Econometrics Journal (2003), volume **6**, pp. 407–419.

Stochastic volatility: Bayesian computation using automatic differentiation and the extended Kalman filter

RENATE MEYER[†§], DAVID A. FOURNIER[‡] AND ANDREAS BERG[†]

[†]*Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand*
E-mail addresses: meyer@stat.auckland.ac.nz; andreas@stat.auckland.ac.nz

[‡]*Otter Research Ltd., P.O. Box 2040, Sidney, B.C., Canada V8L 3S3*
E-mail: otter@otter-rsch.com

Received: September 2002

Summary Stochastic volatility (SV) models provide more realistic and flexible alternatives to ARCH-type models for describing time-varying volatility exhibited in many financial time series. They belong to the wide class of nonlinear state-space models. As classical parameter estimation for SV models is difficult due to the intractable form of the likelihood, Bayesian approaches using Markov chain Monte Carlo (MCMC) techniques for posterior computations have been suggested. In this paper, an efficient MCMC algorithm for posterior computation in SV models is presented. It is related to the integration sampler of Kim *et al.* (1998) but does not need an offset mixture of normals approximation to the likelihood. Instead, the extended Kalman Filter is combined with the Laplace approximation to compute the likelihood function by integrating out all unknown system states. We make use of automatic differentiation in computing the posterior mode and in designing an efficient Metropolis–Hastings algorithm. We compare the new algorithm to the single-update Gibbs sampler and the integration sampler using a well-known time series of pound/dollar exchange rates.

Keywords: *Markov chain Monte Carlo, Extended Kalman filter, Laplace approximation, Automatic differentiation, Heavy-tailed distributions, Non-Gaussian nonlinear state-space models.*

1. INTRODUCTION

Stochastic volatility (SV) models were introduced by Tauchen and Pitts (1983) and Taylor (1982) to describe the evolution of asset returns which typically exhibit changing variances over time. SV models can be cast into the framework of nonlinear state-space models with state-dependent variances. A Bayesian approach to parameter estimation in the SV model relies on Markov chain Monte Carlo (MCMC) techniques (Chen *et al.*, 2000) to sample from the joint posterior distribution. Carlin *et al.* (1992) showed how to implement the Gibbs sampler for nonlinear non-normal state-space models and various MCMC techniques have been suggested that are particularly tailored to SV models (e.g. Jacquier *et al.* (1994), Shephard and Pitt (1997), Kim *et al.* (1998), Stroud *et al.* (2001)). The single-update Gibbs sampler is used, for instance, in Kim *et al.* (1998) and Meyer and Yu (2000) to generate a sample from the joint posterior distribution of unknown

[§]Corresponding author.

parameters *and* unknown states. However, due to dependencies between consecutive states, there are high posterior correlations that cause the Markov chain to traverse the state space in only very tiny steps and thus mix inefficiently. Therefore, convergence of the Markov chain to the equilibrium distribution is slow, a large number of iterations is required to achieve a satisfactory precision of parameter estimates, and the estimation procedure becomes very time-consuming. This led researchers to focus on the development of more efficient MCMC techniques using *block-updating schemes* as in Carter and Kohn (1994), Fruhwirth-Schnatter (1994) and DeJong and Shephard (1995) and *auxiliary discrete mixture models* that approximate the given model as in Kim *et al.* (1998) and Stroud *et al.* (2001).

The auxiliary offset mixture representation in Kim *et al.* (1998) was found to improve the correlation behaviour of the simulations, but only to a small extent. A bigger gain was obtained by employing the Gaussian structure of the mixture representation and integrating out the log-volatilities using an augmented version of the Kalman filter. In this paper, we develop an alternative approach which is also based on the idea of gaining efficiency by first integrating out the unknown states. This reduces the problem of sampling vectors in a high $(n + p)$-dimensional space to that of sampling in a low (p)-dimensional space. If the dependence on states were *linear*, this integration could be performed using the Kalman filter. Due to the nonlinearity and the state-dependent variances in the SV model, however, this is not feasible. Thus we suggest an approach that combines the extended Kalman filter (EKF) (Harvey, 1989) with the Laplace approximation (Laplace, 1986). The EKF has been developed for non-linear non-Gaussian state-space models whereas the Laplace approximation has a long tradition in Bayesian computation as an asymptotic approximation to the posterior distribution (Gamerman, 1997). Here, we use the term 'extended Kalman filter' in a broad sense, referring to extensions of the classical Kalman filter where first- and second-order Taylor series expansions with numerical approximations to posterior modes and curvatures result in normal predictive and filtering densities (Fahrmeir and Tutz, 1994). The proposed technique is not restricted to Gaussian errors but can also be applied to robustify models by allowing for outlying observations through heavy-tailed error distributions. It is also general enough to allow nonlinear state transitions.

Once an analytic approximation to the likelihood function has been obtained via the EKF, one could use a standard numerical optimization routine such as the Newton–Raphson algorithm to find the ML estimate, if one pursues a frequentist approach, or to locate the posterior mode for Bayesian inference. However, the performance of numerical maximization techniques depends crucially on information about derivatives. It can be significantly enhanced by specifying the exact gradient rather than using finite differences which are subject to truncation and subtractive cancellation errors. Hand coding, though, is tedious and error prone. Also, packages such as Mathematica for symbolic differentiation would need to be used with great care to deal with loops and branching and to avoid exponential growth in the expressions produced. Here we suggest using *automatic differentiation* (AD) (Griewank and Corliss, 1991) to generate first-order (and possibly second-order) derivatives of the log-posterior density. AD performs computational derivation, i.e. given a computer code for a real-valued function of say p parameters, $f(\theta_1, \ldots, \theta_p)$, it returns code for evaluating its gradient $\nabla f = \left(\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_p}\right)$. An introduction to AD can be found in Rall and Corliss (1996) and Skaug (2002), details are given in Griewank (2000). The webpage of the Argonne National Laboratory Computational Differentiation Project (`http://www-unix.mcs.anl.gov/autodiff/index.html`) provides an overview of available software tools in FORTRAN and C for AD. Its main applications have been for function maximization in disciplines such as Engineering, Computer Science and Chemistry. Only very few statistical applications exist, where AD has been employed to

find the maximum of the likelihood function (Hovland *et al.* (1997) and Skaug (2002), in the context of random effects models). Here, we suggest the use of AD to find the posterior mode and to design an efficient Metropolis–Hastings (MH) algorithm for Bayesian posterior computation. It is well known that the mixing behaviour of MH algorithms depends upon a judicious choice of the proposal density which should be close to the target density. We construct a multivariate Gaussian proposal centred at the current state and with covariance matrix equal to the inverse of the Hessian matrix of the log-posterior evaluated at the posterior mode by using the quasi-Newton algorithm combined with AD to generate exact first-order derivatives of the log-posterior density. To this end, we make use of the software AD Model Builder (Fournier, 2000) (`http://otter-rsch.com/admodel.htm`), a C++ software package that integrates AD and a quasi-Newton algorithm for function minimization. This yields an extremely flexible, effective and user-friendly MCMC technique for SV models referred to as EKF–AD in what follows.

The outline of the paper is as follows. In Section 2 we describe the theory underlying the calculation of the likelihood function via extended Kalman filtering and Laplace approximation. The MCMC technique to sample from the joint posterior distribution of the parameters is detailed in Section 3. In Section 4 we illustrate the new technique using a time series of daily pound/dollar exchange rates $\{x_t\}$ from 01/10/81 to 28/6/85. Its performance is compared to that of the single-update Gibbs sampler and the integration sampler of Kim *et al.* (1998). Performance is measured in terms of CPU time, integrated autocorrelation time (IACT), and a variety of other diagnostic measures. We conclude in Section 5 with a discussion on the efficiency of this approach.

2. EXTENDED KALMAN FILTERING FOR NONLINEAR STATE-SPACE MODELS

For illustrative and comparative purposes, we use a dataset that has been previously analyzed by Harvey *et al.* (1994) and more recently by Shephard and Pitt (1997), Kim *et al.* (1998) and Meyer and Yu (2000) using Gibbs sampling and by Durbin and Koopman (2000) using a maximum likelihood as well as a Bayesian approach via importance sampling. The data consist of a time series of daily pound/dollar exchange rates $\{x_t\}$ from 01/10/81 to 28/6/85. The series of interest are the daily mean-corrected returns, $\{y_t\}$, given by the transformation $y_t = \log x_t - \log x_{t-1} - \frac{1}{n}\sum_{i=1}^{n}(\log x_t - \log x_{t-1})$, $t = 1, \ldots, n$. The SV model specifies the *conditional* distributions of the observations *given* unknown states, i.e. the underlying latent volatilities, h_t, in the observation equations

$$y_t \mid h_t = \exp(\tfrac{1}{2}h_t)u_t, \qquad u_t \overset{i.i.d.}{\sim} N(0,1), \qquad t = 1, \ldots, n. \tag{2.1}$$

The unknown states are assumed to follow a Markovian transition over time given by the state equations

$$h_t \mid h_{t-1}, \mu, \phi, \tau^2 = \mu + \phi(h_{t-1} - \mu) + v_t, \qquad v_t \overset{i.i.d.}{\sim} N(0, \tau^2), \qquad t = 2, \ldots, n, \tag{2.2}$$

with $h_1 \sim N(\mu, \frac{\tau^2}{1-\phi^2})$. The state h_t determines the amount of volatility on day t and the value of ϕ, $-1 < \phi < 1$, measures the autocorrelation present in the logged squared data. Thus ϕ can be interpreted as the persistence in the volatility, the constant scaling factor $\beta = \exp(\mu/2)$ as the modal volatility, and τ as the volatility of log-volatilities (cf. Kim *et al.* (1998)). Here, the state transition from h_{t-1} to h_t, $f(h_{t-1}, \boldsymbol{\theta}) = \mu + \phi(h_{t-1} - \mu)$, is linear but can easily be generalized

to a nonlinear function f. The formulae given below are expressed in general terms for nonlinear state transition functions f and not necessarily Gaussian observation and state distributions.

The focus is on estimating the unknown parameters $\boldsymbol{\theta} = (\mu, \phi, \tau^2)$ given the observations y_t, $t = 1, \ldots, n$. A fully Bayesian approach specifies the joint distribution of all $n = 945$ observables ($\mathbf{y} = (y_1, \ldots, y_n)$) and $p = 3$ parameters ($\boldsymbol{\theta} = (\mu, \phi, \tau^2)$). The joint probability density function (PDF) $p(\mathbf{y}, \boldsymbol{\theta})$ can be factorized into the product of the PDF of parameters, $p(\boldsymbol{\theta})$, referred to as the *prior* PDF, and the conditional PDF of the observations given the parameters, $p(\mathbf{y} \mid \boldsymbol{\theta})$, referred to as the *likelihood*, i.e. $p(\boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$. After observing the data, the prior knowledge about the parameters, as quantified through the *prior* PDF of $\boldsymbol{\theta}$, is updated to the *posterior* PDF, $p(\boldsymbol{\theta} \mid \mathbf{y})$, via the Bayesian theorem

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{2.3}$$

where $p(\mathbf{y}) = \int p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal PDF of $\mathbf{y}$. Due to the conditioning on unobserved states in a state-space model, the likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ is not available in closed form but requires n-dimensional integration over the state vector $\mathbf{h} = (h_1, \ldots, h_n)$ as

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{h} \mid \boldsymbol{\theta})d\mathbf{h} = \int p(\mathbf{y} \mid \mathbf{h}, \boldsymbol{\theta})p(\mathbf{h} \mid \boldsymbol{\theta})d\mathbf{h}. \tag{2.4}$$

Taking the temporal structure of the observations into account, we can factorize the likelihood by successive conditioning into

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = p(y_1 \mid \boldsymbol{\theta}) \prod_{t=2}^{n} p(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\theta}), \tag{2.5}$$

where $\mathbf{y}_{t-1} = (y_1, \ldots, y_{t-1})$ collects all the observable information obtained until time $t - 1$. Thus, the n-dimensional integration in equation (2.4) can be reduced to n successive one-dimensional integrations, starting with

$$p(y_1 \mid \boldsymbol{\theta}) = \int p(y_1 \mid h_1, \boldsymbol{\theta})p(h_1 \mid \boldsymbol{\theta})dh_1, \tag{2.6}$$

and subsequently, for $t = 2, \ldots, n$

$$p(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\theta}) = \int p(y_t \mid h_t, \boldsymbol{\theta})p(h_t \mid \mathbf{y}_{t-1}, \boldsymbol{\theta})dh_t. \tag{2.7}$$

This also implies that the data can be processed in a single sweep, updating knowledge about states as we receive more information. We will show now how to approximate the likelihood in equation (2.5) by successively applying the Laplace approximation (Gamerman, 1997) to the integrals in equations (2.6) and (2.7).

For instance, in the light of just the first observation y_1, we update the prior $p(h_1 \mid \boldsymbol{\theta})$ of the unknown state h_1 to the *filtering* PDF via Bayes theorem

$$p(h_1 \mid y_1, \boldsymbol{\theta}) = \frac{p(y_1 \mid h_1, \boldsymbol{\theta})p(h_1 \mid \boldsymbol{\theta})}{p(y_1 \mid \boldsymbol{\theta})}, \tag{2.8}$$

where the denominator is just the first factor in the likelihood decomposition in equation (2.5), given in equation (2.6). If both likelihood and prior, $p(y_1 \mid h_1, \boldsymbol{\theta})$ and $p(h_1 \mid \boldsymbol{\theta})$ respectively,

were Gaussian and linear in h_1, the posterior would again be Gaussian and the normalization constant in the denominator easy to determine. However, $p(y_1 \mid h_1, \boldsymbol{\theta})$ is not linear in h_1. Nevertheless we can approximate the denominator by the so-called *Laplace approximation*, an asymptotic approximation of the posterior distribution that dates back to the work of Laplace in the 18th century (Laplace, 1986). For the integral in equation (2.6) this yields

$$p(y_1 \mid \boldsymbol{\theta}) \approx \sqrt{2\pi}\,\mathrm{e}^{-\psi_1(y_1,\hat{h}_1,\boldsymbol{\theta})}|D^2\psi_1(y_1,\hat{h}_1,\boldsymbol{\theta})|^{-1/2}, \tag{2.9}$$

where

$$\begin{aligned}\psi_1(y_1,h_1,\boldsymbol{\theta}) &= -\log[p(y_1 \mid h_1,\boldsymbol{\theta})p(h_1 \mid \boldsymbol{\theta})] \\ &= \frac{1}{2}\log(2\pi\,\mathrm{e}^{h_1}) + \frac{1}{2}\mathrm{e}^{-h_1}y_1^2 + \frac{1}{2}\log\left(\frac{2\pi\tau^2}{1-\phi^2}\right) + \frac{1-\phi^2}{2\tau^2}(h_1-\mu)^2,\end{aligned}$$

$D^2\psi_1(y_1,h_1,\boldsymbol{\theta})$ denotes the second-order derivative of the function $\psi_1(y_1,h_1,\boldsymbol{\theta})$ with respect to h_1, $\hat{h}_1 = \mathrm{argmin}_{h_1}\psi_1(y_1,h_1,\boldsymbol{\theta})$, and $\hat{\sigma}_1^2 = |D^2\psi_1(y_1,\hat{h}_1,\boldsymbol{\theta})|^{-1}$. This is easily seen by a second-order Taylor series expansion of $\psi_1(y_1,h_1,\boldsymbol{\theta})$ at $\hat{h}_1$.

We now learn about a state at time t, successively for $t = 2,\ldots,n$, given contemporaneously available information. This is done repeatedly in a two-stage procedure by on-line extended Kalman filtering. In the first stage of the EKF, after observing $\mathbf{y}_{t-1}$ but before observing y_t, the *predictive* PDF of $h_t \mid \mathbf{y}_{t-1},\boldsymbol{\theta}$ is approximated by a Normal PDF $\tilde{p}(h_t \mid \mathbf{y}_{t-1},\boldsymbol{\theta})$ with mean and variance given by

$$\beta_t = f(\hat{h}_{t-1},\boldsymbol{\theta}) = \mu + \phi(\hat{h}_{t-1}-\mu) \tag{2.10}$$

and

$$\gamma_t^2 = [f'(\hat{h}_{t-1},\boldsymbol{\theta})]^2\hat{\sigma}_{t-1}^2 + \tau^2 = \phi^2\hat{\sigma}_{t-1}^2 + \tau^2, \tag{2.11}$$

respectively, using a first-order Taylor series expansion of $f(h_{t-1},\boldsymbol{\theta})$ at the approximate mean $\hat{h}_{t-1}$ of $h_{t-1} \mid \mathbf{y}_{t-1},\boldsymbol{\theta}$. Here, $f'(h,\boldsymbol{\theta})$ denotes the first derivative of $f(h,\boldsymbol{\theta})$ with respect to h. In the second stage, after observing y_t, the *filtering* PDF $p(h_t \mid \mathbf{y}_t,\boldsymbol{\theta})$ is updated via Bayes theorem to

$$p(h_t \mid \mathbf{y}_t,\boldsymbol{\theta}) \propto p(y_t \mid h_t,\boldsymbol{\theta})p(h_t \mid \mathbf{y}_{t-1},\boldsymbol{\theta}) \approx p(y_t \mid h_t,\boldsymbol{\theta})\tilde{p}(h_t \mid \mathbf{y}_{t-1},\boldsymbol{\theta}). \tag{2.12}$$

Using the Laplace approximation then yields an approximation to the tth likelihood contribution in equation (2.7)

$$\begin{aligned}p(y_t \mid \mathbf{y}_{t-1},\boldsymbol{\theta}) &= \int p(y_t \mid h_t,\boldsymbol{\theta})p(h_t \mid \mathbf{y}_{t-1},\boldsymbol{\theta})dh_t \\ &\approx \int p(y_t \mid h_t,\boldsymbol{\theta})\tilde{p}(h_t \mid \mathbf{y}_{t-1},\boldsymbol{\theta})dh_t \\ &= \sqrt{2\pi}\,\mathrm{e}^{-\psi_t(y_t,\hat{h}_t,\boldsymbol{\theta})}|D^2\psi_t(y_t,\hat{h}_t,\boldsymbol{\theta})|^{-1/2},\end{aligned} \tag{2.13}$$

where

$$\begin{aligned}\psi_t(y_t,h_t,\boldsymbol{\theta}) &= -\log(p(y_t \mid h_t,\boldsymbol{\theta})\tilde{p}(h_t \mid \mathbf{y}_{t-1},\boldsymbol{\theta})) \\ &= \frac{1}{2}\log(2\pi\,\mathrm{e}^{h_t}) + \frac{1}{2}\mathrm{e}^{-h_t}y_t^2 + \frac{1}{2}\log(2\pi\gamma_t^2) + \frac{1}{2\gamma_t^2}(h_t-\beta_t)^2,\end{aligned}$$

$\hat{h}_t = \text{argmin}_{h_t} \psi_t(y_t, h_t, \boldsymbol{\theta})$ and $\hat{\sigma}_t^2 = |D^2 \psi_t(y_t, \hat{h}_t, \boldsymbol{\theta})|^{-1}$. We use the Newton–Raphson algorithm to derive $\hat{h}_t$ and find that a few iterations usually suffice to get close enough to the minimum.

Completion of this sequential two-stage procedure yields a closed-form approximative expression for the likelihood of equation (2.5) that no longer depends on the latent states $\mathbf{h}$. More precisely, this likelihood is given by

$$\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}) = \exp\left\{ -\frac{1}{2}\sum_{t=1}^{n} \log(2\pi e^{\hat{h}_t^2}) - \sum_{t=1}^{n} \frac{1}{2e^{\hat{h}_t}} y_t^2 - \frac{1}{2}\sum_{t=1}^{n} \log(2\pi\gamma_t^2) - \sum_{t=1}^{n} \frac{1}{2\gamma_t^2}(\hat{h}_t - \beta_t)^2 - \frac{1}{2}\sum_{t=1}^{n} \log\left(\frac{1}{\hat{\sigma}_t^2}\right) \right\}, \tag{2.14}$$

with $\beta_1 = \mu$ and $\gamma_1^2 = \frac{\tau^2}{1-\phi^2}$. The reader should note that this approximation is exact for linear Gaussian state-space models. From equation (2.3) we then obtain the posterior PDF up to a normalization constant

$$\tilde{p}(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\boldsymbol{\theta})\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}). \tag{2.15}$$

Due to marginalization over the latent volatilities, EKF–AD can provide inference for the model parameters but not for the state variables. However, it is possible to approximate the likelihood function $p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{h})$ by making use of the 'particle filter' method (Kitagawa, 1996) or by more efficient techniques such as the auxiliary particle filter (Pitt and Shephard, 1999a). In the context of SV models, this particular algorithm has been implemented in Kim *et al.* (1998), Pitt and Shephard (1999b), and Chib *et al.* (2002). Furthermore, due to iterated Laplace approximations, the inference is based on an approximation $\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})$ (equation (2.14)) to the exact likelihood, which is exact only for linear Gaussian state-space models. Therefore, posterior inference is not exact, which is reflected in the discrepancy between estimates of the parameter β in the example in Section 4.

3. POSTERIOR INFERENCE USING AUTOMATIC DIFFERENTIATION

Once the (approximate) posterior PDF (2.15) has been obtained, traditional Bayesian posterior inference proceeds by using the Bayesian central limit theorem, i.e. finding the multivariate normal approximation $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by calculating the posterior mode $\boldsymbol{\mu}$ and its approximate variance, the latter by evaluating the Hessian matrix of minus the log-posterior at the posterior mode:

$$\begin{aligned} \phi(\boldsymbol{\theta}) &= -\log[p(\boldsymbol{\theta})\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})], \\ \boldsymbol{\mu} &= \text{argmin}_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}), \\ \boldsymbol{\Sigma} &= D^2\phi(\boldsymbol{\mu})^{-1}. \end{aligned}$$

Thus, we need to find the posterior mode, or alternatively minimize $\phi(\boldsymbol{\theta})$. To this end, we employ the quasi-Newton algorithm and make use of AD (Griewank and Corliss, 1991) to calculate the first-order partial derivatives of $\phi(\boldsymbol{\theta})$ as implemented in the software AD Model Builder (Fournier, 2000). This can be done to the same degree of accuracy as the function evaluation itself. AD Model Builder uses AD implemented in a C++ class library which combines an array language with the reverse mode of AD supplemented with precompiled adjoint code for

the derivatives of common array and matrix operations. This results in an extremely fast and accurate function minimization routine.

With the development of efficient simulation techniques for posterior computation, modern Bayesian inference no longer ends with posterior mode estimation, rather it has become possible to draw a sample from the posterior PDF and use this sample to summarize the posterior distribution. Various techniques are feasible to obtain a sample from the posterior (2.15), e.g. sampling/importance resampling, adaptive versions of importance sampling as in Givens and Raftery (1996), and MCMC algorithms such as Gibbs sampling or the MH algorithm (Smith and Roberts, 1993). In this paper, we use the MH algorithm, developed by Metropolis *et al.* (1953) and generalized by Hastings (1970). The MH algorithm is top of the list of the great algorithms of 20th century scientific computing (Dongarra and Sullivan, 2000). It is a MCMC method which means that it generates a Markov chain whose equilibrium distribution is just the target posterior distribution. The MH algorithm shares the concept of a generating PDF with the well-known simulation technique of *rejection sampling* (Gamerman, 1997). However, the *candidate generating PDF* $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_c)$, $\int q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_c)d\boldsymbol{\theta} = 1$, can now depend on the current state $\boldsymbol{\theta}_c$ of the sampling process. A new candidate $\boldsymbol{\theta}^*$ is accepted with a certain *acceptance probability* $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_c)$, also depending on the current state $\boldsymbol{\theta}_c$, and chosen such that the transition probability $p(\boldsymbol{\theta}_c, \boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_c)\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_c)$ satisfies detailed balance. This is met by setting

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_c) = \min\left\{\frac{\tilde{p}(\boldsymbol{\theta}^* \mid \mathbf{y})q(\boldsymbol{\theta}_c \mid \boldsymbol{\theta}^*)}{\tilde{p}(\boldsymbol{\theta}_c \mid \mathbf{y})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_c)}, 1\right\}.$$

The outcomes from the MH algorithm can be regarded as a sample from the invariant PDF only after a certain 'burn-in' period. A menu-driven collection of SPLUS functions, CODA (Best *et al.*, 1995), is available for analysing the samples obtained from MCMC. CODA can be downloaded from `http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`. Besides trace plots and convergence diagnostics based on Cowles and Carlin (1996), CODA calculates statistical summaries of the posterior distributions and kernel density estimates.

The efficiency of the MH algorithm depends crucially on the choice of the proposal PDF. Similar to rejection sampling, the efficiency can be improved by choosing a proposal that is 'close' to the posterior PDF. Having already obtained the Laplace approximation to the posterior PDF via AD, it is obvious to make use of this approximation to $\tilde{p}(\boldsymbol{\theta} \mid \mathbf{y})$ to determine a good proposal PDF. Thus we use a multivariate normal PDF with mean equal to the current state $\boldsymbol{\theta}^{(i)}$ and covariance matrix $\boldsymbol{\Sigma}$ equal to the inverse of the Hessian matrix of minus the log-posterior evaluated at the posterior mode. The covariance matrix is dynamically scaled until a reasonable acceptance rate in the MH algorithm is observed.

But the reader should note that in order to obtain a sample from the posterior, it is not absolutely necessary to employ AD. Once the approximate posterior PDF (2.15) has been obtained by EKF, any common SIR or MCMC technique can be used.

4. APPLICATION

We implemented the EKF–AD technique using the software package AD Model Builder which integrates AD with a quasi-Newton function minimization algorithm. We fitted the SV model specified in equations (2.1) and (2.2) to the time series of daily pound/dollar exchange rates described in Section 2. We assumed prior independence of the parameters μ, ϕ and τ^2, and

used the same priors as in Kim *et al.* (1998). We employed a slightly informative prior for μ, $\mu \sim N(0, 10)$. We set $\phi = 2\phi^* - 1$ and specified a Beta(α, β) prior for ϕ^* with $\alpha = 20$ and $\beta = 1.5$, which gives a prior mean for ϕ of 0.86. A conjugate inverse-gamma prior is chosen for τ^2, i.e. $\tau^2 \sim IG(2.5, 0.025)$, with a prior mean of 0.0167 and prior standard deviation of 0.0236.

Using the EKF followed by AD and the quasi-Newton algorithm, we obtained posterior modes and asymptotic standard deviations (in brackets) for ϕ, τ and β of 0.9830 (0.009,09), 0.1445 (0.030,72), and 0.7932 (0.126,43), respectively.

We then used the MH algorithm as described above (EKF–AD) to obtain samples from the posterior distribution. We also implemented an alternative random walk MH algorithm (EKF–RW) that does not require AD. In EKF–RW, after a pilot run of 10,000 iterations using a multivariate normal proposal with mean equal to the current state and a diagonal covariance matrix with rough variance estimates for each parameter obtained from experimental runs, we performed 100,000 iterations using a multivariate normal proposal with the empirical covariance matrix from the pilot run as proposal covariance. The covariance matrix was scaled using the optimal scaling constant $2.4/\sqrt{d}$ with $d = 3$ of Gelman *et al.* (1995).

For comparison purposes, we also used the single-update Gibbs sampler and the integration sampler as implemented in the software package SVPack, (a freeware dynamic link library for the Ox programming language (Doornik, 1996) discussed by Kim *et al.* (1998)) and state results for the transformed parameters β, ϕ and τ. For each MCMC algorithm, we performed 110,000 iterations, discarded the first 10,000 iterations as burn-in period and thinned the output by taking every 10th sample. This resulted in final chain lengths of 10,000. Table 1, gives an overview of the parameter estimates for ϕ, τ and β obtained by the four different samplers and compares the CPU times (all computations were performed on a Pentium III PC) and the IACT. The IACT (Sokal, 1996), also referred to as the 'inefficiency factor' by Kim *et al.* (1998), gives a measure of the efficiency of the MCMC sampler. As the estimate of the posterior mean of a parameter x is the average of N correlated samples from a Markov chain, its variance is a factor of IACT larger than the variance of the sample mean based on the same number of independent samples, i.e.

$$\text{var}(\bar{x}_{\text{MC}}) = \text{IACT} \cdot \frac{\text{var}(x)}{N}.$$

Hence, IACT is the number of correlated samples with the same variance-reducing power as one independent sample. The estimate of IACT is obtained by multiplying the square of the MC standard error by the number of MCMC iterations N and dividing by the square of the standard deviation of the parameter. The MC standard errors in Table 1 are calculated by Geweke's (1992) method.

The discrepancy in the estimates of the parameter β between EKF–AD/RW and the single-update Gibbs sampler/integration sampler is due to the approximation error incurred by iterated Laplace approximations. However, the approximation affects only the constant scaling factor μ, respectively β, but not the estimates of the persistence of the volatility ϕ nor the volatility of log-volatilities τ. The IACT's of the EKF–AD (and EKF–RW) show a substantial decrease over those of the single-update Gibbs sampler, thus yielding a much improved MCMC technique. Its mixing properties come close to those of the integration sampler with a third of its computation time. However, one should note that the implementation of the integration sampler in SVPack did not use optimised code but rather off the shelf SsfPack algorithms to carry out likelihood evaluations and simulation smoothing. Thus, a speed-up of the integration sampler by about a factor of 3 may well be feasible.

Table 1. Comparison of Bayesian estimates, IACT's and CPU times (in seconds) using the single-update Gibbs sampler, EKF–AD, EKF–RW, and the integration sampler of SVPack as in Kim *et al.* (1998).

	Parameter	Mean	SD	MC SE	IACT
Single-update Gibbs	ϕ	0.9775	0.0106	0.00031	86
	τ	0.1598	0.0304	0.00126	172
	β	0.6496	0.0990	0.00126	16
	Time (s)		249		
EKF–AD	ϕ	0.9797	0.0105	0.00018	28
	τ	0.1596	0.0365	0.00068	35
	β	0.8127	0.1414	0.00198	19
	Time (s)		818		
EKF–RW	ϕ	0.9761	0.0112	0.00016	20
	τ	0.1756	0.0352	0.00048	19
	β	0.7995	0.1315	0.00168	16
	Time (s)		832		
Integration sampler	ϕ	0.9775	0.0105	0.00012	12
	τ	0.1575	0.0313	0.00040	16
	β	0.6517	0.0999	0.00105	11
	Time (s)		2311		

The computation times and IACTs of EKF–AD and EKF–RW are comparable. One should note, though, that the specification of the proposal covariance matrix in the pilot run of EKF–RW required several preliminary experimental runs to get reasonable variance estimates of each parameter. EKF–AD, on the other hand, yields the posterior modes and asymptotic standard errors of parameter estimates. Moreover, its implementation of the MH algorithm makes efficient use of the availability of the posterior mode and the Hessian at the mode for an automated construction of a proposal density that is close to the target density.

Figure 1 shows trace and kernel density estimates of the marginal posterior distribution of the model parameters obtained by EKF–AD. These can be compared to Figures 2 and 5 of Kim *et al.* (1998) for the corresponding trace and kernel density plots of the single-update Gibbs sampler and the integration sampler, respectively. Extensive convergence diagnostics based on Cowles and Carlin (1996) were calculated for this chain using the CODA software of Best *et al.* (1995), a menu-driven collection of SPLUS functions. All parameters passed the Heidelberger and Welch stationarity and halfwidth test. Geweke's Z-scores for ϕ, τ, and β were reasonably low (-0.99, 1.54, and -0.56, respectively).

It should be pointed out that EKF–AD in conjunction with AD can be readily used to obtain ML estimates and their asymptotic standard deviations. This is easily done by choosing a flat prior, i.e. setting $p(\boldsymbol{\theta}) \propto 1$, and using AD and a quasi-Newton algorithm to find the value of $\boldsymbol{\theta}$ where $\phi(\boldsymbol{\theta}) = -\log \tilde{p}(y \mid \boldsymbol{\theta})$ is minimized. The ML-estimates thus obtained for ϕ, τ, and β and their standard deviations (in brackets) are 0.9692 (0.013,95), 0.1935 (0.041,85) and 0.7029 (0.086,07), respectively. This yields a fast technique for ML estimation in SV models and an efficient alternative to the quasi-ML approach proposed by Harvey *et al.* (1994), to the approximate ML approach of Fridman and Harris (1998) that uses recursive numerical integration to calculate

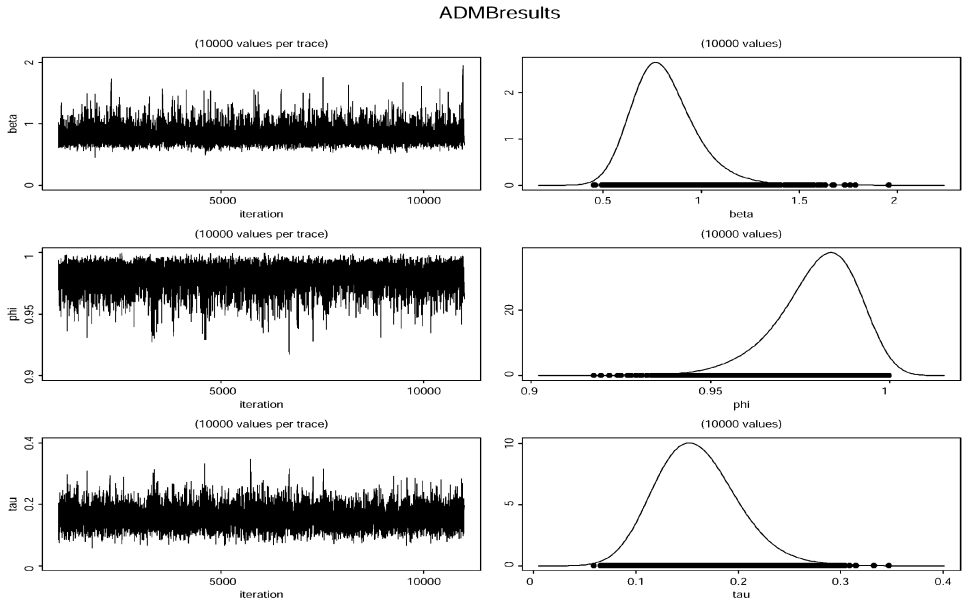

Figure 1. Trace plots and kernel density estimates of the marginal posterior distribution of the model parameters β, ϕ, τ obtained by EKF–AD.

the marginal likelihood, to the Monte Carlo ML approach of Sandmann and Koopman (1998), and to the ML approach proposed by Durbin and Koopman (2000) that employs Monte Carlo simulation using Gaussian importance sampling and antithetic variables.

5. DISCUSSION

Various techniques have been suggested to improve mixing of MCMC algorithms for posterior computation in SV models. In this paper, we introduced a new technique that combines three well-known mathematical and statistical techniques: the EKF, the Laplace approximation, and AD. The EKF combined with the Laplace approximation is used to integrate out the latent volatilities and thus obtain a close approximation to the marginal likelihood of the parameters. This marginal likelihood is needed to find either the maximum likelihood estimate, the posterior mode or other Bayesian posterior estimates. To this end, we use a quasi-Newton function minimization algorithm, employing AD for evaluating first- and second-order partial derivatives which significantly enhances numerical optimization techniques. The use of AD over the classical way of calculating numerical derivatives by 'finite differences' is preferable because AD-derivatives are exact to machine precision and calculated efficiently. Furthermore, we exploit the availability of the posterior mode together with the Hessian matrix at the mode to design a suitable proposal density for the MH algorithm to sample from the posterior distribution of the parameters. This yields a very efficient MCMC technique with good mixing properties as documented by IACT comparable to those of the integration sampler and with a higher computational speed. Alternative MCMC techniques, such as the Gibbs sampler or random walk MH algorithms, that do not make use of AD can be employed as well.

This approach is applicable to the general class of nonlinear non-Gaussian state space models to which the SV models belong. Although the state transition is linear in the basic SV model, the algorithm described here allows for nonlinear state transitions. The error distributions are not restricted to the Gaussian family. To reduce the effect of outlying observations on the posterior modes of the latent variables, for instance, one could use a mixture of a normal and a Cauchy (or other heavy-tailed) distribution for modelling the observation errors conditional on the latent variables.

Even though this technique was presented using a one-dimensional time series and underlying latent volatility process, it lends itself to a straightforward extension to multivariate SV models. With the increased number of parameters in multivariate models, the advantages of the reverse mode of AD should manifest themselves even more clearly in reduced computation times.

ACKNOWLEDGEMENTS

This work was supported by the Royal Society of New Zealand Marsden Fund and the University of Auckland Research Committee.

REFERENCES

Best, N. G., M. K. Cowles and S. K. Vines (1995). *CODA Manual Version 0.30*. New York: MRC Biostatistics Unit.

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

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

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

Chib, S., F. Nardari and N. Shephard (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics 108*, 281–316.

Cowles, M. K. and B. P. Carlin (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of American Statistical Association 91*, 883–905.

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

Dongarra, J. and F. Sullivan (2000). The top 10 algorithms. *Comput. Sci. Eng. 2*.

Doornik, J. A. (1996). *Ox: Object Oriented Matrix Programming, 1.10*. London: Chapman & Hall.

Durbin, J. and S. J. Koopman (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *Journal of the Royal Statistical Society, Series B 62*, 3–56.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

Fournier, D. (2000). *AD Model Builder Version 5.0.1*. Canada: Otter Research Ltd.

Fridman, M. and L. Harris (1998). A maximum likelihood approach for non-Gaussian stochastic volatility models. *Journal of Business and Economic Statistics 16*, 284–91.

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

Gamerman, D. (1997). *Markov Chain Monte Carlo, Stochastic Simulation for Bayesian Inference*. London: Chapman & Hall.

Gelman, A., G. O. Roberts and W. R. Gilks (1995). Efficient metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 5*, pp. 559–607. Oxford: Oxford University Press.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 4*, pp. 169–93. Oxford: Oxford University Press.

Givens, G. H. and A. E. Raftery (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association 91*, 132–41.

Griewank, A. (2000). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Philadelphia: SIAM.

Griewank, A. and G. F. Corliss (1991). *Automatic Differentiation of Algorithms: Theory, Implementation and Application*. Philadelphia: SIAM.

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

Harvey, A. C., E. Ruiz and N. Shephard (1994). Multivariate stochastic variance models. *Review of Economic Studies 61*, 247–64.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Hovland, P., C. Bischof, D. Spiegelman and M. Casella (1997). Efficient derivative codes through automatic differentiation and interface contraction: an application in Biostatistics. *SIAM Journal on Scientific and Statistical Computing 18*, 1056–66.

Jacquier, E., N. G. Polson and P. E. Rossi (1994). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics 12*, 371–89.

Kim, S., N. Shephard and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies 65*, 361–93.

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

Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statististical Science 1*, 364–78.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics 21*, 1087–91.

Meyer, R. and J. Yu (2000). BUGS for a Bayesian analysis of stochastic volatility models. *The Econometrics Journal 3*, 198–215.

Pitt, M. and N. Shephard (1999a). Filtering via simulation: auxiliary particle filter. *The Journal of the American Statistical Association 94*, 590–9.

Pitt, M. and N. Shephard (1999b). Time varying covariances: a factor stochastic volatility approach (with discussion). In J. M. Berardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 6*, pp. 547–70. Oxford: Oxford University Press.

Rall, L. B. and G. F. Corliss (1996). An introduction to automatic differentiation. In M. Berz, C. Bischof, G. Corliss and A. Griewank (eds), *Computational Differentiation, Techniques, Applications, and Tools*, pp. 1–18. Philadelphia, PA: SIAM.

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

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

Skaug, H. J. (2002). Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. *Journal of Computational and Graphical Statistics 11*, 458–70.

Smith, A. F. M. and G. O. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B 55*, 3–24.

Sokal, A. D. (1996). Monte Carlo methods in statistical mechanics: foundations and new algorithms. Lectures at the Cargese Summer School on "Functional Integration: Basics and Applications".

Stroud, J. R., P. Mueller and N. G. Polson (2001). *Nonlinear State-Space Models with State-Dependent Variances*. Technical Report, Department of Statistics, University of Chicago. `http://galton.uchicago.edu/~stroud/`.

Tauchen, G. and M. Pitts (1983). The price variability–volume relationship on speculative markets. *Econometrica 51*, 485–505.

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes—a study of the daily sugar prices 1961–75. In O. D. Anderson (ed.), *Time Series Analysis: Theory and Practice 1*, pp. 203–26. Amsterdam: North-Holland.