View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DSpace at VU

TI 2005-117

Tinbergen Institute Discussion Paper

On Importance Sampling for State Space Models

Borus Jungbacker

*Siem Jan Koopman**

Vrije Universiteit Amsterdam.

** Tinbergen Institute.*

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Please send questions and/or remarks of non-scientific nature to driessen@tinbergen.nl.
Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

On Importance Sampling for State Space Models

Borus Jungbacker and Siem Jan Koopman
Department of Econometrics, Vrije Universiteit Amsterdam,
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Abstract

We consider likelihood inference and state estimation by means of importance sampling for state space models with a nonlinear non-Gaussian observation $y \sim p(y|\alpha)$ and a linear Gaussian state $\alpha \sim p(\alpha)$. The importance density is chosen to be the Laplace approximation of the smoothing density $p(\alpha|y)$. We show that computationally efficient state space methods can be used to perform all necessary computations in all situations. It requires new derivations of the Kalman filter and smoother and the simulation smoother which do not rely on a linear Gaussian observation equation. Furthermore, results are presented that lead to a more effective implementation of importance sampling for state space models. An illustration is given for the stochastic volatility model with leverage.

Some keywords: Kalman filter; Likelihood function; Monte Carlo integration; Newton-Raphson; Posterior mode estimation; Simulation smoothing; Stochastic volatility model.

1 Introduction

In this paper we develop new results for the application of importance sampling methods to a class of nonlinear and non-Gaussian state space models. The general ideas of importance sampling are well established in statistics and econometrics, see Kloek and Van Dijk (1978), Ripley (1987) and Geweke (1989). The importance sampling techniques for state space models have been explored by Danielsson and Richard (1993), Shephard and Pitt (1997), Durbin and Koopman (1997) and So (2003). A textbook treatment is given in part II of Durbin and Koopman (2001).

Denote the vector of observables by y and the latent state vector by α. In this paper we consider the state space model

$$y \sim p(y|\alpha), \qquad \alpha \sim p_G(\alpha), \tag{1}$$

where density $p(y|\alpha)$ is typically non-Gaussian and density $p_G(\alpha)$ is Gaussian. The latent states in α are modelled by a linear Markovian process with Gaussian innovations. The relationship between α and y in $p(y|\alpha)$ can be nonlinear. Different methods and approaches are considered for time series analysis based on model (1). The standard numerical integration method is considered by Kitagawa (1987) but is not feasible when the dimensions of y and α are high. Therefore, alternative approaches are developed based on simulation. Markov chain Monte Carlo (MCMC) methods for state space models are considered by Carlin, Polson, and Stoffer (1992), Fruhwirth-Schnatter (1994), Carter and Kohn (1994), Shephard (1994), Shephard and Pitt (1997), Berzuini, Best, Gilks, and Larizza (1997) and Gamerman (1998). The particle filtering method of Gordon, Salmond, and Smith (1993) and the modification of Pitt and Shephard (1999), which are reviewed in Doucet, deFreitas, and Gordon (2000), can also be used for the analysis of (1) although these methods are not designed for likelihood-based inference.

Importance sampling methods are in particular employed for the evaluation of the integral

$$\bar{x} = \int x(\alpha) p(\alpha|y)\, \mathrm{d}\alpha, \tag{2}$$

where $x(\alpha)$ can be any function of α and $p(\alpha|y) = p(y)^{-1} p(y|\alpha) p_G(\alpha)$ with $p(y|\alpha)$ and $p_G(\alpha)$ as in (1). It is assumed that an analytical expression for (2) is usually not available while direct numerical integration of (2) is not feasible when dimensions of α and/or y are high. The basic idea of importance sampling is to evaluate (2) by generating M samples from an importance density $f(\alpha; y)$ and by computing

$$\widehat{\bar{x}} = \left\{ \sum_{i=1}^{M} \frac{p(\alpha^i, y)}{f(\alpha^i; y)} \right\}^{-1} \sum_{i=1}^{M} x(\alpha^i) \frac{p(\alpha^i, y)}{f(\alpha^i; y)}, \qquad \alpha^i \sim f(\alpha; y), \tag{3}$$

where the importance function $f(\alpha; y)$ depends on y. The three main issues for this paper are the construction of importance function $f(\alpha; y)$, the construction of a simulation device for sampling from $f(\alpha; y)$ and the evaluation of $f(\alpha; y)$ for any α.

Several choices for the importance density $f(\alpha; y)$ have been proposed in the literature, see Danielsson and Richard (1993), Shephard and Pitt (1997) and Durbin and Koopman (1997). Here we focus on an importance function $f(\alpha; y)$ that is based on a linear Gaussian state space model with the same mode for α and the same curvature at the mode of smoothing density $p(\alpha|y)$. The resulting model represents the Laplace approximation of the smoothing density. When this approximating linear Gaussian model exists, standard Kalman filter and smoother methods can be used to carry out the computations for the constructions of the importance function and its simulation device. For example, de Jong and Shephard (1995) and Durbin and Koopman (2002) have developed simulation smoothing algorithms that are associated with the Kalman filter. When the approximating linear Gaussian model does not exist, the model-based Kalman filter methods can not be used as a result. In this paper, however, we show that an importance function $f(\alpha; y)$ based on the mode can still be constructed using the computationally efficient Kalman filter and smoother recursions. Furthermore, we show that it is still possible to simulate from $f(\alpha; y)$ in a computationally efficient way without the consideration of a linear Gaussian state space model. The relevant proofs and derivations are presented. These results justify the use of state space methods for the construction of $f(\alpha; y)$ and for the sampling from $f(\alpha; y)$ in all circumstances as long as the model for α has a Markovian structure. The range of nonlinear non-Gaussian time series models that can be analysed by importance sampling has increased considerably as a result.

As a consequence of these developments, this paper presents various new results. In particular, we present (i) a new derivation of the Kalman filter based on the LU decomposition of a symmetric matrix, (ii) a new formulation and derivation of the de Jong and Shephard (1995) simulation smoother, (iii) a general method of computing importance weights and (iv) a device of simulating the state vector α based on a simulated signal vector, both conditional on y. The methods are implemented for a stochastic volatility model with leverage. The new modifications are necessary for the maximum likelihood estimation of parameters and the signal extraction of functions of the state vector α.

The paper is organised as follows. The model with a linear Gaussian state equation and a nonlinear non-Gaussian observation equation is presented in §2. The construction of the importance function for the different state space models and for different circumstances is derived in §3. A method for simulating from the importance function and devices related to importance sampling are developed in §4. The methods are illustrated using a stochastic volatility model with leverage in §5. The proofs and derivations of the results are given in the Appendix.

2 State space models

2.1 The linear Gaussian state and signal vectors

The linear Gaussian state vector α_t is modelled by the Markov process

$$\alpha_{t+1} = d_t + T_t\alpha_t + \eta_t, \qquad \eta_t \sim NID(0, Q_t), \qquad t = 1, \ldots, n, \tag{4}$$

where system vector d_t and system matrices T_t and Q_t are fixed and known for $t = 1, \ldots, n$. The state vector α_t and the disturbance vector η_t have dimension $q \times 1$. The system matrices have appropriate dimensions while the variance matrix Q_t is positive semi-definite. The initial state vector is normally distributed with mean a and variance matrix P, that is $\alpha_1 \sim N(a, P)$. The disturbances η_t $(t = 1, \ldots, n)$ are serially independent and are independent of the initial state vector. The joint property of a sequence of n state vectors can be expressed by the multivariate normal density

$$\alpha \sim N(d, \Omega), \tag{5}$$

where

$$\alpha = (\alpha_1', \ldots, \alpha_n')', \qquad d = T\left(a', d_1', \ldots, d_{n-1}'\right)', \qquad \Omega = T\mathrm{diag}(P_1, Q_1, \ldots, Q_{n-1})T',$$

with

$$T = \begin{bmatrix} I & 0 & 0 & \cdots & 0 & 0 \\ T_1 & I & 0 & \ldots & 0 & 0 \\ T_2T_1 & T_2 & I & & 0 & 0 \\ & & & \ddots & & \vdots \\ T_{n-2}\ldots T_1 & T_{n-2}\ldots T_2 & T_{n-2}\ldots T_3 & & I & 0 \\ T_{n-1}\ldots T_1 & T_{n-1}\ldots T_2 & T_{n-1}\ldots T_3 & \cdots & T_{n-1} & I \end{bmatrix}, \tag{6}$$

for $t = 1, \ldots, n$. It follows that the log-density of the state vector α is given by

$$\log p_G(\alpha) = -\frac{qn}{2}\log 2\pi - \frac{1}{2}|\Omega| - \frac{1}{2}(\alpha - d)'\Omega^{-1}(\alpha - d). \tag{7}$$

The $m \times 1$ signal vector θ_t connects the state vector α_t with the observations via

$$\theta_t = c_t + Z_t\alpha_t, \qquad t = 1, \ldots, n, \tag{8}$$

where vector c_t and matrix Z_t are fixed and known. Clearly,

$$\theta \sim N(\mu, \Psi), \qquad \mu = c + Zd, \qquad \Psi = Z\Omega Z', \tag{9}$$

with $\theta = c + Z\alpha$ and where

$$\theta = (\theta_1', \ldots, \theta_n')', \qquad c = (c_1', \ldots, c_n')', \qquad Z = \mathrm{diag}(Z_1, \ldots, Z_n).$$

Finally, the log-density of the signal is given by

$$\log p_G(\theta) = -\frac{mn}{2}\log 2\pi - \frac{1}{2}|\Psi| - \frac{1}{2}(\theta-\mu)'\Psi^{-1}(\theta-\mu). \tag{10}$$

The prediction error decomposition can be applied to (10). Since θ is a linear function of α, the Kalman filter evaluates $\log p_G(\theta)$. Almost all linear Gaussian time series processes can be represented in state space form with a nonsingular variance matrix Ψ.

2.2 The linear Gaussian observation vector

The linear Gaussian observation model for the $m \times 1$ observation vectors y_t is given by

$$y_t = c_t + Z_t\alpha_t + \varepsilon_t = \theta_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, H_t), \qquad t = 1, \ldots, n, \tag{11}$$

where variance matrix H_t is fixed and known. Stacking $y = (y_1', \ldots, y_n')'$, the model becomes

$$y = c + Z\alpha + \varepsilon = \theta + \varepsilon, \qquad \varepsilon \sim N(0, H), \tag{12}$$

with $H = \mathrm{diag}(H_1, \ldots, H_n)$. The linear Gaussian observation density is given by

$$p_G(y|\alpha) = p_G(y|\theta) = N(\theta, H) = \prod_{t=1}^{n} p_G(y_t|\theta_t). \tag{13}$$

The likelihood function $p_G(y)$ can be evaluated by the Kalman filter since y is a linear function of α.

2.3 The nonlinear non-Gaussian observation vector

In this paper we focus primarily on the nonlinear non-Gaussian observation model with density $p(y|\theta)$ for which the conditional independence assumption applies, that is

$$p(y|\theta) = \prod_{t=1}^{n} p(y_t|\theta_t), \qquad t = 1, \ldots, n. \tag{14}$$

Examples of densities $p(y_t|\theta_t)$ are the classes of the exponential family densities and the stochastic volatility models, see Durbin and Koopman (2001, Chapter 10).

3 Computing the posterior mode

In this section we construct the importance function $f(\theta; y)$ that can generate draws of $\theta = c + Z\alpha$ around the location of the mode of smoothing density $p(\theta|y)$ and scaled by the curvature of $p(\theta|y)$ at the mode. It will be argued in §4 that a draw of $\alpha^i \sim f(\alpha; y)$ can be obtained from a draw of $\theta^i \sim f(\theta; y)$ and that the importance function $f(\theta; y)$ is sufficient for the evaluation of (3).

First, the linear Gaussian model $p_G(y|\theta)$ is considered for which standard results are used to obtain the importance function $f(\theta; y) = p_G(\theta|y)$. Second, the general nonlinear non-Gaussian case of $p(y|\theta)$ is treated where the importance function is based on the mode estimate $\widehat{\theta}$, obtained by maximising the smoothing density $p(\theta|y)$ with respect to θ, and its Hessian matrix G, that is $f(\theta; y) = N(\widehat{\theta}, -G^{-1})$. The main results are presented in this section while the derivations are given in the Appendix. Finally, some numerical issues are discussed for a successful implementation of the methods.

3.1 The mode of the linear Gaussian model

Consider the linear Gaussian signal and observation vectors of §§2.1 and 2.2. The unconditional mean, variance and covariance of observation y and signal θ are given by

$$E(y) = \mu, \qquad Var(y) = \Sigma = \Psi + H, \qquad Cov(\theta, y) = \Psi. \tag{15}$$

It follows from the standard lemma of the multivariate normal density that the conditional means and variances are given by

$$E(\theta|y) = \mu + \Psi\Sigma^{-1}(y - \mu), \qquad Var(\theta|y) = \Psi - \Psi\Sigma^{-1}\Psi. \tag{16}$$

The Kalman filter and smoother evaluate the conditional mean $E(\theta_t|y)$ and variance $Var(\theta_t|y)$ in a recursive and computationally efficient way for a linear Gaussian state space model, see Durbin and Koopman (2001, Chapter 4). Since all densities are Gaussian, the conditional or posterior mode of $p_G(\theta|y)$, denoted by $\widehat{\theta}$, is equivalent to the conditional mean of $p_G(\theta|y)$, that is $\widehat{\theta} = E(\theta|y)$. After some minor manipulations, it follows from the first equation in (16) that

$$\widehat{\theta} = \left(\Psi^{-1} + H^{-1}\right)^{-1}\left(H^{-1}y + \Psi^{-1}\mu\right). \tag{17}$$

It should be emphasized that the Kalman filter and smoother effectively computes $\widehat{\theta}$ for the linear Gaussian state space model.

3.2 The mode of the nonlinear non-Gaussian model

Consider the nonlinear non-Gaussian observation model of §2.3 and the linear Gaussian state vector of §2.1. For this class of models, an analytical expression for the posterior mode $\widehat{\theta}$ of

$p(\theta|y)$ is usually not available. We therefore obtain the mode by maximising $p(\theta|y)$ with respect to θ using the Newton-Raphson method of optimisation, see Nocedal and Wright (1999) for a treatment of numerical optimisation methods. The dimension of θ is $mn \times 1$ so that matrix dimensions can be high and straightforward matrix manipulations become infeasable. Therefore efficient algorithms need to be considered.

For a given guess g of the solution for $\widehat{\theta}$, the Newton-Raphson method proposes the new guess as

$$g^+ = g - \left[\ddot{p}(\theta|y)|_{\theta=g}\right]^{-1} \dot{p}(\theta|y)|_{\theta=g}\,, \tag{18}$$

where we define

$$\dot{p}(\cdot|\cdot) = \frac{\partial \log p(\cdot|\cdot)}{\partial\theta}, \qquad \ddot{p}(\cdot|\cdot) = \frac{\partial^2 \log p(\cdot|\cdot)}{\partial\theta\partial\theta'}. \tag{19}$$

Since $\log p(\theta|y) = \log p(y|\theta) + \log p_G(\theta) - \log p(y)$, we have

$$\dot{p}(\theta|y) = \dot{p}(y|\theta) - \Psi^{-1}(\theta - \mu), \qquad \ddot{p}(\theta|y) = \ddot{p}(y|\theta) - \Psi^{-1}. \tag{20}$$

The conditional independence assumption of observation model (14) implies that $\ddot{p}(y|\theta)$ is a block diagonal matrix.

The Newton-Raphson updating step reduces to

$$\begin{aligned} g^+ &= g - \left[\ddot{p}(y|\theta)|_{\theta=g} - \Psi^{-1}\right]^{-1} \left(\dot{p}(y|\theta)|_{\theta=g} - \Psi^{-1}\{g-\mu\}\right) \\ &= \left[\Psi^{-1} - \ddot{p}(y|\theta)|_{\theta=g}\right]^{-1} \left(\dot{p}(y|\theta)|_{\theta=g} - \ddot{p}(y|\theta)|_{\theta=g}\, g + \Psi^{-1}\mu\right) \\ &= \left(\Psi^{-1} + A^{-1}\right)^{-1} \left(A^{-1}x + \Psi^{-1}\mu\right), \end{aligned} \tag{21}$$

where

$$A = -\left[\ddot{p}(y|\theta)|_{\theta=g}\right]^{-1}, \qquad x = g + A\,\dot{p}(y|\theta)|_{\theta=g}\,. \tag{22}$$

We note the similarity of (21) compared to (17). In case $\ddot{p}(y|\theta)$ is negative semi-definite for all θ, it follows that the Kalman filter and smoother can be used to compute g^+ by applying it to the Gaussian state space model (12) with

$$y = x, \qquad H = A,$$

The computation of $E(\theta|y)$ returns g^+. This approach is taken by Shephard and Pitt (1997), Durbin and Koopman (1997) and So (2003, §2). The mode $\widehat{\theta}$ for a non-Gaussian nonlinear observation model is obtained by the Newton-Raphson method where for each step the Kalman filter and smoother computes the new guess g^+. The Hessian matrix of the mode estimator $\widehat{\theta}$ is given by $G = \ddot{p}(\theta|y) = -\Psi^{-1} - A^{-1}$.

This approach of finding the mode $\widehat{\theta}$ is clearly not valid when $\ddot{p}(y|\theta)$ is not negative definite since this will imply that the variance matrix H of the linear Gaussian model (12) is not positive definite. In cases where $\ddot{p}(y|\theta)$ is not negative definite, the following Theorem can be adopted for the computation of (21).

Theorem 1.

Consider the linear Gaussian signal vector of §2.1 with $\theta \sim N(\mu, \Psi)$ *and define matrix* $A = diag(A_1, \ldots, A_n)$ *where* A_i *can be any* $m \times m$ *nonsingular symmetric matrix for* $i = 1, \ldots, n$. *The matrix expression*

$$\mu + \Psi (\Psi + A)^{-1} (x - \mu) \equiv \left(\Psi^{-1} + A^{-1}\right)^{-1} \left(A^{-1}x + \Psi^{-1}\mu\right), \tag{23}$$

can be computed for any $mn \times 1$ *vector* x *by the standard Kalman filter and smoother without the consideration of the linear Gaussian observation equation* (12).

The proof of Theorem 1 is presented in the Appendix. The Kalman filter and smoother are given by the equations (34), (43) and (44) in the Appendix. Theorem 1 does not refer to the existence of an observation model for x. Therefore, we can not use the standard derivation of the Kalman filter and smoother that relies on statistical concepts such as conditional means and variances, see Anderson and Moore (1979) and Durbin and Koopman (2001, Chapter 3). However, the special structure of variance matrix $\Psi = Z\Omega Z'$ of the signal θ does allow us to use decompositions based on triangular matrices such as T in §2.1. The derivation of Theorem 1 in the Appendix is based on the LU decomposition of a possibly indefinite but symmetric matrix $\Psi + A$. Theorem 1 does not take $A + \Psi$ as a variance matrix. It shows that the LU decomposition leads to the Kalman filter and smoother equations for solving x in $\Psi (\Psi + A)^{-1} x = y$ with y given. We note that the reverse argument for the linear Gaussian state space model is a well-known result, that is, the Kalman filter effectively diagonalises the variance matrix of the observation vector y in §2.2. Theorem 1 implies that this result applies to any positive or negative definite matrix $\Psi + A$.

In the context of finding the mode of $p(\theta|y)$ by the Newton-Raphson method, it is concluded that when $\ddot{p}(y|\theta)$ is not negative definite, the Kalman filter and smoother can still be used for the computation of g^+ in (21). The justification is provided by Theorem 1 for which x and A are given by (22).

3.3 Global and local convergence

Although matrix $\Psi + A$ can be indefinite, the Hessian matrix $-\Psi^{-1} - A^{-1}$ should always be semi-negative definite for θ at or in the neighbourhood of $\widehat{\theta}$ by construction. In cases the Hessian matrix is not negative definite, the Newton-Raphson step does not progresses to the maximum of $p(\theta|y)$ with respect to θ. To enforce global convergence, the algorithm can be modified by line-search and other numerical methods, see Nocedal and Wright (1999). In general, line-search strategies often speed up the maximisation and stabilise the algorithm. A line-search

can be implemented by introducing the scalar $0 < \lambda \leq 1$ in (18) and defining

$$g_\lambda^+ = g - \lambda \left[\ddot{p}(\theta|y)|_{\theta=g} \right]^{-1} \dot{p}(\theta|y)|_{\theta=g} \,. \tag{24}$$

The line-search consists of finding a value for λ so that $p(\theta|y)|_{\theta=g_\lambda^+} > p(\theta|y)|_{\theta=g}$. By combining (24) and (18), the line search computations are straightforward and given by

$$g_\lambda^+ = g + \lambda(g^+ - g),$$

where $g^+ = g_\lambda^+\big|_{\lambda=1}$ is computed by (21) only once for different values of $0 < \lambda \leq 1$. When an appropriate value for λ is found, a new guess of the mode $\widehat{\theta}$ can be computed at the location g_λ^+. Global convergence is ensured when an appropriate set of regularity conditions for the line-search is fulfilled, see Nocedal and Wright (1999) for a detailed discussion. To check these conditions, it is usually necessary to evaluate the score function.

The score vector of $p(\theta|y)$ is defined in (19) and is given by (20), that is

$$\frac{\partial \log p(\theta|y)}{\partial \theta} = \dot{p}(\theta|y) = \dot{p}(y|\theta) - \Psi^{-1}(\theta - \mu).$$

Propositions 1 and 2 in the Appendix imply that the term $\Psi^{-1}(\theta - \mu)$ can be evaluated by the Kalman filter and smoother algorithms of Theorem 1 with $x = \theta$ and $A = 0$. This result follows immediately since $\Sigma = \Psi$ when $A = 0$. An analytical expression for the term $\dot{p}(y|\theta)$ is usually straightforward to derive.

Given this computational device for evaluating the score, other maximisation methods may be considered to obtain the mode of $p(\theta|y)$. It is noted that different numerical problems can occur during the maximisation of $p(\theta|y)$ with respect to the high dimensional vector θ. Although line-search methods can stabilise the Newton-Raphson method, it may be necessary to switch to other score-based or quasi-Newton optimisation methods. Therefore this computationally efficient method of computing the score is important in practical work.

4 Importance sampling

When the importance function can be represented as a linear Gaussian state space model, the simulation smoothers of de Jong and Shephard (1995) and Durbin and Koopman (2002) can be used to simulate from $f(\theta; y)$ in a computationally efficient way. In other cases, simulations need to be carried out directly from the importance function $f(\theta; y)$. In this section we show that such simulations can also be carried out by the computationally efficient simulation smoother. A new formulation of the simulation smoother is given and its consistency with the de Jong and Shephard (1995) recursions is shown. Further, it is argued that the computation of $f(\theta; y)$ for a realisation of θ can be carried out within the simulation smoother. Finally, it is shown that a draw from $f(\theta; y)$ can be used to obtain a draw from $f(\alpha; y)$ using the Kalman filter and smoother algorithm. A special simulation smoother is not required for the state vector.

4.1 Simulation smoothing

The importance function under consideration is the multivariate Gaussian density with its mean equal to the mode of $p(\theta|y)$ and with its variance matrix V equal to minus the inverse of the Hessian matrix G, that is

$$f(\theta; y) = N(\widehat{\theta}, V), \qquad V = -G^{-1}, \qquad G = \ddot{p}(\theta|y) = -A^{-1} - \Psi^{-1},$$

where $\widehat{\theta}$ is estimated as described in §3 and A is given by (22) evaluated at $\theta = \widehat{\theta}$. When the maximum of $p(\theta|y)$ is obtained at $\theta = \widehat{\theta}$, matrix G is guaranteed to be negative definite and V is positive definite as a result. By some minor matrix manipulations, it can be shown that

$$V \;=\; (A^{-1} + \Psi^{-1})^{-1} \;=\; \Psi - \Psi\Sigma^{-1}\Psi \;=\; A - A\Sigma^{-1}A, \tag{25}$$

where $\Sigma = \Psi + A$. In case the Gaussian density $p_G(\theta, y) = p_G(\theta)p_G(y|\theta)$ can be properly defined, matrix V is the conditional variance matrix $\mathrm{Var}(\theta|y)$. Given the framework of §2.1, the simulation smoothing algorithm of de Jong and Shephard (1995) or Durbin and Koopman (2002) can be used for generating draws from the importance function $N(\widehat{\theta}, V)$ in a computationally efficient way. However, the derivations of these methods rely on a properly defined linear Gaussian observation equation (12) and positive definite matrices for Σ and $H = A$. These conditions are not necessary for Theorem 2.

Theorem 2.

Consider the linear Gaussian signal vector of §2.1 with $\theta \sim N(\mu, \Psi)$. Further, consider x and A as defined by (22) evaluated at $\theta = \widehat{\theta}$. Notice that both matrices A and $\Sigma = \Psi + A$ can be

negative definite while $V = A - A\Sigma^{-1}A$ is positive definite by construction. Sampling from

$$f(\theta; y) = N\left(\widehat{\theta}\ ,\ A - A\Sigma^{-1}A\right),$$

can be carried out by the Kalman filter (34) *and the simulation smoothing equations*

$$\begin{array}{rclrcl} C_t & = & A_t^{-1} - F_t^{-1} - K_t' N_t K_t, & R_t & = & C_t^{-1}(A_t^{-1} Z_t - K_t' N_t T_t), \\ w_t & \sim & N(0, C_t), & u_t & = & A_t(w_t + F_t^{-1} v_t - K_t' r_t), \\ r_{t-1} & = & Z_t' A_t^{-1} u_t - R_t' w_t + T_t' r_t, & N_{t-1} & = & R_t' C_t R_t - Z_t' A_t^{-1} Z_t + T_t' N_t T_t, \end{array} \tag{26}$$

for $t = n, n-1, \ldots, 1$ and with the initialisations $r_n = 0$ and $N_n = 0$. Matrix A_t is the t-th diagonal block of the block diagonal matrix A. It can be shown that these equations are equivalent to the ones of de Jong and Shephard (1995).

The proof of Theorem 2 is presented in the Appendix. When matrix A is not positive definite, the simulation smoothing method of Durbin and Koopman (2002) can not be used since it requires simulating from the unconditional distribution implied by (12) with $y = x$ and variance matrix $H = A$ given by (22).

4.2 Evaluation of importance weights

The importance sampling estimator of

$$\bar{\theta} = \int \theta p(\theta|y)\,\mathrm{d}\theta = p(y)^{-1} \int \theta p(\theta, y)\,\mathrm{d}\theta, \quad \text{with} \quad p(y) = \int p(\theta, y)\,\mathrm{d}\theta,$$

is given by

$$\widehat{\bar{\theta}} = \left\{ \sum_{i=1}^{M} \frac{p(\theta^i, y)}{f(\theta^i; y)} \right\}^{-1} \sum_{i=1}^{M} \theta^i \frac{p(\theta^i, y)}{f(\theta^i; y)}, \qquad \theta^i \sim f(\theta; y). \tag{27}$$

The computation of (27) requires simulating $\theta^i \sim f(\theta; y)$ and evaluating the importance weight $p(\theta^i, y)\ /\ f(\theta^i; y)$ for $i = 1, \ldots, M$. Given the linear Gaussian state and signal vectors of §2.1, the evaluation of the nominator is based on the identity $p(\theta, y) = p(y|\theta)p(\theta)$ where $p(y|\theta)$ is defined by the model and is usually straightforward to compute. The density of the signal $p(\theta)$ for $\theta = \theta^i$ is evaluated by the Kalman filter since $\theta = c + Z\alpha$ has the Markov property and the prediction error decomposition can be applied to $p(\theta)$.

Given the draw $\theta^i \sim f(\theta; y)$ with $f(\theta; y) = N(\widehat{\theta}, V)$ obtained from the simulation smoother (26), the denominator $f(\theta^i; y)$ of the importance weight is evaluated by

$$f(\theta^i; y) = \exp\left(-\frac{mn}{2} \log 2\pi - \sum_{t=1}^{n} \log |A_t| - \sum_{t=1}^{n} \log |B_t| - \frac{1}{2} \sum_{t=1}^{n} o_t^i\ {}' o_t^i \right), \tag{28}$$

where $o_t^i \sim N(0, I_m)$ enables the calculation of $w_t^i \sim N(0, C_t)$ in (26) using the relation $w_t^i = B_t o_t^i$. Matrix B_t is the result of a Choleski decomposition applied to $C_t = B_t B_t'$. Matrix A_t is the t-th diagonal block of the block diagonal matrix A. The derivation of (28) is presented in the Appendix. This general method of evaluating importance sampling weights is a new result.

4.3 Likelihood evaluation, state estimation and state simulation

An expression for the likelihood function $\ell(\psi)$ is obtained by

$$\ell(\psi) = p(y) = \int p(\theta, y)\, \mathrm{d}\theta,$$

where ψ is a vector of model coeffients that are unknown and need to be estimated. The importance sampling estimator of the likelihood function is given by

$$\widehat{\ell}(\psi) = M^{-1} \sum_{i=1}^{M} \frac{p(\theta^i, y)}{f(\theta^i; y)}, \qquad \theta^i \sim f(\theta^i; y), \tag{29}$$

where the evaluation of importance weight $p(\theta^i, y)\ /\ f(\theta^i; y)$ is described in §4.2 for $i = 1, \ldots, M$.

Linear and nonlinear functions of the signal vector θ can be estimated in the same way as in (27). Furthermore, the smoothed estimate of the state vector $\mathrm{E}(\alpha_t|y)$, for $t = 1, \ldots, n$, can also be based on the importance sample weights for θ since

$$\begin{aligned} \mathrm{E}(\alpha_t|y) &= \int \mathrm{E}(\alpha_t|y, \theta)\ p(\theta|y)\, \mathrm{d}\theta \\ &= p(y)^{-1} \int \mathrm{E}(\alpha_t|\theta)\ p(\theta, y)\, \mathrm{d}\theta. \end{aligned}$$

It follows that the importance sampling estimator can be computed by

$$\widehat{\mathrm{E}}(\alpha_t|y) = \left\{ \sum_{i=1}^{M} \frac{p(\theta^i, y)}{f(\theta^i; y)} \right\}^{-1} \sum_{i=1}^{M} \mathrm{E}(\alpha_t|\theta^i) \frac{p(\theta^i, y)}{f(\theta^i; y)}, \qquad \theta^i \sim f(\theta; y), \tag{30}$$

where $\mathrm{E}(\alpha_t|\theta^i)$ is evaluated by the Kalman filter and smoother applied to the Gaussian state space model (12) with $y = \theta^i$ and $H = 0$. The same argument applies to the evaluation of $\mathrm{Var}(\alpha_t|y)$ for $t = 1, \ldots, n$. It is shown that state smoothing for a nonlinear non-Gaussian state space model does not require a simulation smoothing algorithm for the state vector α_t. The simulation smoother for the usually lower dimensional signal vector θ_t is sufficient.

Finally, a draw from $f(\theta; y)$ can be extended to a draw from $f(\alpha; y)$ using an existing method of conditional simulation from a multivariate normal density. A special simulation smoother for $f(\alpha; y)$ is therefore not required. Since

$$p(\alpha, \theta|y) = p(\alpha|\theta, y) p(\theta|y) = p(\alpha|\theta) p(\theta|y),$$

it also holds that the joint importance function of $\{\alpha, \theta\}$ based on the posterior mode $\widehat{\theta}$ is given by $f(\alpha, \theta; y) = p_G(\alpha, \theta|y) = p_G(\alpha|\theta) f(\theta; y)$ where $f(\theta; y) = N(\widehat{\theta}, V)$ and $p_G(\alpha|\theta)$ is the multivariate normal density with mean vector $\mathrm{E}(\alpha|\theta)$ and variance matrix $\mathrm{Var}(\alpha|\theta)$. This decomposition implies a two-step procedure for the simulation from $f(\alpha, \theta; y)$: (i) simulate $\theta^i \sim f(\theta; y)$ using Theorem 2; (ii) simulate $\alpha^i \sim p_G(\alpha|\theta^i)$ where the draw θ^i is treated as a realisation (or as a given data vector) for the linear Gaussian state space model (12) with $y = \theta^i$ and $H = 0$. To draw from $\alpha^i \sim p_G(\alpha|\theta^i)$ in a computationally efficient way, we adopt a simulation device that is based on an unconditional draw from $p_G(\alpha)$ and on a conditional mean adjustment. In particular, we have

$$\alpha^i \;=\; \mathrm{E}(\alpha|\theta^i) \;+\; \widetilde{\alpha}^i \;-\; \mathrm{E}(\alpha|\widetilde{\theta}^i), \qquad \widetilde{\alpha}^i \sim p_G(\alpha), \qquad \widetilde{\theta}^i = c + Z\widetilde{\alpha}^i. \tag{31}$$

The draw α^i is by definition consistent with the realisation θ^i since both $\mathrm{E}(\alpha|\theta^i)$ and $\mathrm{E}(\alpha|\widetilde{\theta}^i)$ are constructed by a linear Gaussian observation equation $\theta = c + Z\alpha + \varepsilon$ where the observation noise $\varepsilon \sim N(0, H)$ is set to zero, that is $H = 0$, for $\theta = \theta^i$ and $\theta = \widetilde{\theta}^i$, respectively. The conditional simulation device (31) is proposed by Journel (1974) in the geostatistics literature and by Durbin and Koopman (2002) for the linear Gaussian state space model.

5 Stochastic volatility model with leverage

Consider a time series of asset log-returns y_t that has mean zero and time-varying variance $\exp h_t$. The observations are typically sampled at daily intervals. The basic stochastic volatility (SV) model considers a stochastic dynamic process for the log-variance process h_t and is given by

$$\begin{aligned} y_t &= \sigma \exp(\tfrac{1}{2} h_t)\varepsilon_t, & \varepsilon_t &\sim NID(0,1), \\ h_{t+1} &= \phi h_t + \sigma_\eta \eta_t, & \eta_t &\sim NID(0,1), \qquad h_1 \sim N\left\{0, \sigma_\eta^2 (1-\phi^2)^{-1}\right\}, \end{aligned} \tag{32}$$

for $t = 1, \ldots, n$ and where ε_t and η_s are mutually and serially independent of each other at all time points $t, s = 1, \ldots, n$. This basic SV model has a nonlinear observation equation due to the multiplication $\exp(\frac{1}{2} h_t)\varepsilon_t$. It is usually assumed that the log-volatility process is stationary but persistent, that is, $0 < \phi < 1$ is typically larger than 0.9. More detailed discussions on the SV model can be found in the overview papers of Shephard (2005). To account for the leverage effect in return series (see, for example, Black (1976), Nelson (1991) and Yu (2005)), the basic SV model is extended by having correlation between ε_t and η_t, that is

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim NID\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad |\rho| \le 1, \tag{33}$$

and independent of α_1 for $t = 1, \ldots, n$. The unknown coefficients (scaling variance σ^2, autoregressive coefficient ϕ, variance σ_η^2 and correlation coefficient ρ) are collected in the parameter vector ψ.

Parameter estimation is carried out by maximising the Monte Carlo estimate of the likelihood function $\widehat{\ell}(\psi)$ in (29). The methods of this paper can be adopted but it requires a specific formulation of the state space model to incorporate the correlation between the disturbances and to have a non-singular variance matrix Ψ for the signal vector θ. We therefore consider the formulation as given by

$$y_t = \sigma \exp(\frac{1}{2}h_t)\left\{\varepsilon_t^* + \text{sign}(\rho)\eta_{2,t}\right\}, \qquad \varepsilon_t^* \sim NID(0, 1-|\rho|),$$

where

$$h_{t+1} = \phi h_t + \sigma_\eta \left\{\eta_{1,t} + \eta_{2,t}\right\}, \qquad \eta_{1,t} \sim NID(0, 1-|\rho|), \qquad \eta_{2,t} \sim NID(0, |\rho|),$$

for $t = 1, \ldots, n$ and with $h_1 \sim N\left\{0, \sigma_\eta^2(1-\phi^2)^{-1}\right\}$. The disturbances are all mutually and serially independent. It can be verified that this formulation is consistent with the SV model (32) with leverage (33). In terms of the linear Gaussian state and signal vectors of §2.1, we have $\theta_t = \alpha_t = (h_t \ , \ \sigma_\eta \eta_{2,t})'$ and $\eta_t = \sigma_\eta(\eta_{1,t} \ , \ \eta_{2,t+1})'$ where

$$\alpha_{t+1} = \begin{bmatrix} \phi & 1 \\ 0 & 0 \end{bmatrix} \alpha_t + \eta_t, \qquad \eta_t \sim NID\left(0, \sigma_\eta^2 \begin{bmatrix} 1-|\rho| & 0 \\ 0 & |\rho| \end{bmatrix}\right),$$

for $t = 1, \ldots, n$ and with

$$\alpha_1 \sim NID\left(0, \sigma_\eta^2 \begin{bmatrix} (1-\phi^2)^{-1} & 0 \\ 0 & |\rho| \end{bmatrix}\right).$$

The observations $y_1, \ldots, y_n$ are from the conditional density $\log p(y|\theta) = \sum_{t=1}^n \log p(y_t|\theta_t)$ where

$$\log p(y_t|\theta_t) = \text{constant} - \frac{1}{2}h_t - \frac{1}{2}\sigma^{-2}\exp(-h_t)(1-|\rho|)^{-1}\{y_t - \sigma\exp(\frac{1}{2}h_t)\text{sign}(\rho)\eta_{2,t}\}^2,$$

for $t = 1, \ldots, n$. Expressions for the 2×1 vector $\dot{p}(y_t|\theta_t)$ and the 2×2 matrix $\ddot{p}(y_t|\theta_t)$, as defined in (19), can be obtained straightforwardly. The matrix $\ddot{p}(y_t|\theta_t)$ is indefinite and therefore we require the new methods. The Monte Carlo estimator (29) of the likelihood function is then computed by the following three steps:

1. The importance function $f(\theta; y)$ is based on the mode of $p(\theta|y)$ using Theorem 1.
2. M conditional draws of $\theta_1, \ldots, \theta_n$ are obtained from $f(\theta; y)$ using Theorem 2.
3. The Monte Carlo estimator (29) of the likelihood is computed as described in §4.2.

The parameter vector ψ is estimated for three daily stock index return series (the Frankfurt DAX, the London FTSE and the Tokyo NIKKEI). The data is taken from datastream and the sample for all series is from January 4, 1997 to September 9, 2005 (that is 965 days). The parameter estimates are reported in Table 1 together with their standard errors. The estimation results reveal that leverage is an important feature in daily return series as all correlations are very significant. In all cases estimation was fast and computation of standard errors did not lead to numerical problems.

	ϕ	σ_η^2	σ^2	ρ
DAX	0.978	0.016	1.314	−0.834
	(0.011)	(0.009)	(0.244)	(0.097)
FTSE	0.981	0.0081	0.546	−0.993
	(0.008)	(0.005)	(0.044)	(0.035)
NIKKEI	0.974	0.028	1.255	−0.611
	(0.013)	(0.014)	(0.290)	(0.135)

Table 1: Maximum likelihood estimation of parameters in the SV model (32) with leverage (33) for the daily stock index return series Frankfurt DAX, London FTSE and Tokyo NIKKEI from January 4, 1997 to September 9, 2005 (that is 965 days).

6 Conclusion

This paper has shown that standard state space methods can be used for the analysis of time series models with nonlinear non-Gaussian observation equations. The method of importance sampling is considered. The importance density is chosen to be the Laplace approximation of the smoothing density $p(\theta|y)$ where θ is the signal and y is the observation vector. The Kalman filter, smoother and simulation smoother, applied to the approximating Gaussian model, can be used to perform the necessary calculations. The method breaks down when $\ddot{p}(\theta|y)$, the second derivative of $p(\theta|y)$ with respect to θ, is positive definite since an approximating model can not be based on negative variances. It is argued that standard state space methods can still be used when $\ddot{p}(\theta|y)$ is positive definite but that the existing derivations of the methods are no longer valid. We therefore have presented new derivations of the Kalman filter, smoothing and simulation smoothing algorithms that are not based on a linear Gaussian observation model. The new derivations lead to a new variant of the simulation smoothing recursions of de Jong and Shephard (1995). The two recursions are numerically equivalent but the new variant is computationally more efficient. Other results of this paper include the introduction of a line-search strategy within the Newton-Raphson method to find the posterior mode, an efficient method of computing the score $\dot{p}(\theta|y)$, a general method of computing importance weights and a novel treatment of estimating and simulating the state vector α conditional on y that is based on the (usually) lower-dimensional signal vector θ. An illustration for the stochastic volatility model is presented to show that the methods work in practice. The empirical results confirm that leverage in the SV model is a clear salient feature in daily financial return series. The negative correlation between financial returns and innovations in the underlying log-volatility process is estimated significantly for three major stock index return series.

Appendix

The proofs of Theorems 1 and 2 are presented as a series of propositions with derivations. For all propositions in this Appendix, we consider the framework and definitions of §2.1. An important matrix is $\Sigma = \Psi + A$ where $\Psi = Z\Omega Z'$ is the positive definite symmetric matrix from §2.1 and A is a block diagonal matrix of invertible symmetric matrices. Matrices Ψ and A consist of n^2 blocks of $m \times m$ matrices. Matrices Σ and A are possibly not positive definite.

Proof of Theorem 1

Theorem 1 states that $g^+ = \mu + \Psi\Sigma^{-1}(x - \mu)$ can be computed efficiently by the Kalman filter and smoothing equations (34), (43) and (44) for both positive and non-positive definite matrices Σ. The proof is presented as the sequence of Propositions 1, 2 and 3 which are given below together with their derivations.

Proposition 1

Assume that an LU decomposition for the symmetric matrix $\Sigma = DU$ exists where D is a lower block unity triangular matrix and U is an upper block triangular matrix (for sufficient conditions, see Golub and Van Loan, 1997, §3.2). The Kalman filter equations

$$\begin{aligned} v_t &= x_t - c_t - Z_t a_t, & F_t &= A_t + Z_t P_t Z_t', \\ & & K_t &= T_t P_t Z_t' F_t^{-1}, & & t = 1, \ldots, n, \\ a_{t+1} &= d_t + T_t a_t + K_t v_t, & P_{t+1} &= T_t P_t T_t' - K_t F_t K_t' + Q_t, \end{aligned} \tag{34}$$

solve the set of linear equations $Dv = x - \mu$ for $v = (v_1', \ldots, v_n')'$ with $x = (x_1', \ldots, x_n')'$ given. The initialisations a_1 and P_1 are defined in section 2.1. Apart from the fact that matrices A_t and F_t are not necessarily positive definite for $t = 1, \ldots, n$, the equations (34) are the same as the Kalman filter in Anderson and Moore (1979).

Derivation of Proposition 1

The (i, j) block of a matrix is labelled by subscript $_{ij}$ for the range of $i, j = 1, \ldots, n$. The $m \times m$ matrix block (i, j) of Σ is given by

$$\Sigma_{ij} = \begin{cases} \sum_{k=1}^{i-1} D_{ik} U_{ki} + U_{ij}, & i = j, \\ \sum_{k=1}^{j} D_{ik} U_{kj}, & i > j, \\ \sum_{k=1}^{i-1} D_{ik} U_{kj} + U_{ij}, & i < j. \end{cases} \tag{35}$$

From (6), we have

$$\Omega_{ij} = \begin{cases} T_{i-1}\Omega_{i-1,i-1}T'_{i-1} + Q_i, & i = j > 1, \\ T_{i-1}T_{i-2}\dots T_j\Omega_{jj}, & i > j, \\ \Omega_{ii}(T_{j-1}T_{j-2}\dots T_i)', & i < j, \end{cases} \tag{36}$$

and $\Omega_{11} = P_1$. From (9) and the definition $\Sigma = \Psi + A = Z\Omega Z' + A$, we have further

$$\Sigma_{ij} = \begin{cases} Z_i\Omega_{ii}Z'_i + A_i, & i = j, \\ Z_i\Omega_{ij}Z'_j, & i \neq j. \end{cases} \tag{37}$$

Equating (35) with (37) and considering (36), the following expressions for the block element matrices of U and D are obtained,

$$\begin{array}{rcll} U_{ij} & = & Z_i\Omega_{ii}Z'_i + A_i - \sum_{k=1}^{i-1} D_{ik}U_{ki}, & i = j, \\ D_{ij} & = & (Z_iT_{i-1}T_{i-2}\dots T_j\Omega_{jj}Z'_j - \sum_{k=1}^{j-1} D_{ik}U_{kj})U_{jj}^{-1}, & i > j, \\ U_{ij} & = & Z_i\Omega_{ii}(T_{j-1}T_{j-2}\dots T_i)'Z'_j - \sum_{k=1}^{i-1} D_{ik}U_{kj}, & i < j, \end{array} \tag{38}$$

which describe the typical triangular system for which solutions are obtained by forward and backward substitution algorithms, see Golub and Van Loan (1997, Chapters 3 and 4). The implications of these solutions for the special structures of involved matrices are given below. At the end of this derivation it is acknowledged that Σ is symmetric and therefore the LU decomposition reduces to the $\mathrm{LDL}^{\mathrm{T}}$ decomposition of Golub and Van Loan (1997). However, the presented derivation is more straightforward for an LU decomposition and it leads to the same computationally efficient solution.

Given the block diagonal structure of Z and the definition $\Sigma = DU = Z\Omega Z' + A$, the lower block matrix D_{ij} equals a matrix that is premultiplied by Z_i for $i > j$ and the upper block matrix U_{ij} equals a matrix that is postmultiplied by Z_j for $i < j$ with $i, j = 1, \dots, n$. Therefore, we obtain

$$\begin{array}{rcll} U_{ij} & = & Z_iY_{ii}Z'_i + A_i, & i = j, \\ D_{ij} & = & Z_iX_{ij}U_{jj}^{-1}, & i > j, \\ U_{ij} & = & Z_iY_{ij}Z'_j, & i < j. \end{array} \tag{39}$$

To ensure consistency between (38) and (39), matrices X_{ij} and Y_{ij} are defined by

$$\begin{array}{rcll} Y_{ij} & = & \Omega_{ii} - \sum_{k=1}^{i-1} X_{ik}U_{kk}^{-1}Z_kY_{ki}, & i = j, \\ X_{ij} & = & T_{i-1}T_{i-2}\dots T_j\Omega_{jj}Z'_j - \sum_{k=1}^{j-1} X_{ik}U_{kk}^{-1}U_{kj}, & i > j, \\ Y_{ij} & = & \Omega_{ii}(T_{j-1}T_{j-2}\dots T_i)' - \sum_{k=1}^{i-1} X_{ik}U_{kk}^{-1}Z_kY_{kj}, & i < j. \end{array} \tag{40}$$

For a particular value of $i = t+1$, we have

$$\begin{aligned}
X_{t+1,t} &= T_t \Omega_{tt} Z_t' - \textstyle\sum_{k=1}^{t-1} X_{t+1,k} U_{kk}^{-1} U_{kt} \\
&= T_t \left(\Omega_{tt} - \textstyle\sum_{k=1}^{t-1} X_{tk} U_{kk}^{-1} Z_k Y_{kt} \right) Z_t' \\
&= T_t Y_{tt} Z_t', \\
X_{t+1,j} &= T_t T_{t-1} \ldots T_j Y_{jj} Z_j', \\
D_{t+1,j} &= Z_{t+1} T_t T_{t-1} \ldots T_j Y_{jj} Z_j' U_{jj}^{-1} \\
&= Z_{t+1} T_t T_{t-1} \ldots T_{j+1} M_j,
\end{aligned} \tag{41}$$

where we define $M_j = T_j Y_{jj} Z_j' U_{jj}^{-1} \equiv Y_{j,j+1} Z_j' U_{jj}^{-1}$ for $j = 1, \ldots, t+1$. Note that $X_{t+1,t} = M_t U_{tt}$. The matrices X_{ij} and D_{ij} depend on state space matrices and on Y_{ij} for $i > j$.

Given the earlier definitions and results, we continue the proof by deriving a recursion for Y_{ij} for $i = t+1$ and $j = t+2, \ldots, n$. It follows that

$$\begin{aligned}
Y_{t+1,t+1} &= \Omega_{t+1,t+1} - \textstyle\sum_{k=1}^{t} X_{t+1,k} U_{kk}^{-1} Z_k Y_{k,t+1} \\
&= T_t \left(\Omega_{tt} - \textstyle\sum_{k=1}^{t-1} X_{t,k} U_{kk}^{-1} Z_k Y_{kt} \right) T_t' + Q_t - X_{t+1,t} U_{tt}^{-1} Z_t Y_{t,t+1} \\
&= T_t Y_{tt} T_t' + Q_t - M_t U_{tt} M_t', \\
Y_{t+1,j} &= Y_{t+1,t+1} (T_{j-1} T_{j-2} \ldots T_{t+1})', \\
U_{t+1,j}' &= Z_j T_{j-1} \ldots T_{t+2} M_{t+1} U_{t+1,t+1},
\end{aligned} \tag{42}$$

for $j = t+2, \ldots, n$. By defining $P_t \equiv Y_{tt}$, $K_t \equiv M_t$ and $F_t \equiv U_{tt}$ and by some minor manipulations, it is shown that part of the Kalman filter equations (34) carry out the LU decomposition of Σ in a complete recursive way. Since $U_{ij}' = D_{ji} U_{ii}$, we have $U' = DF$ and the LU decomposition can be rewritten by $\Sigma = DU = DFD'$ where F is a block diagonal matrix with blocks $F_1, \ldots, F_n$.

Given the block structure of matrix D, it is straightforward to show that the solution of $Dv = x - \mu$ for v can be obtained by forward substitution, that is

$$v_1 = x_1 - c_1, \qquad v_t = x_t - c_t - \sum_{i=1}^{t-1} Z_t T_{t-1} \ldots T_{i+1} (d_i + M_i v_i),$$

for $t = 2, \ldots, n$. A recursion is obtained by defining $a_t = \sum_{i=1}^{t-1} T_{t-1} \ldots T_{i+1} (d_i + M_i v_i)$ for $t = 2, \ldots, n$. Some minor manipulation completes the derivation of all equations in (34).

Proposition 2

Assume that an LU decomposition for the symmetric matrix $\Sigma = DU$ exists where D is a lower block unity triangular matrix and U is an upper block triangular matrix. The solution $e = \Sigma^{-1}(x - \mu)$ for any given vector x is obtained by back substitution and leads to the smoothing equations

$$e_t = F_t^{-1} v_t - K_t' s_t, \qquad s_{t-1} = Z_t' e_t + T_t' s_t, \tag{43}$$

for $t = n, \ldots, 1$ with the initialisation $s_n = 0$. Similar equations have appeared in Kohn and Ansley (1989), de Jong (1989) and Koopman (1993) although these results do not allow for a negative definite matrix F_t.

Derivation of Proposition 2

Given the result in Proposition 1, we have $e = \Sigma^{-1}(x - \mu) = U^{-1}D^{-1}(x - \mu) = U^{-1}v$ so that we need to solve $Ue = v$ for e. The upper triangular block structure of $U = FD'$ is derived in Proposition 1 and it follows from the back substitution that

$$e_n = F_n^{-1}v_n, \qquad e_t = F_t^{-1}v_t - \sum_{i=t+1}^{n} K_t'T_{t+1}'\ldots T_{i-1}'Z_i'e_i, \qquad t = n-1, n-2, \ldots, 1.$$

The backward recursion (43) relies on the definition $s_t = \sum_{i=t+1}^{n} T_{t+1}'\ldots T_{i-1}'Z_i'e_i$ that can be evaluated recursively as in (43). This completes the derivation of Proposition 2.

Proposition 3

Assume that an LU decomposition for the symmetric matrix $\Sigma = DU$ exists where D is a lower block unity triangular matrix and U is an upper block triangular matrix. The expression of $g^+ = \mu + \Psi\Sigma^{-1}(x - \mu)$ in Theorem 1 is equivalent to $g^+ = x - Ae$ and its t-th block element is computed by

$$g_t^+ = x_t - A_t e_t, \tag{44}$$

for $t = n, \ldots, 1$. This expression can be merged with the backward recursion (43).

Derivation of Proposition 3

The equivalence $x - Ae \equiv \mu + \Psi\Sigma^{-1}(x - \mu)$ follows from

$$\begin{aligned} x - Ae &= (\Sigma - A)\Sigma^{-1}(x - \mu) + \mu \\ &= \mu + \Psi\Sigma^{-1}(x - \mu). \end{aligned}$$

The block structure of A leads to (44).

Proof of Theorem 2

Theorem 2 states that a draw from $N(\widehat{\theta}, V)$ can be computed by the equations (34) and (26). From proposition 3 we have $\widehat{\theta} = \mu + \Psi\Sigma^{-1}(x - \mu) = x - b$ where $b = Ae = A\Sigma^{-1}(x - \mu)$. The variance matrix V is expressed in (25) as $V = A - A\Sigma^{-1}A$ and from Proposition 1 we have

$\Sigma = DU = DFD'$ where D is a lower block unity triangular matrix and F is a block diagonal matrix. Furthermore, it follows from the derivation of Proposition 2 that

$$b = AD'^{\,-1}F^{-1}v, \qquad V = A - AD'^{\,-1}F^{-1}D^{-1}A = AD'^{\,-1}(D'A^{-1}D - F^{-1})D^{-1}A. \tag{45}$$

Suppose that the draw $u^+ \sim N(0, D'A^{-1}D - F^{-1})$ is available, then $u = AD'^{\,-1}(u^+ + F^{-1}v)$ is a draw from $u \sim N(b, V)$. We concentrate on $u \sim N(b, V)$ since we have that $x - u$ is a draw from $N(\widehat{\theta}, V)$. The proof of Theorem 2 is presented as the sequence of Propositions 4 and 5 for which the derivations are given below. Proposition 6 shows that the equations (26) of Theorem 2 are equivalent to the simulation smoothing equations of de Jong and Shephard (1995).

Proposition 4

Assume that the LU decomposition of Proposition 1 applies to the symmetric matrix $\Sigma = DU = DFD'$. Further, assume that the Kalman filter equations of Theorem 1 has been applied and the matrices F_i and K_i are stored for $i = 1, \ldots, n$. Consider the LDL^{T} decomposition of Golub and Van Loan (1997, §3.3) for matrix

$$D'A^{-1}D - F^{-1} = E'CE.$$

where E is a lower block unity triangular matrix and C is a block diagonal matrix. The block elements of C and E can be evaluated by

$$C_i = A_i^{-1} - F_i^{-1} - K_i'N_iK_i, \qquad E_{ij} = R_iT_{i-1}\ldots T_{j+1}K_j, \tag{46}$$

with $R_i = C_i^{-1}(A_i^{-1}Z_i - K_i'N_iT_i)$ and where N_i is computed by the backward recursion

$$N_{i-1} = R_i'C_iR_i - Z_i'A_i^{-1}Z_i + T_i'N_iT_i, \tag{47}$$

with $N_n = 0$ for $i, j = n, \ldots, 1$ and $j < i$.

Derivation of Proposition 4

Since the variance matrix $D'A^{-1}D - F^{-1}$ is symmetric, we restrict ourselves to the lower diagonal blocks. From (41) in Proposition 1 and the definition $K_i = M_i$, we have

$$D_{i,i-1} = Z_iK_{i-1}, \qquad D_{ij} = Z_iT_{i-1}\ldots T_{j+1}K_j, \qquad i = 2, \ldots, n, \quad j = 1, \ldots, i-2.$$

Given the block structures of the matrices E and C, the (i, j) block of $E'CE$ is given by

$$(E'CE)_{ij} = \begin{cases} C_i + \sum_{k=i+1}^{n} E_{ki}'C_kE_{ki} & i = j, \\ C_iE_{ij} + \sum_{k=i+1}^{n} E_{ki}'C_kE_{kj} & i > j. \end{cases} \tag{48}$$

Given the definitions of D_i and $F_i = U_{ii}$ from equations (41) and (42) in Proposition 1, respectively, the block (i,j) of matrix $D'A^{-1}D - F^{-1}$ is given by

$$(D'A^{-1}D - F^{-1})_{ij} =$$

$$\begin{cases} A_i^{-1} - F_i^{-1} + \sum_{k=i+1}^{n} K_i' T_{i+1}' \dots T_{k-1}' Z_k' A_k^{-1} Z_k T_{k-1} \dots T_{i+1} K_i = \\ A_i^{-1} - F_i^{-1} + K_i' N_i^{[a]} K_i, & i = j, \\ \left(A_i^{-1} Z_i + \sum_{k=i+1}^{n} K_i' T_{i+1}' \dots T_{k-1}' Z_k' A_k^{-1} Z_k T_{k-1} \dots T_{i+1} T_i\right) T_{i-1} \dots T_{j+1} K_j = \\ (A_i^{-1} Z_i + K_i' N_i^{[a]} T_i) T_{i-1} \dots T_{j+1} K_j, & i > j, \end{cases} \tag{49}$$

since the summation $\sum_{k=i+1}^{n} T_{i+1}' \dots T_{k-1}' Z_k' A_k^{-1} Z_k T_{k-1} \dots T_{i+1}$ can be evaluated by the backward recursion

$$N_{i-1}^{[a]} = Z_i' A_i^{-1} Z_i + T_i' N_i^{[a]} T_i, \tag{50}$$

for $i = n, \dots, 1$ and with $N_n^{[a]} = 0$. The induction argument allows us to assume that equations (46) and (47) hold for $i = m+1, \dots, n$ and $j < i$. We need to verify (46) and (47) for $i = m$ and $j < m$. By substituting the equation for E_{ij} of (46) into (48), the two summations in (48) can be given by

$$\sum_{k=i+1}^{n} E_{ki}' C_k E_{ki} = K_i' N_i^{[c]} K_i, \qquad \sum_{k=i+1}^{n} E_{ki}' C_k E_{kj} = K_i' N_i^{[c]} T_i T_{i-1} \dots T_{j+1} K_j, \tag{51}$$

for $i = m+1, \dots, n$ and $j > i$, since the summation $\sum_{k=i+1}^{n} T_{i+1}' \dots T_{k-1}' R_k' C_k R_k T_{k-1} \dots T_{i+1}$ can be evaluated by the backward recursion

$$N_{i-1}^{[c]} = R_i' C_i R_i + T_i' N_i^{[c]} T_i, \tag{52}$$

for $i = n, \dots, 1$ and with $N_n^{[c]} = 0$. By substituting the first equation of (51) into the first row of (48) and by equating (48) and (49) for $i = j$, the first equation in (46) is established for $i = m$, that is C_m. In a similar way, we can establish the second equation in (46) for $i = m$ and $j < m$ by

$$\begin{aligned} E_{mj} &= C_m^{-1}[(A_m^{-1} Z_m + K_m' N_m^{[a]} T_m) T_{m-1} \dots T_{j+1} K_j - \textstyle\sum_{k=m+1}^{n} E_{km}' C_k E_{kj}] \\ &= C_m^{-1}[A_m^{-1} Z_m - K_m' (N_m^{[c]} - N_m^{[a]}) T_m] T_{m-1} \dots T_{j+1} K_j \\ &= R_m T_{m-1} \dots T_{j+1} K_j, \end{aligned}$$

where R_m is defined as $R_m = C_m^{-1}(A_m^{-1} Z_m - K_m' N_m T_m)$. The definition $N_i \equiv N_i^{[c]} - N_i^{[a]}$ completes the derivation of Proposition 4.

Proposition 5

The draw $u \sim N(c, V)$ is generated by the backward recursion

$$w_t \sim N(0, C_t), \qquad u_t = A_t(F_t^{-1} v_t + w_t - K_t' r_t), \qquad r_{t-1} = Z_t' A_t^{-1} u_t - R_t' w_t + T_t' r_t, \tag{53}$$

for $t = n, \ldots, 1$ and $r_n = 0$. The recursion (53) is carried out alongside the algorithm of Proposition 4.

Derivation of Proposition 5

The draw $u^+ \sim N(0, E'CE)$ is computed by $u^+ = E'w$ where $w \sim N(0, C)$. Given the block diagonal structure of C and the lower triangular block structure of E with the blocks given by Proposition 4, we have

$$w_t \sim N(0, C_t), \qquad u_t^+ = w_t + \sum_{i=t+1}^{n} E'_{it} w_i = w_t + \sum_{i=t+1}^{n} K'_t T'_{t+1} \ldots T'_{i-1} R'_i w_i,$$

for $t = n, \ldots, 1$. It follows that u_t^+ can be evaluated by the backward recursion

$$u_t^+ = w_t + K'_t r_t^{[c]}, \qquad r_{t-1}^{[c]} = R'_t w_t + T'_t r_t^{[c]},$$

with $r_n^{[c]} = 0$ and $t = n, \ldots, 1$.

The draw $u \sim N(b, V)$ is equivalent to $u = AD'^{\,-1}(u^+ + F^{-1}v)$ and can be computed by solving the system $D'u^x = u^+ + F^{-1}v$ for u^x such that $u = Au^x$. The solution for u^x follows from Proposition 2 where a similar system is solved. We therefore obtain the recursion

$$u_t = A_t u_t^x, \qquad u_t^x = F_t^{-1} v_t + u_t^+ - K'_t r_t^{[a]}, \qquad r_{t-1}^{[a]} = Z'_t u_t^x + T'_t r_t^{[a]},$$

with $r_n^{[a]} = 0$. By defining $r_t = r_t^{[a]} - r_t^{[c]}$ and by re-ordering the equation, the proof of Proposition 5 is complete.

Proposition 6

Propositions 4 and 5 constitute the proof of Theorem 2 that consists of efficient recursions for generating draws from the importance function. These equations are consistent with the simulation smoother of de Jong and Shephard (1995) as given by

$$\begin{array}{rclrcl}
L_t & = & T_t - K_t Z_t, & & & \\
C_t & = & A_t^{-1} - F_t^{-1} - K'_t N_t K_t, & W_t & = & F_t^{-1} Z_t - K'_t N_t L_t, \\
w_t & \sim & N(0, C_t), & u_t & = & A_t(w_t + F_t^{-1} v_t - K'_t r_t), \\
r_{t-1} & = & Z'_t F_t^{-1} v_t - W'_t C_t^{-1} u_t + L'_t r_t, & N_{t-1} & = & Z'_t F_t^{-1} Z_t + W'_t C_t^{-1} W_t + L'_t N_t L_t,
\end{array} \tag{54}$$

for $t = n, n-1, \ldots, 1$ and with the initialisations $r_n = 0$ and $N_n = 0$. Apart from the fact that matrices A_t and F_t can be negative definite for $t = 1, \ldots, n$, the equations (54) are the same as the simulation smoother of de Jong and Shephard (1995).

Derivation of Proposition 6

The definition $L_t = T_t - K_t Z_t$ is introduced and the equations for C_t, u_t^c and u_t are the same as in Propositions 4 and 5. Define $S_t = F_t^{-1} + K_t' N_t K_t$ for $t = 1, \ldots, n$. From the definition of C_t it follows that $A_t^{-1} = C_t + S_t$. A linear relation between R_t and W_t is derived by

$$\begin{aligned} R_t &= C_t^{-1}(A_t^{-1} Z_t - K_t' N_t T_t) \\ &= C_t^{-1}[(C_t + S_t) Z_t - K_t' N_t T_t] \\ &= C_t^{-1}[(C_t + F_t^{-1}) Z_t - K_t' N_t L_t] \\ &= C_t^{-1}(C_t Z_t + W_t), \end{aligned}$$

for $t = 1, \ldots, n$. It follows that $W_t = C_t(R_t - Z_t)$ and $(C_t + S_t)Z_t - C_t R_t - K_t' N_t T_t = 0$. The equivalence of the recursions for N_t in (47) and (54) is shown by

$$\begin{aligned} N_{t-1} &= R_t' C_t R_t - Z_t' A_t^{-1} Z_t + T_t' N_t T_t \\ &= R_t' C_t R_t - Z_t'(C_t + S_t) Z_t + T_t' N_t T_t \\ &+ Z_t'[(C_t + S_t) Z_t - C_t R_t - K_t' N_t T_t] + [Z_t'(C_t + S_t) - R_t' C_t - T_t' N_t K_t] Z_t \\ &= (R_t - Z_t)' C_t (R_t - Z_t) + Z_t' F_t^{-1} Z_t + L_t' N_t L_t \\ &= W_t' C_t^{-1} W_t + Z_t' F_t^{-1} Z_t + L_t' N_t L_t. \end{aligned}$$

The equivalence of the recursions for r_t in (53) and (54) is shown by

$$\begin{aligned} r_{t-1} &= Z_t' A_t^{-1} u_t - R_t' w_t + T_t' r_t \\ &= Z_t' F_t^{-1} v_t + (Z_t - R_t)' w_t - Z_t' K_t' r_t + T_t' r_t \\ &= Z_t' F_t^{-1} v_t - W_t' C_t^{-1} w_t + L_t' r_t. \end{aligned}$$

This completes the proof.

Derivation of Equation (28)

The log of the importance function $f(\theta; y) = N(\widehat{\theta}, V)$ is given by

$$\log f(\theta; y) = -\frac{mn}{2} \log 2\pi - \frac{1}{2} \log |V| - \frac{1}{2}(\theta - \widehat{\theta})' V^{-1} (\theta - \widehat{\theta}).$$

From equation (45) and Proposition 4 we have $V = AD'^{\,-1} E' C E D^{-1} A$ where matrices A and C are block diagonal and matrices D and E are lower block unity triangular matrices. From Proposition 5 we have $\theta = x - u$ with $u \sim N(0, V)$ and where $u = AD'^{\,-1}(E'w + F^{-1}v)$. From Propositions 2 and 3 we have $\widehat{\theta} = g^+ = x - A\Sigma^{-1}(x - \mu) = x - AD'^{\,-1} F^{-1} v$. It follows that

$$\theta - \widehat{\theta} = AD'^{\,-1} F^{-1} v - u = AD'^{\,-1} F^{-1} v - AD'^{\,-1}(E'w + F^{-1}v) = -AD'^{\,-1} E' w,$$

and

$$(\theta - \widehat{\theta})' V^{-1} (\theta - \widehat{\theta}) = w' E D^{-1} A A^{-1} D E^{-1} C^{-1} E'^{\,-1} D' A^{-1} A D'^{\,-1} E' w = w' C^{-1} w.$$

Given the lower block unity triangular matrices D and E, the log determinental term is

$$\log|V| = \log|AD'^{-1}E'CED^{-1}A| = \log(|A||D||E||C||E||D||A|) = 2\log|A| + \log|C|.$$

Since matrices A and C are block diagonal and the t-th diagonal block C_t of C is positive definite, we have

$$w'C^{-1}w = \sum_{t=1}^{n} o_t' \, o_t, \qquad \log|V| = 2\sum_{t=1}^{n}(\log|A_t| + \log|B_t|),$$

where $C_t = B_t B_t'$ and $o_t \sim N(0, I_m)$ so that $w_t = B_t o_t$ for $t = 1, \ldots, n$. This completes the derivation of (28).

References

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

Berzuini, C., N. G. Best, W. R. Gilks, and C. Larizza (1997). Dynamic conditional independence models and Markov chain Monte Carlo methods. *J. American Statistical Association 92*, 140312.

Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the Business and Economic Statistics Section*, 177–181.

Carlin, B. P., N. G. Polson, and D. S. Stoffer (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modelling. *J. American Statistical Association 87*, 493–500.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*, 541–53.

Danielsson, J. and J. F. Richard (1993). Accelerated Gaussian importance sampler with application to dynamic latent variable models. *J. Applied Econometrics 8*, S153–S174.

de Jong, P. (1989). Smoothing and interpolation with the state space model. *J. American Statistical Association 84*, 1085–8.

de Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*, 339–50.

Doucet, A., J. F. G. deFreitas, and N. J. Gordon (Eds.) (2000). *Sequential Monte Carlo methods in practice*. New York: Springer-Verlag.

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

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*, 603–16.

Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Analysis 15*, 183–202.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika 85*, 215–27.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–39.

Golub, G. H. and C. F. Van Loan (1997). *Matrix Computations* (2nd ed.). Baltimore: The Johns Hopkins University Press.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). A novel approach to non-linear and non-Gaussian Bayesian state estimation. *IEE-Proceedings F 140*, 107–13.

Journel, A. (1974). Geostatistics for conditional simulation of ore bodies. *Economic Geology 69*, 673–687.

Kitagawa, G. (1987). Non-Gaussian state space modelling of non-stationary time series. *J. American Statistical Association 82*, 503–14.

Kloek, T. and H. K. Van Dijk (1978). Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica 46*, 1–20.

Kohn, R. and C. F. Ansley (1989). A fast algorithm for signal extraction, influence and cross-validation. *Biometrika 76*, 65–79.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*, 117–26.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset pricing: a new approach. *Econometrica 59*, 347–70.

Nocedal, J. and S. J. Wright (1999). *Numerical Optimization*. New York: Springer verlag.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: auxiliary particle filter. *J. American Statistical Association 94*, 590–9.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*, 115–31.

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

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–67.

So, M. K. P. (2003). Posterior mode estimation for nonlinear and non-gaussian state space models. *Statistica Sinica 13*, 255–274.

Yu, J. (2005). On leverage in a stochastic volatility model. *J. Econometrics 44*, forthcoming.