Statistica Sinica **15**(2005), 381-406

ESTIMATION FOR STATE-SPACE MODELS BASED ON A LIKELIHOOD APPROXIMATION

Richard A. Davis and Gabriel Rodriguez-Yam

Colorado State University

Abstract: Typically, the likelihood function for non-Gaussian state-space models cannot be computed explicitly and simulation-based procedures, such as importance sampling or MCMC, are commonly used to estimate model parameters. In this paper we consider two alternative estimation procedures, each based on an approximation to the likelihood function. In the first approach, the approximation is computed and maximized directly, and this results in a fast estimation procedure without resort to simulation. Moreover, estimates are competitive with those produced using simulation-based procedures. The speed of the procedure makes it viable to fit a wide range of potential models to the data, and it allows for bootstrapping parameter estimates. In the second approach, importance sampling is used to estimate the error in the approximation to the likelihood. This particular simulation-based method is extremely quick and accurate, since the error term is well-approximated by a linear function.

Key words and phrases: Approximate likelihood, importance sampling, non-linear state space models, stochastic volatility models.

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 $Y_1, Y_2, \ldots,$ is a time series of observations and $\alpha_1, \alpha_2, \ldots$ are the respective "state variables", then it is assumed that

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

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

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

where p is a non-negative integer and $\eta_t \sim$ i.i.d. $N(0,\ \sigma^2)$, $t = 1, 2, \ldots$. Perhaps the most important special case is when the conditional distribution $p(y_t|\alpha_t)$ is a member of the exponential family, an extremely rich class of distributions. Durbin and Koopman (1997) and Kuk (1999) consider this in the form

$$p(y_t|\alpha_t) = e^{(\mathbf{x}_t^T\boldsymbol{\beta}+\alpha_t)y_t - b(\mathbf{x}_t^T\boldsymbol{\beta}+\alpha_t)+c(y_t)}, \tag{2}$$

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

One special application that we 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 (2). 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) and Davis, Dunsmuir and Wang (1998).

Another noteworthy application of the SSM that we 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, Polson and Rossi (1994), Briedt and Carriquiry (1996), Harvey and Streibel (1998), Sandmann and Koopman (1998), Geweke and Tanizaki (1999) and 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 (1). With this specification, the likelihood based on the "complete data" $(\mathbf{y}, \boldsymbol{\alpha})$ of the SSM is

$$\begin{aligned} L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) &= p(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\alpha}|\boldsymbol{\lambda}) \\ &= \left(\prod_{t=1}^{n} p(y_t|\alpha_t,\boldsymbol{\theta})\right)|\mathbf{V}|^{1/2}e^{-(\boldsymbol{\alpha}-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}-\boldsymbol{\mu})/2}/(2\pi)^{n/2}, \end{aligned} \tag{3}$$

where $\mathbf{V}^{-1} := \mathrm{Cov}\,\{\boldsymbol{\alpha}\}$, $\boldsymbol{\mu} = \gamma/(1-\phi_1-\ldots-\phi_p)\mathbf{1}$ is the vector of means of the state process, and $\mathbf{1}$ is a vector of ones. From (3) it follows that the likelihood of the observed data is given by the n-fold integral

$$L(\boldsymbol{\psi};\mathbf{y}) = \int L(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})d\boldsymbol{\alpha}. \tag{4}$$

Except in simple cases, the integral in (4) cannot 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. 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 $g(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ using the same noise sequence. Kuk (1999) advocates a "single-sample" approach in 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 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 $\boldsymbol{\psi}_0$ with the new maximizer at each iteration.

While the formulation of the importance density based on the working model in the Durbin and Koopman setup is straightforward when $p(y_t|\alpha_t;\boldsymbol{\theta})$ is a member of the "standard" exponential family of distributions, it can be tedious and difficult to formulate the working model 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 log returns. Since the time series of log returns may have values near zero, this logarithmic tranformation may create additional modeling obstacles.

A Monte Carlo EM algorithm treating the unobserved $\boldsymbol{\alpha}$'s as missing values was proposed by Chan and Ledolter (1995). At the ith 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 follow a different approach to obtain an approximation $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ to the distribution $p(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$. Unlike Durbin and Koopman's method, our 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, yet it coincides with Durbin and Koopman's importance distribution for this family. This approximation $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ is obtained in Section 2 and it is used to obtain an analytical approximation to (4) which can be maximized to produce an estimate of $\boldsymbol{\psi}$. In a second method, the approximation error is computed using a first order Taylor series expansion. The estimates obtained when this linear approximation is used approximate very well the importance sampling estimates, as is shown in an example in Section 3. The innovations algorithm (Brockwell and Davis (1991)) can be used to speed up the computation of these estimates. In Section 4 we demonstrate the good performance of these procedures via simulation experiments. Illustration use two time series: the monthly number of U.S. cases of poliomyelitis for 1970 to 1983 (Zeger (1988)) is analyzed using a Poisson state-space model; historical pound to dollar exchange rates (Harvey et al. (1994)) are analyzed using a stochastic volatility model.

The quality of the analytical 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 4. Application of the innovations algorithm to the problems considered in Sections 2, 3 and 4 is given in the Appendix.

2. Parameter Estimation

In this section we find a factorization of the observed likelihood $L(\boldsymbol{\psi};\mathbf{y})$ (4) 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 likelihood based on $\mathbf{y},\boldsymbol{\alpha}$ is given by

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

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 $(\partial/\partial\boldsymbol{\alpha})\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha}) = \mathbf{0}$. From (3), 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 $R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*) := \ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha}) - T(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)$ is the corresponding remainder, then

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

where

$$h^* := \ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*} \text{ and } \mathbf{K}^* := -\frac{\partial^2}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}. \tag{5}$$

Thus,

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

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 $\ell(\boldsymbol{\psi};\mathbf{y},\boldsymbol{\alpha})$ when the term $R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)$ is omitted. It follows that

$$p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi}) = \boldsymbol{\phi}(\boldsymbol{\alpha};\boldsymbol{\alpha}^*,(\mathbf{K}^*+\mathbf{V})^{-1}), \tag{6}$$

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

$$L(\boldsymbol{\psi};\mathbf{y}) = L_a(\boldsymbol{\psi};\mathbf{y})\mathrm{Er}_a(\boldsymbol{\psi}), \tag{7}$$

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

$$L_a(\boldsymbol{\psi};\mathbf{y}) := \frac{|\mathbf{V}|^{1/2}}{|\mathbf{K}^*+\mathbf{V}|^{1/2}}e^{h^*-\frac{1}{2}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})^T\mathbf{V}(\boldsymbol{\alpha}^*-\boldsymbol{\mu})}, \tag{8}$$

obtained when $e^{R(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)}$ is ignored in $\mathrm{Er}_a(\boldsymbol{\psi})$. If $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ is highly concentrated around $\boldsymbol{\alpha}^*$, the approximation error should be close to 1.

The approximation $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ can be used to implement a Monte Carlo estimation of the likelihood $L(\boldsymbol{\psi};\mathbf{y})$. Suppose $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(N)}$ are draws from $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$. In the Appendix, a quick procedure to sample from this distribution is provided. Equation (7) suggests the estimator

$$\hat{L}(\boldsymbol{\psi};\mathbf{y}) = L_a(\boldsymbol{\psi};\mathbf{y})\hat{\mathrm{E}}\mathrm{r}_a(\boldsymbol{\psi}), \tag{9}$$

where

$$\hat{\mathrm{E}}\mathrm{r}_a(\boldsymbol{\psi}) = \frac{1}{N}\sum_{i=1}^{N} e^{R(\boldsymbol{\alpha}^{(i)};\boldsymbol{\alpha}^*)}. \tag{10}$$

We see below that for the standard exponential family of distributions, the estimator in (9) is the estimator in (15) proposed by Durbin and Koopman (1997). We call IS the approximate MLE of $\boldsymbol{\psi}$ obtained when (9) is maximized with respect to $\boldsymbol{\psi}$. Evaluations of this function require Monte Carlo integrations, which can be expensive. We propose two approximations to the MLE of $\boldsymbol{\psi}$ that do not need Monte Carlo integration for every value of $\boldsymbol{\psi}$. The first estimator, AL, is obtained by maximizing the approximate likelihood $L_a(\boldsymbol{\psi};\mathbf{y})$ with respect to $\boldsymbol{\psi}$. Since the evaluation of (8) does not involve simulation, AL is obtained faster than IS. For the second estimator, let $\hat{\mathrm{e}}(\boldsymbol{\psi}) := \log \hat{\mathrm{Er}}_a(\boldsymbol{\psi})$ and let $\hat{\boldsymbol{\psi}}_{AL}$ be the AL estimate of $\boldsymbol{\psi}$. Let $T_{\hat{\mathrm{e}}}(\boldsymbol{\psi};\hat{\boldsymbol{\psi}}_{AL})$ be the linear approximation to $\hat{\mathrm{e}}(\boldsymbol{\psi})$ at $\hat{\boldsymbol{\psi}}_{AL}$, i.e.,

$$T_{\hat{\mathrm{e}}}(\boldsymbol{\psi};\hat{\boldsymbol{\psi}}_{AL}) = \hat{\mathrm{e}}(\hat{\boldsymbol{\psi}}_{AL}) + \mathbf{q}_{AL}^T(\boldsymbol{\psi}-\hat{\boldsymbol{\psi}}_{AL}),$$

where $\mathbf{q}_{AL} := (\partial/\partial\boldsymbol{\psi})\hat{\mathrm{e}}(\boldsymbol{\psi})|_{\hat{\boldsymbol{\psi}}_{AL}}$. Let the estimator of the likelihood be

$$L_c(\boldsymbol{\psi};\mathbf{y},\hat{\boldsymbol{\psi}}_{AL}) = L_a(\boldsymbol{\psi};\mathbf{y})\exp\{T_{\hat{\mathrm{e}}}(\boldsymbol{\psi};\hat{\boldsymbol{\psi}}_{AL})\}. \tag{11}$$

Our second estimator, AIS, results from maximizing $L_c(\boldsymbol{\psi};\mathbf{y},\hat{\boldsymbol{\psi}}_{AL})$ with respect to $\boldsymbol{\psi}$. To evaluate this function for distinct values of $\boldsymbol{\psi}$, $\mathbf{q}_{AL}$ needs to be computed only once. For this reason, AIS is much faster than IS. Of the three estimators, AL is the fastest. In order to compute AIS, the derivative of $\hat{\mathrm{e}}(\boldsymbol{\psi})$ at $\hat{\boldsymbol{\psi}}_{AL}$ is found numerically. That is, for δ small,

$$\frac{\partial}{\partial\psi_j}\hat{\mathrm{e}}(\boldsymbol{\psi})|_{\hat{\boldsymbol{\psi}}_{AL}} \approx \frac{\hat{\mathrm{e}}(\hat{\boldsymbol{\psi}}_{AL}+\boldsymbol{\delta}_j)-\hat{\mathrm{e}}(\hat{\boldsymbol{\psi}}_{AL})}{\delta},$$

where $\boldsymbol{\delta}_j$ is a vector consisting of value δ at position j and 0 elsewhere. This numerical procedure avoids an explicit calculation of the derivative of $\boldsymbol{\alpha}^*$ with respect to $\boldsymbol{\psi}$.

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

$$\mathbf{k}^j := \frac{\partial}{\partial\boldsymbol{\alpha}}\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^j} \quad\text{and}\quad \mathbf{K}^j := -\frac{\partial^2}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\ell(\boldsymbol{\theta};\mathbf{y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^j}, \tag{12}$$

then the Newton-Raphson algorithm gives $\boldsymbol{\alpha}^{j+1} = \boldsymbol{\alpha}^j - (\ddot{\ell}^j)^{-1}\dot{\ell}^j$, where

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

Let

$$\tilde{\mathbf{y}}^j := \mathbf{k}^j + \mathbf{K}^j \boldsymbol{\alpha}^j + \mathbf{V}\boldsymbol{\mu}. \tag{13}$$

Substituting this and the derivatives into the Newton-Raphson iteration, we obtain

$$\boldsymbol{\alpha}^{j+1} = (\mathbf{K}^j + \mathbf{V})^{-1} \tilde{\mathbf{y}}^j. \tag{14}$$

Each iteration of (14) needs the inversion of an $n \times n$ matrix, while each evaluation of (8) 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. Its implementation is described in the Appendix.

Application to the exponential family

Assume the exponential family density

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

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 (5) becomes $\mathbf{K}^* = \text{diag}\{(\partial^2/\partial\alpha_t^2) b(\mathbf{x}_t^T\boldsymbol{\beta} +\alpha_t)|_{\alpha_t^*}\}$ and the approximation to the observed likelihood is

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

From (12), $\mathbf{k}^j = \mathbf{y} - \dot{\mathbf{b}}^j$, where $\dot{\mathbf{b}}^j := (\partial/\partial\boldsymbol{\alpha})\mathbf{1}^T\mathbf{b}(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha})|_{\boldsymbol{\alpha}^j}$. Hence, $\tilde{\mathbf{y}}^j := \mathbf{y} - \dot{\mathbf{b}}^j + \mathbf{K}^j\boldsymbol{\alpha}^j + \mathbf{V}\boldsymbol{\mu}$, where $\mathbf{K}^j$ is defined in (12).

In (9), $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ is used as an importance density to estimate $L(\boldsymbol{\psi};\mathbf{y})$. 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 (4) 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 iteratively solving

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

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

initialized with $\mu_t = 0$ and H_t arbitrary. Here, the $\hat{\alpha}_t$ are found by routine application of the Kalman filtering and smoothing algorithms. Let E_g denote the conditional expectation operator under the approximating model. Durbin and Koopman (1997) found that (4) can be expressed as

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

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 (4) is estimated as

$$\hat{L}(\boldsymbol{\psi}) = L_g(\boldsymbol{\psi}) \frac{1}{N} \sum_{i=1}^{N} \frac{p(\mathbf{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\theta})}{g(\mathbf{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi})}. \tag{15}$$

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, the same noise sequence is used to construct $\boldsymbol{\alpha}^{(i)}$ for values of $\boldsymbol{\psi}$.

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

$$H_t^{-1} = \ddot{b}_t, \qquad \mu_t = y_t - \hat{\alpha}_t - \ddot{b}_t^{-1}(y_t - \dot{b}_t). \tag{16}$$

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 (16), 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 (5). Thus, $p_a(\boldsymbol{\alpha}|\mathbf{y};\boldsymbol{\psi})$ in (6) 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 (16), it follows that $g(\mathbf{y}|\boldsymbol{\alpha};\boldsymbol{\theta}) = A(\boldsymbol{\alpha}^*)e^{T(\boldsymbol{\alpha};\boldsymbol{\alpha}^*)}$, where $A(\boldsymbol{\alpha}^*) := (2\pi)^{-n/2}|\mathbf{K}^*|^{1/2}e^{-h^*-(1/2)\mathbf{k}^{*T}(\mathbf{K}^*)^{-1}\mathbf{k}^*}$, $\mathbf{k}^* = \mathbf{y}-(\partial/\partial\boldsymbol{\alpha})\mathbf{1}^T\mathbf{b}(\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\alpha})|_{\boldsymbol{\alpha}^*}$, and h^* and $\mathbf{K}^*$ are defined in (5). Then, $L_g(\boldsymbol{\psi};\mathbf{y}) = A(\boldsymbol{\alpha}^*)L_a(\boldsymbol{\psi};\mathbf{y})$. Using this factorization of $L_g(\boldsymbol{\psi};\mathbf{y})$, it can be shown that the estimates in (9) and (15) produce identical results.

To get a feel for how these 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

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

where $\eta_t \sim$ i.i.d. $N(0, 0.3)$, $t = 1,\ldots, n = 200$. 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 and 100 replicates of the estimation of the approximation error in (10) were obtained in a grid of points of ϕ for each of the values $N = 10$ and $N = 1,000$. Also, the linear approximation $T_{\hat{\mathrm{e}}}(\boldsymbol{\psi}; \hat{\boldsymbol{\psi}}_{AL})$ to each replicate was computed, where $\hat{\boldsymbol{\psi}}_{AL} = 0.51$. To compute $\hat{\mathrm{e}}(\boldsymbol{\psi})$, $\hat{\mathrm{e}}(\hat{\boldsymbol{\psi}}_{AL})$ and $\mathbf{q}_{AL}$ the same value of N and the same realizations $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(N)}$, described in the last paragraph of the appendix, were used. On the left panel of Figure 1, with $N = 10$, the long dashed line is a typical replicate. The solid lines are the pointwise minimum, mean and maximum of the replicates. Notice that in a neighborhood of $\hat{\phi}_{AL}$, replicates are remarkably linear as a function of ϕ. The short-dashed line is the linear approximation $T_{\hat{\mathrm{e}}}(\boldsymbol{\psi}; \hat{\boldsymbol{\psi}}_{AL})$ of the typical replicate (long-dashed line). The dotted lines are the pointwise minimum, mean and maximum of the linear approximation of the replicates. On the right panel of this figure, $N = 1,000$. Notice that the estimator in (10) has “large” Monte Carlo error even for $N = 1,000$. For each replicate of the approximation error for $N = 1,000$, the estimator to the likelihood (9) and its approximation (11) were computed. In Figure 2, the long-dashed line is the likelihood estimation computed with the typical approximation from Figure 1. The short-dashed line is its corresponding approximation (11). The thick solid line is the AL estimate of the observed likelihood. In this figure, notice that (11) is very close to (9).

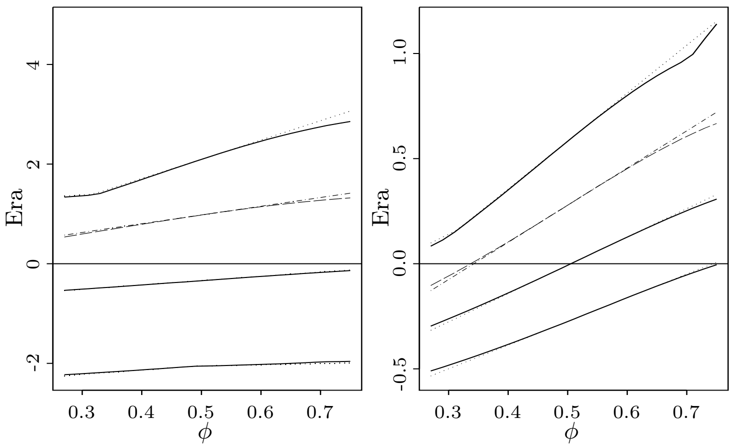

Figure 1. (*Approximation error*) For a grid of values of ϕ, the solid lines are the pointwise minimum, mean and maximum of 100 replicates of (10), using (a) $N = 10$ (left panel) and (b) $N = 1,000$ (right panel), for a Poisson SSM. The long-dashed line is a typical replicate and the short-long dashed line is its linear approximation $T_{\hat{\mathrm{e}}}(\boldsymbol{\psi}; \hat{\boldsymbol{\psi}}_{AL})$. The dotted lines are the pointwise minimum, mean and maximum of the linear approximation of these replicates.

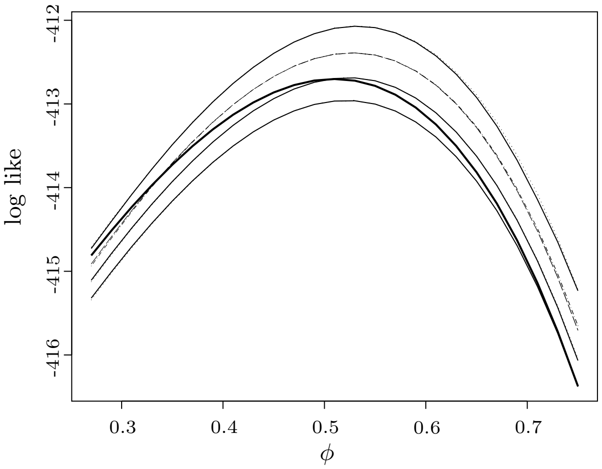

Figure 2. (*Many samples*) For a grid of values of ϕ, distinct estimations of the likelihood of a Poisson SSM are shown. The thick solid line is the AL approximation in (8). The solid lines are the minimum, mean and maximum of 100 replicates of the IS estimator in (9) using $N = 1,000$, and the long dashed line is a typical replicate. The short-dashed lines and dotted-dashed line are these values for the corresponding approximations in (11).

3. Comparison of IS and AIS

In this section we compare the IS and AIS estimators of Section 2 for two state space models for which $p(y_t|\alpha_t; \boldsymbol{\psi})$ is a Poisson distribution with rate $\lambda_t := e^{\beta+\alpha_t}$. For the first model the state process is an AR(2) and, for the second, the state process is an AR(3) process. For both cases, we set $\gamma = 0$ in (1). For the parameter values shown in Table 1, we generated 1,000 realizations of $\mathbf{y}$ of length $n = 200$. IS and AIS estimates were obtained for each realization using $N = 100$. The results are shown in Table 1. For each realization, the IS and AIS estimates were based on the same draws $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(100)}$ defined in the last paragraph of the Appendix. Notice that these estimates are essentially the same. Hence, from now on we use AIS instead of IS estimates.

To compare the speed of the IS and AIS procedures, we provide the computation times of the likelihood for the two models from Table 1. In the AR(2) case with $N = 100$, one importance sampling evaluation of the likelihood function using (9) took 0.0086 seconds. This includes the computation of $\boldsymbol{\alpha}^*$. For the same parameter values, an evaluation of (9) using $N = 1,000$ took 0.0456 seconds. In contrast, once $\mathbf{q}_{AL}$ is available, the evaluation time of the hybrid method based on (11) is approximately that of (8) which, for the fixed parameter values in question, was 0.0005 seconds. We note that $\mathbf{q}_{AL}$ needs to be computed only once and hence further evaluations of the likelihood at other parameter values are very fast. For this model, to obtain $\mathbf{q}_{AL}$ numerically, five Monte Carlo integrations are needed. For $N = 100$ and $N = 1,000$, the time for each Monte Carlo integration has been given above. For the AR(3) model, the evaluation times of (9) were 0.0092 and 0.0866 seconds for $N = 100$ and $N = 1,000$, respectively. An evaluation of (8) took 0.0006 seconds. Of course, to maximize an estimate of the likelihood function, multiple evaluations of the estimated function are required. All the reported times are based on an IBM ThinkPad, with a 1.6 GHz Intel Pentium M processor.

Table 1. Bias comparison of IS and AIS estimates for two Poisson state space models. The estimates are computed using $N = 100$. The bias estimates are based on 1,000 replicates. Root mean square error of estimates are reported below the bias.

Method	β	ϕ_1	ϕ_2	ϕ_3	σ
true	1	1.25	-0.75		0.2
IS	0.0057	-0.0020	0.0036		0.0075
	0.0832	0.0749	0.0668		0.0507
AIS	0.0057	-0.0020	0.0036		0.0075
	0.0832	0.0749	0.0669		0.0507
true	1	1.25	-0.75	0.2	0.2
IS	-0.0009	0.0027	-0.0004	0.0104	0.0031
	0.1139	0.2765	0.3815	0.2011	0.0783
AIS	-0.0009	0.0029	-0.0007	0.0106	0.0031
	0.1139	0.2759	0.3809	0.2009	0.0783

4. 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 datasets. One is a historical dataset of the Pound-Dollar exchange rates, first studied by Harvey, Ruiz and Shepard (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 these data.

4.1. Stochastic volatility model

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

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

where $\xi_t \sim$ i.i.d. $N(0,1)$, $\eta_t \sim$ i.i.d. $N(0,\sigma^2)$, $t = 1,\ldots,n = 1{,}000$, 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 2. 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) and Jacquier et al. (1994)), values of ϕ between 0.9 and 0.98 are of primary interest.

Table 2. Parameter values for a simulation experiment of nine stochastic volatility processes.

		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

The density of the observed series is

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

which differs slightly from the standard representation of the exponential family of distributions given in (2). Equation (13) becomes

$$\tilde{\mathbf{y}}^j = \frac{1}{2}\mathrm{diag}\{(1+\alpha_i^j)y_i^2\}e^{-\boldsymbol{\alpha}^j} - \mathbf{1}/2 + \mathbf{V}\boldsymbol{\mu}.$$

To compare the estimate of $\boldsymbol{\psi}$ obtained by maximizing (8) with those obtained by maximizing (15), 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. Our estimators AL and AIS avoid this transformation.

For our simulation study, we considered $n = 500$ and computed mean and root mean squared errors over 500 simulated realizations for each of the nine parameters given in Table 2. The results for the AL and AIS estimates are shown in Table 3. To attain numerical stability, the same noise was used to generate replicates of $\boldsymbol{\alpha}^{(j)}$'s as a function of the parameters.

For both 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 AIS. For CV=10, the mean squared errors are roughly equal. More importantly, 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 find smaller biases for σ and larger biases for γ.

Table 3. Comparison of AL and AIS estimates based on 500 replications. For each parameter, the bias and root mean square error of each procedure are shown. For the AIS estimates, $N = 100$ was used.

	γ	ϕ	σ	γ	ϕ	σ	γ	ϕ	σ
					CV=10				
true	-0.821	0.900	0.675	-0.411	0.950	0.484	-0.164	0.980	0.308
AL	0.081	0.010	0.012	0.080	0.010	0.005	0.092	0.011	-0.007
rmse	0.299	0.036	0.081	0.210	0.025	0.065	0.176	0.021	0.052
AIS	0.012	0.002	0.036	0.070	0.008	0.007	0.092	0.011	-0.009
rmse	0.238	0.029	0.078	0.195	0.023	0.064	0.172	0.021	0.053
					CV=1				
true	-0.736	0.900	0.363	-0.368	0.950	0.260	-0.147	0.980	0.166
AL	0.193	0.026	-0.013	0.132	0.018	-0.010	0.101	0.014	-0.009
rmse	0.514	0.069	0.091	0.342	0.046	0.068	0.212	0.029	0.048
AIS	0.117	0.016	-0.002	0.115	0.016	-0.009	0.099	0.013	-0.010
rmse	0.395	0.053	0.079	0.299	0.040	0.064	0.199	0.027	0.047
					CV=0.1				
true	-0.706	0.900	0.135	-0.353	0.950	0.096	-0.141	0.980	0.061
AL	0.321	0.045	-0.024	0.419	0.059	-0.040	0.334	0.047	-0.029
rmse	0.809	0.114	0.093	0.841	0.118	0.099	0.723	0.102	0.075
AIS	0.235	0.033	-0.012	0.336	0.047	-0.030	0.304	0.043	-0.025
rmse	0.676	0.095	0.078	0.676	0.095	0.081	0.646	0.091	0.065

4.2. 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 $\alpha_t = \phi\alpha_{t-1} + \eta_t$, $\eta_t \sim$ i.i.d. $N(0,\ \sigma^2)$, $t = 1, \ldots, n$, and $|\phi| < 1$. We again consider 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 4.

For this simulation, we considered realizations of length $n = 500$ and computed mean and root mean squared errors over 500 simulated realizations for each of the nine parameters given in Table 4. The results for the AL and AIS are shown in Table 5. For the AIS estimates, $N = 1,000$ was used. From this table, we notice that the bias for ϕ and σ is small for large D and is large for $D = 0.1$. In general, both methods produce remarkably similar results.

Table 4. Parameter values for a simulation experiment of nine Poisson state-space models.

		ϕ		
D		-0.50	0.50	0.9
10.0	β	-0.6130	-0.6130	-0.6130
	σ	1.2360	1.2360	0.6221
1.0	β	0.1501	0.1501	0.1501
	σ	0.6190	0.6190	0.3115
0.1	β	0.3732	0.3732	0.3732
	σ	0.2200	0.2200	0.1107

Table 5. Comparison of AL and AIS estimates based on 500 replications. For the AIS estimates, $N = 1,000$ was used. Root mean square errors of estimates are reported below each bias estimate.

	β	ϕ	σ	β	ϕ	σ	β	ϕ	σ
					$D = 10$				
true	-0.613	-0.500	1.236	-0.613	0.500	1.236	-0.613	0.900	0.622
AL	0.020	-0.006	-0.019	-0.008	0.042	0.021	0.002	0.014	0.006
	0.098	0.056	0.090	0.141	0.082	0.087	0.296	0.033	0.060
AIS	-0.031	0.022	0.062	-0.056	-0.005	0.094	-0.001	0.010	0.010
	0.093	0.063	0.100	0.143	0.065	0.120	0.294	0.029	0.058
					$D = 1$				
true	0.150	-0.500	0.619	0.150	0.500	0.619	0.150	0.900	0.312
AL	0.006	-0.010	-0.011	0.004	0.045	-0.003	-0.002	0.011	0.002
	0.050	0.084	0.057	0.075	0.107	0.061	0.148	0.039	0.048
AIS	-0.004	0.000	0.009	-0.001	0.012	0.010	-0.001	0.010	0.001
	0.049	0.089	0.059	0.073	0.091	0.061	0.148	0.037	0.048
					$D = 0.1$				
true	0.373	-0.500	0.220	0.373	0.500	0.220	0.373	0.900	0.111
AL	0.011	-0.084	-0.015	0.011	0.159	-0.022	0.004	0.091	-0.023
	0.041	0.360	0.094	0.047	0.393	0.092	0.061	0.249	0.071
AIS	0.005	-0.065	0.012	0.005	0.123	0.003	0.003	0.078	-0.016
	0.039	0.383	0.088	0.045	0.393	0.083	0.060	0.231	0.062

4.3. Bias correction via bootstrap

In the two simulation studies that we considered, the AL estimate of the parameters for the Poisson and stochastic volatility models can be slightly biased. Indeed, we see in the two applications to data, that AL and AIS can be very close to each other. Closeness here is "measured" via the Monte Carlo standard error of the AIS estimates. In this section, we show via simulation that the bias of the estimates can be reduced considerably using the bootstrap. Stoffer and Wall (1991) use 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 (8). Following Efron and Tibshirani (1993), the *bootstrap bias correction* of the estimate $\hat{\boldsymbol{\psi}}_{AL}$ of $\boldsymbol{\psi}$ is given by

$$\bar{\boldsymbol{\psi}}_{AL} = \hat{\boldsymbol{\psi}}_{AL} - \widehat{\text{bias}}, \tag{17}$$

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 (8) 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

$$\widehat{\text{var}(\hat{\boldsymbol{\psi}}_{AL})} = \frac{1}{B-1} \sum_{j=1}^{B} (\hat{\boldsymbol{\psi}}_j^* - \bar{\boldsymbol{\psi}}^*)(\hat{\boldsymbol{\psi}}_j^* - \bar{\boldsymbol{\psi}}^*)^T. \tag{18}$$

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 4. As seen in Table 5, ϕ has a moderate bias in these models. The results of the simulation are given in Table 6. BC refers to the average of 500 bias corrected estimates (17), computed with B=200 bootstrap estimates. The standard errors of the 1,000 bias corrected estimates are also shown in the table. The AL estimates were obtained from 500 simulated realizations from the state-space model having true parameters given in the second row of Table 4. The row labeled AL is the average of the 500 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 little alteration of the standard errors.

In Figure 3 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 (0.15, -0.5, 0.619), (0.15, 0.5, 0.619) and (0.15, 0.9, 0.312), respectively. As seen from these graphs, the BC estimates have essentially shifted the location of the AL estimates.

Table 6. Simulation results of bootstrap bias correction of AL estimates for three Poisson state-space models based on 500 replications. The rows labelled AL and BC are the average of the replications. Each BC estimate is the bootstrap bias correction estimate defined in (17) with B=200.

	β	ϕ	σ	β	ϕ	σ	β	ϕ	σ
true	0.150	-0.500	0.619	0.150	0.500	0.619	0.150	0.900	0.312
AL	0.144	-0.490	0.630	0.146	0.455	0.622	0.152	0.889	0.310
S.E.	0.049	0.083	0.056	0.075	0.097	0.061	0.148	0.038	0.048
BC	0.153	-0.500	0.615	0.147	0.501	0.618	0.153	0.906	0.302
S.E.	0.049	0.091	0.060	0.074	0.092	0.065	0.149	0.034	0.050

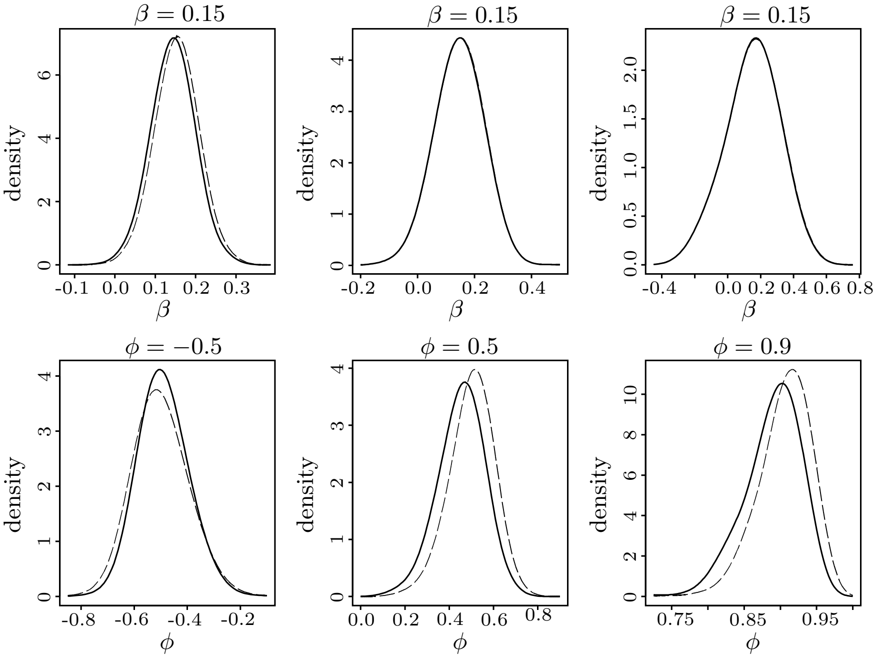

Figure 3. Parameter densities for β (first row) and ϕ (second column) for estimations AL (solid line) and BC (dotted line) for three Poisson state-space models.

4.4. 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 (4.1) to model $y_t := \log(z_t) - \log(z_{t-1})$. Setting the parameter vector $\boldsymbol{\psi} := (\gamma, \phi, \sigma^2)$, Table 7 shows the AL and AIS estimates of $\boldsymbol{\psi}$ and the corresponding bootstrap bias corrections. MCSE denotes Monte Carlo standard error and is obtained as the standard error of 1,000 AIS estimates of $\boldsymbol{\psi}$, using for each estimate the same observations $y_1, \ldots, y_{945}$. The standard error of AL and AIS estimates are obtained using (18). The columns labeled as BC are bootstrap bias corrections of AL and AIS, computed with $B = 500$ bootstrap estimates. Notice that the AL and AIS estimates are remarkably close. In fact, the difference between these estimates is due to the randomness of the AIS estimate. For example, two distinct AIS estimates of σ^2 are unlikely to differ by more than four times the Monte Carlo error, i.e., 0.0028, while the estimates AL and MCE of σ^2 differ by only 0.0006. In other words, we would not be able to differentiate the AL estimate from a "cloud" of AIS replicates.

Table 7. AL and AIS estimates for the Pound-Dollar exchange rates data. BC are bootstrap bias corrected estimates ($B = 500$) and S.E. are bootstrap estimates of the standard errors of AL and AIS, respectively. MCSE is the standard error of 1,000 AIS replicates.

Parameter	AL	S.E.	BC	AIS	MCSE	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

4.5. 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, 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

$$\mathbf{x}_t^T = (1, t/1,000, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6)),$$

and the state process is assumed to follow the AR(p) model in (1). The vector of parameters of this SSM is $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2)$. In the second and fourth columns in Table 8, the estimated likelihoods obtained for various values of p are shown, where AL is the maximum value of (8) and AIS ($N = 1,000$) is the maximum value of (11). The model in the last row ($p = 13$) was included to allow for other montly effects not captured by the deterministic mean function. The AIC values for both methods are shown. Based on the AIC values from the AL estimates, an AR(1) process seems adequate to model the states. However, based on the AIC values from the AIS estimates, an AR(4) model must be selected for the state process. In the last two columns we show the mean and Monte Carlo standard error of the AIC values based on 100 replicates. Based on these columns, an AR(1) model seems to be appropriate for the state process. Table 9 contains the AL and AIS estimates for the case when the state process follows AR(1) model. The Monte Carlo standard error MCSE is based on 1,000 replicates of AIS estimates, using for each replicate the same observations $y_1, \ldots, y_{168}$. BC are bootstrap bias corrections of these estimates based on $B = 1,000$ bootstrap estimates. Notice that only the AL and AIS estimates for β_2 and ϕ differ by more than the expected difference between two AIS estimates (4 times MCSE). In general the AL estimates are very close to the AIS estimates in spite of the fact that the length of the observed time series is not large. We obtained larger MCSE than in Table 7, even when we used the same number of draws ($N = 1,000$) to compute the Monte Carlo standard error in (9). 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.

Table 8. Likelihood and AIC values for various Poisson state space models for the polio data. The simulation is based on 100 replicates of AIC values based on AIS estimates of the likelihood (N=1,000).

	AL		AIS (N=1,000)		simulation	
p	log like	AIC	log like	AIC	AIC	MCSE
0	-252.00	518.00	-252.86	519.76	519.73	0.204
1	-248.14	512.28	-248.29	512.58	512.49	0.210
2	-247.14	512.28	-247.04	512.08	512.23	0.208
3	-246.93	513.86	-246.93	513.86	513.86	0.247
4	-245.15	512.30	-244.75	511.50	512.18	0.272
5	-245.09	514.18	-245.16	514.32	514.10	0.274
13	-243.72	527.43	-243.74	527.47	527.44	0.170

Table 9. AL and AIS estimates for the polio data. BC are bootstrap bias corrected estimates and S.E. are bootstrap estimates of the standard error of AL and AIS, respectively. MCSE is the standard error of 1,000 AIS replicates.

Parameter	AL	S.E.	BC	AIS	MCSE	S.E.	BC
β_1	0.242	0.273	0.260	0.239	0.002	0.285	0.238
β_2	-3.814	2.767	-3.955	-3.746	0.013	2.867	-3.761
β_3	0.162	0.142	0.162	0.161	0.001	0.151	0.161
β_4	-0.482	0.166	-0.480	-0.480	0.001	0.164	-0.467
β_5	0.413	0.128	0.410	0.414	0.001	0.122	0.409
β_6	-0.011	0.129	-0.020	-0.011	0.001	0.127	-0.013
ϕ	0.627	0.229	0.731	0.661	0.006	0.209	0.731
σ^2	0.289	0.122	0.302	0.272	0.008	0.112	0.299

4.6. 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 (6) 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 (20) with the quantiles of a Chi-squared distribution.

For the first method, 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 (4) 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 (9), an estimate of the posterior mean is given by

$$\hat{\boldsymbol{\alpha}} = \frac{1}{N\hat{L}(\boldsymbol{\psi};\mathbf{y})}\sum_{i=1}^{N}\boldsymbol{\alpha}^{(i)}\frac{p(\mathbf{y},\boldsymbol{\alpha}^{(i)}|\boldsymbol{\psi})}{p_a(\boldsymbol{\alpha}^{(i)}|\mathbf{y};\boldsymbol{\psi})} = \frac{1}{N\hat{\mathrm{Er}}_a(\boldsymbol{\psi};\mathbf{y})}\sum_{i=1}^{N}\boldsymbol{\alpha}^{(i)}e^{R(\boldsymbol{\alpha}^{(i)};\boldsymbol{\alpha}^*)}. \quad (19)$$

As an example, for the Pound-Dollar exchange rates and polio data let $\boldsymbol{\psi}$ be the AL estimate from Tables 7 and 9, respectively. Using $N = 1,000$ in (19), $\hat{\boldsymbol{\alpha}}$ was computed. In Figures 4 and 5 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 (14). 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 can assess the compatibility of the samples with a normal population. Such a sample can be obtained as follows: First generate an independent sample $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(N)}$ from the approximate distribution $p_a(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$. For N large, an i.i.d. sample from the discrete distribution with masses

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

is an (approximate) i.i.d. 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 $\tilde{\boldsymbol{\alpha}}^{(1)},\ldots,\tilde{\boldsymbol{\alpha}}^{(M)}$ is an i.i.d. sample from $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$.

If $p_a(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$ in (6) were a good approximation to $p(\boldsymbol{\alpha}|\mathbf{y},\boldsymbol{\psi})$, for $M-n$ large, the squared generalized distances

$$d_j^2 := (\tilde{\boldsymbol{\alpha}}^{(j)} - \boldsymbol{\alpha}^*)^T(\mathbf{K}^* + \mathbf{V})(\tilde{\boldsymbol{\alpha}}^{(j)} - \boldsymbol{\alpha}^*), \quad j = 1, \ldots, M, \tag{20}$$

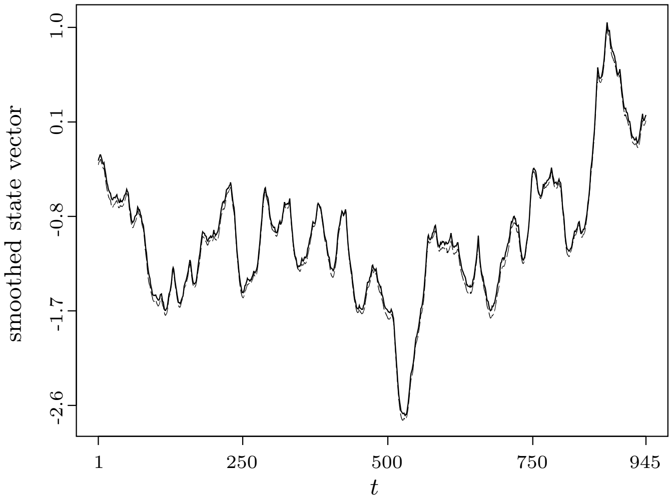

Figure 4. (*smoothed state vector*) For the Pound-Dollar exchange rates data, the solid line shows estimate of the posterior mean of the state vector and the dashed line shows its posterior mode.

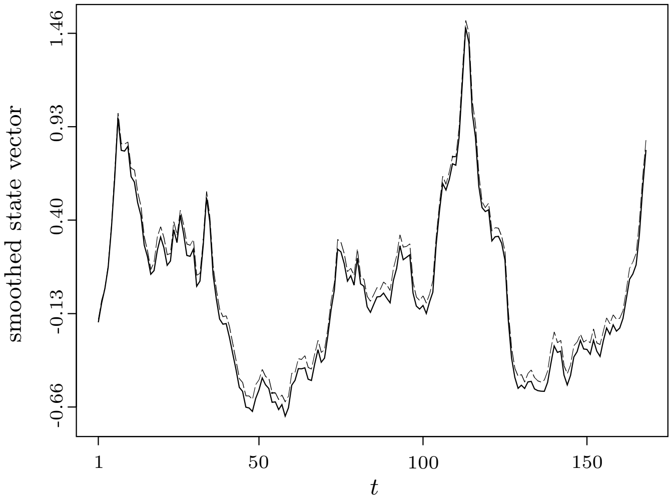

Figure 5. (*smoothed state vector*) For the Polio data, the solid line shows estimate of the posterior mean of the state vector and the dashed line shows its posterior mode.

would resemble an i.i.d. sample from the chi-squared distribution with n degrees of freedom (Johnson and Wichern (2002)). 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 $\lambda_t := e^{\beta+\alpha_t}$, $\alpha_t = \phi\alpha_{t-1}+\eta_t$, $\eta_t \sim$ i.i.d. $N(0, \sigma^2)$, $t = 1, \ldots, n$, and $|\phi| < 1$. The vector of parameters of this process, $\boldsymbol{\psi} = (\beta, \phi, \sigma^2)$, is fixed at $(0.373, 0.9, 0.012)$. Chi-squared QQ-plots of $d_1^2, \ldots, d_M^2$ are shown in Figure 6. With a sample of size N=5,000 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 jth column of this figure corresponds to the parameter value $\boldsymbol{\psi} = \boldsymbol{\psi}_j$, 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.

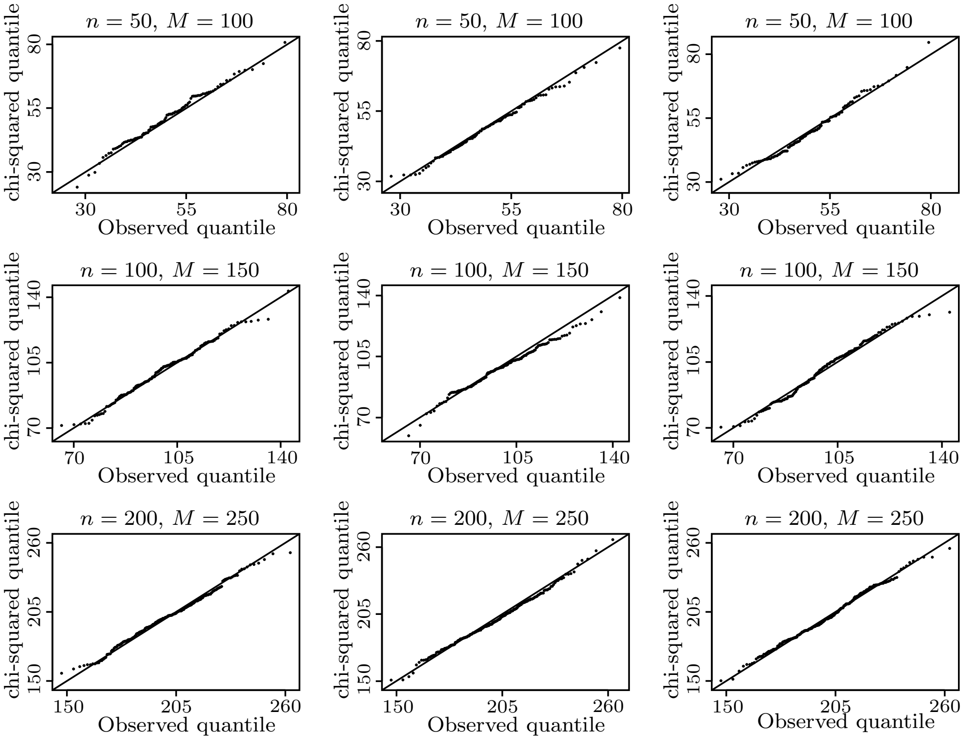

Figure 6. (*Chi-squared QQ-plots*) The QQ-plot from ith row and jth column was obtained using a SIR sample $\tilde{\boldsymbol{\alpha}}^{(1)}, \ldots, \tilde{\boldsymbol{\alpha}}^{(M)}$ from $p(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi}_j)$ by resampling a sample of size 5,000 from the approximation $p_a(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi}_j)$.

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 (2002)). The nine correlations r_Q for the data used to create Figure 6 are shown in columns 3 through 5 of Table 10. 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% significance, are given in the last column of this table. In all cases, normality is not rejected. This provides some evidence that the distribution in (6) is a reasonable approximation for the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\psi})$.

Table 10. Correlation coefficients of the points in the QQ-plots from figure 6.

			r_Q		
N	M	$\boldsymbol{\psi}_1$	$\boldsymbol{\psi}_2$	$\boldsymbol{\psi}_3$	$r_{0.05}$
50	100	0.9952	0.9978	0.9925	0.9873
100	150	0.9957	0.9952	0.9926	0.9913
200	250	0.9974	0.9974	0.9973	0.9920

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 (14) and the determinant needed in (8). 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 $\mathbf{\Gamma} = \{\gamma_{i,j}\}_{i,j=1}^n$. 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 (2002, pp.71-72))

$$\mathbf{X} = \mathbf{C}(\mathbf{X} - \hat{\mathbf{X}}), \tag{21}$$

where

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

The entries θ_{ij} of this matrix can be found recursively as in Proposition 5.2.2. from Brockwell and Davis (1991). Equating the covariance matrices of $\mathbf{X}$ and $\mathbf{C}(\mathbf{X} - \hat{\mathbf{X}})$, it follows that

$$\mathbf{\Gamma} = \mathbf{C}\mathbf{D}\mathbf{C}^T, \tag{22}$$

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 equality comes from the fact that the components of $\mathbf{X}-\hat{\mathbf{X}}$ are uncorrelated.

For example, consider the SSM for which the observations $y_1, \ldots, y_n$ are realizations from a Poisson distribution with rates $\lambda_t = e^{\beta+\alpha_t}$, and the state process follows the AR(p) model in (1). For identifiability of the parameters, we set $\gamma = 0$. Notice that the distribution of the observations has the format of the exponential family in (2) where $b(\alpha_t) = e^{\alpha_t+\beta}$.

To implement the innovations algorithm, the matrix $\mathbf{V} = \{v_{ij}\}_{n\times n}$ is needed. To start, set $\mathbf{V}_0^{-1} = \mathrm{Cov}\,(\boldsymbol{\alpha}_0)$, where $\boldsymbol{\alpha}_0 := (\alpha_1, \alpha_2, \ldots, \alpha_p)$. The p components $\gamma(0), \gamma(1), \ldots, \gamma(p-1)$ of this matrix can be obtained by solving the $p+1$ linear equations (Brockwell and Davis (2002, p.90))

$$\gamma(k) - \phi_1\gamma(k-1) - \ldots - \phi_p\gamma(k-p) = \sigma^2 \mathrm{I}_{\{0\}}(k), \quad k = 0, 1, \ldots, p,$$

where $\mathrm{I}_{\{0\}}(k)$ is the indicator function. Now, set $\mathbf{Z} = (\boldsymbol{\alpha}_0, \eta_{p+1}, \eta_{p+2}, \ldots, \eta_n)$, where η_t is the tth error term of the AR(p) process in (1). Thus,

$$\mathrm{Cov}\,(\mathbf{Z}) = \begin{pmatrix} \mathbf{V}_0^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_{n-p} \end{pmatrix}.$$

Since $\eta_t = -\phi_p\alpha_{t-p} - \ldots - \phi_1\alpha_{t-1} + \alpha_t$, then $\mathbf{Z} = \mathbf{A}\boldsymbol{\alpha}$, where the square matrix $\mathbf{A}$ of dimension $n \times n$ is given by

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

$$\mathbf{A}_{21} = \begin{pmatrix} -\phi_p & -\phi_{p-1} & \ldots & -\phi_1 \\ 0 & -\phi_p & \ldots & -\phi_2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -\phi_p \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}, \qquad \mathbf{A}_{22} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -\phi_1 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\phi_{p-1} & -\phi_{p-2} & \ldots & 0 \\ -\phi_p & -\phi_{p-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

Hence, $\mathrm{Cov}\,(\mathbf{Z}) = \mathbf{A}\,\mathrm{Cov}\,(\boldsymbol{\alpha})\mathbf{A}^T = \mathbf{A}\mathbf{V}^{-1}\mathbf{A}^T$. From the two expressions for $\mathrm{Cov}\,(\mathbf{Z})$, it follows that

$$\mathbf{V} = \mathbf{A}^T \begin{pmatrix} \mathbf{V}_0 & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sigma^2}\mathbf{I} \end{pmatrix} \mathbf{A}.$$

The determinant of the matrix $\mathbf{V}$, needed in (8), can be now computed. Since $|\mathbf{A}| = 1$, it follows that $|\mathbf{V}| = (1/(\sigma^2)^{n-p})|\mathbf{V}_0|$. The determinant of $\mathbf{V}_0$ is computed numerically. For $p = 1$, $\mathbf{V}_0 = (1-\phi_1^2)/\sigma^2$. Then $|\mathbf{V}| = (1-\phi_1^2)/(\sigma^2)^n$.

Now, substituting the partitioned version of $\mathbf{A}$ in $\mathbf{V}$, we obtain

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_0 + \frac{1}{\sigma^2}\mathbf{A}_{21}^T\mathbf{A}_{21} & \frac{1}{\sigma^2}\mathbf{A}_{21}^T\mathbf{A}_{22} \\ \frac{1}{\sigma^2}\mathbf{A}_{22}^T\mathbf{A}_{21} & \frac{1}{\sigma^2}\mathbf{A}_{22}^T\mathbf{A}_{22} \end{pmatrix}.$$

It can be shown that $v_{ij} = 0$ if $|j-i| > p$. If $n > 2p$, there is no need to store v_{ij}, $i, j = p+1, \ldots, n-p-1$. For large n, this is a substantial saving in memory. Now, let $\boldsymbol{\alpha}^j$ be the current iterate to the value of $\boldsymbol{\alpha}^*$. From (12),

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

Since no intercept is included in the AR(p) process, $\boldsymbol{\mu} = \mathbf{0}$. Thus, $\tilde{\mathbf{y}}^j$ defined in (13) 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, the entry θ_{ik} of matrix $\mathbf{C}$ is zero if $k > p$. Hence, $\mathbf{C}$ can be stored in a $n \times p$ matrix. For large n, this is again a substantial memory saving. Start with $v_0 = \gamma_{1,1}$ and $\hat{X}_1 = 0$. From Proposition 5.2.2 of Brockwell and Davis (1991), for $i = 1, \ldots, n-1$,

$$\begin{aligned} \theta_{i,i-k} &= v_k^{-1}[\gamma_{i+1,k+1} - \sum_{m=\max\{0,i-p,k-p\}}^{k-1} \theta_{k,k-m}\theta_{i,i-m}v_m], \\ &\qquad k = \max\{0, i-p\}, \ldots, i-1, \\ v_i &= \gamma_{i+1,i+1} - \sum_{m=\max\{0,i-p\}}^{i-1} \theta_{i,i-m}^2 v_m, \\ \hat{X}_{i+1} &= \sum_{m=1}^{\min\{p,i\}} \theta_{im}(X_{i+1-m} - \hat{X}_{i+1-m}). \end{aligned}$$

Now, using (21) and (22), it follows that $\boldsymbol{\Gamma}^{-1}\mathbf{X} = \mathbf{C}^{-T}\mathbf{e}$, where the entries e_j of the vector $\mathbf{e}$ are the "normalized" residuals $(X_j - \hat{X}_j)/\nu_{j-1}$. Therefore, the iteration in (14) becomes

$$\boldsymbol{\alpha}^{j+1} = (\mathbf{K}^j + \mathbf{V})^{-1}\tilde{\mathbf{y}}^j = \boldsymbol{\Gamma}^{-1}\mathbf{X} = \mathbf{C}^{-T}\mathbf{e}. \tag{23}$$

Since $\mathbf{C}$ is a triangular matrix, there is no need to invert $\mathbf{C}$ to compute $\boldsymbol{\alpha}^{j+1}$. To see this, notice that $\mathbf{e} = \mathbf{C}^T\boldsymbol{\alpha}^{j+1}$. Equating the nth components of $\mathbf{e}$ and

$\mathbf{C}^T\boldsymbol{\alpha}^{j+1}$ we obtain $\alpha_n^{j+1} = e_n$. And equating their $(n-i)$th entries the "reversed" recursion

$$\alpha_{n-i}^{j+1} = e_{n-i} - \sum_{k=1}^{\min\{p,j\}} \theta_{n+k-i-1,k}\alpha_{n+k-i}^{j+1}, \quad i = 1, 2, \ldots, n-1,$$

is obtained. The iteration in (23) tends to converge quite rapidly -only a few steps are required. To compute the determinant of the matrix $\mathbf{K}^* + \mathbf{V}$ needed in (8), set $\boldsymbol{\Gamma} := \mathbf{K}^* + \mathbf{V}$, where $\mathbf{K}^* = e^{\beta}\text{diag}\{e^{\boldsymbol{\alpha}^*}\}$ -see (5), 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 (23). Because the determinant of the matrix $\mathbf{C}$ is 1, taking determinants on both sides of (22), we obtain

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

where ν_j, $j = 0, \ldots, n-1$, comes from the last iteration in (23).

The innovations algorithm allows us to sample from $p_a(\boldsymbol{\alpha}; \boldsymbol{\alpha}^*, (\mathbf{K}^* + \mathbf{V})^{-1})$, useful for implementing the importance sampling procedure. If $\mathbf{u}$ is a draw from $N(\mathbf{0}, \mathbf{I}_n)$, then $\boldsymbol{\alpha} := \boldsymbol{\alpha}^* + \mathbf{C}^{-T}\mathbf{D}^{-1/2}\mathbf{u}$ is a draw from $p_a(\boldsymbol{\alpha}; \boldsymbol{\alpha}^*, (\mathbf{K}^* + \mathbf{V})^{-1})$. To show this, notice that $\text{E}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^*$ and $\text{Cov}\,\{\boldsymbol{\alpha}\} = \mathbf{C}^{-T}\mathbf{D}^{-1/2}\mathbf{D}^{-1/2}\mathbf{C}^{-1} = (\mathbf{C}\mathbf{D}\mathbf{C}^T)^{-1} = (\mathbf{K}^* + \mathbf{V})^{-1}$. To compute $\mathbf{C}^{-T}\mathbf{D}^{-1/2}\mathbf{u}$, a "reversed" recursion can be used.

Acknowledgements

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 to thank W.T.M Dunsmuir and the referees for their comments and helpful suggestions.

References

Bernardo, J. M. and Smith, A. F. M (1994). *Bayesian Theory.* Wiley, New York.

Breidt, F. J. and Carriquiry, A. L. (1996). Improved quasi-maximum likelihood estimation for stochastic volatility models. In *Modeling and Prediction: Honouring Seymour Geisser* (Edited by A. Zellner and J. S. Lee). Springer, New York.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* 2nd Edition. Springer-Verlag, New York.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting.* 2nd edition. Springer-Verlag, New York.

Campbell, M. J. (1994). Time series regression for counts: an investigation into the relationship between sudden infant death syndrome and environmental temperature. *J. R. Statist. Soc. Ser. A* **157**, 191-208.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *J. Amer. Statist. Assoc.* **90**, 242-252.

Davis, R. A., Dunsmuir, W. T. M. and Wang, Y. (1998). Modelling time series of count data. In *Asymptotics, Nonparametrics and Time Series* (Edited by S. Ghosh), 63-112. Marcel Dekker, New York.

Durbin, J. and Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84**, 669-684.

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods.* Oxford, New York.

Efron, B. and Tibshirani R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

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.

Geyer, C. J. (1996). Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice* (Edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 89-114. Chapman and Hall, London.

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

Harvey, A. C. and Fernandes, C. (1989). Time series models for count or qualitative observations. *J. Amer. Statist. Assoc.* **7**, 407-417.

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

Harvey, A. C. and Shepard, N. (1993). Estimation and testing of stochastic variance models. Unpublished manuscript, The London School of Economics.

Harvey, A. C., Ruiz, E. and Shepard, N. (1994). Multivariate stochastic variance models. *Rev. Econom. Stud.* **61**, 247-264.

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.

Johnson, R. A. and Wichern, W. (2002). *Applied Multivariate Statistical Analysis.* 5th edition. Prentice Hall, New Jersey.

Kuk, A. Y. (1999). The use of approximating models in Monte Carlo maximum likelihood estimation. *Statist. Probab. Lett.* **45**, 325-333.

Kuk, A. Y. and Cheng, Y. W. (1997). The Monte Carlo Newton-Raphson algorithm. *J. Statist. Comput. Simulation* **59**, 233-250.

Pitt, M. K and Shepard N. (1999). Filtering via simulation: auxiliary particle filters. *J. Amer. Statist. Assoc.* **94**, 590-599.

Sandmann, G. and Koopman, S. J. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *J. Econometrics* **87**, 271-301.

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.

Zeger, S. L. (1988). A regresion model for time series of counts. *Biometrika* **75**, 621-629.

Department of Statistics, Colorado State University, 102A Statistics Building, Fort Collins, Colorado, CO 80523, U.S.A.

E-mail: rdavis@stat.colostate.edu

Department of Statistics, Colorado State University, 102A Statistics Building, Fort Collins, Colorado, CO 80523, U.S.A.

E-mail: rodrigue@stat.colostate.edu

(Received October 2003; accepted September 2004)