

Simulated Maximum Likelihood in Nonlinear Continuous-Discrete State Space Models: Importance Sampling by Approximate Smoothing

Hermann Singer

FernUniversität Hagen, Lehrstuhl für angewandte Statistik und Methoden der empirischen Sozialforschung, D-58084 Hagen, Germany

Summary

The likelihood function of a continuous-discrete state space model is computed recursively by Monte Carlo integration, using importance sampling techniques. A functional integral representation of the transition density is utilized and importance densities are obtained by smoothing. Examples are the likelihood surfaces of an AR(2) process, a Ginzburg-Landau model and stock price models with stochastic volatilities.

Keywords: Stochastic differential equations; Nonlinear filtering; Discrete noisy measurements; Maximum likelihood estimation; Monte Carlo simulation; Importance sampling.

1 Introduction

The continuous-discrete state space model is a convenient specification for the dynamic modeling of quantitative variables in continuous time, which are subject to random disturbances both in the dynamics and in the discrete process of observation (Jazwinski, 1970). In order to meet requirements of empirical data analysis, the dynamics of the state vector is given in continuous time (system equation), whereas the measurements are assumed to be given only at discrete, possibly unequally spaced time points (measurement model). Moreover, analogously to structural equation models (SEM) or factor analysis, the state is only incompletely observable and subject to measurement error.

In the linear case with gaussian random errors, the system can be estimated efficiently by maximum likelihood (ML) using the Kalman filter algorithm (cf. Jones, 1984, Harvey and Stock, 1985, Jones and Tryon, 1987, Zadrozny, 1988, Jones and Ackerson, 1990, Singer, 1993, 1995), but in nonlinear systems complicated equations for the transition density arise, which must be solved numerically. One approach to do this is the Monte Carlo simulation method, where many sample trajectories are simulated and unknown probability densities and integrals are estimated from these data. In order to keep close contact to the linear Kalman filter, a sequence of time and measurement updates (continuous-discrete filter) is utilized, and the resulting integral expressions (expectation values) are approximated by statistical averages. To reduce the simulation error of the likelihood function, importance sampling and other variance reduction techniques (such as antithetical sampling) are used.

Simulation based filtering methods in discrete time have been used in the literature such as Markov chain Monte Carlo (MCMC; Carlin et al., 1992, Kim et al., 1998), rejection sampling using density estimators (Tanizaki, 1996, Tanizaki and Mariano, 1995, Hürzeler and Künsch, 1998), importance sampling and antithetic variables (Durbin and Koopman, 1997, 2000) and recursive bootstrap resampling (Gordon et al., 1993, Kitagawa, 1996).

In this paper the time update is generalized to the continuous time case by using the Chapman-Kolmogorov equation and importance sampling is implemented by approximate smoothing in order to reduce the variance of the simulated likelihood function. For this purpose the Gaussian sum filter of Alspach and Sorenson (1972) is used. In linear systems, the smoothing is exact and the simulation error of the likelihood estimate is zero, given the data (cf. section 7.1).

Section 2 defines the continuous-discrete state space model and section 3 presents the recursive computation of the filter densities and the likelihood. Sections 4 and 5 derive the variance reduction and its implementation by

smoothing, whereas sections 6 and 7 discuss practical issues and present 3 examples.

2 Nonlinear continuous-discrete state space models

We discuss the nonlinear *continuous-discrete state space model* (Jazwinski, 1970)

$$dy(t) = f(y(t), t, \psi)dt + g(y(t), t, \psi)dW(t) \quad (1)$$

$$z_i = h(y(t_i), t_i, \psi) + \epsilon_i, \quad (2)$$

where discrete time measurements z_i are taken at times $\{t_0, t_1, \dots, t_T\}$, $t_0 \leq t \leq t_T$. In state equation (1), the process error $W(t)$ is a r -dimensional Wiener process and the state is described by the p -dimensional state vector $y(t)$. It fulfils a system of *stochastic differential equations* (SDE) in the sense of Itô (cf. Arnold, 1974) with random initial condition $y(t_0) \sim p_0(y, t_0)$ (prior distribution). The functions $f : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^u \rightarrow \mathbb{R}^p$ and $g : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^u \rightarrow \mathbb{R}^p \times \mathbb{R}^r$ are called drift and diffusion coefficients, respectively. In measurement equation (2), $\epsilon_i \sim N(0, R(t_i, \psi))$ is a k -dimensional discrete time white noise process (measurement error) and $h : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^u \rightarrow \mathbb{R}^k$ is the output function. It is assumed that the error processes $dW(t)$, ϵ_i and the initial state $y(t_0)$ are mutually independent. Parametric estimation is based on the u -dimensional parameter vector ψ . The key quantity for the computation of the likelihood function is the transition probability $p(y, t|x, s)$ between states y and x at times t and s , respectively, which is a solution of the *Fokker-Planck equation*

$$\begin{aligned} \frac{\partial p(y, t|x, s)}{\partial t} &= - \sum_i \frac{\partial}{\partial y_i} [f_i(y, t, \psi)p(y, t|x, s)] \\ &\quad + \frac{1}{2} \sum_{ij} \frac{\partial^2}{\partial y_i \partial y_j} [\Omega_{ij}(y, t, \psi)p(y, t|x, s)] \\ &:= F(y, t, \psi)p(y, t|x, s) \end{aligned} \quad (3)$$

subject to the initial condition $p(y, s|x, s) = \delta(y - x)$ (Dirac delta function). The symbol $F(y, t, \psi)$ denotes the Fokker-Planck operator. The diffusion matrix is given by $\Omega = gg' : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^u \rightarrow \mathbb{R}^p \times \mathbb{R}^p$. Under certain technical conditions the solution of (3) is the conditional density of $y(t)$ given $y(s) = x$ (see, e.g. Wong and Hajek, 1985, ch. 4).

In order to model exogenous influences, f , g , h and R are assumed to depend on deterministic regressor variables $x(t) : \mathbb{R} \rightarrow \mathbb{R}^q$, i.e. $f(y, t, \psi) =$

$f(y, t, x(t), \psi)$ etc. For notational simplicity, the dependence on $x(t)$ and on ψ will be suppressed.

It may be noted that state space model (1, 2) allows the modeling of ARIMA systems, since unobserved higher order derivatives can be accommodated in an extended state vector $\eta = \{y, \dot{y}, \ddot{y}, \dots\}$.

Furthermore, the functions f , g and h , R may depend on earlier measurements $Z^t = \{z(t_j); t_j \leq t\}$ and $Z^{t_{i-1}} = \{z(t_j); t_j \leq t_{i-1}\}$, respectively, which allows the modeling of (G)ARCH effects ((generalized) autoregressive conditional heteroskedasticity). For example, the diffusion matrix $g(y, t)$ may depend on earlier innovations $\nu_j = z_j - E[z_j | z_{j-1}, \dots, z_0]; t_j \leq t$ and if the functions are linear in the state y , the state space model is conditionally gaussian (cf. Liptser and Shirayev, 1978, vol. II, ch. 13). Again, for notational simplicity, the dependence on the Z^t will be suppressed.

3 Computation of the likelihood function

The *exact* time and measurement updates of the *continuous-discrete filter* are given by the recursive scheme (Jazwinski, 1970) for the a priori and a posteriori densities $p_{i+1|i}, p_{i|i}$:

time update:

$$\begin{aligned} \frac{\partial p(y, t | Z^i)}{\partial t} &= F(y, t) p(y, t | Z^i); t \in [t_i, t_{i+1}] \\ p(y, t_i | Z^i) &:= p(y_i | Z^i) := p_{i|i} \\ p(y, t_{i+1} | Z^i) &:= p(y_{i+1} | Z^i) := p_{i+1|i} \end{aligned} \quad (4)$$

measurement update:

$$\begin{aligned} p(y_{i+1} | Z^{i+1}) &= \frac{p(z_{i+1} | y_{i+1}, Z^i) p(y_{i+1} | Z^i)}{p(z_{i+1} | Z^i)} \\ &:= p_{i+1|i+1} \end{aligned} \quad (5)$$

$$p(z_{i+1} | Z^i) = \int p(z_{i+1} | y_{i+1}, Z^i) p(y_{i+1} | Z^i) dy_{i+1}, \quad (6)$$

$i = 0, \dots, T-1$, where F is the Fokker-Planck operator, $Z^i = \{z(t) | t \leq t_i\}$ are the observations up to time t_i and $L_{i+1} := p(z_{i+1} | Z^i)$ is the *likelihood function* of observation z_{i+1} . The time update describes the time evolution of the conditional density $p(y, t | Z^i)$ given information up to the last measurement and the measurement update is a discontinuous change due to new

information z_{i+1} using the Bayes formula. Thus the likelihood of the complete observation $Z^T = \{z_T, \dots, z_0\}$ can be computed sequentially and new observations z_{T+1} can be processed with only one more update step.

Some remarks may be in order:

1. In the linear case, the conditional densities are gaussian and the recursive steps can be implemented for the conditional moments $y(t|t_i) = E[y(t)|Z^i]$ and $P(t|t_i) = \text{Var}[y(t)|Z^i]$. Instead of the Fokker-Planck equation, only linear ordinary differential equations must be solved. Furthermore, the measurement update can be computed analytically since all involved quantities are jointly (conditionally) gaussian. This is the celebrated *Kalman filter* algorithm extensively used in engineering, control theory, statistics, economics and the social sciences (cf. Jazwinski, 1970, Gelb, 1974, Liptser and Shirayev, 1977, 1978, Harvey, 1989, Fahrmeir and Kaufmann, 1991, Singer, 1993).
2. In the general nonlinear case, the time and measurement updates require the solution of partial differential equations and integrals (likelihood function) which can be obtained only numerically by several approximation methods. Linearizing the system one obtains the extended Kalman filter (EKF), but more elaborate methods such as the second order nonlinear filter (SNF; Jazwinski, 1970), the gaussian sum filter (Alspach and Sorenson, 1972), numerical integration (Kitagawa, 1987), and simulation methods have been used (Kitagawa, 1996, Tanizaki, 1996, Singer, 1997, Kim et al. 1998, Hürzeler and Künsch, 1998). Usually, the filters are formulated in discrete time, however.

In order to compute the solution of Fokker-Planck equation (4), a Monte Carlo approach was utilized (Wagner, 1988, Kloeden and Platen, 1992). We use an integral representation based on the Chapman-Kolmogorov-equation for Markov processes

$$p(y_{i+1}|y_i) = \int p(y_{i+1}|\eta)p(\eta|y_i)d\eta, \quad (7)$$

which can be iterated to express

$$p(y_{i+1}|y_i) = \int p(y_{i+1}|\eta_{J_i-1})p(\eta_{J_i-1}|\eta_{J_i-2}) \dots p(\eta_1|y_i)d\eta_{J_i-1} \dots d\eta_1 \quad (8)$$

as a product of transition densities inserted into the interval $[t_i, t_{i+1}]$.

The auxiliary variables on the grid are defined as $\eta_j = y(\tau_j)$; $\tau_j = t_i + j\delta t$; $j = 0, \dots, J_i = \Delta t_i / \delta t$, $y_i = \eta_0, \dots, \eta_{J_i} = y_{i+1}$, so that $\delta t = \Delta t_i / J_i$ can be chosen

so small, that

$$p(\eta_{j+1}|\eta_j) \approx \phi(\eta_{j+1}; \eta_j + f(\eta_j, \tau_j)\delta t, \Omega(\eta_j, \tau_j)\delta t), \quad (9)$$

$\Omega := gg'$, can be approximated by a Gaussian density ϕ (cf. equation (1)). The J_i -fold product of Gaussian densities is called an Euler density (cf. Kloeden and Platen, 1992, ch. 16.3). In the limit $J_i \rightarrow \infty$ the so called *path integral (functional integral)* representation

$$\begin{aligned} p(y_{i+1}|y_i) &= \lim_{J_i \rightarrow \infty} \int \exp \left[-\frac{1}{2} \sum_{j=0}^{J_i-1} (\eta_{j+1} - \eta_j - f_j \delta t)' (\Omega_j \delta t)^{-1} \times \right. \\ &\quad \left. \times (\eta_{j+1} - \eta_j - f_j \delta t) \right] \prod_{j=0}^{J_i-1} |2\pi \Omega_j \delta t|^{-1/2} d\eta_{J_i-1} \dots d\eta_1 \\ &:= \int \exp(-\frac{1}{2} O[y]) Dy(t) \end{aligned} \quad (10)$$

of the transition density is obtained (Haken, 1977, ch. 6.6, Risken, 1989, ch. 4.4.2). The exponent

$$O[y] = \int_{t_i}^{t_{i+1}} [\dot{y}(t) - f(y, t)]' \Omega(y, t)^{-1} [\dot{y}(t) - f(y, t)] dt \quad (11)$$

is the *Onsager-Machlup functional*. The expression is only formal since $y(t)$ is not differentiable. Analogous expressions are obtained when computing likelihood functionals, which can be transformed to formally existing limits (likelihood ratios) on dividing by a reference density and using Itô or Stratonovich integrals (cf. Wong and Hajek, 1985, ch. 6, p. 216 and the remark in ch. 7, p. 257; Stratonovich, 1989).

In numerical computations one does not go to the limit, but uses a δt small enough (a so called ϵ -Version in the sense of Stratonovich). The resulting $(J_i - 1)$ -dimensional integral (8) can be estimated by the mean value

$$\hat{p}(y_{i+1}|y_i) = N^{-1} \sum_{n=1}^N p(y_{i+1}|\eta_{n, J_i-1}), \quad (12)$$

where N is the Monte Carlo sample size. Here $\eta_n = \{\eta_{n, J_i-1}, \dots, \eta_{n, 1}, \eta_0\}$ are replications of the vector $\{y(t_i + (J_i - 1)\delta t), \dots, y(t_i + \delta t), y(t_i)\}$, which represents the path of the Itô process $y(t)$ on the time grid (conditioned on the initial value $y_i = \eta_0$). The approximation errors are controlled by the parameters δt and N , where the first corresponds to the approximation of the SDE and the second reflects the accuracy of the Monte Carlo integration.

The SDE is simulated by using the Euler-Maruyama scheme

$$\eta_{j+1} = \eta_j + f(\eta_j, \tau_j)\delta t + g(\eta_j, \tau_j)\sqrt{\delta t}z_j; j = 0, \dots, J_i - 2 \quad (13)$$

$$z_j \sim N(0, I), i.i.d. \quad (14)$$

(cf. Kloeden and Platen, 1992, ch. 9 and 14).

In the desired extrapolation integral (time update)

$$\begin{aligned} p_{i+1|i} := p(y_{i+1}|Z^i) &= \int p(y_{i+1}|y_i)p(y_i|Z^i)dy_i \\ &= \int p(\eta_{J_i}|\eta_{J_i-1})p(\eta_{J_i-1}|\eta_{J_i-2})\dots p(\eta_1|\eta_0) \times \\ &\quad \times p_{i|i}(\eta_0)d\eta_{J_i-1}\dots d\eta_1d\eta_0 \end{aligned} \quad (15)$$

an additional integration over the initial condition $\eta_0 = y_i$ is required, which can be simulated by drawing $y_{n,i} = \eta_{n,0} \sim p_{i|i}$. The result is an estimator of delta type, similar to a kernel density estimator with variable band widths (variances) $\Omega(\eta_{n,J_i-1}, \tau_{J_i-1})\delta t$ (cf. Silverman, 1986):

$$\begin{aligned} \hat{p}_{i+1|i} &= N^{-1} \sum p(y_{i+1}|\eta_{n,J_i-1}) \\ &\approx N^{-1} \sum \phi(y_{i+1}; y_{n,i+1|i}, P_{n,i+1|i}), \end{aligned} \quad (16)$$

where $y_{n,i+1|i} := \eta_{n,J_i-1} + f(\eta_{n,J_i-1}, \tau_{J_i-1})\delta t$ and $P_{n,i+1|i} := \Omega(\eta_{n,J_i-1}, \tau_{J_i-1})\delta t$. The Gauss form occurs only if the process errors dW are gaussian, however. In contrast, a kernel density estimator can be chosen of gaussian form even if nongaussian error terms (e.g. Poisson processes) are used, as in some finance applications (cf. Lo, 1988). See Singer (1997) for a comparison of several filter algorithms.

4 Importance sampling

The integral representation (15) can be rewritten to reduce the the variance of the estimate (16). In general, the integral

$$\begin{aligned} E_1[g] &= \int g(y)p_1(y)dy = \int g(y)\frac{p_1(y)}{p_2(y)}p_2(y)dy \\ &= E_2\left[g\frac{p_1}{p_2}\right] \end{aligned} \quad (17)$$

can be approximated by a variance reduced unbiased estimate

$$\widehat{E_1[g]} = N^{-1} \sum_{n=1}^N g(y_n) \frac{p_1(y_n)}{p_2(y_n)}, \quad (18)$$

$y_n \sim p_2$, if the density p_2 (importance density) is chosen appropriately. One can show that the optimal density is given by

$$p_{2,opt} = \frac{|g(y)|p_1(y)}{E_1|g(y)|} \quad (19)$$

and the variance of (18) is zero if g is positive (cf. Kloeden and Platen, 1992, ch. 16.3). Unfortunately, the definition involves the desired quantity $E_1|g(y)|$ and $p_{2,opt}$ must be approximated (see below). Setting $g(\eta_{J_i-1}) = p(y_{i+1}|\eta_{J_i-1})$ and $p_1 = p(\eta_{J_i-1}|\eta_{J_i-2}) \dots p(\eta_1|\eta_0)p(\eta_0|Z^i)$ leads to the optimal importance density

$$p_{2,opt} = p(\eta_{J_i-1}|\eta_{J_i-2}, y_{i+1}) \dots p(\eta_1|\eta_0, y_{i+1})p(\eta_0|Z^i, y_{i+1}), \quad (20)$$

where the transition densities are conditioned on future states y_{i+1} , which are not observed. Replacing y_{i+1} by z_{i+1} yields a modified density $\tilde{p}_{2,opt}$ and an estimate

$$\begin{aligned} \tilde{p}(y_{i+1}|Z^i) &= N^{-1} \sum_n p(y_{i+1}|\eta_{n,J_i-1}) \times \\ &\quad \times \frac{p(\eta_{n,J_i-1}|\eta_{n,J_i-2}) \dots p(\eta_{n,1}|\eta_{n,0})}{p(\eta_{n,J_i-1}|\eta_{n,J_i-2}, z_{i+1}) \dots p(\eta_{n,1}|\eta_{n,0}, z_{i+1})} \\ &\quad \times \frac{p(\eta_{n,0}|Z^i)}{p(\eta_{n,0}|Z^i, z_{i+1})} \\ &:= \sum_n p(y_{i+1}|\eta_{n,J_i-1}) \alpha_{n,i+1|i} \\ &\approx \sum_n \phi(y_{i+1}; y_{n,i+1|i}, P_{n,i+1|i}) \alpha_{n,i+1|i}, \end{aligned} \quad (21)$$

which can be shown to imply zero variance (given the data Z^i) for the *unbiased estimated likelihood*

$$\begin{aligned} \tilde{p}(z_{i+1}|Z^i) &= \int p(z_{i+1}|y_{i+1}) \tilde{p}(y_{i+1}|Z^i) dy_{i+1} \\ &= \sum_n p(z_{i+1}|\eta_{n,J_i-1}) \alpha_{n,i+1|i}. \end{aligned} \quad (22)$$

Furthermore the expressions (22) and (25) coincide for $\delta t \rightarrow 0$ (see appendix).

Since the time update $p(y_{i+1}|Z^i)$ (15) can be approximated by a sum of gaussian densities (21) of small band width $\propto \delta t$, the usual measurement updates of the extended Kalman filter (EKF) can be applied to each element in the superposition (21) (cf. Anderson and Moore, 1979, theorem 2.1). One obtains the *estimated a posteriori density*

$$\tilde{p}(y_{i+1}|Z^{i+1}) := \sum_n \phi(y_{i+1}; y_{n,i+1|i+1}, P_{n,i+1|i+1}) \alpha_{n,i+1|i+1} \quad (23)$$

$$\alpha_{n,i+1|i+1} := \alpha_{n,i+1|i} \frac{\tilde{L}_{n,i+1}}{\tilde{L}_{i+1}} \quad (24)$$

$$\tilde{L}_{i+1} = \sum_n \tilde{L}_{n,i+1} \alpha_{n,i+1|i}, \quad (25)$$

where the EKF updates are given by

$$y_{n,i+1|i} = \eta_{n,J_i-1} + f(\eta_{n,J_i-1}, \tau_{J_i-1}) \delta t \quad (26)$$

$$P_{n,i+1|i} = \Omega(\eta_{n,J_i-1}, \tau_{J_i-1}) \delta t \quad (27)$$

$$y_{n,i+1|i+1} = y_{n,i+1|i} + K_{n,i+1|i} \nu_{n,i+1} \quad (28)$$

$$P_{n,i+1|i+1} = (I - K_{n,i+1|i} H'_{n,i+1}) P_{n,i+1|i} \quad (29)$$

$$K_{n,i+1|i} = P_{n,i+1|i} H'_{n,i+1} \Gamma_{n,i+1|i}^{-1} \text{ (Kalman gain)} \quad (30)$$

$$\nu_{n,i+1} = z_{i+1} - h(y_{n,i+1|i}, t_{i+1}) \text{ (innovation)} \quad (31)$$

$$\Gamma_{n,i+1|i} = H_{n,i+1} P_{n,i+1|i} H'_{n,i+1} + R_{i+1} \text{ (innovation covariance)} \quad (32)$$

$$\tilde{L}_{n,i+1} = (\det 2\pi \Gamma_{n,i+1|i})^{-\frac{1}{2}} \exp[-\frac{1}{2} \nu'_{n,i+1} \Gamma_{n,i+1|i}^{-1} \nu_{n,i+1}] \quad (33)$$

$$= \phi(z_{i+1}; h(y_{n,i+1|i}, t_{i+1}), H_{n,i+1} P_{n,i+1|i} H'_{n,i+1} + R_{i+1}) \quad (34)$$

$$H_{n,i+1} = h_y(y_{n,i+1|i}, t_{i+1}) \text{ (Jacobian)}. \quad (35)$$

The sequence of estimated a priori and a posteriori densities $\tilde{p}(y_{i+1}|Z^i)$, $\tilde{p}(y_{i+1}|Z^{i+1})$ (21, 23) yields a numerical implementation of the continuous-discrete filter (4, 5) and permits the *variance reduced computation of the likelihood function* $\tilde{L} = \prod_{i=0}^{T-1} \tilde{L}_{i+1} L_0$, where L_0 is the likelihood of the first observation z_0 . Since a functional integral representation of the density $p(y_{i+1}|Z^i)$ is utilized the algorithm will be called *functional integral filter (FIF)*.

5 Implementation of the importance density by smoothing

In general the optimal weights (likelihood ratio)

$$\begin{aligned}\alpha_{n,i+1|i} &:= \frac{1}{N} \prod_{j=0}^{J_i-2} \frac{p(\eta_{n,j+1}|\eta_{n,j})}{p(\eta_{n,j+1}|\eta_{n,j}, z_{i+1})} \frac{p(\eta_{n,0}|Z^i)}{p(\eta_{n,0}|Z^i, z_{i+1})} \\ &= \frac{1}{N} \frac{p_1}{p_{2,opt}}(\eta_n|Z^i, z_{i+1}),\end{aligned}\tag{36}$$

$i = 0, \dots, T-1$, determined by the importance density $p_{2,opt}$ are difficult to compute, since they involve the unknown conditional densities $p(\eta_{j+1}|\eta_j, z_{i+1})$ and $p(y_i|Z^i, z_{i+1})$, but for the linear system an exact result is available which can be generalized to the nonlinear case.

Thus in linear systems the likelihood estimate (22,25) is dispersion free and the mean is exact with one trajectory. In nonlinear systems, the approximate importance density p_2 leads to a suboptimal estimate \tilde{L}_{i+1} with variance > 0 but still to a variance reduction.

Since in the linear case all densities are gaussian, one obtains the conditional mean and variance ($\eta_j = y(\tau_j)$; $\tau_j = t_i + j\delta t$; $j = 0, \dots, J_i - 1$)

$$\begin{aligned}E[y_{i+1}|\eta_j, z_{i+1}] &= E[y_{i+1}|\eta_j] + K_{i+1}(z_{i+1} - h(E[y_{i+1}|\eta_j], t_{i+1})) \\ \text{Var}[y_{i+1}|\eta_j, z_{i+1}] &= (I - K_{i+1}H_{i+1})\text{Var}[y_{i+1}|\eta_j] \\ K_{i+1} &:= \text{Var}(y_{i+1}|\eta_j)H'_{i+1} \times \\ &\quad \times (H_{i+1}\text{Var}(y_{i+1}|\eta_j)H'_{i+1} + R_{i+1})^{-1} \\ E[\eta_{j+1}|\eta_j, z_{i+1}] &= E[\eta_{j+1}|\eta_j] + F_j(E[y_{i+1}|\eta_j, z_{i+1}] - E[y_{i+1}|\eta_j]) \\ \text{Var}(\eta_{j+1}|\eta_j, z_{i+1}) &= \text{Var}(\eta_{j+1}|\eta_j) + \\ &\quad + F_j(\text{Var}(y_{i+1}|\eta_j, z_{i+1}) - \text{Var}(y_{i+1}|\eta_j))F'_j \\ F_j &:= \text{Var}(\eta_{j+1}|\eta_j)\Phi'(t_{i+1}, \tau_{j+1})\text{Var}(y_{i+1}|\eta_j)^{-1},\end{aligned}$$

which characterize the density $p(\eta_{j+1}|\eta_j, z_{i+1})$ in $p_{2,opt}$.

The smoother gain F_j and the update formulas are similar to the fixed interval smoother (cf. Anderson and Moore, 1979). The quantities in the update formulas can be obtained by solving the differential equations ($\tau_j \leq t \leq t_{i+1}$)

$$\dot{y}(t|\tau_j) = f(y(t|\tau_j), t); y(\tau_j|\tau_j) = \eta_j \tag{37}$$

$$\dot{P}(t|\tau_j) = A(t)P(t|\tau_j) + P(t|\tau_j)A'(t) + \Omega(t); P(\tau_j|\tau_j) = 0 \tag{38}$$

$$\dot{\Phi}(t|\tau_{j+1}) = A(t)\Phi(t|\tau_{j+1}); \Phi(\tau_{j+1}|\tau_{j+1}) = I; \tau_{j+1} \leq t \leq t_{i+1} \tag{39}$$

$$f(y, t) := A(t)y + b(t) \text{ (linear system)} \quad (40)$$

$$h(y, t) := H(t)y + d(t) \quad (41)$$

and setting

$$E[\eta_{j+1}|\eta_j] = y(\tau_{j+1}|\tau_j) \quad (42)$$

$$E[y_{i+1}|\eta_j] = y(t_{i+1}|\tau_j) \quad (43)$$

$$\text{Var}(\eta_{j+1}|\eta_j) = P(\tau_{j+1}|\tau_j) \quad (44)$$

$$\text{Var}(y_{i+1}|\eta_j) = P(t_{i+1}|\tau_j). \quad (45)$$

For the density $p(y_i|Z^i, z_{i+1})$ in $p_{2,opt}$ one obtains the moments:

$$\begin{aligned} E[y_{i+1}|Z^{i+1}] &= E[y_{i+1}|Z^i] + K_{i+1}(z_{i+1} - h(E[y_{i+1}|Z^i], t_{i+1})) \\ \text{Var}(y_{i+1}|Z^{i+1}) &= (I - K_{i+1}H'_{i+1})\text{Var}(y_{i+1}|Z^i) \\ K_{i+1} &= \text{Var}(y_{i+1}|Z^i)H'_{i+1}(H_{i+1}\text{Var}(y_{i+1}|Z^i)H'_{i+1} + R_{i+1})^{-1} \\ E[y_i|Z^i, z_{i+1}] &= E(y_i|Z^i) + F_i(E[y_{i+1}|Z^{i+1}] - E[y_{i+1}|Z^i]) \\ \text{Var}(y_i|Z^i, z_{i+1}) &= \text{Var}(y_i|Z^i) + F_i(\text{Var}(y_{i+1}|Z^{i+1}) - \text{Var}(y_{i+1}|Z^i))F'_i \\ F_i &:= \text{Var}(y_i|Z^i)\Phi'(t_{i+1}, t_i)\text{Var}(y_{i+1}|Z^i)^{-1}. \end{aligned}$$

Again the quantities in the update formulas can be obtained by solving the differential equations ($t_i \leq t \leq t_{i+1}$), but with different initial conditions:

$$\dot{y}(t|t_i) = f(y(t|t_i), t); \quad y(t_i|t_i) = E[y_i|Z^i] \quad (46)$$

$$\dot{P}(t|t_i) = A(t)P(t|t_i) + P(t|t_i)A'(t) + \Omega(t); \quad P(t_i|t_i) = \text{Var}[y_i|Z^i] \quad (47)$$

$$\dot{\Phi}(t|t_i) = A(t)\Phi(t|t_i); \quad \Phi(t_i|t_i) = I \quad (48)$$

and setting

$$E[y_{i+1}|Z^i] = y(t_{i+1}|t_i) \quad (49)$$

$$\text{Var}(y_{i+1}|Z^i) = P(t_{i+1}|t_i). \quad (50)$$

In the limit of small δt one can write

$$\begin{aligned} E[\eta_{j+1}|\eta_j, z_{i+1}] &= \eta_j + f(\eta_j, \tau_j)\delta t + F_j(E[y_{i+1}|\eta_j, z_{i+1}] - E[y_{i+1}|\eta_j]) \\ &:= \eta_j + f_2(\eta_j, \tau_j)\delta t \\ \text{Var}(\eta_{j+1}|\eta_j, z_{i+1}) &= \Omega(\tau_j)\delta t + F_j(\text{Var}(y_{i+1}|\eta_j, z_{i+1}) - \text{Var}(y_{i+1}|\eta_j))F'_j \\ &:= \Omega_2(\tau_j)\delta t, \end{aligned}$$

which may be interpreted as a correction to the drift f and the diffusion matrix Ω . Therefore the optimal density $p_{2,opt}$ and trajectories $\eta_n = \{\eta_n, J_{i-1}, \dots,$

$\eta_{n,0}$ drawn from it can be obtained by using a modified drift f_2 and diffusion coefficient Ω_2 . More precisely, a stochastic Euler-Maruyama scheme (13) with drift f_2 and diffusion coefficient Ω_2 is used to simulate $\eta_n \sim p_{2,opt}$.

Since the moment equations (37-41, 46-48) can be generalized to nonlinear systems by replacing $A(t) \rightarrow f_y(y, t)$, $\Omega(t) \rightarrow \Omega(y, t)$, and $H(t) \rightarrow h_y(y, t)$, one obtains a sampling scheme where the (sub)optimal density p_2 is implemented by means of the EKF updates and trajectories $\{\eta_{J_i-1}, \dots, \eta_0\}_n \sim p_2$ can be simulated using f_2 and Ω_2 .

6 Practical implementation

6.1 Smoothing with Gaussian Sums

If the density $p_{0|-1} = p_0(y_0, t_0)$ (initial condition) is represented by a gaussian mixture distribution of N populations

$$p_{0|-1} := \sum_{n=1}^N \phi(y_0; \mu_n, \Sigma_n) \alpha_{n,0|-1} \quad (51)$$

using appropriate weights $\alpha_{n,0|-1}$, all updates preserve the structure of a gaussian sum and the computation of the importance density proceeds by an N -fold solution of the smoother equations and related trajectories drawn from the density p_2 . More precisely, the measurement update

$$p_{0|0} = p(y_0|Z^0) := \sum_n \phi(y_0; y_{n,0|0}, P_{n,0|0}) \alpha_{n,0|0} \quad (52)$$

is again a gaussian sum and the smoothed density $p(y_0|Z^0, z_1) = p_{0|1}$ required in p_2 may be represented by the moments $E[y_{n,0}|Z^0, z_1]$, $\text{Var}(y_{n,0}|Z^0, z_1)$ of population n and an updated weight $\alpha_{n,0|1} = \alpha_{n,0|0} \frac{L_{n,1}}{L_1}$, $L_1 = \sum \alpha_{n,0|0} L_{n,1}$, i.e.

$$p_{0|1} = p(y_0|Z^0, z_1) := \sum_n \phi(y_0; y_{n,0|1}, P_{n,0|1}) \alpha_{n,0|1}. \quad (53)$$

Therefore, the EKF's and smoothers computing the optimal weights and the importance density run on a gaussian sum which is called a *gaussian sum filter* (Anderson and Moore, 1979, chapter 8). This filter does not involve any simulation and computes an approximate time update via deterministic moment equations using N EKF's. Whereas these are only valid for small sampling intervals Δt_i , the stochastic simulation of trajectories via (13) leads to an estimate of the a priori density valid for arbitrary sampling intervals.

From the density $p_{0|1}$ N random initial conditions $\eta_{n,0|1}$ can be drawn and used to simulate the trajectories $\{\eta_{n,J_0-1}, \dots, \eta_{n,0}\} \sim p_2$ using f_2 and Ω_2 in (13). From this the a priori density $\tilde{p}_{1|0} = \tilde{p}(y_1|Z^0)$ (21) can be estimated and the a posteriori density $\tilde{p}_{1|1} = \tilde{p}(y_1|Z^1)$ and the likelihood estimate \tilde{L}_1 in (23) is obtained. $\tilde{p}_{1|1}$ is again a gaussian sum and the update $\tilde{p}_{1|2} = \tilde{p}(y_1|Z^2)$ may be computed as before etc. The algorithm runs recursively from $i = 0, \dots, T$ and yields a sequence of likelihood contributions \tilde{L}_i .

6.2 Resampling Strategies and Antithetical Sampling

Drawing $y_j, j = 1, \dots, N$ from the mixture distribution (a posteriori density)

$$\tilde{p}(y) := \sum_{n=1}^N \phi(y; y_n, P_n) \alpha_n \quad (54)$$

can be accomplished by drawing a population n with probability α_n and then setting $y_j = y_n + P_n^{1/2} z_j$ where $z_j \sim N(0, I)$ and $P_n^{1/2}$ is a Cholesky root (or another matrix square root). The drawing of n may be implemented by drawing $u_j \sim U[0, 1]$ (uniform distribution) and solving $\sum_{m=1}^{n-1} \alpha_m < u_j \leq \sum_{m=1}^n \alpha_m$. Alternatively, the deterministic values $u_j = (j - c)/N; c \in (0, 1)$ or stratified values $u_j = (j + U_j - 0.5 - c)/N; U_j \sim U[0, 1], c \in (0, 1)$ could be used. According to Kitagawa (1996, appendix), when using deterministic or stratified drawing, it is preferable in advance to sort the mixture in order of magnitude, i.e. $\|\tilde{y}_n\| < \|\tilde{y}_{n+1}\|$ and draw from the sorted \tilde{y}_n, \tilde{P}_n and $\tilde{\alpha}_n$. In my experience, sorting improves the smoothness of the simulated likelihood surface as a function of the parameter vector ψ (cf. example 7.3, figs. (10–15)).

Another device in reduction of sampling error is antithetical sampling (Hammersley and Handscomb, 1964, p. 60). Instead of simulating $z_j \sim N(0, I); j = 1, \dots, N$, pairs $\{z_j, -z_j\}; j = 1, \dots, N/2$, are drawn. The negatively correlated sample leads to estimators with smaller variance. When simulating the Euler scheme (13), the i.i.d. sequence $-z_j; j = 0, \dots, J_i - 2$ can be used to simulate a trajectory $\eta_j(-z)$ which is anticorrelated with $\eta_j(+z)$.

6.3 Implementation Details

The algorithm was programmed with *Mathematica* (Wolfram Research, 1992) and the *MPW C* compiler (Apple Computer, 2001) using the *Mathlink* communication library and run on Apple Power PC 604e and G3 computers. The *Mathlink* routines allow the calling of C programs from within *Mathematica*. For numerical computations (Cholesky roots, random numbers, sorting), the C algorithms in *Numerical Recipes in C* (Press et al., 1992) have been used.

7 Examples

7.1 AR(2) process

In order to test the performance of the importance sampling algorithm, a linear AR(2) model was simulated. The sampler should give a variance free estimate of the likelihood which must coincide with the exact result obtained by the Kalman filter. I used the state space model (equivalent to a 2nd order differential equation)

$$\begin{aligned} d \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} &:= \begin{bmatrix} 0 & 1 \\ -16 & -4 \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} dt + \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} d \begin{bmatrix} W_1(t) \\ W_2(t) \end{bmatrix} \\ z_i &:= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} y_1(t_i) \\ y_2(t_i) \end{bmatrix} + \epsilon_i, \end{aligned} \quad (55)$$

where $\text{Var}(\epsilon_i) := R = 0.1$, the data were equispaced with $\Delta t = 2$, $t_0 = 0 \leq t \leq t_T = 50$, and the discretization interval was chosen as $\delta t = 0.01$, $J_i = \Delta t / \delta t = 200$, $i = 0, \dots, T = 25$. A time series was computed according to (55) and using this data the likelihood function $l(\psi)$ was simulated.

Results: The results are displayed in figures (1–4), where the variance reduction in $M = 10$ replications of the likelihood surface is summarized. Also shown is the exact result using the (linear) Kalman filter. The likelihoods and scores are plotted as a function of the parameter $\psi_3 = -16$ in the interval $[-20, -14]$. Even in the case $N = 1$ (fig. 2), the sampling error is very small. Using larger discretization intervals $\delta t = 0.1, 0.05$, one can show numerically that the variance of the estimated likelihood increases. Therefore, approximation errors in the simulation (13) and in the transition density (9) lead to deviations from the (theoretically) exact variance free estimate $\tilde{p}(z_{i+1}|Z^i)$ (22).

7.2 Ginzburg-Landau model

The Ginzburg-Landau equation is a nonlinear diffusion equation where the drift coefficient is the gradient of a double well potential $\Phi(y, \{\alpha, \beta\}) = \frac{\alpha}{2}y^2 + \frac{\beta}{4}y^4$, $f = -\partial\Phi/\partial y$:

$$dy = -[\alpha y + \beta y^3]dt + \sigma dW(t) \quad (56)$$

$$z_i = y(t_i) + \epsilon_i. \quad (57)$$

Models of this kind have been used to model limit cycles, bifurcations, phase transitions and normal forms of nonlinear systems (cf. V.I. Arnold, 1973,

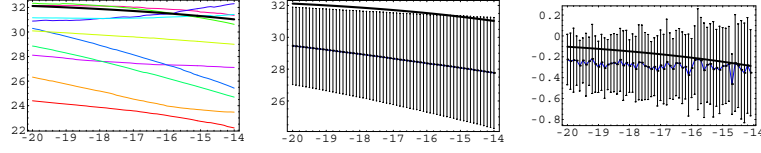


Figure 1: Simulated likelihood surface $\tilde{l}(\psi_3)$: AR(2) *without* importance sampling (sample size $N = 10$). $M = 10$ replications (left), means and standard deviations (middle) and score (right). Interval $-20 \leq \psi_3 \leq -14$. Bold line: exact likelihood $l(\psi_3)$ and score (right).

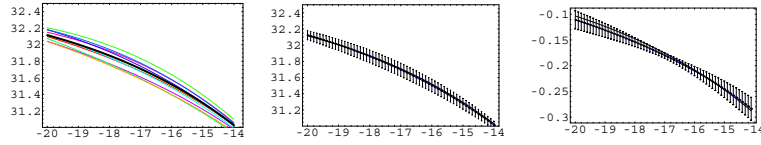


Figure 2: Simulated likelihood surface: AR(2) *with* importance sampling (sample size $N = 1$). Bold line: exact likelihood and score (right).

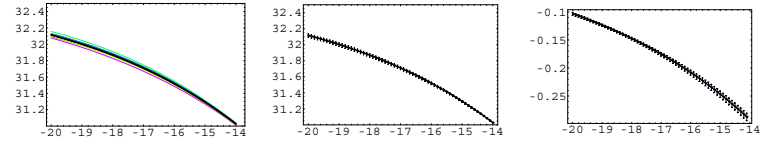


Figure 3: Likelihood surface: AR(2) *with* importance sampling (sample size $N = 10$).

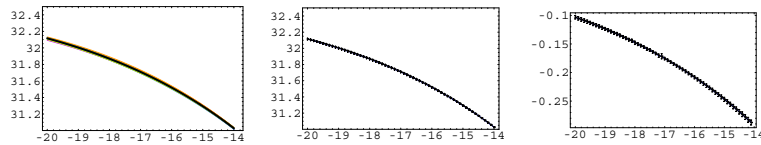


Figure 4: Likelihood surface: AR(2) *with* importance sampling (sample size $N = 20$).

1986, Haken, 1977, Holmes, 1981, normal form theorem 4.4). Other applications are in the modelling of equilibrium states of an economy (Herings, 1996) and the theory of system failure (Frey, 1996).

In the present context a parameter constellation of $\psi = \{\alpha, \beta, \sigma, R\} = \{-1., 0.1, 2., 0.1\}$, $R = \text{Var}(\epsilon_i)$ was chosen, which corresponds to a potential with two minima and noisy sampled measurements ($\Delta t = 2; 0 \leq t \leq 50; \delta t = 0.1$).

The convergence of the simulated likelihood as a function of sample size N is shown in figure (5). Again the variance of the estimates is considerably reduced by importance sampling. The form of the likelihood surface and of the score as a function of $\psi_1 = \alpha$ is compared in figures (6-7).

Finally, figs. (8-9) explain the effect of importance sampling on the trajectories drawn from p_2 . As shown in section 5, the actual drift and diffusion coefficients f, g are modified to f_2, g_2 in order to draw from the importance density, which yields more random numbers near the points in state space, where $p(z_i|y_i)$ is averaged in the likelihood expression $L_{i+1} = \int p(z_{i+1}|y_{i+1})p(y_{i+1}|Z^i)dy_{i+1}$ (cf. eq. 6). Clearly, without importance sampling, only few trajectories are near the measurement at the end of the interval (see fig. 6) and the mean shows high dispersion.

Simulation studies (Singer, 1999b) compared the performance of the functional integral filter (FIF) with a filter based on kernel density estimates and with approximations based on Taylor expansions (EKF, 2nd order nonlinear filter SNF and local linearization (LL), cf. Shoji and Ozaki (1997, 1998)). It was shown that for large sampling intervals, the FIF with importance sampling exhibits the smallest bias even for small Monte Carlo sample sizes ($N = 10$), whereas without importance sampling, sample sizes of at least $N = 50$ are required. Moreover, Taylor expansion methods (EKF, SNF, LL) only yield good results for small measurement intervals.

7.3 Stochastic Volatility

Stochastic volatility models such as

$$\begin{aligned} dS(t) &= \mu S(t)dt + \sigma(t)S(t)dW(t) \\ d\sigma(t) &= \lambda[\sigma(t) - \bar{\sigma}]dt + \gamma dV(t), \end{aligned} \tag{58}$$

$\text{Var}(dW, dV) = \rho dt$ (Scott, 1987, Hull and White, 1987), where the volatility process $\sigma(t)$ is not observable, can account for the fact that the returns

$$r(t) = dS/S = \mu dt + \sigma(t)dW(t) \tag{59}$$

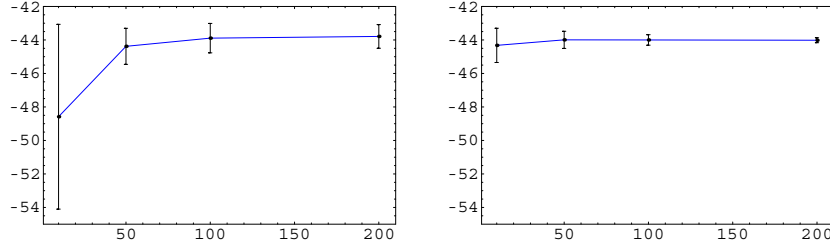


Figure 5: Convergence of the simulated likelihood (Ginzburg-Landau model). Means \pm standard deviations in $M = 10$ replications. Right picture: with importance sampling. Sample size $N = 10, 50, 100, 200$.

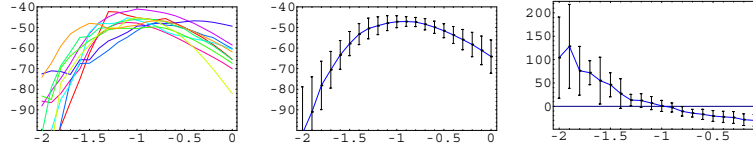


Figure 6: Likelihood surface (Ginzburg-Landau model) as a function of $\psi_1 = \alpha$. $M = 10$ replications (left), means and standard deviations (middle) and score (right). Sample size $N = 10$ (without importance sampling).

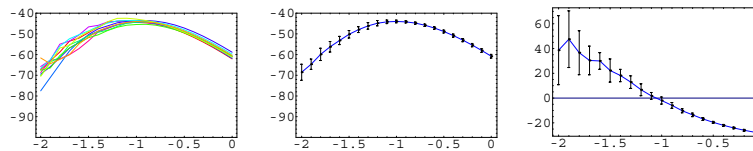


Figure 7: Likelihood surface (Ginzburg-Landau model) as a function of $\psi_1 = \alpha$. Sample size $N = 10$ (with importance sampling).

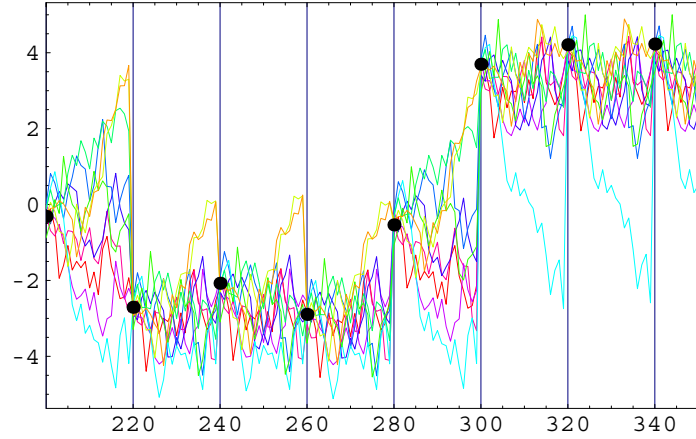


Figure 8: Sample trajectories from p_1 (Ginzburg-Landau model). Sample size $N = 10$ (without importance sampling).

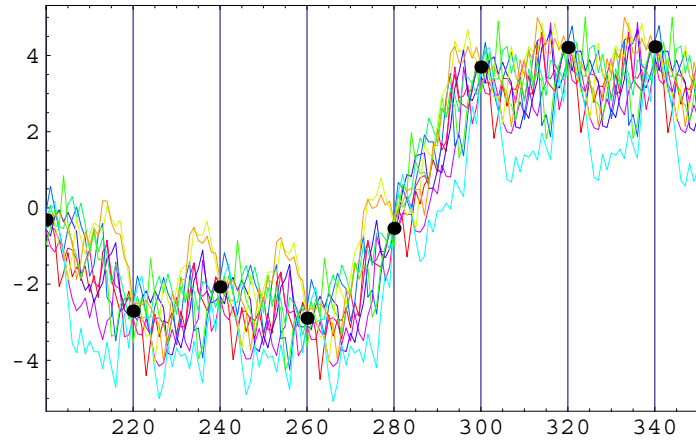


Figure 9: Sample trajectories from p_2 (Ginzburg-Landau model). Sample size $N = 10$ (with importance sampling).

on financial time series exhibit a time dependent variance and for the leptokurtosis of the return distribution. In contrast to ARCH and GARCH models exhibiting conditional heteroscedasticity, too, the variance equation is driven by a separate Wiener process and the variance cannot be eliminated. For example, the discrete time GARCH(1,1) process

$$\begin{aligned}\epsilon_i &= \sigma_i z_i \\ \sigma_i^2 &= \omega + \alpha \epsilon_{i-1}^2 + \beta \sigma_{i-1}^2\end{aligned}$$

permits the recursive computation of σ_i given measurements of the innovation process ϵ_i (which corresponds to σdW) and an initial value σ_0 . It has been shown by Nelson (1990) however, that a continuous time limit of the GARCH(1,1)-M model (Engle and Bollerslev, 1986) in the mean corrected log returns $\log(S_{i+1}/S_i) := y_{i+1} - y_i$

$$\begin{aligned}y_{i+1} &= y_i + c\sigma_i^2 + \sigma_i z_i \\ \sigma_i^2 &= \omega + \alpha \epsilon_{i-1}^2 + \beta \sigma_{i-1}^2\end{aligned}$$

leads to the system of stochastic differential equations

$$\begin{aligned}dy(t) &= c\sigma(t)^2 dt + \sigma(t)dW(t) \\ d\sigma(t)^2 &= [\omega - \theta\sigma(t)^2]dt + \alpha\sigma(t)^2 dV(t),\end{aligned}$$

where W and V are independent standard Wiener processes and the coefficients are scaled as $dt \rightarrow 0$ ($\omega \rightarrow \omega dt$, $\beta \rightarrow 1 - \alpha\sqrt{dt/2} - \theta dt$, $\alpha \rightarrow \alpha\sqrt{dt/2}$). This differs somewhat from equation (58), where the volatility satisfies an Ornstein-Uhlenbeck process.

Stochastic volatility models in discrete time have been used as approximations to the stochastic differential equation (58) or some variants such as

$$\begin{aligned}dS(t) &= \mu S(t)dt + \sigma(t)S(t)dW(t) \\ d\log \sigma^2(t) &= \lambda[\log \sigma^2(t) - \log \bar{\sigma}^2]dt + \gamma dV(t) \\ (dh(t) &= \lambda[h(t) - \bar{h}]dt + \gamma dV(t)),\end{aligned}\tag{60}$$

where the log-volatility $h(t)$ is modeled by an Ornstein-Uhlenbeck-process to ensure a positive σ (cf. Wiggins, 1987, Nelson, 1990). Taking logarithms and using Itô's lemma, $y = \log S$ fulfils

$$\begin{aligned}dy(t) &= [\mu - \sigma(t)^2/2]dt + \sigma(t)dW(t) \\ d\log \sigma^2(t) &= \lambda[\log \sigma^2(t) - \log \bar{\sigma}^2]dt + \gamma dV(t).\end{aligned}\tag{61}$$

This has been shown to be the continuous time limit of an AR(1)-EGARCH model (Nelson, 1990, sect. 3.3) and corresponds to the discrete time model

$$\begin{aligned} y_i &= \exp(h_i/2)\epsilon_i \\ h_{i+1} - \bar{h} &= \lambda[h_i - \bar{h}] + \gamma\eta_i \\ (h_i &= \log \sigma_i^2) \end{aligned}$$

used for the mean corrected returns by Kim et al. (1998).

Since available data are measured in discrete time (daily or weekly), but the models in the option pricing literature are mostly formulated in continuous time, times series formulations are only approximations to the sampled stochastic processes. The continuous time asymptotics and asymptotic GARCH filters developed by Nelson (1990, 1992) are only valid in the limit of small sampling interval. In analogy to linear theory, the differential equations should be filtered and estimated using discrete data with arbitrary time interval. This involves sampled diffusion processes with latent variables, since the volatility is not observed. In contrast to linear models, where exact discrete time series can be derived explicitly (cf. Bergstrom, 1990), the exact analogs of nonlinear systems involve transition densities which are solutions of the Fokker-Planck equation.

The following example serves to illustrate the simulated properties of the continuous time stochastic log volatility model (61). We assume that $T = 365$ daily data are simulated ($y_0 = \log 100, h_0 = \log 0.2^2$), but measurements are taken only weekly, i.e. $\delta t = 1/365; \Delta t = 7/365$. Parameters were chosen as $\psi = \{\mu, \lambda, \bar{h}, \gamma, \rho\} = \{.07, -1, \log(0.2^2) = -3.21888, 2, 0\}$ and the prior distribution of the state $\eta(0) = \{y(0), h(0)\}$ was set to $p_{0|-1} \sim N(\{4, -3\}, \text{diag}(1, 1))$. Thus the correlation ρ between the Wiener processes W and V is zero in accordance with Kim et al. In the measurement model the measurement error was set to a variance of $R = 0.0001$. Figures (10-15) show the simulated likelihood surface as a function of $\psi_3 = \bar{h}$ in the interval $[-6, 1]$. It is seen that sorting of the posterior distribution (cf. section 6.2) improves the smoothness of the likelihood and lowers the sampling error of the score function $\partial l / \partial \psi$. In figure (12), antithetical sampling still further improves the smoothness as seen in the score (right picture). Figures (13 - 15) demonstrate that importance sampling permits the usage of much smaller Monte Carlo sample size N .

8 Conclusion

We have shown how the likelihood function of a continuous-discrete state space model can be simulated using Monte Carlo integration. The variance

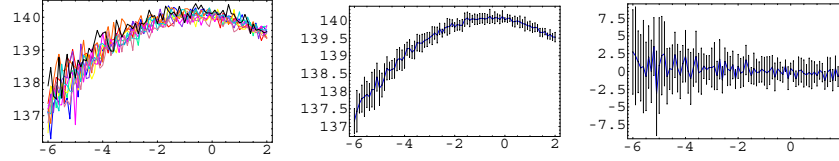


Figure 10: Stochastic volatility model. Likelihood surface simulated without importance sampling. Deterministic resampling without ordering (sample size $N = 10000$).

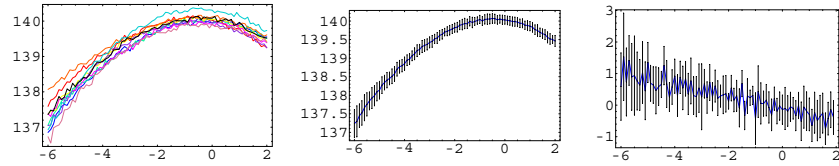


Figure 11: Stochastic volatility model. Likelihood surface simulated without importance sampling. Deterministic resampling with ordering (sample size $N = 10000$).

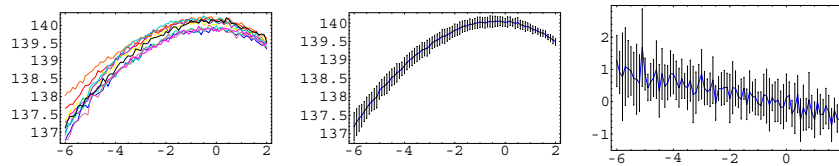


Figure 12: Stochastic volatility model. Likelihood surface simulated without importance sampling. Deterministic resampling with ordering and antithetical variates (sample size $N = 10000$).

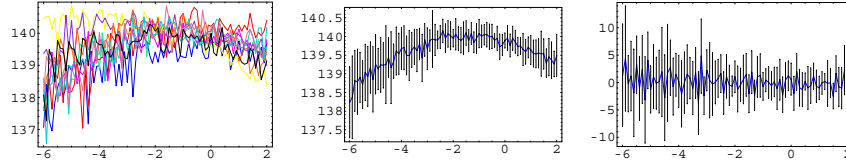


Figure 13: Stochastic volatility model. Likelihood surface simulated with importance sampling. Deterministic resampling without ordering (sample size $N = 200$).

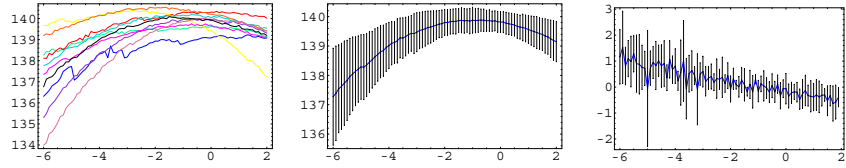


Figure 14: Stochastic volatility model. Likelihood surface simulated with importance sampling. Deterministic resampling with ordering (sample size $N = 200$).

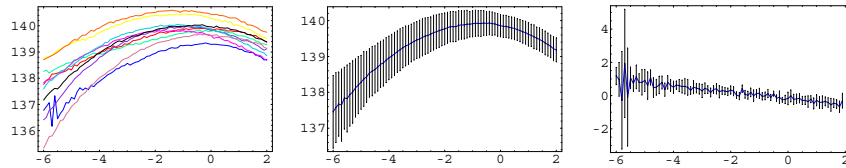


Figure 15: Stochastic volatility model. Likelihood surface simulated with importance sampling. Deterministic resampling with ordering (sample size $N = 500$).

of the estimate is considerably reduced by using importance sampling. The importance density was computed by approximate smoothing algorithms, which run on gaussian sums and are only suboptimal in general nonlinear systems. Nevertheless, a strong reduction in dispersion is achieved. Currently the algorithms are tested in estimating the parameters of stochastic volatility models and the Lorenz model.

Appendix

In this appendix it is shown, that the EKF update of the a priori density (21) leads to the correct variance reduced Monte Carlo estimator of the likelihood

$$L_{i+1} = p(z_{i+1}|Z^i) = \int p(z_{i+1}|y_{i+1})p(y_{i+1}|Z^i)dy_{i+1} \quad (62)$$

Using the Bayes formula this can be rewritten as

$$p(z_{i+1}|Z^i) = \int p(z_{i+1}|\eta_{J_i-1})p(\eta_{J_i-1}, \dots, \eta_0|Z^i)d\eta$$

where $d\eta = d\eta_{J_i-1} \dots d\eta_0$ and for this the optimal variance reducing density is

$$\begin{aligned} p_{2,opt} &= p(z_{i+1}|\eta_{J_i-1})p(\eta|Z^i)/p(z_{i+1}|Z^i) \\ &= p(\eta|Z^{i+1}) \end{aligned}$$

Therefore the optimal estimator is

$$\begin{aligned} \hat{p}(z_{i+1}|Z^i) &= N^{-1} \sum_n p(z_{i+1}|\eta_{n,J_i-1}) \frac{p(\eta|Z^i)}{p(\eta|Z^{i+1})} \\ &= N^{-1} \sum_n p(z_{i+1}|\eta_{n,J_i-1}) \frac{p(\eta|Z^i)}{p(\eta|Z^{i+1})} \\ &= \sum_n p(z_{i+1}|\eta_{n,J_i-1}) \alpha_{n,i+1|i} \end{aligned}$$

which is the same as inserting (21) into (62). It remains to note that

$$\begin{aligned} p(z_{i+1}|\eta_{J_i-1}) &\approx \phi(z_{i+1}; h(y_{i+1|i}, t_{i+1}), H_{i+1}P_{i+1|i}H'_{i+1} + R_{i+1}) \\ y_{i+1|i} &= \eta_{J_i-1} + f(\eta_{J_i-1}, \tau_{J_i-1})\delta t \\ P_{i+1|i} &= \Omega(\eta_{J_i-1}, \tau_{J_i-1})\delta t, \end{aligned}$$

since

$$\begin{aligned} z_{i+1} &= h(y_{i+1}, t_{i+1}) + \epsilon_{i+1} \\ &\approx h(\eta_{J_i-1} + f(\eta_{J_i-1}, \tau_{J_i-1})\delta t + \\ &\quad g(\eta_{J_i-1}, \tau_{J_i-1})\delta W(\tau_{J_i-1}), t_{i+1}) + \epsilon_{i+1} \\ &\approx h(\eta_{J_i-1} + f(\eta_{J_i-1}, \tau_{J_i-1})\delta t, t_{i+1}) + \\ &\quad H_{i+1}g(\eta_{J_i-1}, \tau_{J_i-1})\delta W(\tau_{J_i-1}) + \epsilon_{i+1} \\ H_{i+1} &= h_y(\eta_{J_i-1} + f(\eta_{J_i-1}, \tau_{J_i-1})\delta t, t_{i+1}). \end{aligned}$$

References

- Alspach, D. & Sorenson, H. (1972), ‘Nonlinear Bayesian estimation using Gaussian sum approximations’, *IEEE Transactions on Automatic Control* **17**, 439–448.
- Anderson, B. & Moore, J. (1979), *Optimal Filtering*, Prentice Hall, Englewood Cliffs.
- Apple Computer (2001), *Macintosh Programmer’s Workshop*, <http://developer.apple.com/tools/mpw-tools/>.
- Arnold, L. (1974), *Stochastic Differential Equations*, John Wiley, New York.
- Arnold, V. (1973), *Ordinary Differential Equations*, MIT Press, Cambridge (Mass.), London.
- Arnold, V. (1986), *Catastrophy Theory*, Springer, Berlin.
- Bergstrom, A. (1990), *Continuous Time Econometric Modelling*, Oxford University Press, Oxford.
- Carlin, B., Polson, N. & Stoffer, D. (1992), ‘A Monte Carlo Approach to Non-normal and Nonlinear State-Space Modeling’, *Journal of the American Statistical Association* **87**, 493–500.
- Durbin, J. & Koopman, S. (1997), ‘Monte Carlo maximum likelihood estimation for non-Gaussian state space models’, *Biometrika* **84**, 3, 669–684.
- Durbin, J. & Koopman, S. (2000), ‘Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives.’, *Journal of the Royal Statistical Association B* **62**, 1, 3–56.
- Engle, R. & Bollerslev, L. (1986), ‘Modelling the persistence of conditional variances’, *Econometric Reviews* **5**, 1–50.
- Fahrmeir, L. & Kaufmann, H. (1991), ‘On Kalman Filtering, Posterior Mode Estimation and Fisher Scoring in Dynamic Exponential Family Regression’, *Metrika* **38**, 37–60.
- Frey, M. (1996), ‘A Wiener Filter, State-Space Flux-Optimal Control Against Escape from a Potential Well’, *IEEE Transactions on Automatic Control* **41**, 2, 216–223.
- Gelb, A., ed. (1974), *Applied Optimal Estimation*, MIT Press, Cambridge, Mass.

- Gordon, N., Salmond, D. & Smith, A. (1993), ‘Novel approach to nonlinear/non-Gaussian Bayesian state estimation’, *IEEE Transactions on Radar and Signal Processing* **140**, **2**, 107–113.
- Haken, H. (1977), *Synergetics*, Springer, Berlin.
- Hammersley, J. & Handscomb, D. (1964), *Monte Carlo Methods*, Methuen, London.
- Harvey, A. (1989), *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Cambridge.
- Harvey, A. & Stock, J. (1985), ‘The estimation of higher order continuous time autoregressive models’, *Econometric Theory* **1**, 97–112.
- Herings, J. (1996), *Static and Dynamic Aspects of General Disequilibrium Theory*, Kluwer, Boston, London, Dordrecht.
- Holmes, P. J. (1981), ‘Center manifolds, normal forms and bifurcations of vector fields’, *Physica D* **2**, 449–481.
- Hull, J. & White, A. (1987), ‘The Pricing of Options with Stochastic Volatilities’, *Journal of Finance* **XLII**, **2**, 281–300.
- Hürzeler, M. & Künsch, H. (1998), ‘Monte Carlo Approximations for General State-Space Models’, *Journal of Computational and Graphical Statistics* **7**, **2**, 175–193.
- Jazwinski, A. (1970), *Stochastic Processes and Filtering Theory*, Academic Press, New York.
- Jones, R. (1984), Fitting multivariate models to unequally spaced data, in E. Parzen, ed., ‘Time Series Analysis of Irregularly Observed Data’, Springer, New York, pp. 158–188.
- Jones, R. & Ackerson, L. (1990), ‘Serial correlation in unequally spaced longitudinal data’, *Biometrika* **77**, 721–731.
- Jones, R. & Tryon, P. (1987), ‘Continuous time series models for unequally spaced data applied to modeling atomic clocks’, *SIAM J. Sci. Stat. Comput.* **8**, 71–81.
- Kim, S., Shephard, N. & Chib, S. (1998), ‘Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models’, *Review of Economic Studies* **45**, 361–393.
- Kitagawa, G. (1987), ‘Non-Gaussian state space modeling of nonstationary time series’, *Journal of the American Statistical Association* **82**, 1032 – 1063.

- Kitagawa, G. (1996), ‘Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models’, *Journal of Computational and Graphical Statistics* **5**, 1, 1–25.
- Kloeden, P. & Platen, E. (1992), *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin.
- Liptser, R. & Shiriyayev, A. (1977, 1978), *Statistics of Random Processes, Volumes I and II*, Springer, New York, Heidelberg, Berlin.
- Lo, A. (1988), ‘Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data’, *Econometric Theory* **4**, 231–247.
- Nelson, D. (1990), ‘ARCH models a diffusion approximations’, *Journal of Econometrics* **45**, 7–38.
- Nelson, D. (1992), ‘Filtering and forecasting with misspecified ARCH models I’, *Journal of Econometrics* **52**, 61–90.
- Press, W., Teukolsky, S., Vetterling, W. & Flannery, B. (1992), *Numerical Recipes in C*, second edn, Cambridge University Press, Cambridge.
- Risken, H. (1989), *The Fokker-Planck Equation*, second edn, Springer, Berlin, Heidelberg, New York.
- Scott, L. (1987), ‘Option pricing when the variance changes randomly: Theory, estimation, and an application’, *Journal of Financial and Quantitative Analysis* **22**, 419–438.
- Shoji, I. & Ozaki, T. (1997), ‘Comparative Study of Estimation Methods for Continuous Time Stochastic Processes’, *Journal of Time Series Analysis* **18**, 5, 485–506.
- Shoji, I. & Ozaki, T. (1998), ‘A statistical method of estimation and simulation for systems of stochastic differential equations’, *Biometrika* **85**, 1, 240–243.
- Silverman, B. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, London.
- Singer, H. (1993b), ‘Continuous-time dynamical systems with sampled data, errors of measurement and unobserved components’, *Journal of Time Series Analysis* **14**, 5, 527–545.
- Singer, H. (1995), ‘Analytical score function for irregularly sampled continuous time stochastic processes with control variables and missing values’, *Econometric Theory* **11**, 721–735.

- Singer, H. (1997a), Nonlinear Continuous-Discrete Filtering and ML Estimation using Kernel Density Estimates and Functional Integrals, Regensburger Beiträge zur Statistik und Ökonometrie 40, Universität Regensburg.
- Singer, H. (1999b), Parameter Estimation of Nonlinear Stochastic Differential Equations: Simulated Maximum Likelihood vs. Extended Kalman Filter and Itô-Taylor Expansion, Regensburger Beiträge zur Statistik und Ökonometrie 41, Universität Regensburg.
- Stratonovich, R. (1989), Some Markov methods in the theory of stochastic processes in nonlinear dynamic systems, in F. Moss & P. McClintock, eds, 'Noise in nonlinear dynamic systems', Cambridge University Press, pp. 16–71.
- Tanizaki, H. (1996), *Nonlinear filters: estimation and applications*, second edn, Springer, Berlin.
- Tanizaki, H. & Mariano, R. (1995), Prediction, Filtering and Smoothing in Nonlinear and Non-normal Cases using Monte-Carlo Integration, in H. Van Dijk, A. Monfort & B. Brown, eds, 'Econometric Inference using Simulation Techniques', John Wiley, pp. 245–261.
- Wagner, W. (1988), 'Monte Carlo Evaluation of Functionals of Solutions of Stochastic Differential Equations. Variance Reduction and Numerical Examples', *Stochastic Analysis and Applications* **6**, 447–468.
- Wiggins, J. (1987), 'Option values under stochastic volatility', *Journal of Financial Economics* **19**, 351–372.
- Wolfram, S. (1992), *Mathematica*, 2nd edn, Addison-Wesley, Redwood City.
- Wong, E. & Hajek, B. (1985), *Stochastic Processes in Engineering Systems*, Springer, New York.
- Zadrozny, P. (1988), 'Gaussian likelihood of continuous-time armax models when data are stocks and flows at different frequencies', *Econometric Theory* **4**, 108–124.