.g., Zeger (1988); Harvey and Fernandes (1989); Campbell (1994); Chan and Ledolter (1995); Davis et al. (1998).

Another noteworthy application of the SSM that we will consider, is the stochastic volatility model (SVM), a frequently used model for returns of financial assets. In the basic SVM, the distribution of $Y_t|\alpha_t$ is Gaussian with mean 0 and variance e^{α_t}. Applications, together with estimation for SVMs, can be found in Jacquier, et al. (1994); Briedt and Carriquiry (1996); Harvey and Streibel (1998); Sandmann and Koopman (1998); Geweke and Tanizaki (1999); Pitt and Shepard (1999).

Let $\mathbf{y} := (y_1, \ldots, y_n)$ denote the vector of observations, $\boldsymbol{\alpha} := (\alpha_1, \ldots,\ \alpha_n)$ the vector of states and $\boldsymbol{\psi} := (\boldsymbol{\theta},\ \boldsymbol{\lambda})$ the parameters in the state-space model. Here $\boldsymbol{\theta}$

is the vector of the parameters associated with the specification of $p(y_t|\alpha_t)$, which may include the regression parameter β, and $\lambda := (\phi_1, \ldots, \phi_p, \gamma, \sigma^2)$ is the parameter vector associated with the AR model in (2.2). With this specification, the likelihood based on the "complete data" $(\mathbf{y}, \boldsymbol{\alpha})$ of the SSM becomes

$$\begin{aligned} L(\boldsymbol{\psi}; \mathbf{y}, \boldsymbol{\alpha}) &= p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda}) \\ &= \left(\prod_{t=1}^{n} p(y_t|\alpha_t, \boldsymbol{\theta}) \right) |\mathbf{V}|^{1/2} e^{-(\alpha-\mu)^T \mathbf{V}(\alpha-\mu)/2} / (2\pi)^{n/2}, \end{aligned} \tag{2.4}$$

where $\mathbf{V}^{-1} := \text{cov}\{\boldsymbol{\alpha}\}$, $\mu = \gamma/(1 - \phi_1 - \ldots - \phi_p)\mathbf{1}$ is the vector of means of the state process, and $\mathbf{1}$ is a vector of ones. It follows that the likelihood of the observed data is

$$L(\boldsymbol{\psi}; \mathbf{y}) = \int L(\boldsymbol{\psi}; \mathbf{y}, \boldsymbol{\alpha}) d\boldsymbol{\alpha}. \tag{2.5}$$

Except in simple cases, the integral in (2.5) can not be computed explicitly, which makes maximum likelihood estimation difficult. There are several simulation approaches in the literature for estimating and ultimately maximizing this likelihood. For example, Durbin and Koopman (1997, 2001) use importance sampling to estimate (2.5). The observation density $p(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta})$ is approximated by selecting a Gaussian density $g(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta})$ that best approximates $p(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta})$. The Monte Carlo integration is computed using $g(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$, the conditional density of $\boldsymbol{\alpha}$ relative to the working model, as the importance density. This approach is known as "many samples" because for distinct values of $\boldsymbol{\psi}$, the importance function $g(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$ is updated during the optimization of the approximate observed likelihood. To overcome the instability problem inherent with the "many samples" approach, Durbin and Koopman generate from the noise only once. Kuk (1999) advocates a "single-sample" approach, in which a sample is drawn from the importance density $g(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi}_0)$ for a fixed $\boldsymbol{\psi}_0$, and then the relative likelihood function is optimized using this sample. To get better

approximations of the relative likelihood near the true maximum likelihood estimate, Geyer (1996) suggests repeating the process several times, updating ψ_0 with the new maximizer at each iteration.

A Monte Carlo EM algorithm treating the unobserved α's as missing values was proposed by Chan and Ledolter (1995). At the i-th iteration of the algorithm, the M-step is performed by Monte Carlo integration drawing a sample from the conditional distribution $p(\alpha|\mathbf{y}, \psi^{(i-1)})$, where $\psi^{(i-1)}$ is the maximizer obtained in the previous iteration. Kuk and Cheng (1997) proposed a Monte Carlo implementation of the Newton-Raphson (MCNR) as a viable alternative to the MCEM algorithm. All of these simulation based procedures can be computationally demanding.

In this chapter we will follow a different approach to obtain an approximation to the distribution $p(\alpha|\mathbf{y};\psi)$. In Section 2.1 we will produce an analytical approximation to (2.5) by obtaining an approximation $p_a(\alpha|\mathbf{y};\psi)$ to the posterior distribution $p(\alpha|\mathbf{y};\psi)$. The innovations algorithm (Brockwell and Davis, 1991) can be used to speed up the computation of these approximations. The approximation to the observed likelihood can then be maximized to produce an estimate of ψ. In Section 2.2 we demonstrate the good performance of this procedure via simulation studies. This procedure will also be applied to analyze two datasets: the monthly number of U.S. cases of poliomyelitis for 1970 to 1983 (Zeger, 1988) is analyzed using a Poisson state-space model and a historical pound to dollar exchange rates (Harvey, et al., 1994) is analyzed using a stochastic volatility model.

The quality of our approximation depends, to a large extent, on the normal approximation to the posterior, $p(\alpha|\mathbf{y};\psi)$. In a numerical example we assess this approximation in two ways. First, we notice the closeness between the posterior mode and posterior mean of $p(\alpha|\mathbf{y};\psi)$. As a second check of closeness we com-

pare samples generated from $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ using sampling importance resampling (SIR) with the approximating normal distribution via a Chi-squared QQ-plot and a correlation test. These topics, together with bootstrap bias correction are considered in Sub-section 2.2.6. In Section 2.3 we summarize our findings. Application of the innovations algorithm to the problems considered in Sections 2.1 and 2.2 is given in the appendix.

2.1 Parameter Estimation

In this section we find an approximation to the observed likelihood $L(\boldsymbol{\psi};\mathbf{y})$ given in (2.5) that is based on an approximation $L_a(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})$ to the likelihood $L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})$ using the complete data. For the latter, a Taylor series expansion of $\log p(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta})$ in a neighborhood of the posterior mode of $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ is used.

To begin, let $\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha}) := \log p(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta})$. Note that the log of the observed likelihood is given by

$$\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{V}| + \ell(\boldsymbol{\psi};\mathbf{y}|\boldsymbol{\alpha}) - \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}-\boldsymbol{\mu}). \quad (2.6)$$

Now, let

$$\mathbf{k}^* := \frac{\partial}{\partial\boldsymbol{\alpha}}\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}$$

where $\boldsymbol{\alpha}^*$ is the mode of $\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})$, which solves $\frac{\partial}{\partial\boldsymbol{\alpha}}\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) = \mathbf{0}$. From (2.4), it follows that

$$\mathbf{k}^* = \mathbf{V}(\boldsymbol{\alpha}^* - \boldsymbol{\mu}).$$

Hence, if $T(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)$ denotes the second order Taylor expansion of $\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})$ around $\boldsymbol{\alpha}^*$ and $R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)$ the corresponding remainder, i.e.,

$$R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*) := \ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha}) - T(\boldsymbol{\alpha};\boldsymbol{\alpha}^*),$$

then

$$\begin{aligned}\ell(\boldsymbol{\psi};\mathbf{y}|\boldsymbol{\alpha}) &= T(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)+R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*),\\ &= h^*+(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\mathbf{k}^*-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\mathbf{K}^*(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)+R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*),\\ &= h^*+(\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\mathbf{K}^*(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \qquad (2.7)\\ &\quad +R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*),\end{aligned}$$

where

$$h^* := \ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*} \text{ and } \mathbf{K}^* := -\frac{\partial^2}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}. \qquad (2.8)$$

Thus, substituting (2.7) in (2.6), it follows that

$$\begin{aligned}\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) &= -\frac{n}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{V}|+h^*-\frac{1}{2}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}^*-\boldsymbol{\mu}) \qquad (2.9)\\ &\quad -\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T(\mathbf{K}^*+\mathbf{V})(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)+R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*).\end{aligned}$$

We note that the posterior $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ satisfies $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) \propto L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})$. Let $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ be the posterior based on the log likelihood given in (2.9) when the term $R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)$ is omitted, it follows that

$$p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) = \boldsymbol{\phi}(\boldsymbol{\alpha};\boldsymbol{\alpha}^*,(\mathbf{K}^*+\mathbf{V})^{-1}), \qquad (2.10)$$

where $\boldsymbol{\phi}(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Hence

$$\begin{aligned}L(\boldsymbol{\psi};\mathbf{y}) &= \frac{|\mathbf{V}|^{1/2}}{|\mathbf{K}^*+\mathbf{V}|^{1/2}}e^{h^*-\frac{1}{2}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})}\int e^{R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)}p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})d\boldsymbol{\alpha}, \qquad (2.11)\\ &= L_a(\boldsymbol{\psi};\mathbf{y})\mathrm{Er}_a(\boldsymbol{\psi}),\end{aligned}$$

where $L_a(\boldsymbol{\psi};\mathbf{y})$ is the approximation to $L(\boldsymbol{\psi};\mathbf{y})$

$$L_a(\boldsymbol{\psi};\mathbf{y}) := \frac{|\mathbf{V}|^{1/2}}{|\mathbf{K}^*+\mathbf{V}|^{1/2}}e^{h^*-\frac{1}{2}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})}, \qquad (2.12)$$

that is obtained when the factor $e^{R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)}$ is ignored in the integral in (2.11); and $\mathrm{Er}_a(\boldsymbol{\psi})$ is the *approximation error*

$$\mathrm{Er}_a(\boldsymbol{\psi}) := \int e^{R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)} p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) d\boldsymbol{\alpha}. \tag{2.13}$$

Thus, if $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ is highly concentrated around $\boldsymbol{\alpha}^*$, the integral in (2.13) should be close to 1.

Since the evaluation of (2.12) does not involve simulation, it can be maximized to obtain an approximate MLE of $\boldsymbol{\psi}$. In fact, we will see later that both the computation of $\boldsymbol{\alpha}^*$ and the evaluation of (2.12) can be accelerated with the aid of the innovations algorithm (Brockwell and Davis, 1991).

A second way to motivate our approximation $L_a(\boldsymbol{\psi};\mathbf{y})$ is based on a Bayesian viewpoint. If we treat $\boldsymbol{\alpha}$ as the parameters of the system with prior $p(\boldsymbol{\alpha}|\boldsymbol{\lambda})$, then under regularity conditions and a fixed number of parameters, the posterior $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ can be approximated by a normal density function for n large (e.g., Bernardo and Smith, 1994; page 287). This normal density matches the mode of the posterior $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ and has covariance matrix equal to the inverse of the information matrix of the posterior evaluated at the posterior's mode. Notice that $\boldsymbol{\alpha}^*$ is the mode of the posterior $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ and the observed information matrix is given by $\mathbf{K}^* + \mathbf{V}$. Both assertions can be obtained from the fact that $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) \propto L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})$. Thus, in this context, the normal approximation is, in fact, the same as $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ given in (2.10).

We note that

$$\begin{aligned} L(\boldsymbol{\psi};\mathbf{y}) &= \int p(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda}) d\boldsymbol{\alpha}, \\ &= \int \frac{p(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda})}{p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})} p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) d\boldsymbol{\alpha}. \end{aligned} \tag{2.14}$$

So, $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ in (2.10) can be viewed as an *importance density.*

Now, we provide a recursive algorithm to find $\boldsymbol{\alpha}^*$, the mode of $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$. Let $\boldsymbol{\alpha}^j$ be the current iterate to the value of $\boldsymbol{\alpha}^*$. If

$$\mathbf{k}^j := \frac{\partial}{\partial \boldsymbol{\alpha}}\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\alpha=\alpha^j} \quad \text{and} \quad \mathbf{K}^j := -\frac{\partial^2}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\alpha=\alpha^j}, \tag{2.15}$$

then the Newton-Raphson algorithm gives

$$\boldsymbol{\alpha}^{j+1} = \boldsymbol{\alpha}^j - (\ddot{\ell}^j)^{-1}\dot{\ell}^j, \tag{2.16}$$

where

$$\begin{aligned}
\dot{\ell}^j &:= \frac{\partial}{\partial \boldsymbol{\alpha}}\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})|_{\alpha=\alpha^j} \\
&= \mathbf{k}^j - \mathbf{V}(\boldsymbol{\alpha}^j - \boldsymbol{\mu}) \\
&= \mathbf{k}^j + \mathbf{K}^j\boldsymbol{\alpha}^j + \mathbf{V}\boldsymbol{\mu} - (\mathbf{K}^j + \mathbf{V})\boldsymbol{\alpha}^j, \qquad (2.17) \\
\ddot{\ell}^j &:= \left(\frac{\partial^2}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})\right)^{-1}|_{\alpha=\alpha^j} \\
&= -\mathbf{K}^j - \mathbf{V}. \qquad (2.18)
\end{aligned}$$

Let

$$\tilde{\mathbf{y}}^j := \mathbf{k}^j + \mathbf{K}^j\boldsymbol{\alpha}^j + \mathbf{V}\boldsymbol{\mu}. \tag{2.19}$$

Substituting this, (2.17) and (2.18) into (2.16), we obtain

$$\boldsymbol{\alpha}^{j+1} = (\mathbf{K}^j + \mathbf{V})^{-1}\tilde{\mathbf{y}}^j. \tag{2.20}$$

Application to the exponential family

Assume that the observation density function is from the exponential family given by

$$p(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta}) = \prod_{t=1}^{n} p(y_t|\alpha_t,\boldsymbol{\theta}) = e^{(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha})^T\mathbf{y} - \mathbf{1}^T\{\mathbf{b}(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha}) - \mathbf{c}(\mathbf{y})\}}, \tag{2.21}$$

where $\mathbf{b}(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha}) := [b(\mathbf{x}_1^T\boldsymbol{\beta}+\alpha_1),\ldots,b(\mathbf{x}_n^T\boldsymbol{\beta}+\alpha_n)]^T$ and $\mathbf{c}(\mathbf{y}) := [c(y_1),\ldots,c(y_n)]^T$. In this setting, the matrix $\mathbf{K}^*$ in (2.8) becomes

$$\mathbf{K}^* = \text{diag}\{\frac{\partial^2}{\partial \alpha_t^2}b(\mathbf{x}_t^T\boldsymbol{\beta}+\alpha_t)|_{\alpha_t^*}\}. \tag{2.22}$$

The approximation to the observed likelihood is then

$$L_a(\psi;\mathbf{y}) = \frac{|\mathbf{V}|^{1/2}}{|\mathbf{K}^* + \mathbf{V}|^{1/2}} e^{\mathbf{y}^T(\mathbf{x}\beta+\alpha^*)-\mathbf{1}^T\{\mathbf{b}(\mathbf{x}\beta+\alpha^*)-\mathbf{c}(\mathbf{y})\}-(\alpha^*-\mu)^T\mathbf{V}(\alpha^*-\mu)/2}. \tag{2.23}$$

From (2.15) and (2.21), $\mathbf{k}^j = \mathbf{y} - \dot{\mathbf{b}}^j$, where

$$\dot{\mathbf{b}}^j := \frac{\partial}{\partial\alpha}\mathbf{1}^T\mathbf{b}(\mathbf{x}\beta+\alpha)|_{\alpha^j}. \tag{2.24}$$

Hence,

$$\tilde{\mathbf{y}}^j := \mathbf{y} - \dot{\mathbf{b}}^j + \mathbf{K}^j\alpha^j + \mathbf{V}\mu, \tag{2.25}$$

where $\mathbf{K}^j$ is defined in (2.15). □

Although at this point we can find an approximation to the likelihood, each iteration of (2.20) requires the inversion of a matrix of dimension $n \times n$, while each evaluation of (2.12) requires calculation of the determinant of a matrix of similar dimension. For small values of n, these computations can be carried out directly, but for large values, direct computations are impractical. Recursive prediction algorithms, such as the Kalman recursions or the innovations algorithm accelerate these calculations. Here we use the innovations algorithm, which seems to be ideally suited for this problem. The implementation of the innovation algorithm in this context is described in the Appendix.

As we noted from (2.14), $p_a(\alpha|\mathbf{y};\psi)$ in (2.10) can be used as an importance density. In fact, as we show below for the case of the exponential family of distributions, $p_a(\alpha|\mathbf{y};\psi)$ coincides with the importance density function of Durbin and Koopman (1997) to estimate the likelihood in (2.5) via simulation. In order to describe their method, let $L_g(\psi;\mathbf{y})$ denote the likelihood of the Gaussian approximating model of the state-space model proposed by Durbin and Koopman (1997). Such an approximation is obtained when $p(y_t|\alpha_t;\theta)$ is replaced by a Gaussian distribution

$g(y_t|\alpha_t;\boldsymbol{\theta}) = \phi(y_t;\alpha_t+\mu_t,H_t)$, where μ_t and H_t are found by solving iteratively

$$\frac{\partial}{\partial\alpha_t}\log p(y_t|\alpha_t;\boldsymbol{\theta})|_{\alpha_t=\hat{\alpha}_t} - H_t^{-1}(y_t-\hat{\alpha}_t-\mu_t) = 0 \tag{2.26}$$

$$\frac{\partial^2}{\partial\alpha_t^2}\log p(y_t|\alpha_t;\boldsymbol{\psi})|_{\alpha_t=\hat{\alpha}_t} + H_t^{-1} = 0. \tag{2.27}$$

Here, the $\hat{\alpha}_t$ are found by routine application of the Kalman filtering and smoothing algorithms. The iterations, initialized with $\mu_t = 0$ and H_t arbitrary, must be stopped until convergence of μ_t and H_t. Let E_g denote the conditional expectation operator under the approximating model. Durbin and Koopman (1997) found that the likelihood (2.5) can be expressed as

$$L(\boldsymbol{\psi};\mathbf{y}) = L_g(\boldsymbol{\psi};\mathbf{y})E_g\left\{\frac{p(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\theta})}{g(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\psi})}|\mathbf{y},\boldsymbol{\psi}\right\}. \tag{2.28}$$

Hence, with simulated values $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(N)}$ from the conditional density $g(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ under the approximating model, the integral in (2.5) is estimated as

$$\hat{L}(\boldsymbol{\psi};\mathbf{y}) = L_g(\boldsymbol{\psi};\mathbf{y})\frac{1}{N}\sum_{i=1}^{N}\frac{p(\mathbf{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\theta})}{g(\mathbf{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi})}. \tag{2.29}$$

This method is called a "many samples" approach, because new simulated values of the $\boldsymbol{\alpha}^{(i)}$'s are needed for each value of $\boldsymbol{\psi}$. To ensure stability in their numerical process, they generate from the noise only once.

Alternatively, Kuk (1999) proposes using the relative likelihood

$$\frac{L(\boldsymbol{\psi};\mathbf{y})}{L_g(\boldsymbol{\psi}_0;\mathbf{y})} = E_g\left\{\frac{p(\mathbf{y},\boldsymbol{\alpha}|\boldsymbol{\psi})}{g(\mathbf{y},\boldsymbol{\alpha}|\boldsymbol{\psi}_0)}|\mathbf{y},\boldsymbol{\psi}_0\right\}, \tag{2.30}$$

where the conditional expectation is computed relative to the conditional density $g(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}_0)$ under the approximating model. Using simulated values $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(N)}$ from the conditional density $g(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}_0)$, an estimate of $L(\boldsymbol{\psi};\mathbf{y})$ using (2.30) is

$$\hat{L}(\boldsymbol{\psi};\mathbf{y}) = L_g(\boldsymbol{\psi}_0;\mathbf{y})\frac{1}{N}\sum_{i=1}^{N}\frac{p(\mathbf{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\theta})p(\boldsymbol{\alpha}^{(i)}|\boldsymbol{\lambda})}{g(\mathbf{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\theta}_0)p(\boldsymbol{\alpha}^{(i)}|\boldsymbol{\lambda}_0)}. \tag{2.31}$$

This approach is known as a "single-sample" procedure, since it involves simulating from $g(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi}_0)$ instead of $g(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$. In order for this method to work, a few updatings of $\boldsymbol{\psi}_0$ to the optimizer of $\hat{L}(\boldsymbol{\psi}; \mathbf{y})$ in (2.31) is recommended (Geyer, 1996; Kuk, 1999).

If $p(y_t|\alpha_t; \boldsymbol{\theta})$ is a member of the exponential family of distributions as given in (2.3), then using the notation $\dot{b}_t := \frac{\partial}{\partial \alpha_t} b(\mathbf{x}_t^T \boldsymbol{\beta} + \alpha_t)|_{\alpha_t = \hat{\alpha}_t}$ and $\ddot{b}_t := \frac{\partial^2}{\partial \alpha_t^2} b(\mathbf{x}_t^T \boldsymbol{\beta} + \alpha_t)|_{\alpha_t = \hat{\alpha}_t}$, Durbin and Koopman (1997) find that

$$H_t^{-1} = \ddot{b}_t, \qquad \mu_t = y_t - \hat{\alpha}_t - \ddot{b}_t^{-1}(y_t - \dot{b}_t). \tag{2.32}$$

They comment that $\hat{\boldsymbol{\alpha}} := [\hat{\alpha}_1, \ldots, \hat{\alpha}_n]^T$, obtained using the iterative procedure described above, is the posterior mode of $p(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$. We conclude that $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*$. Furthermore, from (2.32), it follows that the variance of the distribution $g(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$ computed under the approximating model until convergence is achieved is given by $(\mathbf{K}^* + \mathbf{V})^{-1}$, where $\mathbf{K}^*$ is given in (2.22). Thus, $p_a(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$ in (2.10) and $g(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi})$ are identical. Notice that $g(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta}) = \prod_{t=1}^n g(y_t|\alpha_t; \boldsymbol{\theta}) = \prod_{t=1}^n \phi(y_t; \alpha_t + \mu_t, H_t)$. From (2.32), it follows that

$$g(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta}) = (2\pi)^{-n/2} |\mathbf{K}^*|^{1/2} e^{-h^* - \frac{1}{2}\mathbf{k}^{*T}(\mathbf{K}^*)^{-1}\mathbf{k}^*} e^{T(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)},$$

where $\mathbf{k}^* = \mathbf{y} - \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{1}^T \mathbf{b}(\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\alpha})|_{\boldsymbol{\alpha}^*}$, and h^* and $\mathbf{K}^*$ are defined in (2.8). A similar procedure used to obtain (2.11) shows that $\int e^{T(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)} p(\boldsymbol{\alpha}|\boldsymbol{\lambda}) d\boldsymbol{\alpha} = L_a(\boldsymbol{\psi}; \mathbf{y})$. Hence, the observed likelihood $L_g(\boldsymbol{\psi}; \mathbf{y})$ of the Durbin and Koopman's approximate Gaussian model is given by

$$\begin{aligned} L_g(\boldsymbol{\psi}; \mathbf{y}) &= \int g(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda}) \\ &= (2\pi)^{-n/2} |\mathbf{K}^*|^{1/2} e^{-h^* + \mathbf{k}^{*T}(\mathbf{K}^*)^{-1}\mathbf{k}^*/2} L_a(\boldsymbol{\psi}; \mathbf{y}). \end{aligned}$$

To get a feel for how these two procedures perform, we consider the case when the observation density is Poisson with rate $\lambda_t = e^{0.7+\alpha_t}$ and the state process follows

the AR(1) model

$$\alpha_t = \phi\alpha_{t-1} + \eta_t, \tag{2.33}$$

where $\eta_t \sim$ iid $N(0, 0.3)$, $t = 1, \ldots, n = 200$. In this example, the state-space model has only one parameter, i.e., $\psi = \phi$. Using $\phi = 0.5$, one realization $y_1, \ldots, y_{200}$ from this process was generated. In Figure 2.1 we show two estimates of the observed likelihood of this process. In this figure, the solid line is the approximation to the

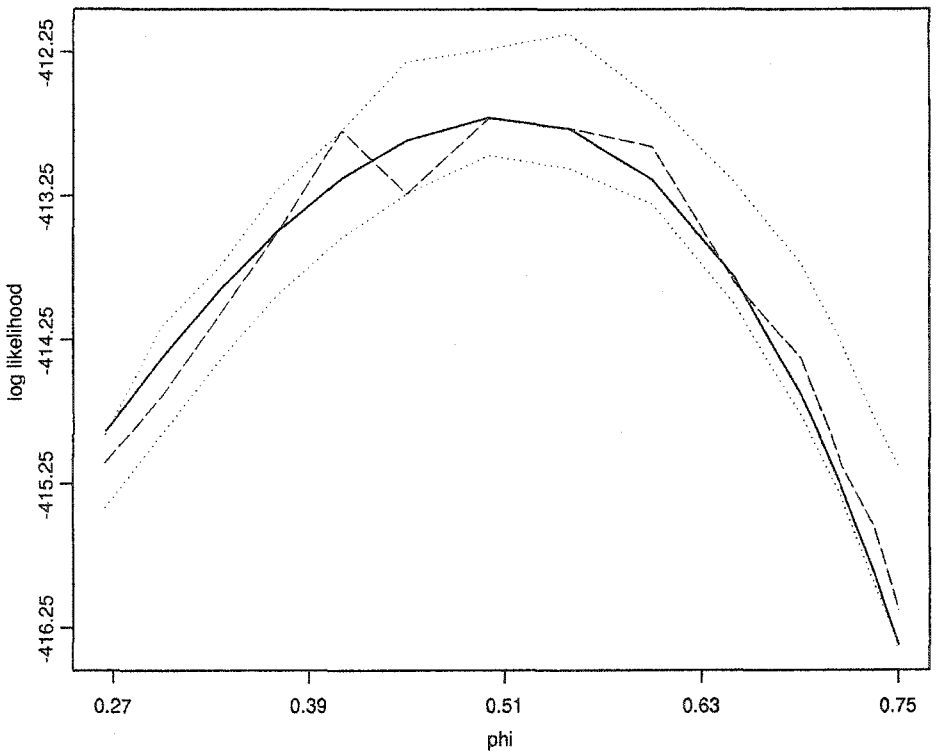

Figure 2.1: (*Many samples*) For a grid of values of ϕ, the logarithm of the estimation of the likelihood of a Poisson SSM are shown. For the solid line, estimates were obtained using (2.23) while for the dashed line, (2.29) was used. The dotted lines are the minimum and maximum respectively, of 100 replicates using (2.29).

observed likelihood given in (2.23). Also, the lower and upper dotted lines, computed for each value of ϕ in a grid of points, are the minimum and maximum, respectively, of 100 replicates of the estimated likelihood given in (2.29) using $N = 1000$. The dashed line is an estimation of the likelihood using (2.29) for one of these realizations.

The pair of dotted lines in this figure illustrate the randomness of the estimation of ϕ that is obtained by the maximization of (2.29). The shape of the dashed line in Figure 2.1, typical of the estimator in (2.29), comes from the "many samples" effect. The maximization of this random function to obtain an estimator of ϕ requires additional effort. In contrast, the approximation in (2.23) is smooth and can be computed much faster.

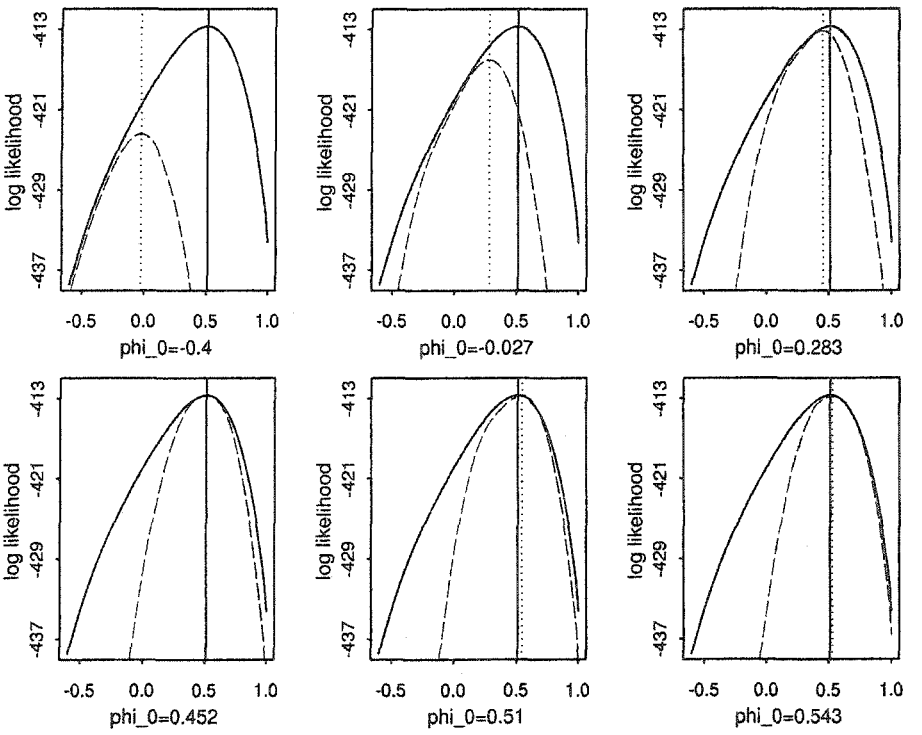

Figure 2.2: (*Single-sample*) From left to right and top to bottom, in each panel we show for a grid of values of ϕ, the logarithm of the estimation of the likelihood of a Poisson SSM. For the solid line, the estimates were obtained using (2.23) while for the dashed line (2.31) was used. ψ_0 is the optimizer (shown by the dotted vertical line) of (2.31) from the preceding panel. The solid vertical line shows the optimizer of (2.23).

To compute the estimator of the observed likelihood in (2.31) using the approach described by Kuk (1999), we need an initial value ψ_0 and update it to the maximizer of (2.29) a "few times". In Figure 2.2, we use an initial value of $\phi_0 = -0.4$, and perform six updatings of this parameter using $N = 1000$. In each panel of this

figure, the solid line is the approximation (2.23) of the observed likelihood and the solid vertical line shows the maximizer 0.5098 of this function, i.e., the (approximate) ML estimate of ϕ. In the upper left panel, the long dashed line is the estimation given in (2.31) of the observed likelihood, while the vertical dotted line shows its maximizer -0.027. This value is then used as ϕ_0 in the middle panel in the top row, and so on. As ϕ_0 is updated, the current maximizer of (2.31) moves "quickly" toward an estimate that is close to the true value. As expected, for given ϕ_0, (2.31) approximates well the observed likelihood only in a neighborhood of ϕ_0. Unlike the estimator in (2.29), the estimator in (2.31) is smooth, but there is a price to pay for this gain in terms of imposing a stopping rule. Note that in a vicinity of ϕ_0, the estimate in (2.31) is close to the approximate likelihood in (2.23).

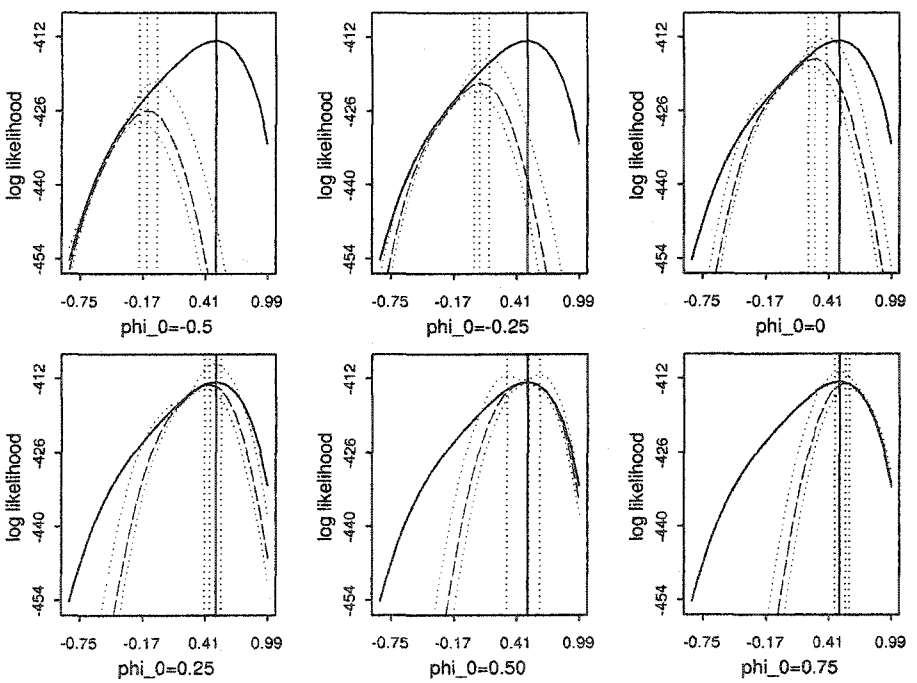

Figure 2.3: (*Single-sample*) From left to right and top to bottom, in each panel we show for a grid of values of ϕ, the logarithm of the estimation of the likelihood of a Poisson SSM. For the solid line, estimates were obtained using (2.23). The solid vertical line shows its optimizer. For the dashed line, estimations were obtained using (2.31) with ϕ_0 shown in the x axis. The dotted lines are the minimum and maximum respectively, of 100 replicates using (2.31). From left to right, the dotted vertical lines are the minimum, mean and maximum of the optimizers of these replicates.

In Figure 2.3, we show the randomness feature of (2.31). In each panel, a fixed value of ϕ_0 is used. The solid line and vertical solid line are as in Figure 2.2. The lower and upper dotted lines are the minimum and maximum, respectively, of one hundred replicates of (2.31) while from left to right, the dotted vertical lines are the minimum, mean and maximum of their optimizers. The long dashed line is one replicate of (2.31).

2.2 Numerical Results

In this section, we perform two simulation studies; one based on the basic stochastic volatility model and the second based on a Poisson observation density for modeling a time series of counts. Also, we analyze two real datasets. One is a historical dataset of the Pound-Dollar exchange rates, first studied by Harvey, et al. (1994) using a basic stochastic volatility model. The other is the polio incidence data analyzed by Zeger (1988) who used estimating equations to fit the model. Kuk and Cheng (1997) use the Monte Carlo Newton Raphson algorithm to analyze this data.

2.2.1 Stochastic Volatility Model

The stochastic volatility process that is often used for modeling log-returns of financial assets is defined by

$$y_t = \sigma_t \xi_t = e^{\alpha_t/2}\xi_t, \qquad \alpha_t = \gamma + \phi\alpha_{t-1} + \eta_t, \tag{2.34}$$

where $\xi_t \sim$ iid $N(0,1)$, $\eta_t \sim$ iid $N(0,\sigma^2)$, $t = 1,\ldots,n$, and $|\phi| < 1$. In this case, $\psi = (\gamma,\ \phi,\ \sigma^2)$. The format for this simulation study is the same as the layout considered in Jacquier, et al. (1994). They considered nine models, indexed by the coefficient of variation CV of the conditional variance $\sigma_t^2 := e^{\alpha_t}$. For convenience,

the parameters of these models are reproduced in Table 2.1. Jacquier, et al. (1994) point out that the nine models are calibrated so that $\mathrm{E}(\sigma_t^2) = 0.0009$. Also, from empirical studies (e.g., Harvey and Shepard, 1993; Jacquier, et al. 1994) values of ϕ between 0.9 and 0.98 are of primary interest.

		ϕ		
CV		0.90	0.95	0.98
10.0	γ	-0.821	-0.4106	-0.1642
	σ	0.6750	0.4835	0.308
1.0	γ	-0.736	-0.368	-0.1472
	σ	0.363	0.260	0.1657
0.1	γ	-0.706	-0.353	-0.1412
	σ	0.135	0.0964	0.0614

Table 2.1: Parameter values for a simulation experiment of nine stochastic volatility processes.

The density of the observed series is given by

$$p(y_t|\alpha_t;\boldsymbol{\psi}) = e^{-\{y_t^2 e^{-\alpha_t}+\alpha_t+\log(2\pi)\}/2},$$

which differs slightly from the standard representation of the exponential family of distributions given in (2.3). Equation (2.19) becomes

$$\tilde{\mathbf{y}}^j = \mathrm{diag}\{\mathbf{1}/2 + \boldsymbol{\alpha}^j/2\}\mathrm{diag}\{\mathbf{y}^2\}e^{-\boldsymbol{\alpha}^j} - \mathbf{1}/2 + \mathbf{V}\boldsymbol{\mu}. \tag{2.35}$$

To compare the estimate of $\boldsymbol{\psi}$ obtained by maximizing (2.12) with those obtained by maximizing either (2.29) or (2.31), the normal approximation $g(y_t|\alpha_t;\boldsymbol{\theta})$, $t = 1,\ldots,n$ proposed by Durbin and Koopman is required. Working with the distribution of the log of the squared observations, Sandmann and Koopman (1998) obtain this approximation and comment that this tranformation may cause problems when zero or small values are encountered. To avoid this "inlier" problem we use the general importance sampling procedure proposed in (2.14). Thus, if $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(N)}$ are draws from $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$, an estimate of $L(\boldsymbol{\psi};\mathbf{y})$ is given by

$$\hat{L}(\boldsymbol{\psi};\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(\mathbf{y},\boldsymbol{\alpha}^{(i)}|\boldsymbol{\psi})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y};\boldsymbol{\psi})}. \tag{2.36}$$

For a fixed value ψ_0, estimate $L(\psi; \mathbf{y})$ by

$$\hat{L}(\psi; \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \frac{p(\mathbf{y}, \alpha^{(i)}|\psi)}{p_a(\alpha^{(i)}|\mathbf{y}, \psi_0)}, \tag{2.37}$$

where $\alpha^{(1)}, \ldots, \alpha^{(N)}$ is a sample from $p_a(\alpha|\mathbf{y}, \psi_0)$. As in (2.31), to estimate ψ by maximizing (2.37), a few updatings of ψ_0 is recommended.

For our simulation study, we consider realizations of length $n = 500$ and compute mean and root mean squared errors over 500 simulated realizations for each of the nine parameters given in Table 2.1. The results are shown in Table 2.2. In this table, AL denotes the estimates obtained by maximizing the approximating likelihood given in (2.12) and MCL denotes estimates obtained by maximizing the estimate of the

CV	Method	γ	ϕ	σ	γ	ϕ	σ	γ	ϕ	σ
	true	-0.821	0.900	0.675	-0.411	0.950	0.484	-0.164	0.980	0.308
	AL	-0.902	0.890	0.663	-0.491	0.940	0.478	-0.257	0.969	0.315
		0.299	0.036	0.081	0.210	0.025	0.065	0.176	0.021	0.052
10	MCL	-0.866	0.894	0.657	-0.491	0.940	0.484	-0.260	0.968	0.320
		0.255	0.031	0.075	0.203	0.024	0.064	0.176	0.021	0.054
	MCL0	-0.878	0.894	0.661	-0.490	0.940	0.481	-0.257	0.967	0.317
		0.283	0.034	0.092	0.216	0.026	0.073	0.175	0.049	0.058
	true	-0.736	0.900	0.363	-0.368	0.950	0.260	-0.147	0.980	0.166
	AL	-0.956	0.870	0.377	-0.499	0.932	0.270	-0.260	0.965	0.176
		0.685	0.092	0.093	0.341	0.046	0.068	0.341	0.046	0.052
1	MCL	-0.894	0.879	0.372	-0.484	0.934	0.270	-0.271	0.963	0.178
		0.597	0.081	0.085	0.296	0.040	0.065	0.518	0.070	0.051
	MCL0	-0.883	0.880	0.367	-0.485	0.934	0.268	-0.263	0.964	0.176
		0.536	0.072	0.086	0.324	0.043	0.068	0.399	0.054	0.053
	true	-0.706	0.900	0.135	-0.353	0.950	0.096	-0.141	0.980	0.061
	AL	-1.848	0.740	0.188	-1.260	0.823	0.151	-0.830	0.883	0.104
		2.524	0.354	0.156	2.240	0.314	0.137	1.860	0.260	0.113
0.1	MCL	-1.918	0.729	0.172	-1.569	0.779	0.147	-1.258	0.823	0.115
		2.748	0.387	0.126	2.898	0.407	0.116	2.682	0.375	0.114
	MCL0	-2.184	0.692	0.169	-1.555	0.780	0.140	-1.097	0.845	0.096
		2.784	0.392	0.127	2.506	0.353	0.117	2.074	0.291	0.098

Table 2.2: Comparison of AL, MCL and MCL0 estimates based on 500 replications. Root mean square errors of estimates are reported below each estimate.

likelihood in (2.36). To attain numerical stability, the same noise was used to generate replicates of $\alpha^{(j)}$'s as a function of the AR parameters. MCL0 denotes estimates obtained by maximizing the single-sample estimate of the likelihood in (2.37). For

this case, we start ψ_0 with the AL estimate and the updating scheme is as follows: 10 updates with N=100, 5 updates with N=500 and 5 updates with N=1000. We notice that MCL and MCL0 essentially produce the same estimates, but with a few exceptions MCL gives smaller mean square errors. Because of this, we focus only on the MCL estimator. For all methods, the estimates become more biased as CV decreases. The large bias for CV=0.1 comes from the fact that the data appear almost indistinguishable from a constant volatility model (Breidt and Carriquiry, 1996; Sandmann and Koopman, 1998). For the remaining cases, the bias for ϕ and σ are small, while the bias for γ is large even for large CV. Also, for this parameter, AL has larger bias than MCL. For CV=10, MCL and AL have roughly equal mean squared errors. For CV=1, MCL has smaller mean squared errors for the first two values of ϕ. More importantly, is that the two estimation procedures have comparable performance throughout the range of parameter values. The setup of the models in the simulation study by Sandmann and Koopman (1998) is similar to ours. They obtain parameter estimates following the Durbin and Koopman procedure by working the log of the squared observations. The bias and root mean square errors of ϕ for the models for which CV is 10 or 1, are comparable with ours. For most of the cases we obtain smaller bias for σ and larger bias for γ.

2.2.2 Poisson Model

For the second simulation example, we assume that $p(y_t|\alpha_t;\psi)$ is a Poisson distribution with rate $\lambda_t := e^{\beta+\alpha_t}$, where $\alpha_t = \phi\alpha_{t-1} + \eta_t$, $\eta_t \sim$ iid $N(0,\ \sigma^2)$, $t = 1, \ldots, n$, and $|\phi| < 1$. We consider again nine models. This time, to classify the models, the index of dispersion D of the conditional variance of the observations $\sigma_t^2 = e^{\beta+\alpha_t}$ appears to be a more useful characterization of the ability to extract information in

the signal α_t than its coefficient of variation. The mean of σ_t^2 is held fixed at 1.5. The parameters of the models that result with this set up are shown in Table 2.3.

		ϕ		
D		0.90	0.95	0.98
10.0	β	-0.6130	-0.6130	-0.6130
	σ	0.6221	0.4456	0.2840
1.0	β	0.1501	0.1501	0.1501
	σ	0.3115	0.2232	0.1422
0.1	β	0.3732	0.3732	0.3732
	σ	0.1107	0.0793	0.0506

Table 2.3: Parameter values for a simulation experiment of nine Poisson state-space models.

For this simulation, we consider samples of size $n = 500$ and compute mean and root mean squared errors over 1000 simulated realizations for each of the nine parameters given in Table 2.3. The results are shown in Table 2.4. In this table, AL denotes the estimates obtained by maximizing the approximated likelihood

D	Method	β	ϕ	σ	β	ϕ	σ	β	ϕ	σ
	true	-0.613	0.900	0.622	-0.613	0.950	0.446	-0.613	0.980	0.284
	AL	-0.593	0.889	0.617	-0.629	0.940	0.444	-0.599	0.969	0.288
10		0.288	0.033	0.061	0.390	0.023	0.055	0.605	0.037	0.061
	MCL	-0.592	0.892	0.614	-0.630	0.941	0.445	-0.600	0.969	0.289
		0.287	0.030	0.059	0.390	0.022	0.054	0.603	0.030	0.049
	true	0.150	0.900	0.312	0.150	0.950	0.223	0.150	0.980	0.142
	AL	0.152	0.888	0.312	0.143	0.938	0.229	0.142	0.968	0.150
1		0.143	0.039	0.046	0.201	0.028	0.039	0.317	0.030	0.033
	MCL	0.151	0.889	0.313	0.142	0.938	0.230	0.142	0.968	0.150
		0.143	0.037	0.046	0.201	0.027	0.039	0.317	0.022	0.031
	true	0.373	0.900	0.111	0.373	0.950	0.079	0.373	0.980	0.051
	AL	0.369	0.759	0.146	0.369	0.868	0.103	0.370	0.873	0.075
0.1		0.064	0.336	0.083	0.081	0.242	0.066	0.114	0.329	0.060
	MCL	0.371	0.774	0.136	0.369	0.864	0.102	0.370	0.855	0.076
		0.063	0.327	0.070	0.080	0.248	0.063	0.114	0.353	0.060

Table 2.4: Comparison of AL and MCL estimates based on 500 replications. Root mean square errors of estimates are reported below each estimate.

given in (2.23) and MCL denotes estimates obtained by maximizing the estimate of the likelihood in (2.29). From this table, we notice that the estimates of ϕ and σ^2 deteriorate as D decreases, with large bias for these parameters when $D = 0.1$.

Except for a couple of cases, AL and MCL produce remarkably similar results.

2.2.3 Bias Correction via Bootstrap

In the two simulation studies that we considered, the approximate MLE of the parameters for the Poisson and stochastic volatility models can be slightly biased. Indeed, we will see in the two applications to real data, that the approximate likelihood and importance sampling estimates can be very close to each other. Closeness here is "measured" via the Monte Carlo error. In this section, we will show via simulation that the bias of the estimates can be reduced considerably using the bootstrap. Stoffer and Wall, 1991 uses the bootstrap to reduce the bias of the ML estimates of the parameters of a classical Gaussian state-space model.

To implement the bootstrap in our modeling setup, let $y_1 \ldots, y_n$ be observations from a state-space model and let $\hat{\psi}_{AL}$ be the maximizer of the approximate likelihood in (2.12). Following Efron and Tibshirani (1993), the *bootstrap bias correction* of the estimate $\hat{\psi}_{AL}$ of ψ is given by

$$\bar{\psi}_{AL} = \hat{\psi}_{AL} - \widehat{\text{bias}}, \tag{2.38}$$

where $\widehat{\text{bias}} = \bar{\psi}^* - \hat{\psi}_{AL}$, and $\bar{\psi}^*$ is the average of B bootstrap estimates $\hat{\psi}^*_1, \ldots, \hat{\psi}^*_B$. Here, the bootstrap estimate $\hat{\psi}^*_j$ is the maximizer of the approximate likelihood in (2.12) computed with a realization $y^*_1 \ldots, y^*_n$ drawn from the state-space model that has true parameters $\hat{\psi}_{AL}$. The *bootstrap estimate of the variance* of the estimator $\hat{\psi}_{AL}$ is

$$\widehat{\text{var}(\hat{\psi}_{AL})} = \frac{1}{B-1}\sum_{j=1}^{B}(\hat{\psi}^*_j - \bar{\psi}^*)(\hat{\psi}^*_j - \bar{\psi}^*)^T. \tag{2.39}$$

To assess the performance of the bootstrap bias correction, we conducted a simulation study on three Poisson models with parameters given in the middle section of

Table 2.3 (i.e., $D = 1$). As seen in Table 2.4, ϕ has a moderate bias in these models. The results of the simulation are given in Table 2.5. BC refers to the average of 1000 bias corrected estimates defined in (2.38) computed with $B = 100$ bootstrap

estimate	β	ϕ	σ	β	ϕ	σ	β	ϕ	σ
true	0.150	0.900	0.312	0.150	0.950	0.223	0.150	0.980	0.142
AL	0.153	0.887	0.313	0.147	0.938	0.227	0.140	0.967	0.147
S.E.	0.144	0.038	0.047	0.201	0.026	0.038	0.302	0.029	0.033
BC	0.154	0.904	0.305	0.147	0.953	0.217	0.141	0.985	0.133
S.E.	0.144	0.034	0.048	0.202	0.023	0.040	0.303	0.025	0.036

Table 2.5: Simulation results of bias correction for three Poisson state-space models based on 1000 replications. The rows labelled AL and BC are the average of the replications. Each AL estimate is the optimizer of the approximate likelihood in (2.23) and each BC estimate is the bootstrap bias correction estimate defined in (2.38).

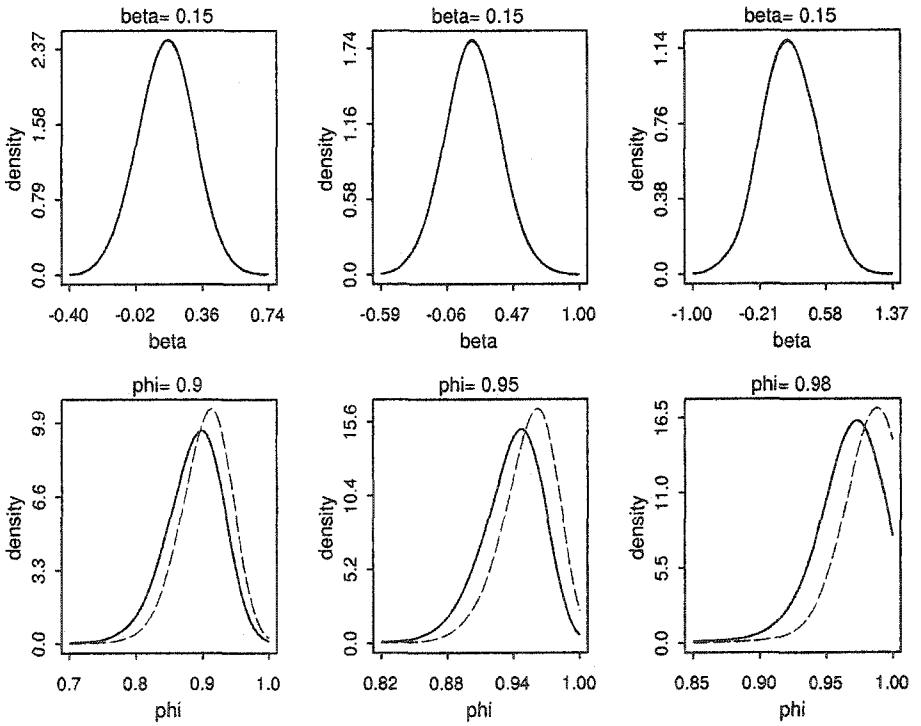

Figure 2.4: Parameter densities for β (first row) and ϕ (second row) for estimations AL (solid line) and BC (dotted line) for three Poisson state-space models.

estimates. The standard errors of the 1000 bias corrected estimates are also shown in the table. The AL estimates were obtained from 1000 simulated realizations from the state-space model having true parameters given in the middle section of Table 2.3.

The row labeled AL is the average of the 1000 simulated $\hat{\psi}_{AL}$ estimates. Inspecting this table, the bootstrap bias correction has done a good job in reducing the bias of the AL estimate of ϕ with only little alteration of the standard errors.

In Figure 2.4 we compare the estimated densities of the AL and BC estimates of the parameters β and ϕ. Each column in this figure corresponds to the models with parameters (0.150, 0.900, 0.312), (0.150, 0.950, 0.223) and (0.150, 0.980, 0.142) respectively. As seen from these graphs, the BC estimates have essentially shifted the location of the AL estimates.

2.2.4 Pound-Dollar Exchange Rates

The first dataset that we analyze is the Pound/Dollar exchange rates. The data, taken from the site http://staff.feweb.vu.nl/koopman/sv/ consists of the log differences y_t of the daily observations of weekdays closing pound to dollar exchange rates z_t, $t = 1, \ldots, 946$ from 10/1/81 to 6/28/85. We use the basic stochastic volatility model (2.34) to model $y_t := \log(z_t) - \log(z_{t-1})$. Setting the parameter vector $\psi := (\gamma, \phi, \sigma^2)$, Table 2.6 shows various estimates of ψ. The second column, labeled as AL, contains the estimate of ψ obtained by maximizing (2.12). The column labeled MCL contains the estimate of ψ obtained by maximizing (2.36). MCE denotes Monte Carlo error and is obtained as the standard error of 1000 estimates of

Parameter	AL	S.E.	BC	MCL	MCE	S.E.	BC
γ	-0.0227	0.0198	-0.0140	-0.0230	0.0004	0.0173	-0.0153
ϕ	0.9750	0.0194	0.9845	0.9747	0.0004	0.0166	0.9832
σ^2	0.0267	0.0141	0.0228	0.0273	0.0007	0.0138	0.0228

Table 2.6: Parameter estimates for the Pound-Dollar exchange rates data. AL and MCE are the maximizers of (2.12) and (2.36), respectively. BC are bootstrap bias corrected estimates ($B = 500$) and S.E. are bootstrap estimates of the standard errors of AL and MCL, respectively. MCE is the standard error of 1000 MCL replicates.

ψ, using for each estimate the same observations $y_1, \ldots, y_{945}$. The standard error of the estimates AL and MCL are obtained using (2.39). The columns labeled as BC are bootstrap bias corrections of AL and MCL computed with $B = 500$ bootstrap estimates. Notice that the AL and MCL estimates are remarkably close. In fact, the difference between these estimates is due to the randomness of the MCL estimate. For example, two distinct MCL estimates of σ^2 are unlikely to differ more than four times the Monte Carlo error, i.e., 0.0028, while the estimates AL and MCE of σ^2 differ only by 0.0006. In other words, we would not be able to differentiate the AL estimate from a "cloud" of MCL replicates.

2.2.5 Polio data

The second dataset consists of the observed time series $y_1, \ldots, y_{168}$ of the monthly number of U.S. cases of poliomyelitis for 1970 to 1983 that was first considered by Zeger (1988). We adopt the same model used by Zeger in which the distribution of Y_t, given the state α_t, is Poisson with rate $\lambda_t := e^{\alpha_t + \mathbf{x}_t^T \beta}$. Here, $\beta^T := (\beta_1, \ldots, \beta_6)$, $\mathbf{x}_t$ is the vector of covariates given by

$$\mathbf{x}_t^T = (1, t/1000, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6)),$$

and the state process is assumed to follow the AR(1) model given by, $\alpha_t = \phi\alpha_{t-1} + \eta_t$, where $\eta_t \sim$ iid $N(0, \sigma^2)$, $t = 1, \ldots, n$, and $|\phi| < 1$. The vector of parameters of this SSM is $\psi = (\beta, \phi, \sigma^2)$. Table 2.7 contains the results of two estimation procedures. Columns 2 and 5 labeled as AL and MCL respectively, contain the estimates of ψ obtained by maximizing (2.12) and (2.29), respectively. As in the previous example, MCE denotes Monte Carlo error, based on 1000 replicates of the MCL estimates using for each replicate the same observations $y_1, \ldots, y_{168}$.

Parameter	AL	S.E.	BC	MCL	MCE	S.E.	BC
β_1	0.202	0.332	0.043	0.200	0.0010	0.345	0.220
β_2	-2.691	3.376	-2.484	-2.647	0.0064	3.551	-2.820
β_3	0.113	0.124	0.105	0.112	0.0003	0.121	0.108
β_4	-0.454	0.142	-0.451	-0.454	0.0003	0.142	-0.445
β_5	0.396	0.108	0.392	0.396	0.0003	0.109	0.392
β_6	0.017	0.108	0.011	0.017	0.0003	0.110	0.014
ϕ	0.845	0.212	0.945	0.850	0.0018	0.181	0.936
σ^2	0.104	0.074	0.094	0.102	0.0020	0.067	0.095

Table 2.7: Parameter estimates for the polio data. AL and MCE are the maximizers of (2.12) and (2.29), respectively. BC are boostrap bias corrected estimates and S.E. are bootstrap estimates of the standar errors of AL and MCL, respectively. MCE is the standard error of 1000 MCL replicates.

Notice that only the AL and DK estimates for β_2 differ more than the expected difference between two DK estimates (4 times MCE). In general the AL estimates are very close to the DK estimates in spite of the fact that the length n of the observed time series is not large. We obtain here larger Monte Carlo error than in Table 2.6 even when we have used the same number of draws ($N = 1000$) to compute the Monte Carlo integration in (2.36) and (2.29) respectively. This may not be surprisingly since the polio data set has far fewer observations than the Pound-Dollar exchange rate data. Moreover, the model fitted to the latter has fewer parameters.

2.2.6 How good is the posterior approximation?

As seen in the simulation studies considered above, the use of $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ in (2.10) as the normal approximation to the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$ gives good results. The quality of the likelihood approximation is due largely to the closeness of the normal approximation to the posterior. In this subsection we provide two methods for examining the closeness of this normal approximation. The first method compares the posterior mean with the posterior mode. The second method is a statistical test based on the correlation between the *generalized squared distances*

defined in (2.42) above with the quantiles of a Chi-squared distribution.

For the first method, first recall that the posterior mode is given by $\boldsymbol{\alpha}^*$. We now provide an estimate $\hat{\boldsymbol{\alpha}}$, also known as the *smoothed state vector*, of the posterior mean of the state vector. From (2.5) and the fact that $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) \propto L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})$

$$\mathrm{E}(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}) = \int \boldsymbol{\alpha} p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}) d\boldsymbol{\alpha} = \frac{1}{L(\boldsymbol{\psi};\mathbf{y})} \int \boldsymbol{\alpha} L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) d\boldsymbol{\alpha}.$$

Hence, if $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(N)}$ are draws from $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ and $\hat{L}(\boldsymbol{\psi};\mathbf{y})$ is the estimate of the likelihood given in (2.36), an estimate of the posterior mean is given by

$$\hat{\boldsymbol{\alpha}} = \frac{1}{N\hat{L}(\boldsymbol{\psi};\mathbf{y})} \sum_{i=1}^{N} \boldsymbol{\alpha}^{(i)} \frac{p(\mathbf{y},\boldsymbol{\alpha}^{(i)}|\boldsymbol{\psi})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y};\boldsymbol{\psi})}. \tag{2.40}$$

As an example, for the Pound-Dollar exchange rates and polio data let $\boldsymbol{\psi}$ be the AL estimate from Tables 2.6 and 2.7 respectively. Using $N = 1000$ in (2.40), $\hat{\boldsymbol{\alpha}}$ was computed. In Figures 2.5 and 2.6 the solid line shows the smoothed state vector, and the dashed line shows the posterior mode $\boldsymbol{\alpha}^*$ of $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$ obtained as in (2.16).

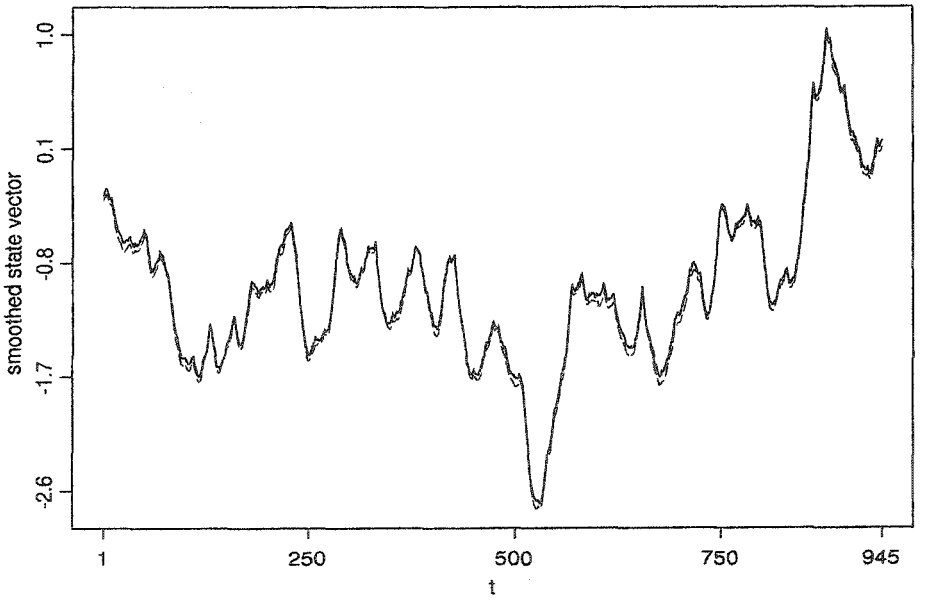

Figure 2.5: (*Smoothed state vector*) For the Pound-Dollar exchange rates data, the solid line shows estimate of the posterior mean of the state vector and the dashed line shows its posterior mode.

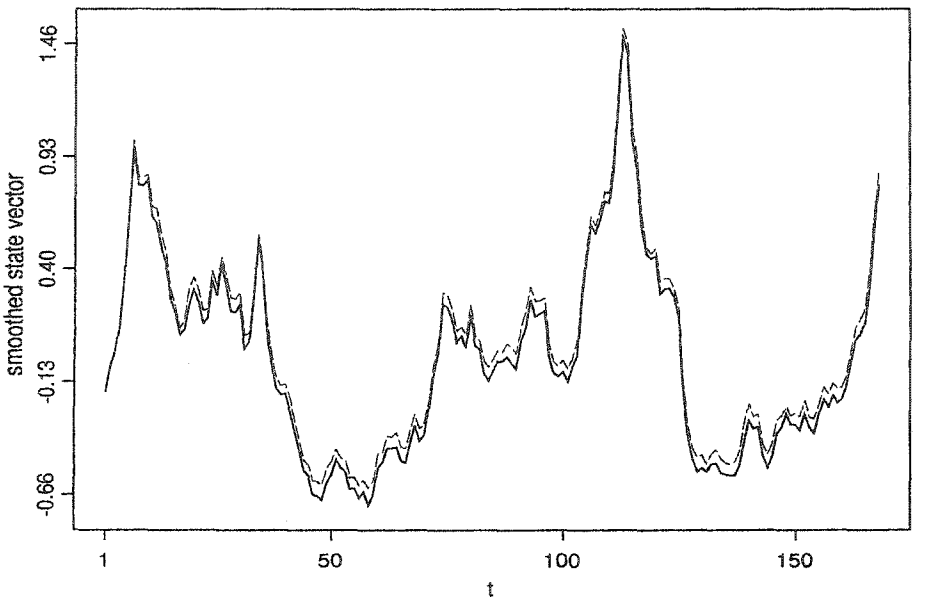

Figure 2.6: (*Smoothed state vector*) For the Polio data, the solid line shows estimate of the posterior mean of the state vector and the dashed line shows its posterior mode.

In both cases, the posterior mode and smoothed state vector are relatively close even though the number of observations of the polio data (n=168) is not large. This adds support to the goodness of the approximation to the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ by a multivariate normal density.

For the second method, if an independent sample from $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$ can be generated, then we could assess the compatibility of the samples with a normal population. Such a sample can be obtained as follows: First generate an independent sample $\boldsymbol{\alpha}^{(1)},\boldsymbol{\alpha}^{(2)},\ldots,\boldsymbol{\alpha}^{(N)}$ from the approximate distribution $p_a(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$. For N large, an iid sample from the discrete distribution that puts mass p_i given by

$$p_i := \frac{w_i}{\sum_{i=1}^{N} w_i}, \quad w_i = \frac{p(\boldsymbol{\alpha}^{(i)}|\mathbf{y},\boldsymbol{\psi})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y},\boldsymbol{\psi})} \propto \frac{L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}^{i})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y},\boldsymbol{\psi})}, \tag{2.41}$$

is an (approximate) iid sample from $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$. In the Bayesian literature, this method is known as *sampling importance-resampling* (SIR), e.g., Bernardo and Smith (1994). Assume now that $\tilde{\boldsymbol{\alpha}}^{(1)},\tilde{\boldsymbol{\alpha}}^{(2)},\ldots,\tilde{\boldsymbol{\alpha}}^{(M)}$ is an iid sample from $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$. If $p_a(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$ in (2.10) were a good approximation to $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$, for $M-n$ large, the

squared generalized distances

$$d_j^2 := (\tilde{\boldsymbol{\alpha}}^{(j)} - \boldsymbol{\alpha}^*)^T(\mathbf{K}^* + \mathbf{V})(\tilde{\boldsymbol{\alpha}}^{(j)} - \boldsymbol{\alpha}^*), \quad j = 1, \ldots, M, \tag{2.42}$$

would resemble an iid sample from the chi-squared distribution with n degrees of freedom (Johnson and Wichern, 1998). Thus, a chi-squared QQ-plot of $d_1^2, \ldots, d_M^2$, should resemble a straight line through the origin with slope 1.

To illustrate this technique, consider the state-space model for which $p(y_t|\alpha_t; \boldsymbol{\psi})$ is the Poisson distribution with rate $\lambda_t := e^{\beta+\alpha_t}$; $\alpha_t = \phi\alpha_{t-1} + \eta_t$, $\eta_t \sim$ iid $N(0,\ \sigma^2)$, $t = 1, \ldots, n$; and $|\phi| < 1$. The vector of parameters of this process, $\boldsymbol{\psi} = (\beta, \phi, \sigma^2)$, is fixed to $(0.373, 0.9, 0.012)$. Chi-squared QQ-plots of $d_1^2, \ldots, d_M^2$ are shown in Figure 2.7. With a sample of size N=5000 from $p_a(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$, a sample of size M from $p(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$ was obtained via SIR. The j-th column of this figure corresponds to the

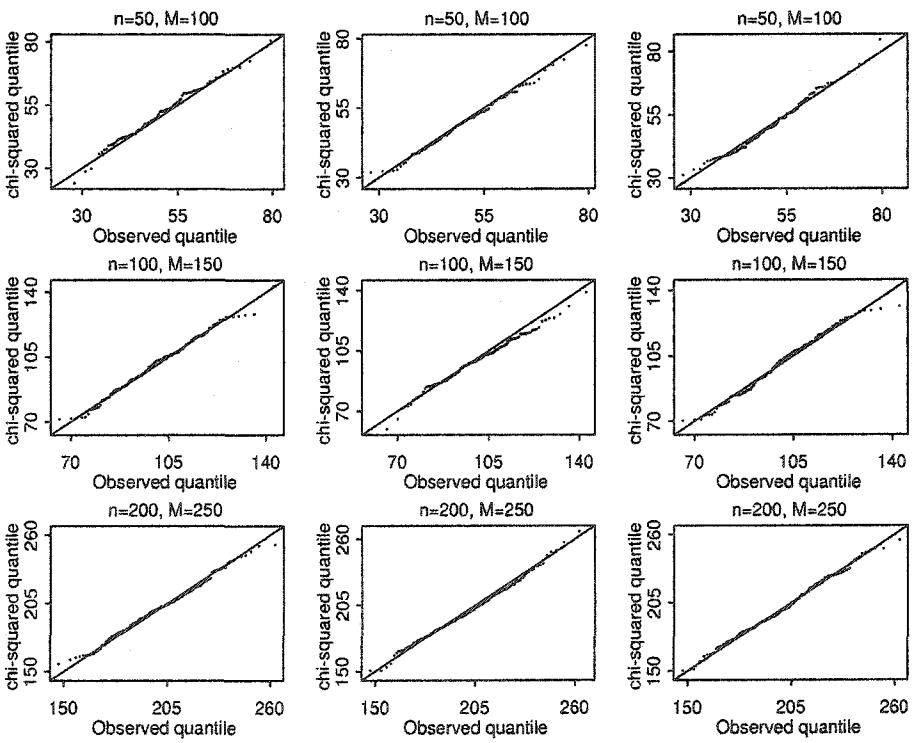

Figure 2.7: (*Chi-squared QQ-plots*) The QQ-plot from i-th row and j-th column was obtained using a SIR sample $\tilde{\boldsymbol{\alpha}}^{(1)}, \tilde{\boldsymbol{\alpha}}^{(2)}, \ldots, \tilde{\boldsymbol{\alpha}}^{(M)}$ from $p(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi}_j)$ by resampling a sample of size 5000 from the approximation $p_a(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi}_j)$.

parameter value of $\psi = \psi_j$, where $\psi_1 := (0.2, 0.8, 0.002)$, $\psi_2 := (0.373, 0.9, 0.012)$ and $\psi_3 := (0.5, 0.95, 0.02)$. From this figure, we notice that even for a small sample ($n = 50$), the squared generalized distances closely resemble the chi-squared distribution with n degrees of freedom.

The correlation coefficient r_Q between the ordered distances $d^2_{(j)}, j = 1, \ldots, M$ and the Chi-squared quantiles can be used to test any departure from normality of $p_a(\alpha|\mathbf{y}, \psi)$ (Johnson and Wichern, 1998). The nine correlations r_Q for the data used to create Figure 2.7 are shown in the last three columns of Table 2.8. The hypothesis must be rejected at level $\alpha\%$ if the correlation falls below r_α. The critical points $r_{0.05}$ for each M, needed to test the null hypothesis of normality with 5% of significance level are given in the third column of this table. In all cases, normality is not rejected. This provides some evidence that the distribution in (2.10) may be a reasonable approximation for the posterior distribution $p(\alpha|\mathbf{y}, \psi)$.

				r_Q	
N	M	$r_{0.05}$	ψ_1	ψ_2	ψ_3
50	100	0.9873	0.9952	0.9978	0.9925
100	150	0.9913	0.9957	0.9952	0.9926
200	250	0.9920	0.9974	0.9974	0.9973

Table 2.8: Correlation coefficients of the points in the QQ-plots from figure 2.7.

2.3 Conclusions

For the state-space model, a second order Taylor series expansion of the log of the conditional likelihood gives an approximation to the observed likelihood of the state-space model. An approximate MLE of the parameters of the state-space model can be obtained from this function. Because no simulation is involved, this procedure is fast. The Taylor series expansion gives also a normal approximation $p_a(\alpha|\mathbf{y}; \psi)$

to the posterior distribution of the states. For the exponential family of distributions in standard form, the approximate distribution $p_a(\alpha|\mathbf{y};\psi)$ coincides with the approximation to the conditional distribution of α found by Durbin and Koopman (1997). This approximation can be used to implement existing estimation procedures based on the Monte Carlo approximation to the likelihood, as it is the case of the procedure given by Kuk (1999) and Durbin and Koopman (1997). In various simulation studies, the results obtained with our approach are close to other simulation based approximations of the MLE. Although the (approximate) likelihood estimates may have some bias, the speed of this procedure makes bootstrap method for bias correction a viable procedure.

CHAPTER 3

Particle Filters

In Chapter 2, the posterior distribution of the state vector given the data, $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$, played a key role in the approximate likelihood approach for parameter estimation of the SSM. In this chapter we consider another estimation approach that consists in updating the posterior distributions of the states as the observations arrive. Such an approach, which has been widely adopted in the literature, is based on the *general Kalman recursions.* Starting from an initial distribution $p(\alpha_0)$, the Kalman recursions provide a recursive algorithm for computing the *filtering density* $p(\alpha_t|y_{1:t-1};\boldsymbol{\psi})$, $t = 1,\ldots,n$, where $y_{1:t-1} := [y_1,\ldots,y_{t-1}]^T$. At time t, the (*one-step ahead*) *prediction density* and filtering densities are updated as follows (Brockwell and Davis, 1996):

prediction density

$$\begin{aligned} p(\alpha_t|y_{1:t-1};\boldsymbol{\psi}) &= \int p(\alpha_{t-1},\alpha_t|y_{1:t-1};\boldsymbol{\psi})d\alpha_{t-1} \\ &= \int p(\alpha_t|\alpha_{t-1};\boldsymbol{\lambda})p(\alpha_{t-1}|y_{1:t-1};\boldsymbol{\psi})d\alpha_{t-1}, \end{aligned} \tag{3.1}$$

filter density

$$p(\alpha_t|y_{1:t};\boldsymbol{\psi}) = \frac{p(y_t|\alpha_t;\boldsymbol{\theta})p(\alpha_t|y_{1:t-1};\boldsymbol{\psi})}{p(y_t|y_{1:t-1};\boldsymbol{\psi})}. \tag{3.2}$$

Except for some simple cases, the prediction density in (3.1) can not be computed explicitly. In the literature, the filtering densities are frequently approximated using

Sequential Monte Carlo (SMC) methods, a set of simulation based procedures. Depending on the field of study, these procedures have appeared as bootstrap filters, condensation filters, particle filters, Monte Carlo filters, and so on. See for example Kitagawa (1996), Pitt and Shepard (1999), Doucet, et al. (2001).

In this chapter we compare estimations based on the approximation (2.12) to the likelihood of the state space model with those obtained via particle filtering. In Section 3.1 a brief introduction to particle filters is given. In the remaining sections of this chapter we estimate the likelihood of the SSM using various particle filters. In Section 3.2, the particle filters are obtained using the accept-reject procedure given in Hurzeler (1998). In the last two sections we give two implementations that are faster than the accept-reject method. The first, a new implementation based on the *Griddy Gibbs sampler* introduced by Ritter and Tanner (1992) as a tool to generate draws in Gibbs sampling implementations, is presented in Section 3.3. The second, given in Section 3.4, is the auxiliary particle filters procedure of Pitt and Shepard (1999).

3.1 Particle filters

Pitt and Shepard (1999) give the following definition of particle filters: "particle filters are the class of simulation filters that recursively approximate the filtering random variable $\alpha_t|y_{1:t}$ by "particles" $\alpha_t^1, \ldots, \alpha_t^M$, with discrete probability mass of $\pi_t^1, \ldots, \pi_t^M$". The particles and masses are obtained iteratively. With $\alpha_{t-1}^1, \ldots, \alpha_{t-1}^M$, $\pi_{t-1}^1, \ldots$, and π_{t-1}^M available from iteration (t-1), the particles $\alpha_t^1, \ldots, \alpha_t^M$ (and masses) are obtained as described below.

Using the discrete support of the particles $\alpha_{t-1}^1, \ldots, \alpha_{t-1}^M$, the following approxi-

mation to the prediction density in (3.1) is obtained

$$\hat{p}(\alpha_t|y_{1:t-1};\psi) = \sum_{j=1}^{M} p(\alpha_t|\alpha_{t-1}^j;\lambda)\pi_{t-1}^j, \tag{3.3}$$

which is known as the "empirical prediction density". Substituting this into (3.2), the "empirical filtering density" is obtained

$$\hat{p}(\alpha_t|y_{1:t};\psi) \propto p(y_t|\alpha_t,\theta)\sum_{j=1}^{M} p(\alpha_t|\alpha_{t-1}^j;\lambda)\pi_{t-1}^j. \tag{3.4}$$

In the next three sections we describe procedures to generate new particles $\alpha_t^1,\ldots,\alpha_t^M$ and their masses $\pi_t^1,\ldots,\pi_t^M$ from the empirical filtering density.

The likelihood of the SSM can be written as

$$L(\psi;\mathbf{y}) = p(y_1|\psi)\prod_{t=2}^{n} p(y_t|y_{1:t-1};\psi), \tag{3.5}$$

where

$$p(y_1|\psi) = \int p(y_1,\alpha_1|\psi)d\alpha_1 = \int p(y_1|\alpha_1;\theta)p(\alpha_1|\lambda)d\alpha_1,$$

and for $t \geq 2$,

$$p(y_t|y_{1:t-1};\psi) = \int p(y_t,\alpha_t|y_{1:t-1};\psi)d\alpha_t = \int p(y_t|\alpha_t;\theta)p(\alpha_t|y_{1:t-1};\psi)d\alpha_t. \tag{3.6}$$

Once a particle filtering procedure has been implemented to approximate $p(\alpha_t|y_{1:t-1};\psi)$, the integral in (3.6) can be estimated. For example, if $\alpha_{t|t-1}^{(1)},\ldots,\alpha_{t|t-1}^{(N)}$ are draws from the empirical prediction density $\hat{p}(\alpha_t|y_{1:t-1};\psi)$, the integral in (3.6) can be estimated as

$$\hat{p}(y_t|y_{1:t-1};\psi) = \frac{1}{N}\sum_{j=1}^{N} p(y_t|\alpha_{t|t-1}^{(j)};\theta). \tag{3.7}$$

To draw $\alpha_{t|t-1}^{(j)}$ from (3.3), the *composition method* (Ripley, 1987, page 63) can be used: First draw j from the distribution of the discrete random variable J for which

$$Pr(J=j) = \pi_{t-1}^j, \quad j=1,\ldots,M,$$

and then draw $\alpha_{t|t-1}^{(j)}$ from $p(\alpha_t|\alpha_{t-1}^j;\boldsymbol{\lambda})$. Once (3.6) has been estimated, the logarithm of the estimation $\hat{L}(\boldsymbol{\psi};\mathbf{y})$ of the likelihood in (3.5) becomes,

$$\hat{\ell}(\boldsymbol{\psi};\mathbf{y}) = \sum_{t=1}^{n} \log \hat{p}(y_t|y_{1:t-1};\boldsymbol{\psi}) \tag{3.8}$$

3.2 Accept-Reject

In Hurzeler (1998), the particles $\alpha_t^1,\ldots,\alpha_t^M$ were taken as draws from the empirical filtering density at time t. In this approach, $\pi_t^j = 1/M$ for $j = 1,\ldots,M$. If $p(y_t|\alpha_t;\boldsymbol{\theta})$ is bounded in α_t, one can sample from (3.4) by an accept-reject procedure in which the mixture distribution in (3.3) is used as the instrumental density. If c_t is given by

$$c_t \geq \sup_{\alpha_t} p(y_t|\alpha_t;\boldsymbol{\theta}),$$

the draw $\alpha_t^j, j = 1,\ldots,M$ is generated as follows:

1. Generate X from (3.3) and U from $U(0,c_t)$.
2. If $p(y_t|\alpha_t = X;\boldsymbol{\theta}) \geq U$, set $\alpha_t^j = X$, otherwise return to 1. □

Example. Consider the case when the observation density is Poisson with rate $\lambda_t = e^{\beta+\alpha_t}$ and the state process follows the AR(1) model

$$\alpha_t = \phi\alpha_{t-1} + \eta_t, \tag{3.9}$$

where $\eta_t \sim$ iid $N(0,\sigma^2)$, $t = 1,\ldots,n = 200$. Using the same realization $y_1,\ldots,y_{200}$ as in Figure 2.1 we show in Figure 3.1 an estimate of the log likelihood for a grid of 14 points of ϕ. The parameters β and σ^2 are fixed at 0.7 and 0.30, respectively. In this figure, the solid line is the approximation to the likelihood given in (2.23). The lower and upper dotted lines are the minimum and maximum, respectively, of

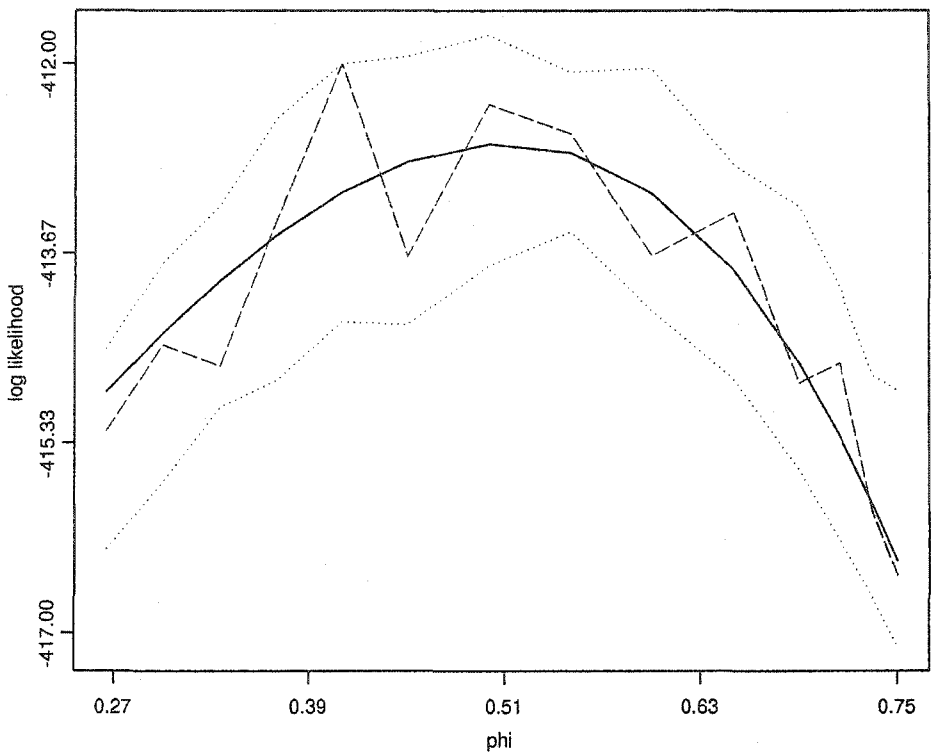

Figure 3.1: (*Accept-reject*) For a grid of values of ϕ, estimations of the likelihood of a Poisson SSM are shown. For the solid line, estimates were obtained using (2.23). The dotted lines are the minimum and maximum respectively, of 100 replicates of (3.8) with N=1000 in (3.7). The dashed line is an element of these replicates.

100 replicates of the estimation in (3.8) using $N = 1000$ in (3.7). The length of the particle filters is $M = 1000$. The dashed line is one element of these replicates.

In Figures 2.1 and 3.1, the Monte Carlo error obtained when (3.8) is used to estimate the likelihood is much larger than that obtained when (2.29) is used.

In Figure 3.2 the dashed line is the average of the 100 replicates from Figure 2.1 and the dotted line is the average of the 100 replicates from Figure 3.1. The solid line shows the AL estimate of the likelihood. As it can be seen in this figure, the log of the estimates of L are very close. Thus, for large N, difference between estimates that maximize (2.29) and (3.8) will be due only to Monte Carlo error. □

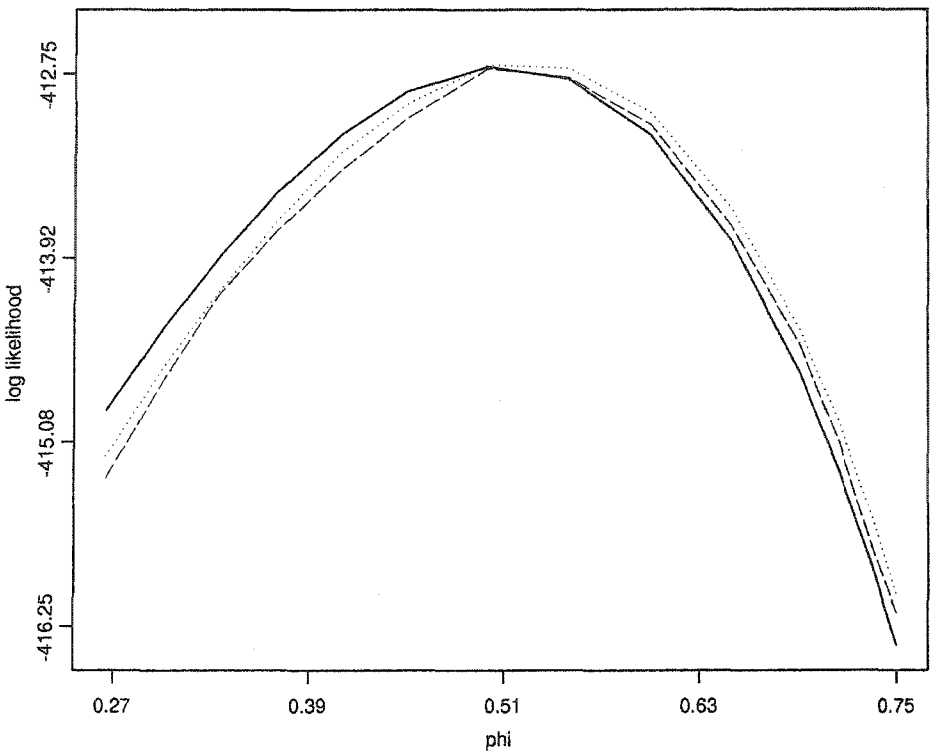

Figure 3.2: The solid line is the log of the AL estimate of the likelihood. The dashed and dotted lines are the average of the 100 replicates of $\hat{\ell}$ from Figures 2.1 and 3.1 respectively.

3.3 Griddy particle filters (GPF)

When the number of rejections in the accept-reject procedure is high, the estimation of the likelihood considered in Section 3.2 can be very slow. To accelerate the estimation of the likelihood, we construct particles obtained with the *Griddy Gibbs sampler* (GGS) introduced by Ritter and Tanner (1992) to facilitate the Gibbs sampler in situations where it is difficult to sample from a univariate distribution. This method approximates the cumulative distribution function of a univariate distribution by a piecewise linear function and then samples from the approximation.

Griddy Gibbs Sampler algorithm

To describe the GGS algorithm, suppose that a draw from the empirical filtering

density in (3.4) is needed. Denote $\hat{p}_c(\alpha_t|y_{1:t};\psi)$ the term on the right hand side of (3.4), i.e.,

$$\hat{p}_c(\alpha_t|y_{1:t};\psi) = p(y_t|\alpha_t;\theta)\hat{p}(\alpha_t|y_{1:t-1};\psi). \tag{3.10}$$

The Ritter and Tanner (1992) algorithm to sample from the empirical density function becomes

Step 1. At a predetermined set of ordered points $\alpha_t^1 < \ldots < \alpha_t^{M_t}$ (to be described below), calculate

$$w_t^j = \hat{p}_c(\alpha_t^j|y_{1:t};\psi), \quad j = 1,\ldots,M_t.$$

Step 2. Use $w_t^1,\ldots,w_t^{M_t}$ to obtain an approximation to the inverse cdf of $\hat{p}(\alpha_t|y_{1:t};\psi)$. Simple approximations are:

a. Piecewise constant corresponding to the discrete distribution for $\alpha_t^1,\ldots,\alpha_t^{M_t}$ with α_t^j having mass $\pi_t^j := w_t^j / \sum_{j=1}^{M_t} w_t^j$.

b. Piecewise linear corresponding to a piecewise uniform distribution on the interval $[a_t^j, a_t^{j+1}]$, $j = 1,\ldots,M_t$, with $z_j \in [a_t^j, a_t^{j+1}]$ having density

$$w_j / \sum_{k=1}^{M_t} w_t^j (a_t^{j+1} - a_t^j).$$

Typically, α_t^j is centered in the interval $[a_t^j, a_t^{j+1}]$.

Step 3. Sample a $u(0,1)$ distribution and transform the observation via the approximate inverse cdf. □

We note that the GGS algorithm needs the empirical filtering density $\hat{p}(\alpha_t|y_{1:t};\psi)$ to be known up to a proportionality constant.

In their Remark 5, Ritter and Tanner (1992) suggest that for unbounded support the grid must be expanded to the left (right) if $w_t^1 > qR$ ($w_t^{M_t} > qR$), where $R :=$

$\max\{w_t^j : j = 1, \ldots, M_t\}$ and q is any user-defined value such that $q \in (0, 1)$. They suggest that $q = 0.1$ can be used, unless interest centers on the tails of the conditional distribution.

For the case of unbounded support, Ritter and Tanner's Remark 5 shows how to construct a regular grid of points. To do this, let α_t^0 be a rough approximation to the mean of the distribution $\hat{p}(\alpha_t|y_{1:t}; \boldsymbol{\psi})$, S a rough approximation to its standard deviation; M_0 the "desired" length of the grid, and set the grid separation at $\Delta_0 := 16S/M_0$. We can start with the pair $\{\alpha_t^0 - \Delta_0, \alpha_t^0\}$ and let it grow until the extremes are not greater than qR. If M_t, the number of grid points when this *grid-grower* stops, is less than M_0, which means that S overestimated the standard deviation of the filter variable α_t, then double the number of grid points as many times as needed until the final grid length is at least M_0.

To compute α_t^0, the empirical filtering density in (3.4) can be used. For example, if d_j^k is an approximation to the integral (see the examples below)

$$\int (\alpha_t)^k p(y_t|\alpha_t, \boldsymbol{\theta}) p(\alpha_t|\alpha_{t-1}^j; \boldsymbol{\lambda}) d\alpha_t,$$

then, an approximation to the mean of the filter variable α_t is given by

$$\alpha_t^0 := \frac{\pi_{t-1}^1 d_1^1 + \ldots + \pi_{t-1}^{M_{t-1}} d_{M_{t-1}}^1}{\pi_{t-1}^1 d_1^0 + \ldots + \pi_{t-1}^{M_{t-1}} d_{M_{t-1}}^0}. \tag{3.11}$$

Particle filters selection

Let $\alpha_t^1, \ldots, \alpha_t^{M_t}$ be the grid of points with separation Δ_t that results from the grid-grower algorithm that starts with the set $\{\alpha_t^0 - \Delta_0, \alpha_t^0\}$. Then, an iid sample of size M from the cdf approximation of the empirical filtering density obtained as described in Step 3 of the GGS are particle filters for $\hat{p}(\alpha_t|y_{1:t}; \boldsymbol{\psi})$ with common mass $1/M$. In particular, when the approximation in (a) of Step 2 of the GGS is used, the

griddy particle filters (GPF) $\alpha_t^1, \ldots, \alpha_t^{M_t}$ with masses π_t^j are obtained. □

In practice, values of M_t around 30 work well. Hence, the speed of this procedure must be faster than the accept-reject procedure from section 3.2, where typically M large is needed to approximate the filtering density well.

To estimate the t-th factor of the likelihood in (3.6), equation (3.7) can be used. However, an approximation based on the output of the GGS can be obtained. If the prediction density in the integrand in (3.6) is replaced by the empirical prediction density, then

$$\begin{aligned} p(y_t|y_{1:t-1};\psi) &\approx \int p(y_t|\alpha_t;\theta)\hat{p}(\alpha_t|y_{1:t-1}\psi)d\alpha_t, \\ &= \int \hat{p}_c(\alpha_t|y_{1:t};\psi)d\alpha_t, \\ &\approx \sum_{j=1}^{M_t} w_t^j \Delta_t. \end{aligned}$$

Hence, the following estimate of the prediction density results

$$\hat{p}(y_t|y_{1:t-1};\psi) := \sum_{j=1}^{M_t} w_t^j \Delta_t. \tag{3.12}$$

Since no draws from the empirical prediction density are needed to obtain this approximation, as it would be the case for the estimate in (3.7), this approach is much faster.

Example 1 (*Poisson*). Consider the Poisson state space model of the example in Section 3.2. Denote $\mu_t^j := \phi\alpha_{t-1}^j$ the mean of $\alpha_t|(\alpha_{t-1} = \alpha_t^j)$. Then,

$$p(y_t|\alpha_t;\beta)p(\alpha_t|\alpha_{t-1}^j;\phi,\sigma^2) = \frac{1}{y_t!\sqrt{2\pi\sigma^2}} e^{-e^{\beta+\alpha_t}+y_t(\beta+\alpha_t)-\frac{1}{2\sigma^2}(\alpha_t-\mu_t^j)^2}.$$

Notice that

$$y_t\alpha_t - \frac{1}{2\sigma^2}(\alpha_t - \mu_t^j)^2 = \mu_t^j y_t + \frac{\sigma^2}{2}y_t^2 - \frac{1}{2\sigma^2}(\alpha_t - \mu_t^j - \sigma^2 y_t)^2.$$

Hence,

$$p(y_t|\alpha_t;\beta)p(\alpha_t|\alpha_{t-1}^j;\phi,\sigma^2) = q_j e^{e^{\beta+\mu_t^j+\sigma^2 y_t}-e^{\beta+\alpha_t}}\phi(\alpha_t;\mu_t^j+\sigma^2 y_t,\sigma^2),$$

where

$$\log(q_j) = -\log y_t! + y_t\beta + \mu_t^j y_t + \sigma^2 y_t/2 - e^{\beta+\mu_t^j+\sigma^2 y_t}.$$

A zero order Taylor expansion of $e^{\beta+\alpha_t}$ around the point $\mu_t^j + \sigma^2 y_t$ gives

$$d_j^0 = q_j, \quad \text{and} \quad d_j^1 = q_j(\mu_t^j + \sigma^2 y_t).$$

Hence, the initial particle in (3.11) becomes

$$\alpha_t^0 = y_t\sigma^2 + \phi \sum_{j=1}^{M_{t-1}} \alpha_{t-1}^j q_j \pi_{t-1}^j / \sum_{j=1}^{M_{t-1}} q_j \pi_{t-1}^j. \tag{3.13}$$

Using the observations from the example of Section 3.2, to estimate the t-th term in (3.12) in a grid of points for ϕ, griddy particle filters were obtained using $q = 0.01$; $M_0 = 25$; and $S^2 = \sigma^2$ (the variance of the noise η_t). The values β and σ^2 are fixed to 0.7 and 0.3, respectively. In Figure 3.3, the solid line is (3.8), where the t-th term of the likelihood is computed using (3.12). The dashed line in Figure 3.1 and the dotted line in Figure 3.2 are also shown. Notice that the estimate considered in this section is remarkably close to the average of the 100 replicates of the estimator from Section 3.2.

Computation time

For the grid of ϕ of length 14, the computation time of the various estimates of the likelihood of the model considered in this example, is as follows: the solid line (AL) in Figure 2.1 took 0.015 seconds; the 100 replicates (DK) of the estimate in (2.29) used in Figure 2.1 took 210 seconds; the 100 replicates (AR) in Figure 3.1 took 3362 seconds and the solid line (GPF) in Figure 3.3 took 6 seconds. All times are based on an IBM ThinkPad, with a 1.6 GHz intel pentium M processor.

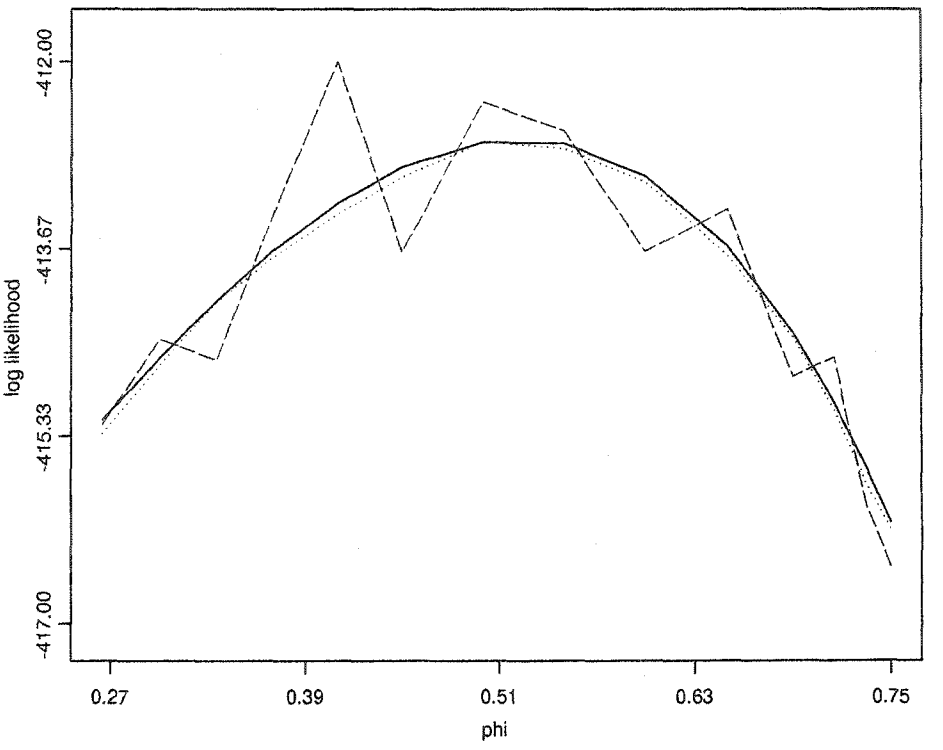

Figure 3.3: (*Griddy particle filtering*) For a grid of values of ϕ, the logarithm of the estimation of the likelihood of a Poisson SSM are shown. The long dashed line is as in Figure 3.1. The dotted line is as in Figure 3.2. The solid line is the estimate using (3.12).

Comparison of particle filters

Now, we "compare" the griddy particle filters obtained in this example, with the particle filters from Section 3.2. For $\beta = 0.7$, $\phi = 0.5$, and $\sigma^2 = 0.3$ denote $\alpha_t^1, \ldots, \alpha_t^{1000}$ one replicate of the particles obtained in Section 3.2, and $\tilde{\alpha}_t^1, \ldots, \tilde{\alpha}_t^{M_t}$ the griddy particle filters obtained in the example of this section. For $2 \leq t \leq 60$, in Figure 3.4, the pair of symbols "$\times$" shows the minimum and maximum, respectively, of $\alpha_t^1, \ldots, \alpha_t^{1000}$; the symbol "$-$" is the average of these particles; the extremes of the upper vertical line correspond to the minimum and maximum respectively, of $\tilde{\alpha}_t^1, \ldots, \tilde{\alpha}_t^{M_t}$; the large square is the approximation α_t^0 to the mean in (3.13); and the square is the average of the particles $\tilde{\alpha}_t^1, \ldots, \tilde{\alpha}_t^{M_t}$. At time t, the average of the two sets of particles are close. The extremes of the griddy particle filters depend to a

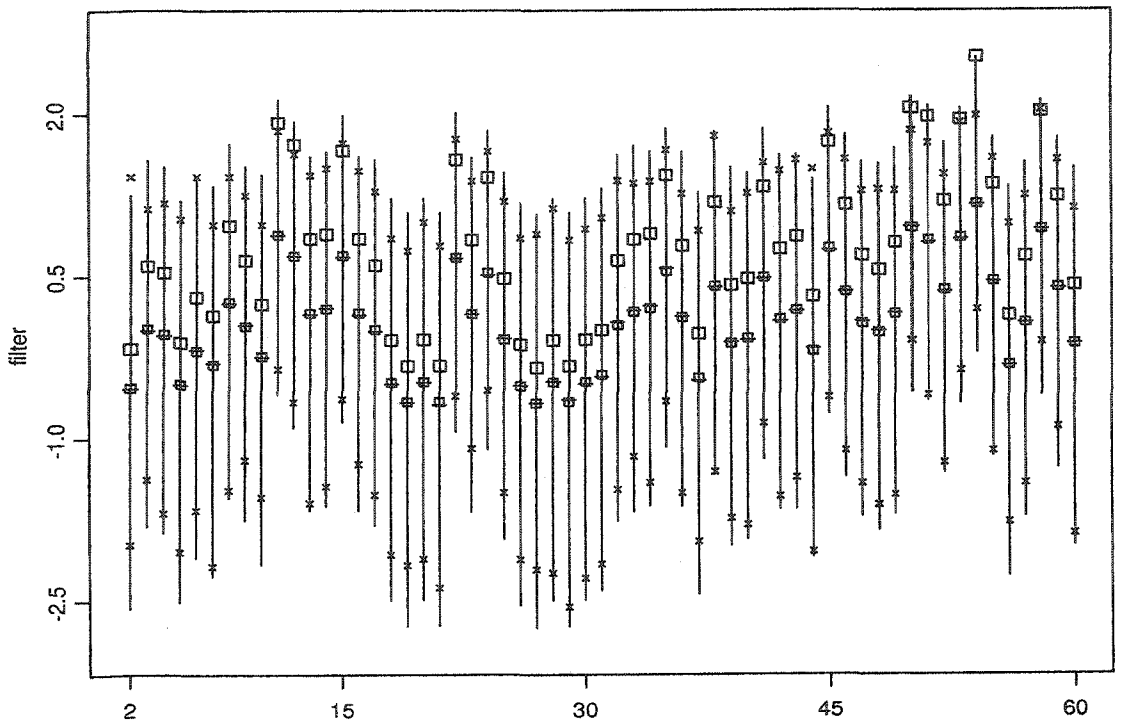

Figure 3.4: Accept-reject and griddy particle filters for $\phi = 0.5$. For each value of $2 \leq t \leq 60$ the symbols "×" are the minimum and maximum of the accept-reject particles; the symbol "–" is the average of these particles; the vertical line shows the range of the griddy particle filters, the small square is the average of these particles and the large square is the approximation to the mean in (3.11).

large extend on the value of q. The "initializer" α_t^0 of the griddy of points has little impact on these extremes. This initializer in fact does a good job. The discrepancy between the extremes of both sets of particles, is due mainly to the randomness of the accept-reject particle filters. Overall, the two set of particle filters give comparable results in describing the filter density at time t, even though the griddy particle filters is much faster. □

Example 2. (*Stochastic volatility model*) Consider the basic stochastic volatility model in (2.34). The mean μ_t^j of the state $\alpha_t|(\alpha_{t-1} = \alpha_{t-1}^j)$ is $\mu_t^j = \gamma + \phi\alpha_{t-1}^j$. Proceeding as in Example 1, we obtain

$$d_j^0 = q_j, \quad \text{and} \quad d_j^1 = q_j(\mu_t^j - \sigma^2/2),$$

where

$$\log(q_j) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\mu_t^j + \frac{1}{8}\sigma^2 - \frac{1}{2}y_t^2 e^{-\mu_t^j + \sigma^2/2}.$$

The initial value in (3.11) becomes

$$\alpha_t^0 = \gamma - \sigma^2/2 + \phi \sum_{j=1}^{M_{t-1}} \alpha_{t-1}^j q_j \pi_{t-1}^j / \sum_{j=1}^{M_{t-1}} q_j \pi_{t-1}^j. \tag{3.14}$$

Using $\gamma = -0.368$, $\phi = 0.95$ and $\sigma^2 = 0.0676$, one realization $y_1, \ldots, y_{200}$ from this process was generated. In Figure 3.5 we show estimations of the observed likelihood of this process computed in a grid of points of ϕ ($\gamma = -0.368$ and $\sigma^2 = 0.0676$). In this figure, the thin solid line is the AL estimate. The dotted lines are the minimum and maximum of 100 replicates of the estimate in (2.36) using $N = 1000$. The dashed line is the average of these replicates and the thick solid line is the estimate obtained

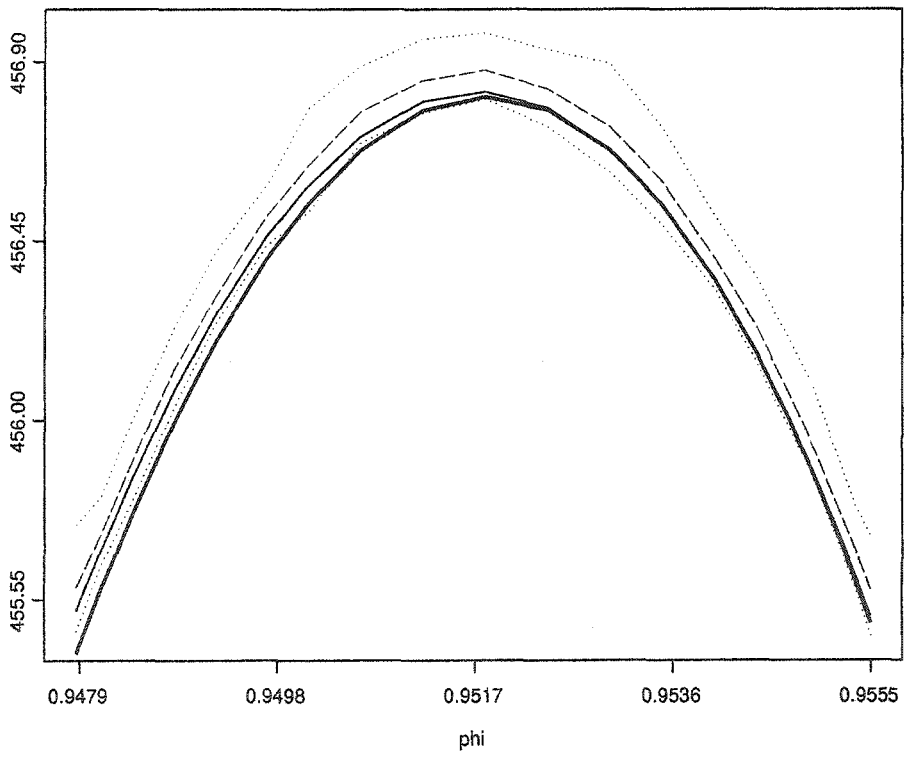

Figure 3.5: (*Griddy particle filters*) For a grid of values of ϕ, the logarithm of the estimation of the likelihood of the SVM are shown. For the thin solid line, estimates were obtained using (2.12). The pair of dotted lines and the dashed line are the minimum, maximum and average of 100 replicates using (2.36). The thick solid line is the estimate obtained using (3.12).

when (3.12) is used. The particle filters were obtained using $q = 0.01$; $M_0 = 25$; and $S^2 = \sigma^2$.

The average of the 100 importance sampling replicates obtained with $N = 1000$ shown in Figure 3.5 is an importance sampling estimate of the likelihood which has a Monte Carlo error 10 times smaller than the Monte Carlo error shown by the pair of dotted lines. Thus, the unknown likelihood must be reasonable close to the dashed line. Since the shapes of the estimations of the likelihood are similar, for N large, estimates that maximize (2.36) and (2.12) are expected to be close. For values of ϕ less than 0.952, the approximate likelihood in (2.12) is closer to the dashed line than the estimate approximation obtained with the griddy particle filters. For the stochastic volatility model, as we will see in the implementation of the auxiliary particle filters in Section 3.4, this is in fact a deficiency of particle filters. □

3.4 Auxiliary Particle Filters

The SIR algorithm described in Subsection 2.2.6 can be used to draw from the empirical filtering density in (3.4). Let $\alpha_t^1, \ldots, \alpha_t^T$ be a sample from $g(\alpha_t|y_{1:t-1}, \psi)$, a proposal density for the empirical filtering density. For T large, a sample $\tilde{\alpha}_t^1, \ldots, \tilde{\alpha}_t^M$ from the discrete distribution α_t^k, $k = 1, \ldots, T$, with mass proportional to

$$\hat{p}(\alpha_t^k|y_{1:t}, \psi)/g(\alpha_t^k|y_{1:t}, \psi)$$

is an (approximate) sample from the empirical filtering density. When the proposal density is the empirical prediction density in (3.3), Pitt and Shepard (1999) define $\alpha_t^1, \ldots, \alpha_t^T$ as *blind proposals*, and because the most recent observation y_t is not incorporated into the proposal density, they call the particle filtering procedure non *adapted*. In this case, the weights are proportional to $p(y_t|\alpha_t^k; \theta)$, $k = 1, \ldots, T$. An

adapted particle filtering can be slow. To see this, notice that the computation of the weights $\hat{p}(\alpha_t^k|y_{1:t},\psi)/g(\alpha_t^k|y_{1:t},\psi)$ needs T evaluations of the empirical prediction density. And so, MT evaluations of the density $p(\alpha_t|\alpha_{t-1};\boldsymbol{\lambda})$ are required. For M and T large, this is generally not feasible.

To make the adaptation of particle filtering procedures feasible, Pitt and Shephard (1999) introduce *auxiliary particle filtering*, a procedure based on the joint distribution

$$p(\alpha_t, j|y_{1:t};\psi) \propto p(y_t|\alpha_t;\boldsymbol{\theta})p(\alpha_t|\alpha_{t-1}=\alpha_{t-1}^j;\boldsymbol{\lambda})\pi_{t-1}^j, \quad j=1,\ldots,M, \tag{3.15}$$

where j is an index of the mixture in (3.3). They generate $\alpha_t^1,\ldots,\alpha_t^M$ by discarding the second component in a sample $(\alpha_t^1,j_1),\ldots,(\alpha_t^M,j_M)$ generated from $p(\alpha_t,j|y_{1:t};\psi)$. In this setup, $\pi_t^j = 1/M, j=1,\ldots,M$.

Example 1. (*Stochastic volatility model*) Consider the basic stochastic volatility model from the Example 2 of Section 3.3. Pitt and Sheppard (1999) implement the auxiliary particle filtering to this model using an "adapted" SIR procedure. For this implementation, the special structure of the model is "exploited."

Denote $g(y_t|\alpha_t,\mu_t^j)$ the exponent of the first-order Taylor expansion of $\log p(y_t|\alpha_t)$ around the mean μ_t^j of $\alpha_t|(\alpha_{t-1}=\alpha_{t-1}^j)$. Then,

$$\log g(y_t|\alpha_t,\mu_t^j) = -\log(2\pi) - \frac{1}{2}y_t^2(1+\mu_t^j)e^{-\mu_t^j} - \frac{1}{2}(1-y_t^2e^{-\mu_t^j})\alpha_t.$$

Hence,

$$g(y_t|\alpha_t,\mu_t^j)p(\alpha_t|\alpha_{t-1}^j;\gamma,\phi,\sigma^2) = g(y_t|\mu_t^j)\phi(\alpha_t;\tilde{\mu}_t^j,\sigma^2),$$

where

$$\begin{aligned}\log g(y_t|\mu_t^j) &= -\log(2\pi) + \{(\tilde{\mu}_t^j)^2 - (\mu_t^j)^2\}/(2\sigma^2) - \frac{1}{2}y_t^2(1+\mu_t^j)e^{-\mu_t^j}\\ \tilde{\mu}_t^j &:= \mu_t^j + \frac{\sigma^2}{2}(y_t^2e^{-\mu_t^j}-1).\end{aligned}$$

To sample from $p(\alpha_t, j|y_{1:t-1}; \gamma, \phi, \sigma^2)$, Pitt and Shephard implement the SIR algorithm using the proposal density $g(\alpha_t, j|y_{1:t}; \gamma, \phi, \sigma^2)$ given, up to a proportionality constant, by $g(y_t|\alpha_t; \mu_t^j)p(\alpha_t|\alpha_{t-1}^j; \gamma, \phi, \sigma^2)$. Then,

$$\begin{aligned} g(\alpha_t, j|y_{1:t}; \gamma, \phi, \sigma^2) &\propto g(y_t|\alpha_t; \mu_t^j)p(\alpha_t|\alpha_{t-1}^j; \gamma, \phi, \sigma^2) \\ &= g(y_t|\mu_t^j)\phi(\alpha_t; \tilde{\mu}_t^j, \sigma^2) \end{aligned} \tag{3.16}$$

In the first phase of the SIR algorithm, a sample $(\alpha_t^{j_k}, j_k), k = 1, \ldots, T$ from the proposal density in (3.16) is drawn and the weights $w_t^k = p(y_t|\alpha_t^{j_k})/g(y_t|\alpha_t^{j_k}; \mu_t^{j_k})$, $k = 1, \ldots, T$ are computed. To draw from (3.16), first draw j with probability proportional to $g(y_t|\mu_t^j)$, and then α_t from $\phi(\alpha_t; \tilde{\mu}_t^j, \sigma^2)$. In the second phase, a sample of size M is drawn from $(\alpha_t^{j_k}, j_k)$ with mass proportional to w_k, $k = 1, \ldots, T$. Then, the first components of this sample are particle filters with common mass $1/M$.

Using the realization $y_1, \ldots, y_{200}$ from Example 2 of Section 3.3, $\gamma = -0.368$, $\phi = 0.955$, and $\sigma^2 = 0.0676$, the SIR algorithm was implemented with $T = 5000$ and $M = 1000$. For $2 \leq t \leq 60$, some results of the algorithm are shown in Figure 3.6. In the middle panel, the pair of solid lines shows the 10-th and 90-th quantiles, respectively, of the weights w_t^j, $j = 1, \ldots, 5000$. In the bottom panel, the lower and upper symbols "×" are the minimum and maximum, respectively, of the auxiliary particle filters. The symbol "−" is the average of these particles. The vertical lines show the range of the griddy particle filters and the square is the average of these particles.

As seen in the bottom panel of Figure 3.6, at $t = 10$, the auxiliary particle filters "collapse" to a small subset of the support of the filter variable. As time increases, the particles gradually "improve", until a collapse occurs again at time $t = 17$.

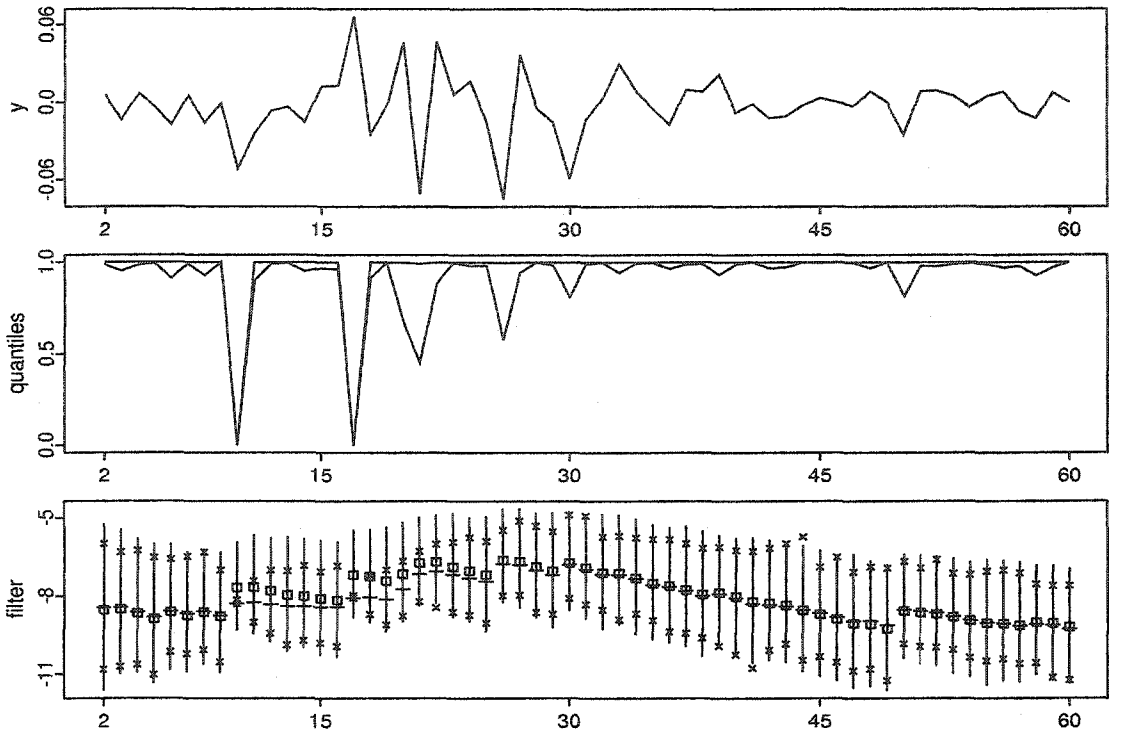

Figure 3.6: *Top*: Scatter plot of the observations for $2 \leq t \leq 60$. *Middle*: The solid lines show the 10-th and 90-th quantiles, respectively, of the SIR weights for $\phi = 0.955$, $T = 5000$ and $M = 1000$. *Bottom*: The pair of symbols "×" is the minimum and maximum of the auxiliary particle filters, the symbol "−" is the average of these particles. The vertical line shows the range of the griddy particle filters and the square is the average of these particles.

Since this particle filtering is adapted, this behavior is "unexpected". As seen from the top panel of Figure 3.6, this behavior does not appear to be connected with outliers. Notice that at $t = 10$ and $t = 17$, a large proportion of the weights w_t^j, $j = 1, \ldots, 5000$ are small. This means that "most of the time", the proposal function $g(\alpha_t, j|y_{1:t}; \gamma, \phi, \sigma^2)$ exceeds $\hat{p}(\alpha_t, j|y_{1:t}; \gamma, \phi, \sigma^2)$. This fact alone would not be a problem (to see this, suppose that the "ideal" proposal function $K\hat{p}(\alpha_t, j|y_{1:t}; \gamma, \phi, \sigma^2)$ is used. Then all the weights are equal to 1/K and for K large, these would be close to zero). In this example, $-\log p(y_t|\alpha_t)$ is convex, thus $g(y_t|\alpha_t; \mu_t^j) \geq p(y_t|\alpha_t)$, with equality at μ_t^j. Hence, the weights can take values in the interval $(0, 1]$. The overall result is that in spite of the fact that the proposal density in (3.16) proposed by Pitt

and Sheppard is adapted, it fails at times $t = 10$ and 17. □

Example 2. The stochastic volatility model from Example 1 is considered again. To select M particle filters at time t with common mass $1/M$, a new adapted SIR algorithm is implemented using a proposal function given below.

To start, let $\hat{\alpha}_t^j$ be the mode of $p(y_t|\alpha_t)p(\alpha_t|\alpha_{t-1} = \alpha_{t-1}^j; \gamma, \phi, \sigma^2)$. A similar procedure used to obtain (2.9) shows that

$$\begin{aligned} p(y_t|\alpha_t)p(\alpha_t|\alpha_{t-1}^j; \gamma, \phi, \sigma^2) &= \frac{\hat{\sigma}_t^j}{\sigma} e^{\hat{h}_t^j - \frac{1}{2\sigma^2}(\hat{\alpha}_t^j - \mu_t^j)^2} \phi(\alpha_t; \hat{\alpha}_t^j, \hat{\sigma}_t^{2j}) e^{R(\alpha_t; \hat{\alpha}_t^j)} \\ &= g(j|y_t)\phi(\alpha_t; \hat{\alpha}_t^j, \hat{\sigma}_t^{2j}) e^{R(\alpha_t; \hat{\alpha}_t^j)}, \end{aligned} \tag{3.17}$$

where $\hat{h}_t^j := \log p(y_t|\alpha_t = \hat{\alpha}_t^j)$; $R(\alpha_t; \hat{\alpha}_t^j)$ is the remainder of the second order Taylor expansion of $\log p(y_t|\alpha_t)$ around $\hat{\alpha}_t^j$; $\mu_t^j = \gamma + \phi\alpha_{t-1}^j$;

$$\hat{\sigma}_t^{2j} := (\frac{1}{\sigma^2} + \frac{1}{2} y_t^2 e^{-\hat{\alpha}_t^j})^{-1}; \quad \text{and} \quad g(j|y_t) := \frac{\hat{\sigma}_t^j}{\sigma} e^{\hat{h}_t^j - \frac{1}{2\sigma^2}(\hat{\alpha}_t^j - \mu_t^j)^2}.$$

Define the new proposal function $g(\alpha_t, j|y_{1:t}; \gamma, \phi, \sigma^2)$ to be the function that results when the last factor in (3.17) is omitted. Then, in the first phase of the SIR algorithm, R draws $(\alpha_t^{j_k}, j_k)$ from this function are obtained and the weights

$$w_t^k = p(y_t|\alpha_t^{j_k})p(\alpha_t|\alpha_{t-1}^{j_k}; \gamma, \phi, \sigma^2)/g(\alpha_t, j_k|y_{1:t}; \gamma, \phi, \sigma^2), \quad k = 1, \ldots, R,$$

are computed. The log of the weight w_t^k is given by

$$\log w_t^k = -\frac{1}{2} y_t^2 e^{-\hat{\alpha}_t^{j_k}} (e^{\hat{\alpha}_t^{j_k} - \alpha_t^{j_k}} - 1 + (\alpha_t^{j_k} - \hat{\alpha}_t^{j_k}) - \frac{1}{2}(\alpha_t^{j_k} - \hat{\alpha}_t^{j_k})^2). \tag{3.18}$$

To draw (α_t, j) from the proposal function, first draw j from the discrete distribution that puts mass proportional to $g(j|y_t)$ to the integers $j = 1, \ldots, T$ and then draw α_t from the normal density with mean $\hat{\alpha}_t^j$ and variance $\hat{\sigma}_t^{j2}$. In the second phase, a sample of size M from the discrete distribution that puts mass proportional to w_t^k to the point $(\alpha_t^{j_k}, j_k)$, $k = 1, \ldots, T$, is obtained.

Using the same realization $y_1, \ldots, y_{200}$ as in the previous example, the SIR algorithm was implemented with $R = 5000$ and $M = 1000$ at $\gamma = -0.368$, $\phi = 0.955$ and $\sigma^2 = 0.0676$. Some results of the implementation are shown in Figure 3.7. In the top panel, the vertical line is the standard error of the griddy particle filters at time t, and the symbol "×" is the standard error of the auxiliary particle filters. In the bottom panel, the large square is the approximation α_t^0 to the mean in (3.14). The remaining components of the middle and bottom panels are as in Figure 3.6. As seen in (3.18), the second-phase weights of the SIR algorithm attain the value 1. In the middle panel of this figure, most of the weights are close to 1. In general, the griddy particle filters and auxiliary particle filters are in agreement. The discrepancy between the estimates of the variance (top panel) is due to (the large) Monte Carlo

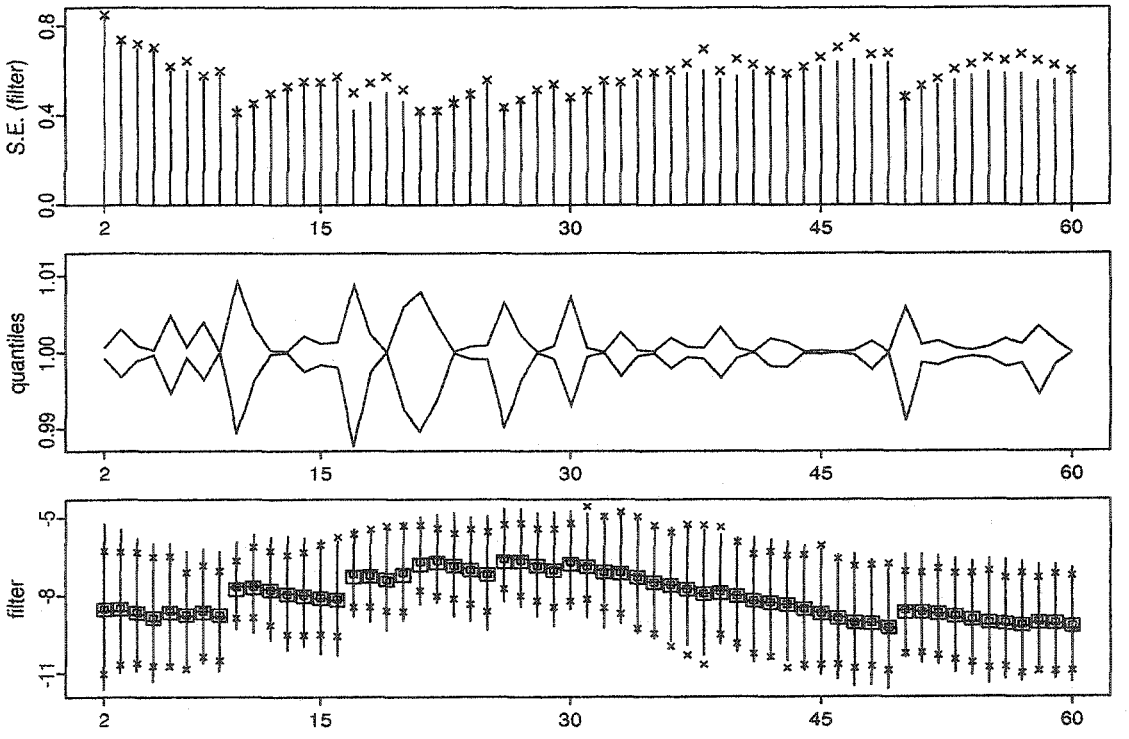

Figure 3.7: *Top*: The vertical lines show the standard error of the griddy particle filters and the symbol "×"" the standard error of the auxiliary particle filters from Example 2. The middle and bottom panels are as in Figure 3.6.

error of the variance computed with the auxiliary particle filters.

In Figure 3.8, the log-likelihood of the SVM is computed in a grid of values of ϕ. The thin and thick solid lines, and dashed line are as in Figure (3.5); the dotted-dashed line is the average of 100 replicates of (3.8) based on auxiliary particle filters obtained via the new SIR implementation with $T = 5000$ and $M = 1000$. The t-th factor of the likelihood is computed using (3.7) with $N = 1000$. This time, the Monte Carlo error (not shown in Figure 3.8) of the estimate of the likelihood based on the auxiliary particle filters is much larger than that computed via (2.36), and larger than the Monte Carlo error obtained when the estimation of the likelihood is based on the accept-reject particle filters from Section 3.2. In the latter case, the increase in the Monte Carlo error is not unexpected since the auxiliary particle filters

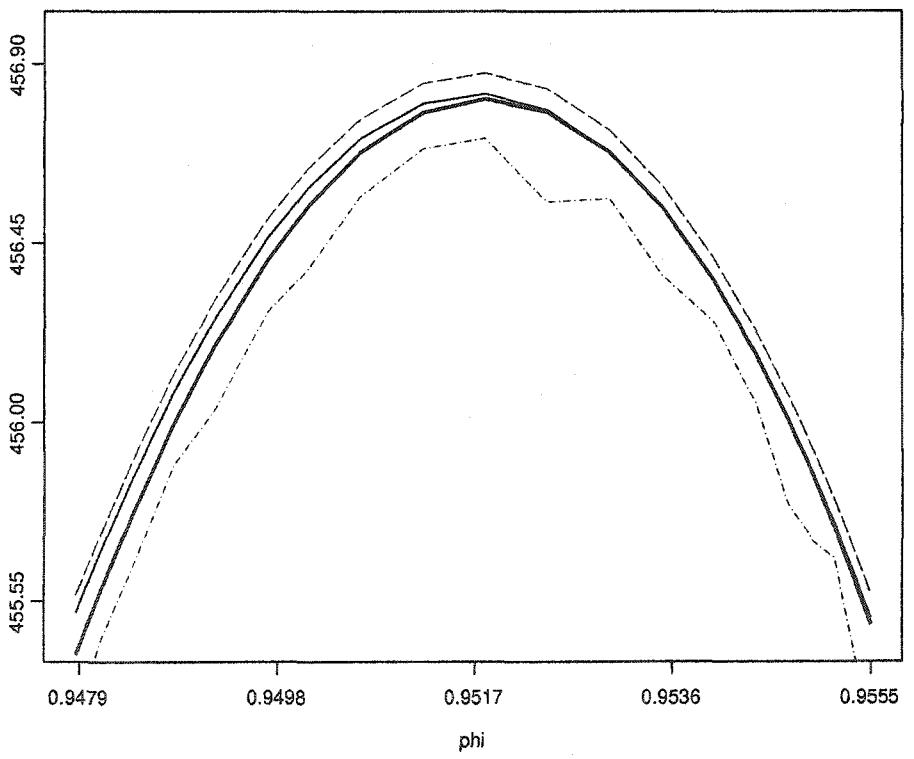

Figure 3.8: (*Auxiliary particle filters*) For the grid of values of ϕ as in Figure 3.5, the log likelihood of the stochastic volatility model is shown. The thin and thick solid lines, and dashed line are as in Figure (3.5), the dotted-dashed line is the average of 100 replicates of (3.8) based on auxiliary particle filters.

are drawn from (3.15), which has one more variable than the number of variables of (3.4), where the accept-reject particle filters are drawn. This additional variable is in fact the auxiliary variable j that appears in (3.15).

3.5 Conclusions

In this chapter, approximations to the likelihood of the state-space model based on the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ of the state vector (Chapter 2) have been compared with approximations based on the sequence of filter densities $p(\alpha_t|y_{1:t};\boldsymbol{\psi})$, $t = 1,\ldots,n$. To approximate the filtering densities, particle filtering, a procedure suggested independently by various authors in the early 90's, is used. Three implementations of particle filtering are provided in the numerical examples of this chapter. The first, implemented via accept-reject, gave comparable results with the importance sampling procedure of Chapter 2. However, the rejection rate can be very high, due to the lack of flexibility in the selection of the proposal density. A faster approximation to the likelihood is obtained with the griddy particle filters, a new particle filtering implementation based on the Griddy Gibbs sampler procedure of Ritter and Tanner (1992). This time, the quality of the estimation depends on the accuracy of the empirical prediction density as an estimate of the prediction density. In the two numerical examples, the griddy particle filters give estimates close to the approximate likelihood from Chapter 2. However, the latter is substantially faster. The last approximation to the likelihood is based on the auxiliary particle filters introduced by Pitt and Shepard (1999) to allow adaptation of existing particle filtering procedures. The approximation to the likelihood obtained with either the accept-reject procedure or auxiliary particle filters have large Monte Carlo error and

it is many times slower than the approximation based on the importance sampling of Chapter 2. For the Poisson state-space model, for large N, the estimates of the likelihood computed via the importance sampling procedure of Chapter 2 and the sequential procedures of this chapter are close. This is not the case, however, for the stochastic volatility model. This discrepancy is because the latter approach "accumulates" error through time, resulting in large Monte Carlo error on the estimated likelihood. Thus, the method of "multivariate" importance sampling from Chapter 2 to estimate the likelihood of the state space models considered in this dissertation, outperforms the particle filtering approach.

CHAPTER 4

Multiple Linear Regression with Inequality Linear Constraints

In the classical linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\mathbf{Y} = [Y_1, ..., Y_n]^T$ is the data vector, $\mathbf{X}$ is an $n \times k$ $(n > k)$ design matrix having full rank, $\boldsymbol{\epsilon}$ is a vector of errors that are independent and $N(0, \sigma^2)$ distributed, and $\boldsymbol{\beta}$ is the vector of regression parameters, the maximum likelihood estimate of $\boldsymbol{\beta}$, which coincides with the least squares estimator, is multivariate normal. Often times, there are applications in which inequality constraints are placed on $\boldsymbol{\beta}$. For example, in hyperspectral imaging, the spectrum signature of a mixed pixel can be analyzed with the model in (4.1), where the columns of $\mathbf{X}$ are the spectra of the k materials in the pixel (see Manolakis and Shaw, 2002). Due to physical considerations, the components of $\boldsymbol{\beta}$, the abundance parameters, are considered to be non-negative, i.e., $\boldsymbol{\beta} \geq 0$. This example fits into the more general framework where the vector $\boldsymbol{\beta}$ is subject to a set of inequality linear constraints which can be written as

$$\mathrm{B}\boldsymbol{\beta} \leq \mathrm{b}. \tag{4.2}$$

As long as the set defined in (4.2) has positive Lebesgue measure, there is a positive probability that the least squares estimator of $\boldsymbol{\beta}$ may not satisfy all these con-

straints. When it does, it coincides with the maximum likelihood estimate as in the unconstrained case. However, except in simple cases, it is very difficult to obtain sampling properties of the inequality restricted least squares estimator of β. Judge and Takayama (1966) and Liew (1976) give the inequality constrained least-squares (ICLS) estimate of β using the Dantzig-Cottle algorithm. The ICLS estimator reduces to the ordinary least squares estimator for a sufficiently large sample. Conditioned on knowing which constraints are binding and which are not they compute an untruncated covariance matrix of the ICLS estimator. Geweke, (1986) points out that this variance matrix is incorrect, since in practice it is not known ahead of time which constraints will be binding.

In this dissertation, we will consider a Bayesian approach to this constrained inference problem. Geweke (1986) uses a prior that is the product of a conventional uninformative distribution and an indicator function representing the inequality constraints. The posterior distribution and expected values of functions of interest are then computed using importance sampling. In this case, an importance function is easy to find due to the simplicity of the prior. This method can be extremely slow especially when the truncation region has small probability.

Gelfand et al. (1992) suggest an approach to routinely analyze problems with constrained parameters using the Gibbs sampler, a Monte Carlo Markov chain (MCMC) technique. Let $\mathcal{D}$ denote the data and θ a parameter vector with some prior distribution, and suppose it is difficult or impossible to draw samples from the posterior distribution $p(\theta|\mathcal{D})$. The Gibbs sampler, introduced by Geman and Geman (1984) in the context of image restoration, provides a method for generating samples from $p(\theta|\mathcal{D})$. Suppose θ can be partitioned as $\theta = (\theta_1, \ldots, \theta_q)$, where the θ_i's are either uni- or multidimensional and that we can simulate from the conditional posterior

densities $p(\boldsymbol{\theta}_i|\mathcal{D}, \boldsymbol{\theta}_j, j \neq i)$. The Gibbs sampler generates a Markov chain by cycling through $p(\boldsymbol{\theta}_i|\mathcal{D}, \boldsymbol{\theta}_j, j \neq i)$. In each cycle, the most recent information updates the posterior conditionals. Starting from some $\boldsymbol{\theta}^{(0)}$, after t cycles we have a realization $\boldsymbol{\theta}^{(t)}$ that under regularity conditions (Gelfand and Smith, 1990), approximates a drawing from $p(\boldsymbol{\theta}|\mathcal{D})$ for large t. O'Hagan (1994), Roberts (1996), Gilks and Roberts (1996) comment that the rate of convergence depends on the degree of posterior correlation in the $\boldsymbol{\theta}$'s.

Geweke (1996) implements the Gibbs sampler to the problem of multiple linear regression with at most k independent inequality linear constraints given by

$$\mathbf{c} \leq \mathbf{B}\boldsymbol{\beta} \leq \mathbf{d}, \tag{4.3}$$

where $\mathbf{B}$ is a square matrix of full rank, $\mathbf{c} < \mathbf{d}$ and the elements of $\mathbf{c}$ and $\mathbf{d}$ are allowed to be $-\infty$ and $+\infty$, respectively. However, this implementation may suffer from poor mixing (i.e., the chain does not move rapidly through the "entire" support of the posterior distribution). In our implementation we do not impose any limitation on the number of constraints given in (4.2). A major difference however, is that our implementation has faster mixing, requiring substantially fewer iterations of the Markov chain. Notice that the constraints given in (4.3) can be easily rewritten in the form given in (4.2).

The organization of this chapter is as follows. In Section 4.1 we provide a Bayesian framework for the multiple linear regression where the regression parameters are subject to the constraints in (4.2). In Section 4.2 we list standard results for the truncated multivariate normal distribution that are used in this chapter and provide an efficient Gibbs sampler implementation to this distribution. Through an example where the constraints can be written as in (4.3) we compare our implementation with that of Geweke's. In Section 4.3 we use the implementation from Section 4.2 to

provide an implementation of the Gibbs sampler algorithm to the model in Section 4.1 and apply the procedure to two datasets. One is the rental data analyzed by Geweke (1986, 1996) where the regression coefficients are subject to a set of inequality linear constraints that can be written as in (4.3). The other is aggregate data of three leading brands of cigarettes. For this problem, equality linear constraints are needed in addition to inequality linear constraints and the number of inequality linear constraints exceeds the number of regression coefficients. In Section 4.4 we summarize our findings.

4.1 Constrained Linear Regression

In this section we construct a Bayesian model to the multiple linear regression given in (4.1) where the parameters satisfy the constraints in (4.2). Before doing this, we introduce some notation. If R is a subset of $\Re^k$ having positive Lebesgue measure, we call the random k-vector $\mathbf{Y}$ *truncated normal* and write $\mathbf{Y} \sim N_R(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its probability density function is proportional to $\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) I_R(\mathbf{x})$, where $\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the k-variate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $I_R(.)$ is the indicator function for R.

Now, the inequality linear constraints in (4.2) define a subset of $\Re^k$ given by

$$T := \{\boldsymbol{\beta} \in \Re^k : \mathbf{B}\boldsymbol{\beta} \leq \mathbf{b}\}. \tag{4.4}$$

Notice that the model in (4.1) describes the conditional distribution of $\mathbf{Y}$ given the vector of parameters $\boldsymbol{\theta} := (\boldsymbol{\beta}, \sigma^2)$, consisting of the coefficients of regression and the common variance of the noise errors. Now assume the prior for $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\beta} \sim N_T(\boldsymbol{\mu}_0, \sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}), \tag{4.5}$$

$$\sigma^2 \sim \mathrm{IG}(\nu, \lambda), \tag{4.6}$$

where β and σ^2 are independently distributed, σ_0^2, ν and λ are known positive scalars and μ_0 is a known vector. If $p(\beta, \sigma^2|\mathbf{y})$ denotes the posterior distribution of θ given the observed vector $\mathbf{y}$, then,

$$p(\beta, \sigma^2|\mathbf{y}) \propto \mathrm{L}(\beta, \sigma^2; \mathbf{y}) p(\beta) p(\sigma^2) \tag{4.7}$$

where $\mathrm{L}(\beta, \sigma^2; \mathbf{y})$ is the likelihood function based on the data $\mathbf{y}$ from the model in (4.1). A sample from the posterior density $p(\beta, \sigma^2|\mathbf{y})$ will allow us to compute posterior quantities, such as means, variances, probabilities, and so on. In Section 4.3 below we describe how to obtain such a sample.

4.2 Truncated Multivariate Normal Distribution

In order to have an efficient Gibbs sampler implementation to the multiple linear regression problem with inequality linear constraints as given in (4.2), it is imperative to have an efficient sampler to the truncated multivariate normal distribution. Before pursuing this objective we begin by developing two properties of the truncated multivariate normal distribution and then propose an implementation of the Gibbs sampler to the multivariate normal distribution subject to a set of inequality linear constraints. A key feature of this implementation is the construction of a set of variables that are independent when the constraints are ignored. Using the first example from Geweke (1991), the performance of our implementation is compared with that of Geweke's.

For a multivariate normal random vector $\mathbf{X}$, all linear transformations and conditional distributions of $\mathbf{X}$ are normal. It turns out that for a truncated normal distributed vector, these closure properties remain valid. That is, a linear transformation of a truncated normal vector is also truncated normal and so are the

one-dimensional conditional distributions. These conditional distributions play a key role in the implementation of the Gibbs sampler. The specifics are as follows:

Result 1 (a) Suppose $\mathbf{X} \sim N_R(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $R \in \Re^k$ has positive Lebesgue measure, and $\boldsymbol{\Sigma}$ is positive definite. Let $\mathbf{Y} := \mathbf{AX}$, where $\mathbf{A}$ is a matrix of full rank of dimension $r \times k$ with $r \leq k$. Then,

$$\mathbf{Y} \sim N_T(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T), \quad T := \{\mathbf{Ax} : \mathbf{x} \in R\}. \tag{4.8}$$

(b) Partition $\mathbf{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ X_k \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \mu_k \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_1^T & \sigma_{kk} \end{bmatrix}. \tag{4.9}$$

Then,

$$X_k | \mathbf{X}_1 = \mathbf{x}_1 \sim N_{R_k}(\mu_k^*, \sigma_{kk}^*), \tag{4.10}$$

where

$$\mu_k^* = \mu_k + \boldsymbol{\Sigma}_1^T \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \tag{4.11}$$

$$\sigma_{kk}^* = \sigma_{kk} - \boldsymbol{\Sigma}_1^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_1, \tag{4.12}$$

$$R_k := \{x_k \in \Re : (\mathbf{x}_1, x_k) \in R\}. \tag{4.13}$$

The proof of (a) is immediate from the form of the density function for truncated normal random vectors. To prove (b) the expressions for the inverse of a partitioned symmetric matrix (e.g., Hocking, 1996) are used from which the result is immediate. □

Gibbs Sampler Algorithm

Suppose $\boldsymbol{\theta}$ is a vector of parameters with posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$, where $\mathcal{D}$ denotes the data. Partition $\boldsymbol{\theta}$ as $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$, where the $\boldsymbol{\theta}_i$'s are either uni- or

multidimensional in such a way that we can simulate from the conditional posterior densities $p(\theta_i|\mathcal{D}, \theta_j, j \neq i)$. The basic Gibbs sampler starts with an initial value $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_q^{(0)})$ from the support of $p(\theta|\mathcal{D})$ and then generates $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_q^{(t)})$, t=1, 2, ..., recursively as follows:

Generate $\theta_1^{(t)}$ from $p(\theta_1|\mathcal{D}, \theta_2^{(t-1)}, \ldots, \theta_q^{(t-1)})$

Generate $\theta_2^{(t)}$ from $p(\theta_2|\mathcal{D}, \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_q^{(t-1)})$

$\vdots$

Generate $\theta_q^{(t)}$ from $p(\theta_q|\mathcal{D}, \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{q-1}^{(t)})$.

Under certain regularity conditions (e.g. Gelfand and Smith, 1990) the Markov chain $\{\theta_0, \theta_1, \theta_2, \ldots\}$ has a stationary distribution which is the posterior distribution $p(\theta|\mathcal{D})$. □

4.2.1 Gibbs sampler implementations

For comparison purposes, we first describe the implementation of the Gibbs sampler algorithm given by Geweke (1991) to a truncated normal random vector of dimension k subject to a set of at most k linearly independent inequality linear constraints. Suppose that $\mathbf{X}$ is a truncated normal random vector of dimension k, such that

$$\mathbf{X} \sim N_T(\mu, \sigma^2\Sigma), \quad T := \{\mathbf{x} \in \Re^k : \mathbf{c} \leq \mathbf{B}\mathbf{x} \leq \mathbf{d}\}, \tag{4.14}$$

where $\mathbf{c}$, $\mathbf{d}$ and $\mathbf{B}$ are as in (4.3).

The Gibbs sampler in Geweke's implementation is applied to the transformed random vector $\mathbf{Y} = \mathbf{B}\mathbf{X}$. Note that

$$\mathbf{Y} \sim N_S(\mathbf{B}\mu, \sigma^2\mathbf{B}\Sigma\mathbf{B}^T), \quad S = \{\mathbf{y} \in \Re^k : \mathbf{c} \leq \mathbf{y} \leq \mathbf{d}\}. \tag{4.15}$$

Thus, using (4.10)

$$Y_j|(Y_1 = y_1, \ldots, Y_{j-1} = y_{j-1}, Y_{j+1} = y_{j+1}, \ldots, Y_k = y_k) \sim N_{S_j}(\mu_j^*, \sigma_{jj}^*), \qquad (4.16)$$

where $S_j = \{y_j \in \Re : c_j \leq y_j \leq d_j\}$, and μ_j^* and σ_{jj}^* must be obtained as in (4.11) and (4.12), respectively. Geweke's implementations, which we call *Algorithm* TN1, is then

Algorithm TN1

Update the last component $\mathbf{y}^{(t)} = [y_1^{(t)}, y_2^{(t)}, \ldots, y_k^{(t)}]^T$ of the current Gibbs path $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_t$, as follows: For $j = 1, \ldots, k$

- draw $y_j^{(t+1)}$ from $p(y_j|y_1^{(t+1)}, \ldots, y_{j-1}^{(t+1)}, y_{j+1}^{(t)}, \ldots, y_k^{(t)})$, (4.17)

where each conditional distribution is as in (4.16). □

The next algorithm allows for the number of constraints to exceed k. Suppose now

$$\mathbf{X} \sim N_T(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}), \quad T := \{\mathbf{x} \in \Re^k : \mathbf{B}\mathbf{x} \leq \mathbf{b}\}, \qquad (4.18)$$

where the rows of the matrix $\mathbf{B}$ are not restricted to be linearly independent.

Let $\mathbf{A}$ be a square matrix of full rank, such that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and set

$$\mathbf{Z} := \mathbf{A}\mathbf{X}. \qquad (4.19)$$

From (a) of result 1, it follows that

$$\mathbf{Z} \sim N_S(\mathbf{A}\boldsymbol{\mu}, \sigma^2\mathbf{I}), \qquad S = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \Re^k, \mathbf{B}\mathbf{x} \leq \mathbf{b}\}. \qquad (4.20)$$

The set S can be rewritten in the more suggestive way,

$$S = \{\mathbf{z} \in \Re^k, \mathbf{D}\mathbf{z} \leq \mathbf{b}\}, \qquad (4.21)$$

where

$$\mathbf{D} := \mathbf{B}\mathbf{A}^{-1}. \tag{4.22}$$

Thus, the transformation in (4.19) simplifies the functional form of the truncated multivariate distribution, but not the constraints.

If $\mathbf{z}_{-j}$ denotes the vector $[z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_k]^T$ and $\boldsymbol{\alpha} := \mathbf{A}\boldsymbol{\mu}$, then by (b) of Result 1,

$$Z_j|\mathbf{z}_{-j} \sim N_{S_j}(\alpha_j,\ \sigma^2), \tag{4.23}$$

where from (4.13) and (4.21)

$$S_j = \{z_j \in \Re : \mathbf{z} \in S\} = \{z_j \in \Re : \mathbf{D}\mathbf{z} \leq \mathbf{b}\}.$$

Let $\mathbf{D}_{-j}$ be the matrix obtained from $\mathbf{D} = [\mathbf{D_1} \ldots \mathbf{D_k}]$ by removing the j-th column. Then the set S_j can be easily computed from the equation

$$S_j = \{z_j \in \Re : \mathbf{D}_j z_j \leq \mathbf{b} - \mathbf{D}_{-j}\mathbf{z}_{-j}\}. \tag{4.24}$$

Although the idea behind the transformation in (4.19) is to obtain an efficient implementation of the Gibbs sampler based on the set of k conditional distributions in (4.23), these distributions have a simple form. That is, once the transformed mean $\boldsymbol{\alpha}$ has been obtained, we do not need to use (4.11) and (4.12) to compute the k means and variances if we had not transformed.

To illustrate this process, consider the following example. Let $\mathbf{X} \sim N_T(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma})$, where

$$T = \{\mathbf{x} \in \Re^2 : \mathbf{x} \geq \mathbf{0}\}, \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 4/5 \\ 4/5 & 1 \end{bmatrix}.$$

Notice that in the notation in (4.4), $\mathbf{B} = -\mathbf{I}$ and $\mathbf{b} = \mathbf{0}$, where $\mathbf{I}$ is the identity

matrix. For the matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ -4/3 & 5/3 \end{bmatrix},$$

we obtain $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T = \mathbf{I}$. The matrix $\mathbf{D} = \mathbf{B}\mathbf{A}^{-1}$ given in (4.22), and the submatrices $\mathbf{D}_{-1}$ and $\mathbf{D}_{-2}$, are

$$\mathbf{D} = \begin{bmatrix} -1 & 0 \\ -4/5 & -3/5 \end{bmatrix}, \quad \mathbf{D}_{-1} = \begin{bmatrix} 0 \\ -3/5 \end{bmatrix}, \quad \mathbf{D}_{-2} = \begin{bmatrix} -1 \\ -4/5 \end{bmatrix}.$$

Then,

$$Z_1|\mathbf{Z}_{-1} = \mathbf{z}_{-1} \sim N_{S_1}(\mu_1,\ \sigma^2), \qquad Z_2|\mathbf{Z}_{-2} = \mathbf{z}_{-2} \sim N_{S_2}(-\frac{4}{3}\mu_1 + \frac{5}{3}\mu_2,\ \sigma^2),$$

where

$$\begin{aligned} S_1 &= \{z_1 \in \Re : \begin{bmatrix} -1 \\ -4/5 \end{bmatrix} z_1 \leq -\begin{bmatrix} 0 \\ -3/5 \end{bmatrix} z_2\} \\ &= \{z_1 \in \Re : z_1 \geq 0; z_1 \geq -\frac{3}{4}z_2\} \\ &= [\max\{0, -\frac{3}{4}z_2\}, \infty) \end{aligned}$$

$$\begin{aligned} S_2 &= \{z_2 \in \Re : \begin{bmatrix} 0 \\ -3/5 \end{bmatrix} z_2 \leq -\begin{bmatrix} -1 \\ -4/5 \end{bmatrix} z_1\} \\ &= \{z_2 \in \Re : z_2 \geq -\frac{4}{3}z_1\} \\ &= [-\frac{4}{3}z_1, \infty). \end{aligned}$$

For this example, the matrix $\mathbf{B}$ needed in (4.14) is the identity. Hence, in Algorithm TN1, the Gibbs sampler is implemented in fact to the random vector $\mathbf{X}$. In particular, using (4.11)-(4.12) in (4.16), it follows that

$$\begin{aligned} Y_1|Y_2 = y_2 &\sim N_{S_1}(\mu_1 + \frac{4}{5}(y_2 - \mu_2)\sigma^2,\ \frac{9}{25}\sigma^2), \\ Y_2|Y_1 = y_1 &\sim N_{S_2}(\mu_2 + \frac{4}{5}(y_1 - \mu_1)\sigma^2,\ \frac{9}{25}\sigma^2) \end{aligned}$$

where $S_1 = S_2 = [0, +\infty)$. □

Now, to obtain a sample from the distribution of $\mathbf{X}$ we obtain first a sample from the transformed vector $\mathbf{Z}$ in (4.19). A sample from $\mathbf{X}$ is then obtained by "undoing" this transformation. For later reference this new implementation will be called *Algorithm* TN2.

Algorithm TN2

Let $\mathbf{z}_0 \in S$ be an initial value of the sampler. The last component $\mathbf{z}^{(t)} = [z_1^{(t)}, z_2^{(t)}, \ldots, z_k^{(t)}]^T$ of the current Gibbs path $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_t$ is updated as follows:

- draw $z_j^{(t+1)}$ from $p(z_j | z_1^{(t+1)}, \ldots, z_{j-1}^{(t+1)}, z_{j+1}^{(t)}, \ldots, z_k^{(t)})$, $j = 1, \ldots, k$,
- set $\mathbf{X}^{(t+1)} = \mathbf{A}^{-1}\mathbf{Z}^{(t+1)}$,

where the conditional distribution $p(z_j | z_1^{(t+1)}, \ldots, z_{j-1}^{(t+1)}, z_{j+1}^{(t)}, \ldots, z_k^{(t)})$ is given in (4.23). □

4.2.2 Performance comparison of Algorithms TN1 and TN2

To compare the performance between Algorithms TN1 and TN2, we consider an example in which $\mathbf{X} \sim N_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & \rho \\ \rho & 0.1 \end{bmatrix}, \tag{4.25}$$

and T is the region determined by the constraints,

$$c_1 \leq X_1 + X_2 \leq d_1, \quad c_2 \leq X_1 - X_2 \leq d_2, \tag{4.26}$$

which can be written in the format in (4.14) with $\mathbf{c} = [c_1\ c_2]^T$, $\mathbf{d} = [d_1\ d_2]^T$ and

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

To provide some indication of the efficiency of his procedure, Geweke (1991) considered five configurations of truncation points of c_1, c_2, d_1 and d_2 (and $\rho = 0$). In this example we consider four configurations of truncation points of c_1, c_2, d_1 and d_2 and three values of ρ. For each configuration we apply the two Gibbs sampler implementations described above and stop after 1600 iterations. As a mean of comparison of the two implementations, the results of the Raftery and Lewis convergence diagnostic procedure for each chain are shown in Tables 4.1 and 4.2. In the general set up of the Gibbs sampler algorithm, this diagnostic, introduced by Raftery and Lewis (1992), determines the total number of iterations required to compute quantiles of functionals of $\boldsymbol{\theta}$. Also, the method gives the number of initial iterations that must be discarded to allow for "burn-in". Some specifics of the method are as follows: Let ξ be a function of the parameter vector $\boldsymbol{\theta}$. For a fixed probability s, a known q and accuracy r, suppose that we want to estimate the value of the quantile u, given by $P(\xi \leq u|\mathcal{D}) = q$ in such a way that $P(|\hat{q} - q| \leq r|\mathcal{D}) = s$, where $\hat{q}$ is an estimator of q based on the sample path of the chain.

In Tables 4.1 and 4.2, the columns labeled as "bound" are the total length needed if the components of the chain were in fact an iid sample. The column labeled as "thinning" means that after the burn-in, every k-th observation is used. In both tables we set $q = 0.5$, $r = 0.025$ and $s = 0.95$. Based on the results from these tables, we note that the convergence of Algorithm TN1 is much slower than that for Algorithm TN2. Also, the untruncated correlation ρ of X_1 and X_2 affects the performance of Algorithm TN1. In general, as the region of truncation gets small, the speed of convergence of Algorithm TN1 improves. One possible explanation of this is that the chain must cover a "small" region faster than a "large" region. On the other hand Algorithm TN2 has the advantage of providing samples that are "close" to iid

ρ	variable	thinning (k)	burn-in	Total	lower bound	dependence factor
		$-\infty < X_1 + X_2 < \infty$,		$-\infty < X_1 - X_2 < \infty$		
-0.7	X_1	23	115	58282	1537	37.92
	X_2	3	12	6726	1537	4.37
0	X_1	18	90	46206	1537	30.06
	X_2	1	2	1551	1537	1.01
0.7	X_1	22	154	72050	1537	23.96
	X_2	3	12	6549	1537	5.89
		$-10 \leq X_1 + X_2 \leq 10$,		$-10 \leq X_1 - X_2 \leq 10$		
-0.7	X_1	25	150	78650	1537	51.17
	X_2	6	24	14160	1537	9.21
0	X_1	10	60	30040	1537	19.54
	X_2	1	2	1558	1537	1.01
0.7	X_1	20	120	60380	1537	39.28
	X_2	6	24	13272	1537	8.64
		$-5 \leq X_1 + X_2 \leq 5$,		$-5 \leq X_1 - X_2 \leq 5$		
-0.7	X_1	15	75	36405	1537	23.69
	X_2	3	12	6036	1537	3.93
0	X_1	15	75	39555	1537	25.74
	X_2	1	2	1490	1537	0.97
0.7	X_1	14	70	36834	1537	23.96
	X_2	4	16	9048	1537	5.89
		$-1 \leq X_1 + X_2 \leq 1$,		$-1 \leq X_1 - X_2 \leq 1$		
-0.7	X_1	3	12	6087	1537	3.96
	X_2	1	3	1397	1537	0.91
0	X_1	2	6	3436	1537	2.24
	X_2	1	3	1312	1537	0.85
0.7	X_1	3	12	5976	1537	3.89
	X_2	1	3	1329	1537	0.86

Table 4.1: Algorithm TN1 implemented to the truncated normal random vector $[X_1, X_2]^T$ with unconstrained mean $\mathbf{0}$, $\text{var}\{X_1\} = 10$, $\text{var}\{X_2\} = 0.10$ and $\text{cor}\{X_1, X_2\} = \rho$, subject to the constraints $c_1 \leq X_1 + X_2 \leq d_1$, $c_2 \leq X_1 - X_2 \leq d_2$ with Raftery and Lewis convergence diagnostics.

samples, regardless of the size of the region of truncation or the correlation of X_1 and X_2. In fact, for the configuration $-\infty < X_1 + X_2 < \infty$, $-\infty < X_1 - X_2 < \infty$, this algorithm provides an iid sample, since the conditional distribution in (4.23) does not depend on the fixed values $\mathbf{z}_{-j}$ (e.g., see O'Hagan, 1994).

The autocorrelations of the output of a Gibbs sampler can be used to measure the performance of a simulation implementation. Chen, et al. (2000) observe that slow decay in the autocorrelations suggests slow mixing within a chain and usually slow convergence to the posterior distribution. For this example, the autocorrelations

ρ	variable	thinning (k)	burn-in	Total	lower bound	dependence factor
		$-\infty < X_1 + X_2 < \infty$,		$-\infty < X_1 - X_2 < \infty$		
-0.7	X_1	1	3	1702	1537	1.11
	X_2	1	2	1501	1537	0.98
0	X_1	1	2	1432	1537	0.93
	X_2	1	2	1582	1537	1.03
0.7	X_1	1	2	1505	1537	0.98
	X_2	1	2	1516	1537	0.99
		$-10 \leq X_1 + X_2 \leq 10$,		$-10 \leq X_1 - X_2 \leq 10$		
-0.7	X_1	1	2	1490	1537	0.97
	X_2	1	2	1566	1537	1.02
0	X_1	1	2	1509	1537	0.98
	X_2	1	2	1490	1537	0.97
0.7	X_1	1	2	1614	1537	1.05
	X_2	1	3	1689	1537	1.10
		$-5 \leq X_1 + X_2 \leq 5$,		$-5 \leq X_1 - X_2 \leq 5$		
-0.7	X_1	1	2	1505	1537	0.98
	X_2	1	2	1446	1537	0.94
0	X_1	1	2	1505	1537	0.98
	X_2	1	2	1610	1537	1.05
0.7	X_1	1	2	1642	1537	1.07
	X_2	1	2	1578	1537	1.03
		$-1 \leq X_1 + X_2 \leq 1$,		$-1 \leq X_1 - X_2 \leq 1$		
-0.7	X_1	1	2	1566	1537	1.02
	X_2	1	2	1450	1537	0.94
0	X_1	1	3	1390	1537	0.90
	X_2	1	2	1520	1537	0.99
0.7	X_1	1	2	1655	1537	1.08
	X_2	1	2	1531	1537	1.00

Table 4.2: Algorithm TN2 implemented to the truncated bivariate normal random vector $[X_1, X_2]^T$ described in Table 4.1 with Raftery and Lewis convergence diagnostics.

of X_1, X_2 and $X_1 + X_2$ for two configurations of values of ρ, $\mathbf{c}$ and $\mathbf{d}$ using the output of the Gibbs sampler obtained with Algorithm TN1 are shown in Figure 4.1. The first row of graphs contains the autocorrelations for $\rho = 0$, $-\infty < X_1 + X_2 < \infty$, and $-\infty < X_1 - X_2 < \infty$ and the second row of graphs contains the autocorrelations for $\rho = -0.7$, $-1 \leq X_1 + X_2 \leq 1$ and $-1 \leq X_1 - X_2 \leq 1$. Figure 4.2 contains the analogous autocorrelations for the output of Algorithm TN2. For the two configurations considered in Figures 4.1 and 4.2, we conclude that the mixing of the sampler from Algorithm TN2 is better than that of the sampler of Algorithm TN1. Notice that the column labeled as "dependence factor" in Tables 4.1 and 4.2

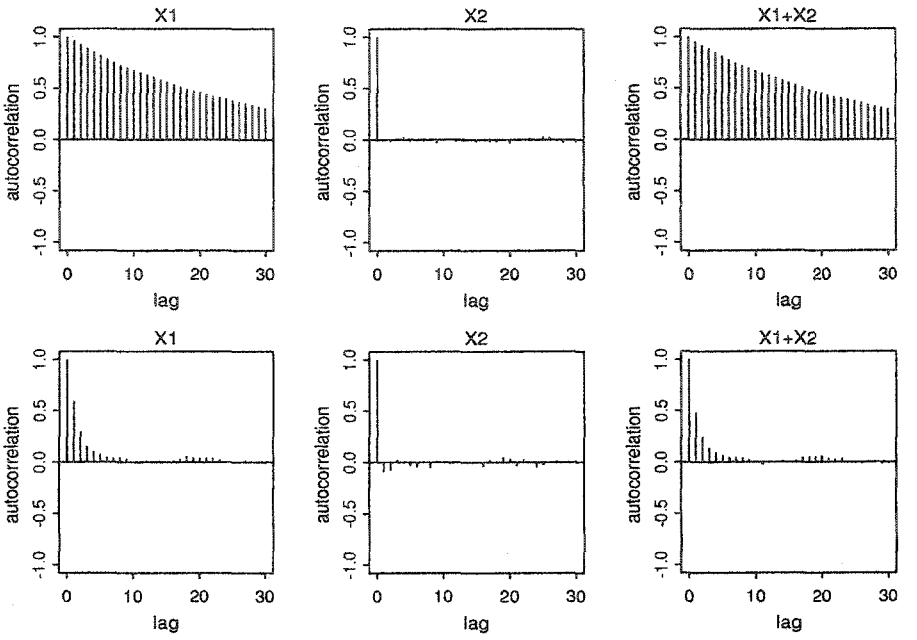

Figure 4.1: Autocorrelation plots of X_1, X_2 and $X_1 + X_2$ for two configurations of values of ρ, **c** and **d** obtained with the Gibbs sampler output of Algorithm TN1.

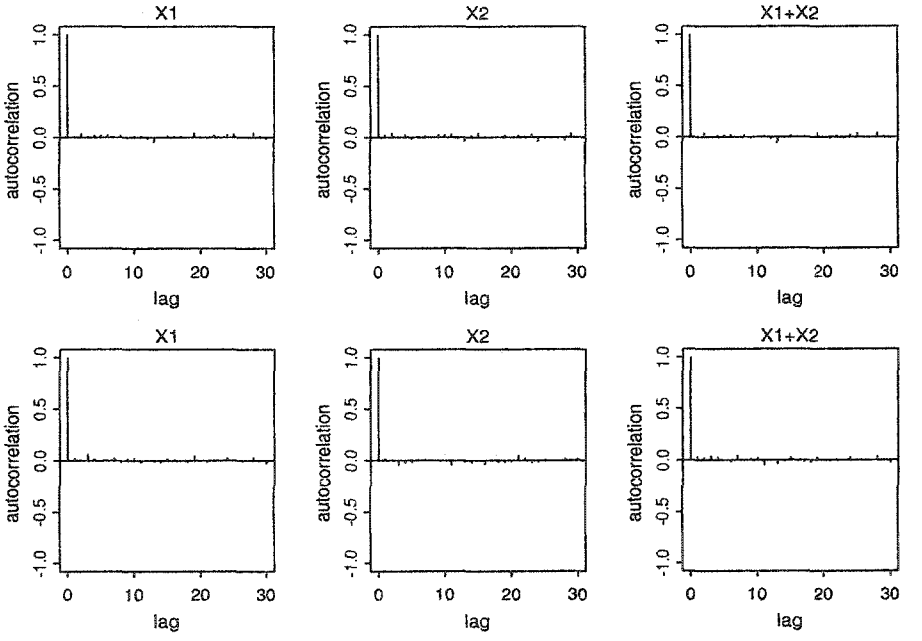

Figure 4.2: Autocorrelation plots of X_1, X_2 and $X_1 + X_2$ for two configurations of values of ρ, **c** and **d** obtained with the output of the Gibbs sampler given in Algorithm TN2.

is related with the information of the mixing provided by the autocorrelation plots in Figures 4.1 and 4.2. That is, the slower the mixing, the higher the dependence factor and vice versa.

A useful graphical tool to assess performance of the simulation procedure consists in monitoring some statistics of the output against iteration. In order to achieve stationarity, the monitored statistics must stabilize at some iteration. Thus, a monitored statistic which has not yet stabilized provides evidence for non convergence to stationarity. In Figure 4.3 we monitor the means of X_1, X_2 and $X_1 + X_2$ against the iteration number using the output of the two implementations with the configurations used in the autocorrelation plots given in Figures 4.1 and 4.2. In this figure, the solid lines are the means obtained using the output of Algorithm TN1, while the dotted lines are the means obtained with the output of Algorithm TN2. Recall that Algorithm TN2 provides an iid sample $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(t)}$ from the distribution of

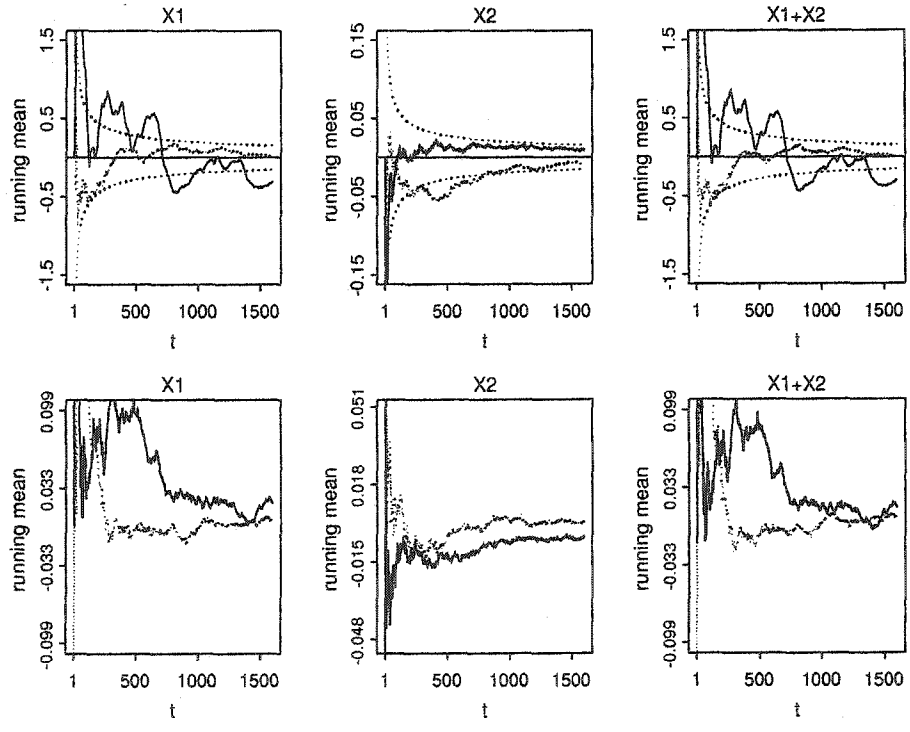

Figure 4.3: Running mean plots of X_1, X_2 and $X_1 + X_2$ for two configurations of values of ρ, **c** and **d**. The solid lines are the running means obtained with the output of sampler from Algorithm TN1. The dotted lines are the running means computed with the output of Gibbs sampler based on Algorithm TN2. In the first row, the horizontal lines show the means of the monitored statistics and the pair of dotted lines is the lower and upper limits of the true 95% confidence intervals for these statistics.

$\mathbf{X}$ when $-\infty < X_1 + X_2 < \infty, \quad -\infty < X_1 - X_2 < \infty$. Thus, the means and variances of $\bar{X}_1$, $\bar{X}_2$ and $\bar{X}_1 + \bar{X}_2$ (which are estimators of the means of X_1, X_2 and $X_1 + X_2$, respectively) are known. For example, $E(\bar{X}_1) = 0$ and $\text{var}(\bar{X}_1) = 10/t$. In the first row of Figure 4.3, the horizontal solid lines show the expected means of the monitored statistics, while the dotted lines show the upper and lower 95% confidence limits $\mp 1.96\sigma$ of the monitored statistics. We note in this figure that the monitored means stabilize earlier for the sampler provided by Algorithm TN2. In particular, in the upper left panel, with algorithm TN2, the monitored means stabilize after 500 iterations, while for Algorithm TN1 they have not yet stabilized even after 1500 iterations.

4.3 Gibbs Sampler Implementations to the Constrained Linear Regression

In this section we implement the Gibbs sampler algorithm to a Bayesian linear regression model in which the regression coefficients satisfy inequality linear constraints. When the number of constraints does not exceed the number of regression coefficients we compare our procedure with the implementation described in Geweke (1996). In addition, we show through an example how the case of equality linear constraints can be handled.

Combining the prior distribution of $(\boldsymbol{\beta}, \sigma^2)$ given in (4.5)-(4.6) with the likelihood of the model in (4.1) we have

$$\boldsymbol{\beta}|(\sigma^2, \mathbf{y}) \sim N_T(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \tag{4.27}$$

$$\sigma^{-2}|(\boldsymbol{\beta}, \mathbf{y}) \sim (SS(\boldsymbol{\beta}) + 2\lambda)^{-1}\chi^2_{n+2\nu}, \tag{4.28}$$

where $\chi^2_{n+2\nu}$ denotes a chi-squared distribution with $n + 2\nu$ degrees of freedom, T is

defined in (4.4), and

$$\begin{aligned}
\mu_1 &= \gamma\hat{\beta} + (1-\gamma)\mu_0 \\
\Sigma_1 &= \sigma^2\gamma(\mathbf{X}^T\mathbf{X})^{-1} \\
SS(\beta) &= (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) \\
\gamma &= \sigma_0^2/(\sigma_0^2+\sigma^2) \\
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.
\end{aligned}$$

For comparison purposes, we describe now the implementation of the Gibbs sampler algorithm given by Geweke (1996) to a multiple linear regression model where the regression coefficients are subject to a set of at most k linearly independent inequality linear constraints, i. e.,

$$T := \{\beta \in \Re^k : \mathbf{c} \leq \mathbf{B}\beta \leq \mathbf{d}\},$$

where $\mathbf{c}$, $\mathbf{d}$ and $\mathbf{B}$ are as in (4.3).

As in Algorithm TN1, the vector of regression coefficients β is transformed to $\eta = \mathbf{B}\beta$. Then,

$$\eta|(\sigma^2,\mathbf{y}) \sim N_S(\mathbf{B}\mu_1, \mathbf{B}\Sigma_1\mathbf{B}^T), \quad S = \{\beta \in \Re^k : \mathbf{c} \leq \beta \leq \mathbf{d}\}. \tag{4.29}$$

The full implementation of the Gibbs sampler to the vector $\theta := (\eta, \sigma^2)$ proposed by Geweke is summarized in the following Algorithm.

Algorithm CLR1

Let $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(t)}$ be the current path of the Gibbs sampler. The last component $\theta^{(t)} = (\eta_1^{(t)}, \ldots, \eta_k^{(t)}, \sigma^{2(t)})$ is updated as follows:

- Generate $\eta_1^{(t+1)}$ from $p(\eta_1|\eta_2^{(t)}, \ldots, \eta_k^{(t)}, \sigma^{2(t)}, \mathbf{y})$

- Generate $\eta_2^{(t+1)}$ from $p(\eta_2|\eta_1^{(t+1)}, \eta_3^{(t)} \ldots, \eta_k^{(t)}, \sigma^{2(t)}, \mathbf{y})$

$\vdots$

- Generate $\eta_k^{(t+1)}$ from $p(\eta_k|\eta_1^{(t+1)}, \eta_2^{(t+1)} \ldots, \eta_{k-1}^{(t+1)}, \sigma^{2(t)}, \mathbf{y})$
- Generate $\sigma^{2(t+1)}$ from $p(\sigma^2|\eta_1^{(t+1)}, \eta_2^{(t+1)} \ldots, \eta_k^{(t+1)}, \mathbf{y})$,

where, due to (4.29), for $j = 1, \ldots, k$, the distribution

$$p(\eta_j|\eta_1^{(t+1)}, \ldots, \eta_{j-1}^{(t+1)} \ldots, \eta_{j+1}^{(t)}, \ldots, \eta_k^{(t)}, \sigma^{2(t)}, \mathbf{y})$$

is univariate normal truncated below by c_i, truncated above by d_i, and its mean and variance can be found using (4.29) along with the expressions in (4.11) and (4.12). Also, $p(\sigma^2|\eta_1^{(t+1)}, \eta_2^{(t+1)} \ldots, \eta_k^{(t+1)}, \mathbf{y})$ can be obtained from (4.28). □

Now, we give a new implementation, similar to Algorithm TN2, to the case when the number of inequality linear constraints can exceed the number of regression parameters. For this case, T is given in (4.4). Let $\mathbf{A}$ be a non-singular matrix for which $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A} = \mathbf{I}$, and set

$$\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\beta}. \tag{4.30}$$

Then, from (4.8) and (4.27)

$$\boldsymbol{\eta}|(\sigma^2, \mathbf{y}) \sim N_S(\mathbf{A}\boldsymbol{\mu}_1, \sigma^2\gamma\mathbf{I}), \qquad S = \{\boldsymbol{\eta} \in \Re^k : \mathbf{D}\boldsymbol{\eta} \leq \mathbf{b}\}, \tag{4.31}$$

where $\mathbf{D} = \mathbf{B}\mathbf{A}^{-1}$ and $\mathbf{B}$ and $\mathbf{b}$ are defined as in (4.2). We implement the Gibbs sampler algorithm to the transformed vector $\boldsymbol{\theta} = (\boldsymbol{\eta}, \sigma^2)$. The details are given in Algorithm CLR2.

Algorithm CLR2

Update the last component $\boldsymbol{\theta}^{(t)} = (\eta_1^{(t)}, \ldots, \eta_k^{(t)}, \sigma^{2(t)})$ of the current path $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(1)}$, …, $\boldsymbol{\theta}^{(t)}$ of the Gibbs sampler as follows

- Generate $\eta_1^{(t+1)}$ from $p(\eta_1|\eta_2^{(t)},\ldots,\eta_k^{(t)},\sigma^{2(t)},\mathbf{y})$
- Generate $\eta_2^{(t+1)}$ from $p(\eta_2|\eta_1^{(t+1)},\eta_3^{(t)}\ldots,\eta_k^{(t)},\sigma^{2(t)},\mathbf{y})$

  $\vdots$

- Generate $\eta_k^{(t+1)}$ from $p(\eta_k|\eta_1^{(t+1)},\eta_2^{(t+1)}\ldots,\eta_{k-1}^{(t+1)},\sigma^{2(t)},\mathbf{y})$
- Generate $\sigma^{2(t+1)}$ from $p(\sigma^2|\eta_1^{(t+1)},\eta_2^{(t+1)}\ldots,\eta_k^{(t+1)},\mathbf{y})$,

where due to (4.31), for $j=1,\ldots,k$, the distribution

$$p(\eta_j|\eta_1^{(t+1)},\ldots,\eta_{j-1}^{(t+1)}\ldots,\eta_{j+1}^{(t)},\ldots,\eta_k^{(t)},\sigma^{2(t)},\mathbf{y})$$

can be obtained similarly as in Algorithm TN2, and $p(\sigma^2|\eta_1^{(t+1)},\eta_2^{(t+1)}\ldots,\eta_k^{(t+1)},\mathbf{y})$ can be obtained from (4.28). □

4.3.1 Example: Rental Data

We consider the 32 observations provided by Pindyck and Rubinfeld (1981; page 44) on rent paid, number of rooms rented, number of occupants, sex and distance from campus in blocks for undergraduates at the University of Michigan. Geweke (1986, 1996) considers the model

$$y_i=\beta_1+\beta_2 s_i r_i+\beta_3(1-s_i)r_i+\beta_4 s_i d_i+\beta_5(1-s_i)d_i+\epsilon_i,$$

where y_i denotes rent paid per person, r_i number of rooms per person, d_i distance from campus in blocks, s_i is a dummy variable representing gender (one for male and zero for female), ϵ_i is normally distributed error with mean 0 and variance σ^2, and the β's are subject to the constraints

$$\beta_2\geq 0,\ \beta_3\geq 0,\ \beta_4\leq 0,\ \beta_5\leq 0. \tag{4.32}$$

Since the number of constraints does not exceed the number of regression coefficients, Algorithm CLR1 can be used to draw a sample from the posterior distribution of $(\boldsymbol{\beta},\sigma^2)$. For this algorithm, the matrix $\mathbf{B}$ is the identity of size 5,

$\mathbf{c} := [-\infty, 0, 0, -\infty, -\infty]^T$ and $\mathbf{d} := [\infty, \infty, \infty, 0, 0]^T$. For Algorithm CLR2, $\mathbf{b} = \mathbf{0}$ and $\mathbf{B}$ is given by

$$\mathbf{B} = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Taking μ_0 to be the constrained MLE of β, $\sigma_0^2 = 1000$, and $\nu = \lambda = 0.001$ as the values needed for the prior distribution of (β, σ^2), we obtained Gibbs paths of length 1600 for the posterior distribution of (β, σ^2) using Algorithms CLR1 and CLR2. In Figure 4.4 we show the autocorrelation function plots of β_2, β_3 and β_4 using these outputs. From this figure, we note that the mixing of the sampler from Algorithm

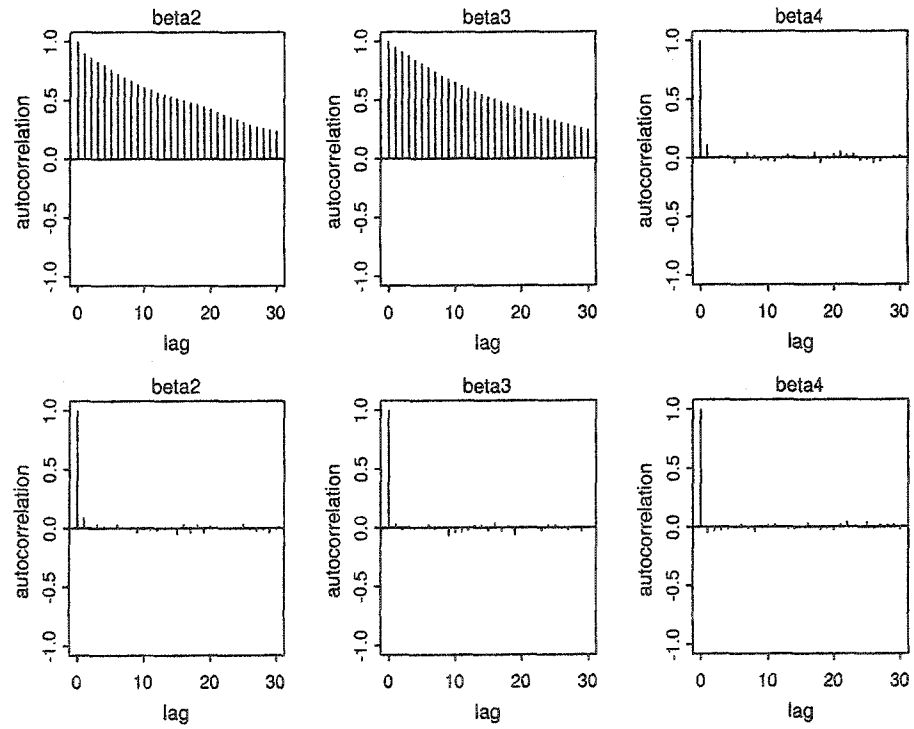

Figure 4.4: Autocorrelation plots of β_2, β_3, and β_4 obtained with Gibbs paths of length 1600. The first row was obtained with the output of Algorithm CLR1. The second was obtained with the output of Algorithm CLR2.

CLR2 is faster than that from Algorithm CLR1.

To compare the speed of convergence of both samplers, we increased the length of both paths to a total length of 20000 (each path). In Figure 4.5 we show two sections

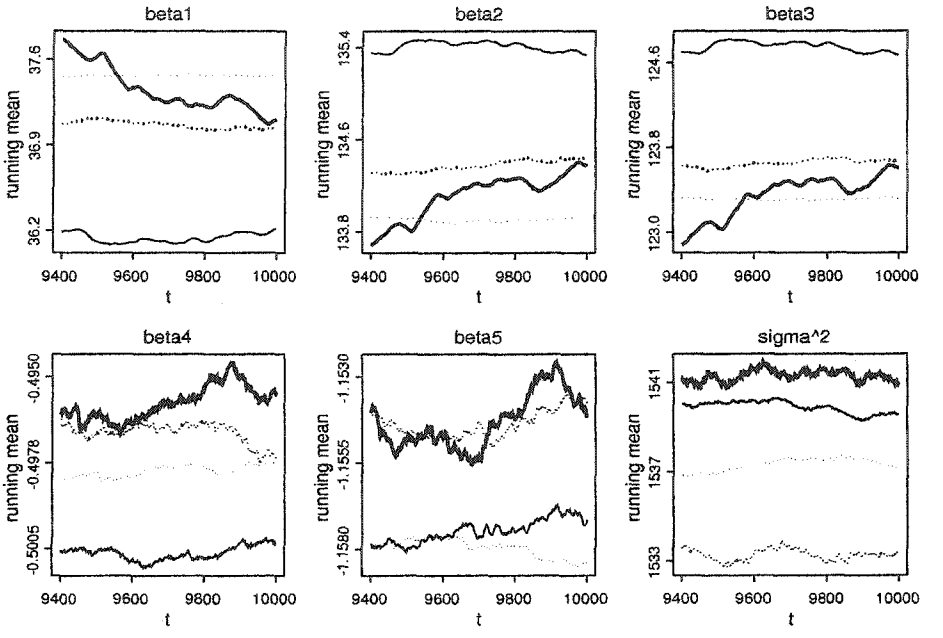

Figure 4.5: Two sections of the running means of $\beta_1, \ldots, \beta_5$ and σ^2. The thick lines show the first sections ($9401 < t < 10000$) and the thin lines the second sections ($19401 < t < 20000$) of these running means. The values obtained with the output of Algorithm CLR1 are shown with solid lines and that obtained with the output of Algorithm CLR2 with dotted lines.

of the running mean plots of σ^2 and all the components of β. In each panel, the thick lines correspond to the section $9401 < t < 10000$ while the thin lines correspond to the section $19401 < t < 20000$. The solid lines were obtained using the output of Algorithm CLR1 and the dotted lines that of Algorithm CLR2. While it appears that for the sampler from Algorithm CLR2, 20000 iterations are enough to stabilize the mean of σ^2 and β, this is not the case for the sampler from Algorithm CLR1.

4.3.2 Example: Least squares estimates of a transition probability matrix

This example considers the estimation of the transitional probability matrix of a finite Markov process when only the time series of the proportions of the sample in each

state is known. To estimate the transition probability matrix, Telser (1963), proposes least-squares estimation based on a set of regression models subject to the sum-to-one constraints for the rows of this matrix and to the non-negativity constraints of all its entries. To illustrate their generalized restricted estimator procedure, this problem is analyzed again by Judge and Takayama (1966), which takes these constraints into account explicitly. The numerical example given by Telser (1963) and Jugde and Takayama (1966), consists of the annual sales in billions of cigarettes for the three leading brands from 1925 to 1943. Given the time ordered market shares of these brands and assuming that the probability of a transition, p_{ij}, from brand i to brand j is constant over time, Telser gives the regression models

$$y_{j,t} = \sum_{i=1}^{3} y_{i,t-1} p_{ij} + u_{jt}, \quad j = 1, 2, 3, \tag{4.33}$$

where y_{jt} is the proportion of individuals in state j at time t and u_{jt}, $t = 1, \ldots, T$ are independent errors. The probabilities p_{ij} are subject to the constraints

$$\sum_{j=1}^{3} p_{ij} = 1, \quad \text{for all } i, \tag{4.34}$$

$$p_{ij} \geq 0, \quad \text{for all } i, \text{and } j. \tag{4.35}$$

For the cigarettes data, the three models in (4.33) can be combined as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{W} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}, \tag{4.36}$$

where $\mathbf{y}_j := [y_{2,j}, \ldots, y_{T,j}]^T$, $\mathbf{W}$ is the common design matrix of dimension $3 \times T - 1$ from the models in (4.33), $\mathbf{p}_j$ is the j-th column of the probability transition matrix $\mathbf{P}$ of the finite Markov process, and $\mathbf{u}_j$ is the vector of errors from the model in (4.33).

We propose to treat the equality constraints in (4.34) as in the frequentist approach. e.g., Hocking (1996; page 70) incorporates equality linear constraints into the so called *full model* to obtain a *reduced model.* In the Bayesian approach, a new feature appears. The equality constraints need to the incorporated in the support of the full model. Denote by $\mathbf{y}$ the response vector of the full model in (4.36), by $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ the matrices having the columns 1 through 3, 4 through 6 and 7 through 9, respectively of the design matrix in (4.36). Substituting $p_{i3} = 1 - p_{i1} - p_{i2}$, $i = 1, 2, 3$, in this model, we obtain

$$\mathbf{y} - \mathbf{W}_3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = [\mathbf{W}_1 - \mathbf{W}_3 \quad \mathbf{W}_2 - \mathbf{W}_3] \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{bmatrix} + \mathbf{u}, \tag{4.37}$$

subject to the constraints

$$p_{i1} + p_{i2} \leq 1, \quad i = 1, 2, 3, \tag{4.38}$$

$$p_{ij} \geq 0, \quad i = 1, 2, 3, \; j = 1, 2, \tag{4.39}$$

where $\mathbf{u}$ is the vector of errors from the model in (4.36). In their method, Judge and Takayama (1966) assume that $\text{var}(\mathbf{u}) = \sigma^2\mathbf{I}$. For simplicity we also assume that $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. However, a more general matrix of variance-covariance for $\mathbf{u}$ can be used, e.g., $\text{var}\{\mathbf{u}_j\} = \sigma_j^2\mathbf{I}$ and $\text{cov}\{\mathbf{u}_{j_1}, \mathbf{u}_{j_2}\} = \mathbf{0}$, $j_1 \neq j_2$. For this case, a prior distribution for the vector $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ needs to be specified.

Notice that the number of constraints in (4.38-4.39) to the regression model in (4.37) exceeds the number of regression coefficients. This time, Algorithm CLR1 can not be carried out. To implement the Gibbs sampler described in Algorithm CLR2, set $\boldsymbol{\mu}_0$ equal to the constrained MLE of $\boldsymbol{\beta} := [\mathbf{p}_1^T \quad \mathbf{p}_2^T]$, σ_0^2 large (100), and $\nu = \lambda = 0.001$ as the values needed for the prior distribution of $(\boldsymbol{\beta}, \sigma^2)$. A path of

length 5000 for the posterior distribution of (β, σ^2) was generated. Based on the last 2500 iterates of this sample, the estimate $\hat{\mathbf{P}}$ of the probability transition matrix and the matrix $\hat{\sigma}_{\hat{\mathbf{P}}}$ having in its entries the estimated standard error of each component of $\hat{\mathbf{P}}$ are

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.690 & 0.118 & 0.192 \\ 0.035 & 0.844 & 0.121 \\ 0.334 & 0.060 & 0.606 \end{bmatrix}, \tag{4.40}$$

$$\hat{\sigma}_{\hat{\mathbf{P}}} = \begin{bmatrix} 0.0016 & 0.0009 & 0.0016 \\ 0.0006 & 0.0008 & 0.0009 \\ 0.0023 & 0.0010 & 0.0024 \end{bmatrix}.$$

The restricted least-squares estimates obtained by Judge and Takayama (1966) are given by

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.6686 & 0.1423 & 0.1891 \\ 0 & 0.8683 & 0.1317 \\ 0.4019 & 0 & 0.5981 \end{bmatrix}. \tag{4.41}$$

The estimates in (4.40) differ slightly from the restricted least-squares in (4.41). Perhaps the most important difference is the fact that the estimates of p_{21} and p_{32} are non zero. The zero estimates of the elements of $\mathbf{P}$ can induce misleading interpretations. For example, because $\hat{p}_{21} = 0$, a smoker of the second brand never tries cigarettes of the first brand, unless he tries cigarettes of the third brand. This unlikely behavior does not show up with the estimates in (4.40).

To get an indication of how the sampler performs, the autocorrelation plots of the components of the matrix $\mathbf{P}$ and the running means of these components are shown in Figures 4.6 and 4.7, respectively. In Figure 4.6 we observe a fast decay on the autocorrelations. Following Chen, et al. (2000), we expect a good mixing and

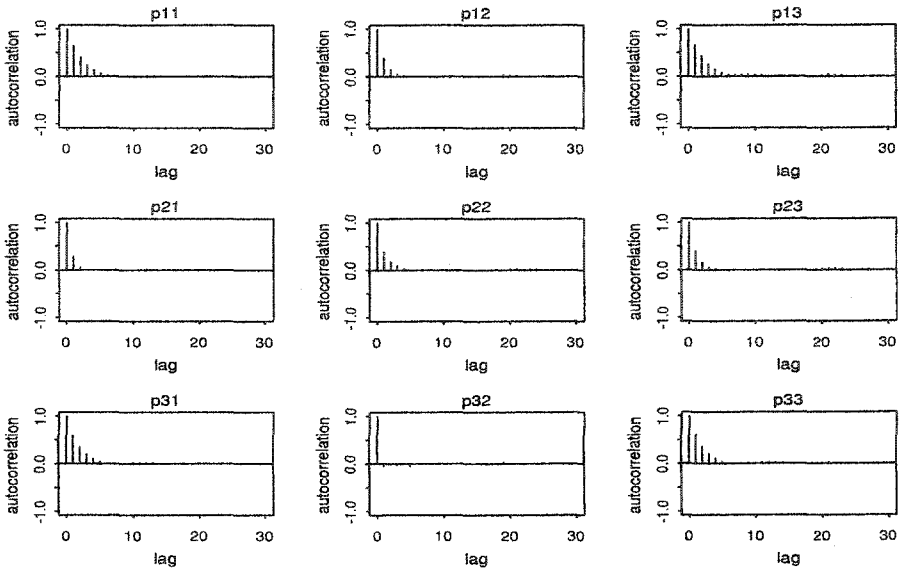

Figure 4.6: Autocorrelation plots of the components of the transition probability matrix $\mathbf{P}$ of the cigarettes data obtained with a Gibbs path of length 5000.

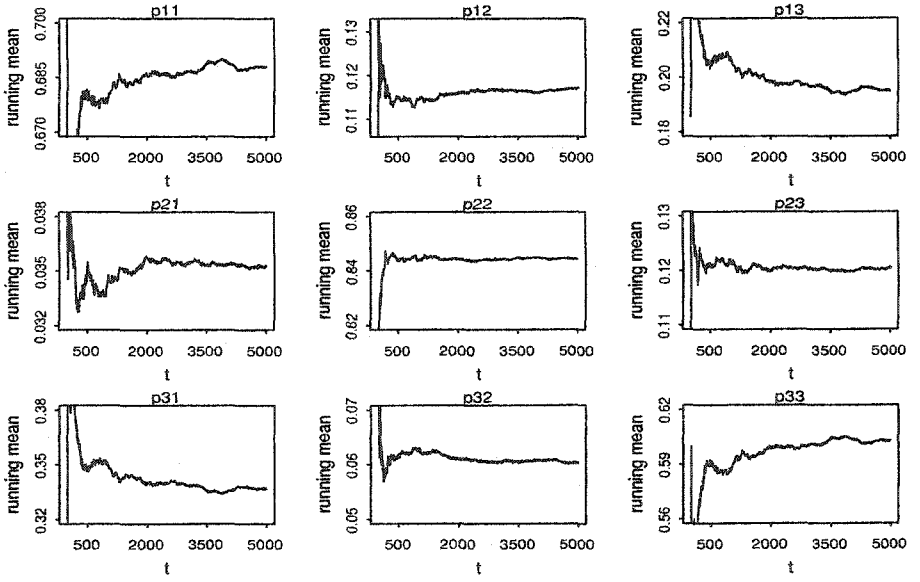

Figure 4.7: Running mean plots of the components of the transition probability matrix $\mathbf{P}$ of the cigarattes data obtained with a Gibbs path of length 5000.

fast convergence. This is in fact corroborated by the results in Figure 4.7, where the monitored statistics seem to be stabilized after a relatively small number of iterations.

4.4 Conclusions

In this chapter, Bayesian analysis of a linear regression model where the parameters are subject to inequality linear constraints has been considered. Our method is based on an efficient Gibbs sampler implementation to the truncated multivariate normal distribution. This sampler mixes faster than others implementations in the literature (e.g. Geweke, 1991; Chen and Deely, 1996; Geweke, 1996.) Although the number of constraints can exceed the number of regression coefficients, the constrained region must have positive Lebesgue measure. Furthermore, we have shown through an example how to use equality linear constraints in addition to inequality linear constraints.

APPENDIX A

The Innovations algorithm

In this appendix, we briefly describe the innovations algorithm (Brockwell and Davis, 1991) and show with an example, how it can be adapted to compute the recursion in (2.16) and the determinant needed in approximation (2.23). This algorithm is applicable to any time series with finite second moments, whether stationary or not.

Suppose that $\{X_t\}_{t=1}^n$ is a time series with finite second moment and covariance matrix $\mathbf{\Gamma}$. Define $\mathbf{X} := (X_1, X_2, \ldots, X_n)$. Let $\hat{\mathbf{X}}$ be the vector of one-step predictors, i.e., $\hat{\mathbf{X}} := (0, \hat{X}_2, \ldots, \hat{X}_n)$ and $\nu_j := \mathrm{E}(X_{j+1} - \hat{X}_{j+1})^2$ be the mean-squared error of the one-step predictor $\hat{X}_{j+1}$. Then (Brockwell and Davis, 1996; pp. 70-71)

$$\mathbf{X} = \mathbf{C}(\mathbf{X} - \hat{\mathbf{X}}), \tag{A.1}$$

where

$$\mathbf{C} := \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ \theta_{11} & 1 & 0 & \ldots & 0 \\ \theta_{22} & \theta_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \ldots & 1 \end{pmatrix}. \tag{A.2}$$

The entries θ_{ij} of this matrix can be found recursively as in Proposition 5.2.2. from Brockwell and Davis (1991). Computing the covariance matrices on both sides of

(A.1), it follows that

$$\boldsymbol{\Gamma} = \mathbf{C}\mathbf{D}\mathbf{C}^T, \tag{A.3}$$

where $\mathbf{D} := E\{(\mathbf{X} - \hat{\mathbf{X}})(\mathbf{X} - \hat{\mathbf{X}})^T\} = \text{diag}\{\nu_0, \nu_1, \ldots \nu_{n-1}\}$. The last equallity comes from the fact that the components of $\mathbf{X} - \hat{\mathbf{X}}$ are uncorrelated. Also, because the determinant of the matrix $\mathbf{C}$ is 1, taking determinants in both sides of (A.3), we obtain

$$|\boldsymbol{\Gamma}| = |\mathbf{C}\mathbf{D}\mathbf{C}^T| = |\mathbf{D}| = \prod_{j=0}^{n-1} \nu_j, \tag{A.4}$$

Now, using using (A.1) and (A.3), we can show that

$$\boldsymbol{\Gamma}^{-1}\mathbf{X} = \mathbf{C}^{-T}\mathbf{e}, \tag{A.5}$$

where the entries e_j of the vector $\mathbf{e}$ are the "normalized" residuals $(X_j - \hat{X}_j)/\nu_{j-1}$.

For example, consider the SSM for which the observations $y_1, \ldots, y_n$ are realizations of a Poisson distributed with rates $\lambda_t = e^{\beta+\alpha_t}$ and the state process follows the AR(1) model

$$\alpha_t = \phi\alpha_{t-1} + \eta_t, \tag{A.6}$$

where $\eta_t \sim$ iid $N(0, \sigma^2)$, $t = 1, \ldots, n$. Notice that the distribution of the observations has the format of the exponential family in (2.3) where $b(\alpha_t) = e^{\alpha_t+\beta}$.

From the fact that $\text{cov}\{\alpha_t, \alpha_{t+h}\} = \sigma^2|\phi|^h/(1-\phi^2)$, we have

$$\mathbf{V} = \text{cov}\{\boldsymbol{\alpha}\}^{-1} = 1/\sigma^2 \begin{pmatrix} 1 & -\phi & 0 & \ldots & 0 & 0 \\ -\phi & 1+\phi^2 & -\phi & \ldots & 0 & 0 \\ 0 & -\phi & 1+\phi^2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & \ldots & -\phi & 1 \end{pmatrix}.$$

Now, let $\boldsymbol{\alpha}^j$ be the current iterate to the value of $\boldsymbol{\alpha}^*$. From (2.15) and (2.24)

$$\begin{aligned}\dot{\mathbf{b}}^j &= \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{1}^T \mathbf{b}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}^j} = e^{\beta} e^{\boldsymbol{\alpha}^j} \\ \mathbf{K}^j &= \frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \mathbf{1}^T \mathbf{b}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}^j} = e^{\beta} \mathrm{diag}\{e^{\boldsymbol{\alpha}^j}\}.\end{aligned}$$

Since no intercept is included in the AR(1) process in (A.6), $\boldsymbol{\mu} = \mathbf{0}$. Thus, $\tilde{\mathbf{y}}^j$ defined in (2.25) is given by

$$\tilde{\mathbf{y}}^j = \mathbf{y} - \dot{\mathbf{b}}^j + \mathbf{K}^j \boldsymbol{\alpha}^j + \mathbf{V}\boldsymbol{\mu} = \mathbf{y} - e^{\beta} e^{\boldsymbol{\alpha}^j} + e^{\beta} \mathrm{diag}\{e^{\boldsymbol{\alpha}^j}\} \boldsymbol{\alpha}^j.$$

Set $\boldsymbol{\Gamma} := \mathbf{K}^j + \mathbf{V}$ and $\mathbf{X} := \tilde{\mathbf{y}}^j$. Since $\boldsymbol{\Gamma}$ is a band-limited matrix, it follows from Proposition 5.2.2. of Brockwell and Davis (1991) that

$$\begin{aligned}\nu_j &= \begin{cases} \gamma_{11}, & \text{if } j = 0, \\ \gamma_{j+1,j+1} - \theta_{j1}^2 \nu_{j-1}, & \text{if } j = 1, \ldots, n-1, \end{cases} \\ \hat{X}_j &= \begin{cases} 0, & \text{if } j = 1, \\ \theta_{j-1,1}(X_{j-1} - \hat{X}_{j-1}), & \text{if } j = 2, \ldots, n \end{cases}\end{aligned} \tag{A.7}$$

and for $m = 1, \ldots, n-1$,

$$\theta_{mj} = \begin{cases} \nu_{j-1}^{-1} \gamma_{j+1,j}, & \text{if } j = 1, \\ 0, & \text{if } j = 2, \ldots, m. \end{cases}$$

Once these values have been computed, then the vector of normalized residuals $\mathbf{e}$ needed in (A.5) is easily obtained, and the iteration in (2.20) becomes

$$\boldsymbol{\alpha}^{j+1} = (\mathbf{K}^j + \mathbf{V})^{-1} \tilde{\mathbf{y}}^j = \boldsymbol{\Gamma}^{-1} \mathbf{X} = \mathbf{C}^{-T} \mathbf{e} \tag{A.8}$$

Due to the fact that $\mathbf{C}$ is a band matrix, rather than inverting it to obtain $\boldsymbol{\alpha}^{j+1}$ we can compute it by a reversed iteration obtained from $\mathbf{e} = \mathbf{C}\boldsymbol{\alpha}^{j+1}$.

The iteration in (2.20) tends to converge quite rapidly -only a few iterations are required. Now, to compute the determinant of the matrix $\mathbf{K}^* + \mathbf{V}$ needed in (2.23),

set $\mathbf{\Gamma} := \mathbf{K}^* + \mathbf{V}$, where $\mathbf{K}^* = e^{\beta}\text{diag}\{e^{\boldsymbol{\alpha}^*}\}$ -see (2.22), and $\mathbf{X} = \mathbf{y} - e^{\beta}e^{\boldsymbol{\alpha}^*} + e^{\beta}\text{diag}\{e^{\boldsymbol{\alpha}^*}\}\boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^*$ is the converged value of the iteration in (A.8). Then, from (A.4),

$$|\mathbf{K}^* + \mathbf{V}| = |\mathbf{\Gamma}| = \prod_{j=0}^{n-1} \nu_j,$$

where ν_j, $j = 0, \ldots, n-1$ must be computed as in (A.7). Extensions to state processes following an AR(p) model can be handled in a similar fashion.

BIBLIOGRAPHY

[1] Bernardo, J. M. and Smith, A. F. M (1994). "Bayesian Theory" J. Wiley, New York.

[2] Breidt, F. J. and Carriquiry, A. L. (1996). "Improved Quasi-Maximum Likelihood Estimation for Stochastic Volatility Models." In: Zellner, A., Lee, J. S. (Eds.), Modeling and Prediction: Honouring Seymour Geisser. Springer, New York.

[3] Brockwell, P. J. and Davis, R. A. (1991). "Time Series: Theory and Methods." (2nd ed.) Springer-Verlag, New York.

[4] Brockwell, P. J. and Davis, R. A. (1996). "Introduction to Time Series and Forecasting." Springer-Verlag, New York.

[5] Campbell, M. J. (1994) "Time Series Regression for Counts: an Investigation Into the Relationship Between Sudden Infant Death Syndrome and Environmental Temperature." *J. R. Stat. Soc. Ser. A*, **157**, 191-208.

[6] Chan, K. S. and Ledolter, J. (1995). "Monte Carlo EM Estimation for Time Series Models Involving Counts." *J. Amer. Statist. Assoc.*, **90**, 242-252.

[7] Chen, M-H. and Deeley, J. J. (1996) "Bayesian Analysis for a Constrained Linear Multiple Regression Problem for Predicting the New Crops of Apples," *J. Agric. Biol. Environ. Stat.*, **1**, 467-89.

[8] Chen, M-H., Shao, Q-M. and Ibrahim, J. G. (2000). "Monte Carlo Methods in Bayesian Computation." Springer, New York, 2000.

[9] Davis, R. A., Dunsmuir, W. T. M. and Wang, Y. (1998). "Modelling Time Series of Count Data." In Asymptotics, Nonparametrics and Time Series (ed Subir Ghosh), Marcel Dekker.

[10] Doucet, A, De Freitas, N. and Gordon, N. (2001). "Sequential Monte Carlo Methods in Practice," Springer-Verlag Inc (Berlin; New York)

[11] Durbin, J. and Koopman, S. J. (1997) "Monte Carlo Maximum Likelihood Estimation for non-Gaussian State Space Models." *Biometrika*, **84**, 669-684.

[12] Durbin, J. and Koopman, S. J. (2001) "Time Series Analysis by State Space Methods." Oxford, NY.

[13] Efron, B. and Tibshirani R. J. (1993) "An Introduction to the Bootstrap." Chapman and Hall, NY.

[14] Gelfand, A. E. and Smith, A. F. M. (1990) "Sampling-based Approaches to Calculating Marginal Densities," *J. Amer. Statist. Assoc.*, **85**, 398-409.

[15] Gelfand, A. E., Smith, A. F. M. and Lee, T. M. (1992) "Bayesian Analysis of Constrained Parameters and Truncated Data Problems." *J. Amer. Statist. Assoc.*, **87**, 523-532.

[16] Geman, S. and Geman, D. (1984) "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE trans. pattern anal. mach. intell*, **6**, 721-741.

[17] Geyer, C. J. (1996) "Estimation and optimization of functions." In Markov Chain Monte Carlo in Practice (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), Chapman & Hall, London, pp. 89-114.

[18] Geweke, J. (1986) "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *J. Appl. Econ.*, **1**, 127-141.

[19] Geweke, J. (1991) "Efficient Simulation From the Multivariate Normal and Student t-Distributions Subject to Linear Constraints," Computer Sciences and Statistics: Proc. 23d Symp. Interface, 571-577.

[20] Geweke, J. (1996) "Bayesian Inference for Linear Models Subject to Linear Inequality Constraints," In: Zellner, A., Lee, J. S. (Eds.), Modeling and Prediction: Honouring Seymour Geisser. Springer, New York.

[21] Geweke, J. and Tanizaki, H. (1999) "On Markov Chain Monte Carlo Methods for Nonlinear and Non-Gaussian State-Space Models." *Comm. Statist. Simulation Comput.*, **28**, 867-894.

[22] Gilks, W. R. and Roberts, G. O. (1996) "Strategies for Improving MCMC," In Markov Chain Monte Carlo in Practice (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), Chapman & Hall, London, 241-258.

[23] Hahivassiliou, V., and McFadden, D. (1997). "The Method of Simulated Scores for the Estimation of LDV Models." Discussion Paper No. EM/97/328, School of Economics and Political Science, London.

[24] Harvey, A. C. (1989) "Forecasting, Structural Time Series Models and the Kalman Filter." Cambridge: Cambridge University Press.

[25] Harvey, A. C. and Fernandes, C. (1989) "Time Series Models for Count or Qualitative Observations." *J. Amer. Statist. Assoc.*, **7**, 407-417.

[26] Harvey, A. C., Ruiz, E. and Shepard, N. (1994) "Multivariate Stochastic Variance Models." *Rev. Econom. Stud.*, **61**, 247-264.

[27] Harvey, A. C. and Shepard, N. (1993) "Estimation and Testing of Stochastic Variance Models." Unpublished manuscript, The London School of Economics.

[28] Harvey, A. C. and Streibel, M. (1998) "Testing for a slowly changing level with special reference to stochastic volatility." *J. Econometrics*, **87**, 167-189.

[29] Hocking, R. R. (1996) "Methods and Applications of Linear Models: Regression and the Analysis of Variance," Wiley, New York.

[30] Hodges, P.E. and Hale, D. F. (1993). "A Computational Method for Estimating Densities of Non-Gaussain Nonstationary Univariate Time series." *J. Time Ser. Anal.*, **14**, 163-178.

[31] Hurzeler, M. (1998). "Statistical Methods for General State-Space Models,." Ph.D. Thesis, Department of Mathematics, ETH Zurich, Zurich.

[32] Jacquier, E., Polson, N. G. and Rossi, P. E. (1994) "Bayesian analysis of stochastic volatility models (with discussion)." *J. Bus. Econom. Statist.*, **12**, 371-417.

[33] Johnson, R. A. and Wichern, D. W. (1998) "Applied Multivariate Statistical Analysis." (Fourth ed.) Prentice Hall, New Jersey.

[34] Jugdge, G. C. and Takayama, T. (1966) "Inequality Restrictions In Regression Analysis," *J. Amer. Statist. Assoc.*, **61**, 166-181.

[35] Kitagawa, G. (1987). "Non-Gaussian State-Space Modeling of Nonstationary Time Series." *J. Amer. Statist. Assoc.*, **82**, 1032-1063.

[36] Kitagawa, G. (1996). "Monte Carlo Filter and Smoother for Non-Gaussian Non-Linear State Space Models." *J. Comput. Graph. Statist.*, **5**, 1-25.

[37] Kuk, A. Y. (1999) "The Use of Approximating Models in Monte Carlo Maximum Likelihood Estimation." *Statist. Probab. Lett.*, **45**, 325-333.

[38] Kuk, A. Y. and Cheng, Y. W. (1997) "The Monte Carlo Newton-Raphson Algorithm." *J. Stat. Comput. Simul.*, **59**, 233-250.

[39] Liew, C. K. (1976) "Inequality Constrained Least-Squares Estimation," *J. Amer. Statist. Assoc.*, **71**, 746-751.

[40] Lovell, M. C. and Prescott, E. (1970). "Multiple Regression with inequality constraints: Pretesting Bias, Hypothesis Testing, and Efficiency." *J. Amer. Statist. Assoc.*, **65**, 913-925.

[41] Manolakis, D. and Shaw, G. (2002) " Detection Algorithms for Hyperspectral Imaging Applications," *IEEE Signal Processing Magazine*, **19**, 29-43.

[42] O'Hagan, A. (1994). "Kendall's Advanced Theory of Statistics 2B: Bayesian Inference." London: Edward Arnold.

[43] Pindyck, R. S. and Rubinfeld, D. L. (1981). "Econometric Models and Economic Forecasts." (2nd ed.), McGraw-Hill, New York.

[44] Pitt, M. K and Shepard N. (1999). "Filtering via Simulation: Auxiliary Particle Filters." *J. Amer. Statist. Assoc.*, **94**, 590-599.

[45] Raftery, A. L. and Lewis, S. (1992). "How Many Iterations in the Gibbs Sampler?" In Bayesian Statistics 4, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. Oxford: Oxford University Press.

[46] Ripley, B. D. (1987). "Stochastic Simulation." Wiley, New York.

[47] Ritter, C. and Tanner, M. A. (1992). "The Gibbs Stopper and the Griddy Gibbs Sampler." *J. Amer. Statist. Assoc.*, **87**, 861-868.

[48] Robert, C. P. (1995) "Simulation of Truncated Normal Variables," *Statist. Comput.*, **5**, 121-125.

[49] Roberts, G. O. (1996) "Markov Chain Concepts Related to Sampling Algorithms." In Markov Chain Monte Carlo in Practice (eds W. R. Gilks, s. Richardson and D. J. Spiegelhalter), 45-57. London: Chapman & Hall.

[50] Rodriguez-Yam, G. A., Davis, R. A. and Scharf, L. L. (2002) "A Bayesian Model and Gibbs Sampler for Hyperspectral Imaging," 2002 IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings, 105-109.

[51] Sandmann, G. and Koopman, S. J. (1998) "Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood." *J. Econometrics*, **87**, 271-301.

[52] Stoffer, D. S. and Wall, K. D. (1991) "Bootstrapping State-Space Models: Gaussian Maximum Likelihood and the Kalman Filter." *J. Amer. Statist. Assoc.*, **86**, 1024-1032.

[53] Telser, L. G. (1963) "Least Squares Estimates of Transition Probabilities," in Christ, C. F., Friedman, M., Goodman, L. A., Griliches, Z., Harberger, A. C., Liviatan, N., Mincer, J. Mundlak, Y. Nerlove, M. Patinkin, D., Telser, L. G. and Theil, H. (Eds.), Measurement in Economics: Studies in Mathematical Economics and Econometrics: In memory of Yehuda Grunfeld. Stanford University Press, Stanford.

[54] Zeger, S. L. (1988) "A Regresion Model for Time Series of Counts." *Biometrika*, **75**, 621-629.