Estimation for State-Space Models; an Approximate Likelihood Approach

Richard A. Davis and Gabriel Rodriguez-Yam
Department of Statistics, Colorado State University, Fort Collins, Colorado

October 22, 2003

Abstract

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 paper, 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.

1 Introduction

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 paper 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

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

belongs to a known parametric family of distributions. In addition, the state process is assumed to follow an AR(p) model, given by

$$\alpha_t = \gamma + \phi_1 \alpha_{t-1} + \ldots + \phi_p \alpha_{t-p} + \eta_t, \tag{2}$$

where p is an integer greater than zero and $\eta_t \sim$ iid $N(0,\ \sigma^2)$, $t = 1, 2, \ldots$. Perhaps the most important special case is when the conditional distribution in (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

$$p(y_t|\alpha_t) = e^{(\mathbf{x}_t^T\boldsymbol{\beta}+\alpha_t)y_t - b(\mathbf{x}_t^T\boldsymbol{\beta}+\alpha_t)+c(y_t)}, \tag{3}$$

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, $p(y_t|\alpha_t; \boldsymbol{\beta})$ is a particular case of (3). Models of this type have been used for modeling counts of individuals infected by a rare disease, e.g., Zeger (1988); Campbell (1994); Chan and Ledolter (1995); Harvey and Fernandes (1989); 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 $\boldsymbol{\beta}$, and $\boldsymbol{\lambda} := (\phi_1, \ldots, \phi_p, \gamma, \sigma^2)$ is the parameter vector associated with the AR model in (2). With this specification, the likelihood based on the "complete data" $(\mathbf{y}, \boldsymbol{\alpha})$ of the SSM becomes

$$\begin{aligned} L(\boldsymbol{\psi}; \mathbf{y}, \boldsymbol{\alpha}) &= p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda}) \\ &= \left(\prod_{t=1}^{n} p(y_t|\alpha_t, \boldsymbol{\theta}) \right) |\mathbf{V}|^{1/2} e^{-(\boldsymbol{\alpha}-\boldsymbol{\mu})^T \mathbf{V} (\boldsymbol{\alpha}-\boldsymbol{\mu})/2} / (2\pi)^{n/2}, \end{aligned} \tag{4}$$

where $\mathbf{V}^{-1} := \text{cov}\{\boldsymbol{\alpha}\}$, $\boldsymbol{\mu} = \gamma/(1 - \phi_1 - \ldots - \phi_p)\mathbf{1}$ is the vector of means of the state process, and $\mathbf{1}$ is a vector of ones. From (4) it follows that the likelihood of the observed data is

$$L(\boldsymbol{\psi}; \mathbf{y}) = \int L(\boldsymbol{\psi}; \mathbf{y}, \boldsymbol{\alpha}) d\boldsymbol{\alpha}. \tag{5}$$

Except in simple cases, the integral in (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 (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 same noise. Kuk (1999) advocates a "single-sample" approach, which for a fixed $\boldsymbol{\psi}_0$, a sample is drawn from the importance density $g(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}_0)$ and then the relative likelihood function is optimized using these samples. To get better approximations of the relative likelihood near the true maximum likelihood estimate, Geyer (1996) suggests repeating the process several times, updating $\boldsymbol{\psi}_0$ with the new maximizer at each iteration.

A Monte Carlo EM algorithm treating the unobserved $\boldsymbol{\alpha}$'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 $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}^{(i-1)})$, where $\boldsymbol{\psi}^{(i-1)}$ 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 intense.

In this paper we will follow a different approach to obtain an approximation to the distribution $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$. In Section 2 we will produce an analytical approximation to (5) by obtaining an approximation $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ to the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$. 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 $\boldsymbol{\psi}$. In Section 3 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, $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$. In a numerical example we assess this approximation in two ways. First, we notice the closeness between the posterior mode and posterior mean of $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$. As a second check of closeness we compare 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 Section 3. In Section 4 we summarize our findings. Application of the innovations algorithm to the problems considered in Sections 2 and 3 is given in the appendix.

2 Parameter Estimation

In this section we find an approximation to the observed likelihood $L(\boldsymbol{\psi};\mathbf{y})$ given in (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

$$\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{V}| + \ell(\boldsymbol{\psi};\mathbf{y}|\boldsymbol{\alpha}) - \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}-\boldsymbol{\mu}). \quad (6)$$

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

$$\begin{aligned}
\ell(\boldsymbol{\psi};\mathbf{y}|\boldsymbol{\alpha}) &= T(\boldsymbol{\alpha};\boldsymbol{\alpha}^*) + R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*),\\
&= h^* + (\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\mathbf{k}^* - \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\mathbf{K}^*(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*),\\
&= h^* + (\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) - \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\mathbf{K}^*(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \quad (7)\\
&\quad + R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*),
\end{aligned}$$

where

$$h^* := \ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*} \text{ and } \mathbf{K}^* := -\frac{\partial^2}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}. \quad (8)$$

Thus, substituting (7) in (6), it follows that

$$\begin{aligned}
\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) &= -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{V}| + h^* - \frac{1}{2}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}^*-\boldsymbol{\mu}) \quad (9)\\
&\quad -\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T(\mathbf{K}^*+\mathbf{V})(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*).
\end{aligned}$$

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 (9) when the term $R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)$ is omitted, it follows that

$$p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) = \boldsymbol{\phi}(\boldsymbol{\alpha};\boldsymbol{\alpha}^*,(\mathbf{K}^*+\mathbf{V})^{-1}), \quad (10)$$

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

$$\begin{aligned}
L(\boldsymbol{\psi};\mathbf{y}) &= \frac{|\mathbf{V}|^{1/2}}{|\mathbf{K}^*+\mathbf{V}|^{1/2}}e^{h^*-\frac{1}{2}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})}\int e^{R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)}p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})d\boldsymbol{\alpha}, \quad (11)\\
&= L_a(\boldsymbol{\psi};\mathbf{y})\mathrm{Er}_a(\boldsymbol{\psi}),
\end{aligned}$$

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

$$L_a(\boldsymbol{\psi};\mathbf{y}) := \frac{|\mathbf{V}|^{1/2}}{|\mathbf{K}^* + \mathbf{V}|^{1/2}} e^{h^* - \frac{1}{2}(\boldsymbol{\alpha}^* - \boldsymbol{\mu})^T \mathbf{V} (\boldsymbol{\alpha}^* - \boldsymbol{\mu})}, \tag{12}$$

that is obtained when the factor $e^{R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)}$ is ignored in the integral in (11); and $\mathrm{Er}_a(\boldsymbol{\psi})$ is the *approximation error*

$$\mathrm{Er}_a(\boldsymbol{\psi}) := \int e^{R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)} p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) d\boldsymbol{\alpha}. \tag{13}$$

Thus, if $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ is highly concentrated around $\boldsymbol{\alpha}^*$, the integral in (13) should be close to 1.

Since the evaluation of (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 (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, is this context, the normal approximation is, in fact, the same as $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ given in (10).

We note that

$$\begin{aligned} L(\boldsymbol{\psi};\mathbf{y}) &= \int p(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda}) d\boldsymbol{\alpha}, \\ &= \int \frac{p(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\lambda})}{p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})} p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) d\boldsymbol{\alpha}. \end{aligned} \tag{14}$$

So, $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ in (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

$$\mathbf{k}^j := \frac{\partial}{\partial \boldsymbol{\alpha}} \ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^j} \quad \text{and} \quad \mathbf{K}^j := -\frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^j}, \tag{15}$$

then the Newton-Raphson algorithm gives

$$\boldsymbol{\alpha}^{j+1} = \boldsymbol{\alpha}^j - (\ddot{\ell}^j)^{-1} \dot{\ell}^j, \tag{16}$$

where

$$\begin{aligned}
\dot{\ell}^j &:= \frac{\partial}{\partial \boldsymbol{\alpha}} \ell(\boldsymbol{\psi}; \mathbf{y}, \boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^j} \\
&= \mathbf{k}^j - \mathbf{V}(\boldsymbol{\alpha}^j - \boldsymbol{\mu}) \\
&= \mathbf{k}^j + \mathbf{K}^j \boldsymbol{\alpha}^j + \mathbf{V}\boldsymbol{\mu} - (\mathbf{K}^j + \mathbf{V})\boldsymbol{\alpha}^j, \qquad (17) \\
\ddot{\ell}^j &:= (\frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \ell(\boldsymbol{\psi}; \mathbf{y}, \boldsymbol{\alpha}))^{-1}|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^j} \\
&= -\mathbf{K}^j - \mathbf{V}. \qquad (18)
\end{aligned}$$

Let

$$\tilde{\mathbf{y}}^j := \mathbf{k}^j + \mathbf{K}^j \boldsymbol{\alpha}^j + \mathbf{V}\boldsymbol{\mu}. \qquad (19)$$

Substituting this, (17) and (18) into (16), we obtain

$$\boldsymbol{\alpha}^{j+1} = (\mathbf{K}^j + \mathbf{V})^{-1} \tilde{\mathbf{y}}^j. \qquad (20)$$

Application to the exponential family

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

$$p(\mathbf{y}|\boldsymbol{\alpha}; \boldsymbol{\theta}) = \prod_{t=1}^{n} p(y_t|\alpha_t, \boldsymbol{\theta}) = e^{(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha})^T \mathbf{y} - \mathbf{1}^T\{\mathbf{b}(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha}) - \mathbf{c}(\mathbf{y})\}}, \qquad (21)$$

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 (8) becomes

$$\mathbf{K}^* = \text{diag}\{\frac{\partial^2}{\partial \alpha_t^2} b(\mathbf{x}_t^T\boldsymbol{\beta} + \alpha_t)|_{\alpha_t^*}\}. \qquad (22)$$

The approximation to the observed likelihood is then

$$L_a(\boldsymbol{\psi}; \mathbf{y}) = \frac{|\mathbf{V}|^{1/2}}{|\mathbf{K}^* + \mathbf{V}|^{1/2}} e^{\mathbf{y}^T(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha}^*) - \mathbf{1}^T\{\mathbf{b}(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha}^*) - \mathbf{c}(\mathbf{y})\} - (\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})/2}. \qquad (23)$$

From (15) and (21), $\mathbf{k}^j = \mathbf{y} - \dot{\mathbf{b}}^j$, where

$$\dot{\mathbf{b}}^j := \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{1}^T \mathbf{b}(\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\alpha})|_{\boldsymbol{\alpha}^j}. \qquad (24)$$

Hence,

$$\tilde{\mathbf{y}}^j := \mathbf{y} - \dot{\mathbf{b}}^j + \mathbf{K}^j \boldsymbol{\alpha}^j + \mathbf{V}\boldsymbol{\mu}, \qquad (25)$$

where $\mathbf{K}^j$ is defined in (15). □

Although at this point we can find an approximation to the likelihood, each iteration of (20) needs the inversion of a matrix of dimension $n \times n$, while each

evaluation of (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 (14), $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ in (10) can be used as an importance density. In fact, as we show below for the case of the exponential family of distributions, $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ coincides with the importance density function of Durbin and Koopman (1997) to estimate the likelihood in (5) via simulation. In order to describe their method, let $L_g(\boldsymbol{\psi})$ 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 $p(y_t|\alpha_t;\boldsymbol{\psi})$ is replaced by a Gaussian distribution $g(y_t|\alpha_t;\boldsymbol{\theta}) = \phi(y_t; Z_t\alpha_t + \mu_t, H_t)$, where μ_t and H_t are found by solving iteratively

$$\frac{\partial}{\partial \alpha_t} \log p(y_t|\alpha_t;\boldsymbol{\psi})|_{\alpha_t=\hat{\alpha_t}} - H_t^{-1}(y_t - \hat{\alpha}_t - \mu_t) = 0 \tag{26}$$

$$\frac{\partial^2}{\partial \alpha_t^2} \log p(y_t|\alpha_t;\boldsymbol{\psi})|_{\alpha_t=\hat{\alpha_t}} + H_t^{-1} = 0. \tag{27}$$

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 (5) can be expressed as

$$L(\boldsymbol{\psi}) = L_g(\boldsymbol{\psi}) E_g \left\{ \frac{p(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\theta})}{g(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\psi})} |\mathbf{y},\boldsymbol{\psi} \right\}. \tag{28}$$

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 (5) is estimated as

$$\hat{L}(\boldsymbol{\psi}) = L_g(\boldsymbol{\psi}) \frac{1}{N} \sum_{i=1}^{N} \frac{p(\mathbf{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\theta})}{g(\mathbf{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi})}. \tag{29}$$

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

$$\frac{L(\boldsymbol{\psi})}{L_g(\boldsymbol{\psi}_0)} = E_g \left\{ \frac{p(\mathbf{y},\boldsymbol{\alpha}|\boldsymbol{\psi})}{g(\mathbf{y},\boldsymbol{\alpha}|\boldsymbol{\psi}_0)} |\mathbf{y},\boldsymbol{\psi}_0 \right\}, \tag{30}$$

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 $L(\boldsymbol{\psi})$ using (30) is

$$\hat{L}(\boldsymbol{\psi}) = L_g(\boldsymbol{\psi}_0)\frac{1}{N}\sum_{i=1}^{N}\frac{p(\mathbf{y}|\boldsymbol{\alpha}^{(i)}, \boldsymbol{\theta})p(\boldsymbol{\alpha}^{(i)}|\boldsymbol{\lambda})}{g(\mathbf{y}|\boldsymbol{\alpha}^{(i)}, \boldsymbol{\theta}_0)p(\boldsymbol{\alpha}^{(i)}|\boldsymbol{\lambda}_0)}. \tag{31}$$

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 $\hat{L}(\boldsymbol{\psi})$ in (31) is recommended (Geyer, 1996; Kuk, 1999).

If $p(y_t|\alpha_t; \boldsymbol{\psi})$ is a member of the exponential family of distributions as given in (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

$$H_t^{-1} = \ddot{b}_t, \qquad \mu_t = y_t - \hat{\alpha}_t - \ddot{b}_t^{-1}(y_t - \dot{b}_t). \tag{32}$$

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 (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 (22). Thus, $p_a(\boldsymbol{\alpha}|\mathbf{y}; \boldsymbol{\psi})$ in (10) and $g(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi})$ are identical. As a result, the observed likelihood $L_g(\boldsymbol{\psi})$ of the Durbin and Koopman's approximate Gaussian model is given by

$$L_g(\boldsymbol{\psi}) = (2\pi)^{-n/2}|\ddot{\mathbf{b}}^*|^{1/2}e^{\mathbf{1}^T\{\mathbf{b}(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha}^*)-\mathbf{c}(\mathbf{y})\}-\mathbf{y}^T\boldsymbol{\alpha}^*-\{\mathbf{y}-\dot{\mathbf{b}}^*\}^T(\ddot{\mathbf{b}}^*)^{-1}\{\mathbf{y}-\dot{\mathbf{b}}^*\}/2}L_a(\boldsymbol{\psi}; \mathbf{y}), \tag{33}$$

where $\dot{\mathbf{b}}^* := \frac{\partial}{\partial\boldsymbol{\alpha}}\mathbf{1}^T\mathbf{b}(\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\alpha})|_{\boldsymbol{\alpha}^*}$, and $\ddot{\mathbf{b}}^* := \frac{\partial^2}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\mathbf{1}^T\mathbf{b}(\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\alpha})|_{\boldsymbol{\alpha}^*}$.

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

$$\alpha_t = \phi\alpha_{t-1} + \eta_t, \tag{34}$$

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., $\boldsymbol{\psi} = \phi$. Using $\phi = 0.5$, one realization $y_1, \ldots, y_{200}$ from this process was generated. In Figure 1 we show two estimates of the observed likelihood of this process. In this figure, the solid line is the approximation to the observed likelihood given in (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 (29) using $N = 1000$. The dashed line is an estimation of the likelihood using (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 (29). The shape of the dashed line in Figure 1, typical of the estimator in (29), comes from the "many samples" effect. The maximization of this random function to obtain an estimator of ϕ requires

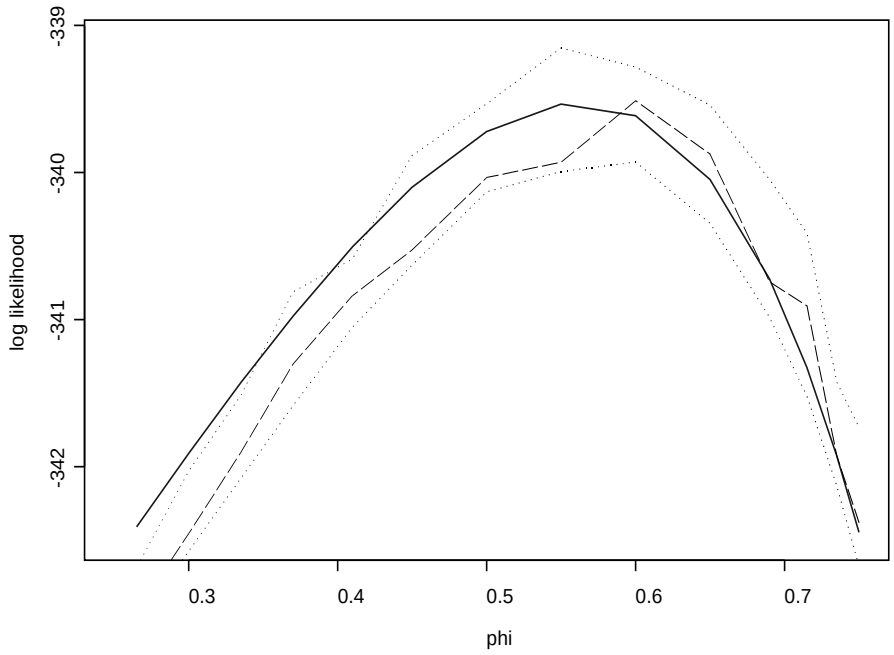

Figure 1: (*Many samples*) For a grid of values of ϕ, the logarithm of the estimation of the likelihood of a Poisson SSM are shown. For the solid line, estimates were obtained using (23) while for the dashed line, (29) was used. The dotted lines are the minimum and maximum respectively, of 100 replicates using (29).

additional effort. In contrast, the approximation in (23) is smooth and can be computed much faster.

To compute the estimator of the observed likelihood in (31) using the approach described by Kuk (1999), we need an initial value $\boldsymbol{\psi}_0$ and update it to the maximizer of (29) a "few times". In Figure 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 (23) of the observed likelihood and the thick 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 (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 (31) moves "quickly" toward an estimate that is close to the true value. As expected, for given ϕ_0, (31) approximates well the observed likelihood only in a neighborhood of ψ_0. Unlike the estimator in (29), the estimator in (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 ψ_0, the estimate in (31) is close to the approximate likelihood in (23).

In Figure 3, we show the randomness feature of (31). In each panel, a fixed value of ϕ_0 is used. The solid line and vertical solid line are as in Figure 2. The

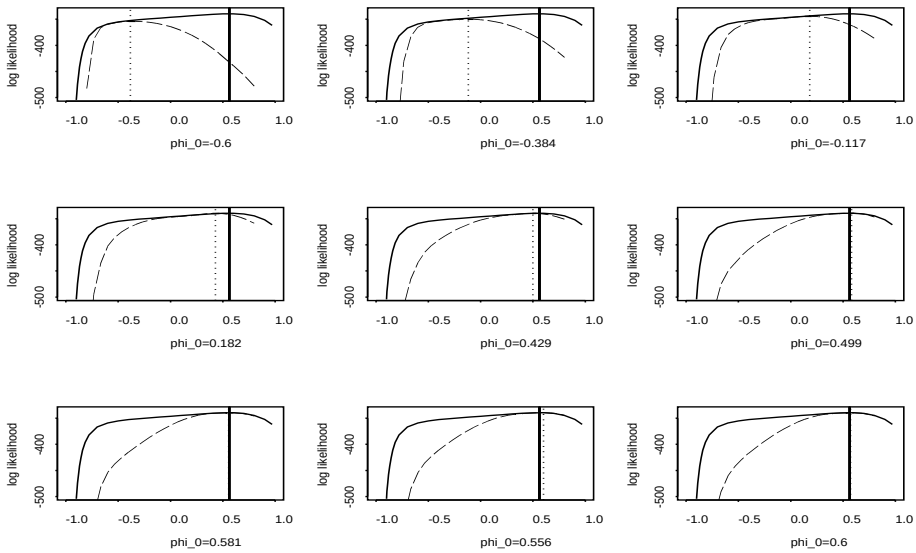

Figure 2: (*Single-sample*) From left to right and top to bottom, in each panel we show for a grid of values of ϕ, the logarithm of the estimation of the likelihood of a Poisson SSM. For the solid line, the estimates were obtained using (23) while for the dashed line (31) was used. ψ_0 is the optimizer (shown by the dotted vertical line) of (31) from the preceding panel. The solid vertical line shows the optimizer of (23).

lower and upper dotted lines are the minimum and maximum, respectively, of one hundred replicates of (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 (31).

3 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.

3.1 Stochastic Volatility Model

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

$$y_t = \sigma_t \xi_t = e^{\alpha_t/2}\xi_t, \qquad \alpha_t = \gamma + \phi\alpha_{t-1} + \eta_t, \tag{35}$$

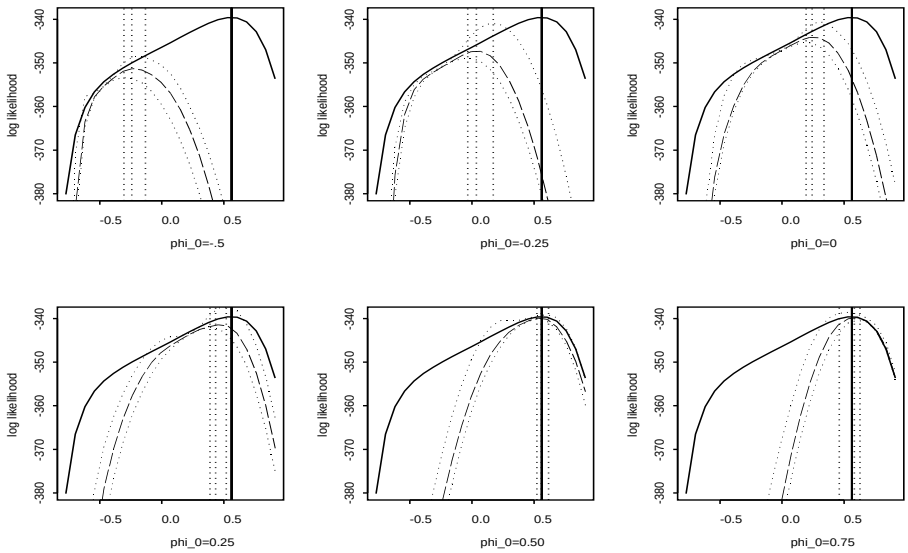

Figure 3: (*Single-sample*) From left to right and top to bottom, in each panel we show for a grid of values of ϕ, the logarithm of the estimation of the likelihood of a Poisson SSM. For the solid line, estimates were obtained using (23). The solid vertical line shows its optimizer. For the dashed line, estimations were obtained using (31) with ϕ_0 shown in the x axis. The dotted lines are the minimum and maximum respectively, of 100 replicates using (31). From left to right, the dotted vertical lines are the minimum, mean and maximum of the optimizers of these replicates.

where $\xi_t \sim$ iid $N(0,1)$, $\eta_t \sim$ iid $N(0,\sigma^2)$, $t = 1,\ldots,n = 1000$, and $|\phi| < 1$. In this case, $\boldsymbol{\psi} = (\gamma,\ \phi,\ \sigma^2)$. 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 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 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 (3). Equation (19) becomes

$$\tilde{\mathbf{y}}^j = \text{diag}\{\mathbf{1}/2 + \boldsymbol{\alpha}^j/2\}\text{diag}\{\mathbf{y}^2\}e^{-\boldsymbol{\alpha}^j} - \mathbf{1}/2 + \mathbf{V}\boldsymbol{\mu}. \tag{36}$$

To compare the estimate of $\boldsymbol{\psi}$ obtained by maximizing (12) with those obtained by maximizing either (29) or (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 (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

$$\hat{L}(\boldsymbol{\psi}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(\mathbf{y},\boldsymbol{\alpha}^{(i)}|\boldsymbol{\psi})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y};\boldsymbol{\psi})}. \tag{37}$$

For a fixed value $\boldsymbol{\psi}_0$, estimate $L(\boldsymbol{\psi};\mathbf{y})$ by

$$\hat{L}(\boldsymbol{\psi}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(\mathbf{y},\boldsymbol{\alpha}^{(i)}|\boldsymbol{\psi})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y},\boldsymbol{\psi}_0)}, \tag{38}$$

where $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(N)}$ is a sample from $p_a(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}_0)$. As in (31), to estimate $\boldsymbol{\psi}$ by maximizing (38), a few updatings of $\boldsymbol{\psi}_0$ is recommended.

For our simulation study, we consider samples of size $n = 500$ and compute mean and root mean squared errors over 500 simulated realizations for each of the nine parameters given in Table 1. The results are shown in Table 2. In this Table, AL denotes the estimates obtained by maximizing the approximating likelihood given in (12) and MCL denotes estimates obtained by maximizing the estimate of the likelihood in (37). To attain numerical stability, the same noise was used to generate replicates of $\boldsymbol{\alpha}^{(j)}$'s as a function of the AR parameters. MCL0 denotes estimates obtained by maximizing the single-sample estimate of the likelihood in (38). For this case, we start $\boldsymbol{\psi}_0$ 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

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

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.

3.2 Poisson Model

For the second simulation example, we assume that $p(y_t|\alpha_t;\boldsymbol{\psi})$ 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 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 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 3. The results are shown in Table 4. In this table, AL denotes the estimates obtained by maximizing the approximated likelihood given in (23) and MCL denotes estimates obtained by maximizing the estimate of the likelihood in (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.

3.3 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 $\hat{\boldsymbol{\psi}}_{AL}$ be the maximizer of the approximate likelihood in (12). Following Efron and Tibshirani (1993), the *bootstrap bias correction* of the estimate $\hat{\boldsymbol{\psi}}_{AL}$ of $\boldsymbol{\psi}$ is given by

$$\bar{\boldsymbol{\psi}}_{AL} = \hat{\boldsymbol{\psi}}_{AL} - \widehat{\text{bias}}, \tag{39}$$

where $\widehat{\text{bias}} = \bar{\boldsymbol{\psi}}^* - \hat{\boldsymbol{\psi}}_{AL}$, and $\bar{\boldsymbol{\psi}}^*$ is the average of B bootstrap estimates $\hat{\boldsymbol{\psi}}_1^*, \ldots, \hat{\boldsymbol{\psi}}_B^*$. Here, the bootstrap estimate $\hat{\boldsymbol{\psi}}_j^*$ is the maximizer of the approximate likelihood in

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

(12) computed with a realization $y_1^* \ldots, y_n^*$ drawn from the state-space model that has true parameters $\hat{\boldsymbol{\psi}}_{AL}$. The *bootstrap estimate of the variance* of the estimator $\hat{\boldsymbol{\psi}}_{AL}$ is

$$\widehat{\text{var}}(\hat{\boldsymbol{\psi}}_{AL}) = \frac{1}{B-1}\sum_{j=1}^{B}(\hat{\boldsymbol{\psi}}_j^* - \bar{\boldsymbol{\psi}}^*)(\hat{\boldsymbol{\psi}}_j^* - \bar{\boldsymbol{\psi}}^*)^T. \tag{40}$$

To assess the performance of the bootstrap bias correction, we conducted a simulation study on three Poisson models with parameters given in the second row of Table 3. As seen in Table 4, ϕ has a moderate bias in these models. The results of the simulation are given in Table 5. BC refers to the average of 1000 bias corrected estimates defined in (39) computed with B=100 bootstrap 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 second row of Table 3. The row labeled AL is the average of the 1000 simulated $\hat{\boldsymbol{\psi}}_{AL}$ 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 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.

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 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 (23) and each BC estimate is the bootstrap bias correction estimate defined in (39).

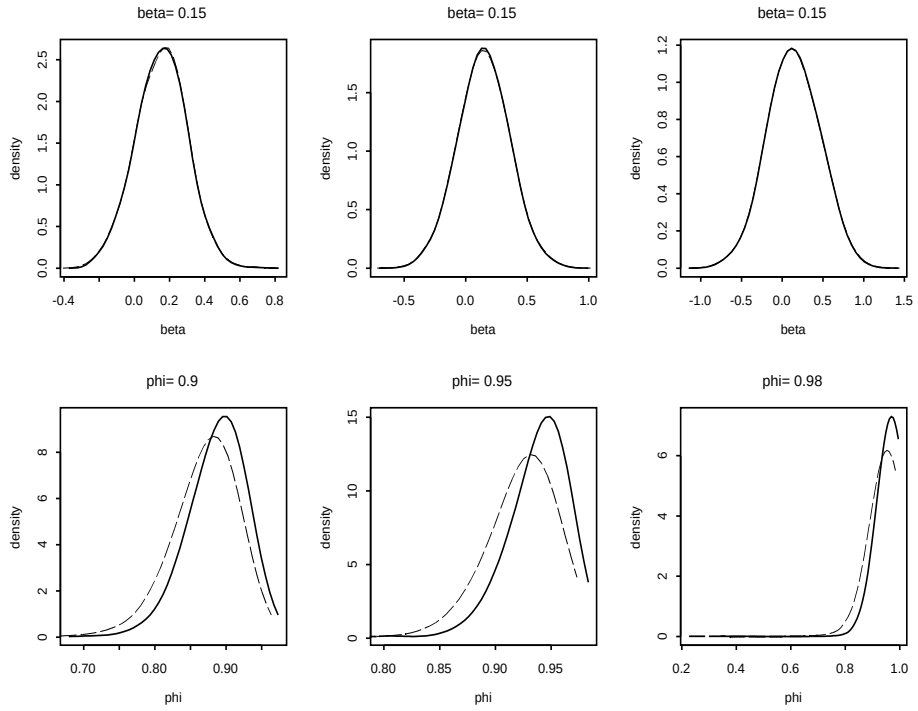

Figure 4: Parameter densities for β (first row) and ϕ (second column) for estimations AL (solid line) and BC (dotted line) for three Poisson state-space models.

3.4 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 (35) to model $y_t := \log(z_t) - \log(z_{t-1})$. Setting the parameter vector $\boldsymbol{\psi} := (\gamma, \phi, \sigma^2)$, Table 6 shows various estimates of $\boldsymbol{\psi}$. The second column, labeled as AL, contains the estimate of $\boldsymbol{\psi}$ obtained by maximizing (12). The column labeled MCL contains the estimate of $\boldsymbol{\psi}$ obtained by maximizing (37). MCE

denotes Monte Carlo error and is obtained as the standard error of 1000 estimates of $\boldsymbol{\psi}$, using for each estimate the same observations $y_1, \ldots, y_{945}$. The standard error of the estimates AL and MCL are obtained using (40). 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.

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 6: Parameter estimates for the Pound-Dollar exchange rates data. AL and MCE are the maximizers of (12) and (37), 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.

3.5 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 $\lambda_t := e^{\alpha_t + \mathbf{x}_t^T \boldsymbol{\beta}}$. Here, $\boldsymbol{\beta}^T := (\beta_1, \ldots, \beta_6)$, $\mathbf{x}_t$ 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 $N(0, \sigma^2)$, $t = 1, \ldots, n = 1000$, and $|\phi| < 1$. The vector of parameters of this SSM is $\boldsymbol{\psi} = (\boldsymbol{\beta}, \phi, \sigma^2)$. Table 7 contains the results of two estimation procedures. Columns 2 and 5 labeled as AL and MCL respectively, contain the estimates of $\boldsymbol{\psi}$ obtained by maximizing (12) and (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 $y_1, \ldots, y_{168}$.

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 6 even when we have used the same number of draws ($N = 1000$) to compute the Monte Carlo integration in (37) and (29) respectively. This may not

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 7: Parameter estimates for the polio data. AL and MCE are the maximizers of (12) and (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.

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.

3.6 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 (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 (43) 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 (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 (37), an estimate of the posterior mean is given by

$$\hat{\boldsymbol{\alpha}} = \frac{1}{N\hat{L}(\boldsymbol{\psi};\mathbf{y})} \sum_{i=1}^{N} \boldsymbol{\alpha}^{(i)} \frac{p(\mathbf{y},\boldsymbol{\alpha}^{(i)}|\boldsymbol{\psi})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y};\boldsymbol{\psi})}. \tag{41}$$

As an example, for the Pound-Dollar exchange rates and polio data let $\boldsymbol{\psi}$ be the AL estimate from Tables 6 and 7 respectively. Using $N = 1000$ in (41), $\hat{\boldsymbol{\alpha}}$ was

computed. In Figures 5 and 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 (16). 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.

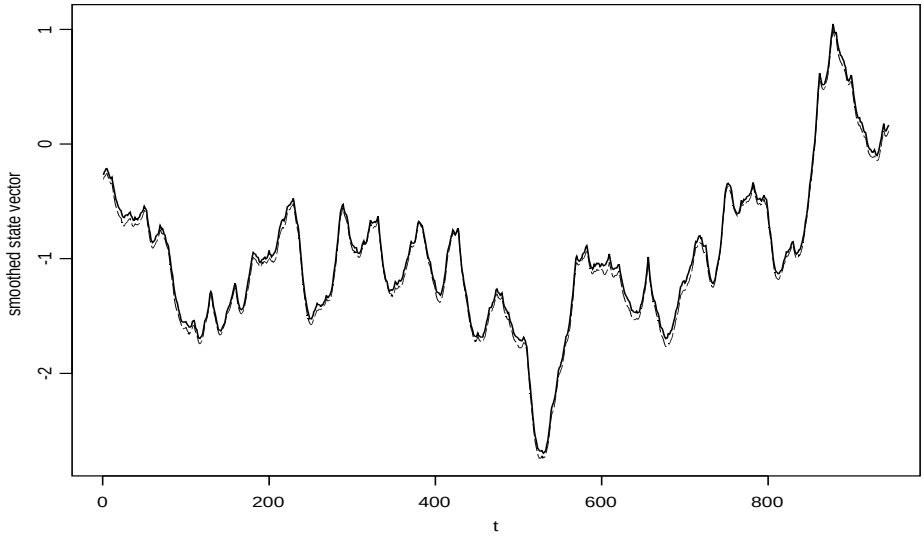

Figure 5: (*smoothed state vector*) For the Pound-Dollar exchange rates data, the solid line shows estimate of the posterior mean of the state vector and the dashed line shows its posterior mode.

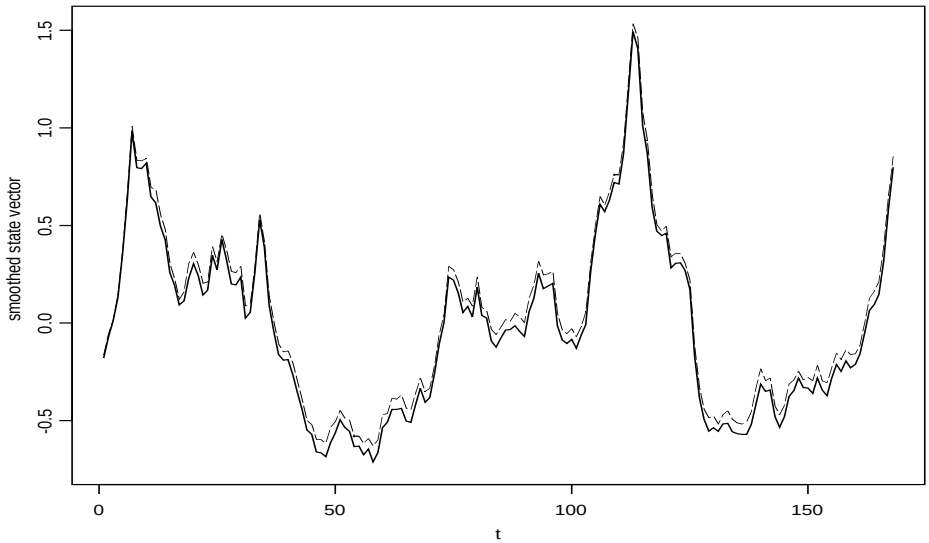

Figure 6: (*smoothed state vector*) For the Polio data, the solid line shows estimate of the posterior mean of the state vector and the dashed line shows its posterior mode.

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

$$p_i := \frac{w_i}{\sum_{i=1}^N w_i}, \quad w_i = \frac{p(\boldsymbol{\alpha}^{(i)}|\mathbf{y},\boldsymbol{\psi})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y},\boldsymbol{\psi})} \propto \frac{L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}^i)}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y},\boldsymbol{\psi})}, \tag{42}$$

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 (10) were a good approximation to $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$, for $M-n$ large, the squared generalized distances

$$d_j^2 := (\tilde{\boldsymbol{\alpha}}^{(j)} - \boldsymbol{\alpha}^*)^T(\mathbf{K}^* + \mathbf{V})(\tilde{\boldsymbol{\alpha}}^{(j)} - \boldsymbol{\alpha}^*), \quad j = 1,\ldots,M, \tag{43}$$

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.

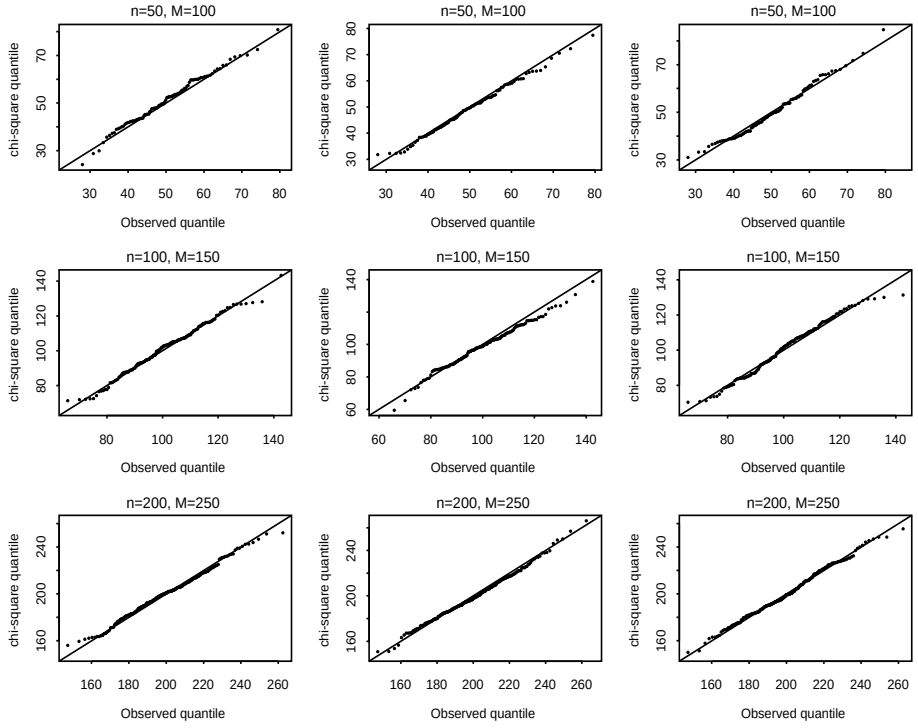

Figure 7: (*Chi-squared QQ-plots*) The QQ-plot from i-th row and j-th column was obtained using a SIR sample $\tilde{\boldsymbol{\alpha}}^{(1)},\tilde{\boldsymbol{\alpha}}^{(2)},\ldots,\tilde{\boldsymbol{\alpha}}^{(M)}$ from $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}_j)$ by resampling a sample of size 5000 from the approximation $p_a(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi}_j)$.

To illustrate this techniques, 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 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 parameter value of $\boldsymbol{\psi} = \boldsymbol{\psi}_j$, where $\boldsymbol{\psi}_1 := (0.2, 0.8, 0.002)$, $\boldsymbol{\psi}_2 := (0.373, 0.9, 0.012)$ and $\boldsymbol{\psi}_3 := (0.5, 0.95, 0.02)$. 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 $d_{(j)}^2, j = 1, \ldots, M$ and the Chi-squared quantiles can be used to test any departure from normality of $p_a(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi})$ (Johnson and Wichern, 1998). The nine correlations r_Q for the data used to create Figure 7 are shown in the last three columns of Table 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 (10) may be a reasonable approximation for the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi})$.

				r_Q	
N	M	$r_{0.05}$	$\boldsymbol{\psi}_1$	$\boldsymbol{\psi}_2$	$\boldsymbol{\psi}_3$
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 8: Correlation coefficients of the points in the QQ-plots from figure 7.

4 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 $p_a(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi})$ to the posterior distribution of the states. For the exponential family of distributions in standard form, the approximate distribution $p_a(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi})$ coincides with the approximation to the conditional distribution of $\boldsymbol{\alpha}$ 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 (1998). 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.

5 Appendix. 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 (16) and the determinant needed in approximation (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)

$$\mathbf{X} = \mathbf{C}(\mathbf{X} - \hat{\mathbf{X}}), \tag{44}$$

where

$$\mathbf{C} := \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ \theta_{11} & 1 & 0 & \ldots & 0 \\ \theta_{22} & \theta_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \ldots & 1 \end{pmatrix}. \tag{45}$$

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 (44), it follows that

$$\mathbf{\Gamma} = \mathbf{C}\mathbf{D}\mathbf{C}^T, \tag{46}$$

where $\mathbf{D} := E\{(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T\} = \mathrm{diag}\{\nu_0, \nu_1, \ldots \nu_{n-1}\}$. 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 (46), we obtain

$$|\mathbf{\Gamma}| = |\mathbf{C}\mathbf{D}\mathbf{C}^T| = |\mathbf{D}| = \prod_{j=0}^{n-1} \nu_j, \tag{47}$$

Now, using using (44) and (46), we can show that

$$\mathbf{\Gamma}^{-1}\mathbf{X} = \mathbf{C}^{-T}\mathbf{e}, \tag{48}$$

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

$$\alpha_t = \phi\alpha_{t-1} + \eta_t, \tag{49}$$

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 (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

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

Now, let $\boldsymbol{\alpha}^j$ be the current iterate to the value of $\boldsymbol{\alpha}^*$. From (15) and (24)

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

Since no intercept is included in the AR(1) process in (49), $\boldsymbol{\mu} = \mathbf{0}$. Thus, $\tilde{\mathbf{y}}^j$ defined in (25) is given by

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

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

$$\nu_j = \begin{cases} \gamma_{11}, & \text{if } j = 0, \\ \gamma_{j+1,j+1} - \theta_{j1}^2 \nu_{j-1}, & \text{if } j = 1, \dots, n-1, \end{cases} \tag{50}$$

$$\hat{X}_j = \begin{cases} 0, & \text{if } j = 1, \\ \theta_{j-1,1}(X_{j-1} - \hat{X}_{j-1}), & \text{if } j = 2, \dots, n \end{cases}$$

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

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

Once these values have been computed, then the vector of normalized residuals $\mathbf{e}$ needed in (48) is easily obtained, and the iteration in (20) becomes

$$\boldsymbol{\alpha}^{j+1} = (\mathbf{K}^j + \mathbf{V})^{-1} \tilde{\mathbf{y}}^j = \boldsymbol{\Gamma}^{-1} \mathbf{X} = \mathbf{C}^{-T} \mathbf{e} \tag{51}$$

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 (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 (23), set $\boldsymbol{\Gamma} := \mathbf{K}^* + \mathbf{V}$, where $\mathbf{K}^* = e^{\beta} \text{diag}\{e^{\boldsymbol{\alpha}^*}\}$ -see (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 (51). Then, from (47),

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

where ν_j, $j = 0, \ldots, n-1$ must be computed as in (50). Extensions to state processes following an AR(p) model can be handled in a similar fashion.

Acknowledgments

This research was supported in part by NSF grant DMS-0308109 and EPA STAR CR-829095, and a scholarship from Consejo Nacional de Ciencia y Tecnologia (CONACYT). The authors also wish thank W.T.M Dunsmuir for his comments and helpful suggestions.

References

[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] Davis, R. A., Dunsmuir, W. T. M. and Wang, Y. (1988). "Modelling Time Series of Count Data." In Asymptotics, Nonparametrics and Time Series (ed Subir Ghosh), Marcel Dekker.

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

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

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

[11] 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.

[12] 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.

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

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

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

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

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

[18] 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.

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

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

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

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

[23] 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.

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

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