Inference methods for discretely observed continuous-time stochastic volatility models: A commented overview

J.C. Jimenez*, R. Biscay* and T. Ozaki †

Abstract

In this paper an overview of inference methods for continuous-time stochastic volatility models observed at discrete times is presented. It includes estimation methods for both parametric and nonparametric models that are completely or partially observed in a variety of situations where the data might be nonlinear functions of the components of the model and/or contaminated with observation noise. In each case, the main reported methods are presented, making emphasis on underlying ideas, theoretical properties of the estimators (bias, consistency, efficient, etc.), and the viability of their implementation to solve actual problems in finance.

Key words: stochastic volatility models, diffusion processes, inference methods

1 Introduction

In the last decade, volatility models of diffusion type have played an important role in better understanding several empirical regularities in various subfields of finance such as derivative securities pricing, term structure of interest rate, asset pricing, optimal consumption and portfolio choices (Ghysels et al. 1996, Sundarresan 2000). Hence statistical inference for such processes has become of great importance in finance not only from the theoretical but also from the practical point of view in model building, estimation, testing and prediction.

The simplest continuous-time stochastic volatility models are the well-known Black-Scholes and Cox-Ingersoll-Ross (CIR) models. They describe the dynamics of an asset price $S$ by the scalar stochastic differential equations

$$dS_t = \alpha S_t dt + \sigma S_t dw,$$

*Instituto de Cibernética, Matemática y Física, Departamento de Matemática Interdisciplinaria, Calle 15, e/ C y D, no. 551, Vedado, La Habana 4, C.P. 10400, Cuba. e-mails: jcarlos, biscay@icmf.inf.cu

†Dept. of Prediction and Control, The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. e-mail: ozaki@ism.ac.jp
and
\[ dS_t = (\alpha S_t + \beta) dt + \sigma \sqrt{S_t} dw, \]
respectively (Black & Scholes, 1973; Cox et al, 1985). Here, \( \alpha, \beta \) and \( \sigma \) are constants. More sophisticated models have also been considered; for instance, the two dimensional time-deformed model (Ghysels et al.1995, 1996; Stock 1988)
\[ ds_t = f_t g_t \tau_t \ dt + \gamma_2 f_t \tau_t d\mathbf{w}_t^2 + \sigma_t \ d\mathbf{w}_t^1, \]
\[ \sigma^2_t = \gamma_1 g_t \tau_t - \gamma_2 f_t^2 \tau_t, \]
\[ d\tau_t = g_t \tau_t \ dt + \gamma_2 \tau_t d\mathbf{w}_t^2, \]
and the tree-dimensional micro-market model (Bouchard & Cont 1998, Iino & Ozaki 2000)
\[ ds_t = \phi_t \sigma^2_t \ dt + \gamma_1 \sigma_t \ d\mathbf{w}_t^1, \]
\[ \sigma^2_t = \exp(\lambda_t), \]
\[ d\phi_t = (\beta_1 + \alpha_1 \phi_t) \ dt + \gamma_2 \ d\mathbf{w}_t^2, \]
\[ d\lambda_t = (\beta_2 + \alpha_2 \lambda_t) \ dt + \gamma_3 \ d\mathbf{w}_t^3. \]
Here, \( s_t = \ln(S_t), f_t = \alpha_1 s_t + \beta_1, g_t = \alpha_2 \log(\tau_t) + \beta_2, \) and \( \alpha_i, \beta_i, \gamma_i \) are constants.

All these models are stochastic volatility models of diffusion type. They can be written in the general form
\[ dS_t = \mu(t, S_t, U_t) dt + \sigma_t d\mathbf{w}_t^1, \]
\[ \sigma^2_t = g(U_t, S_t), \]
\[ dU_t = a(t, U_t) dt + b(t, U_t) d\mathbf{w}_t^2, \]
where \( \mathbf{w}_t = (\mathbf{w}_t^1, \mathbf{w}_t^2) \) is a multidimensional Wiener process, and the functions \( \mu, g, a \) and \( b \) can be nonlinear. In general \( U \) is an unobserved component of the multidimensional diffusion process \( \mathbf{x} = (S, U) \). The function \( g \), called the volatility function, is assumed to be one to one. We will focus attention on this class of models though other kinds of stochastic processes are also subject of current research for modelling volatility; for instance, jump diffusion, fractional, Levy and stable processes.

The present paper deals with methods for estimating stochastic volatility models of diffusion type. It is convenient to classify such methods from several viewpoints. A first useful classification is according the kind of available data. In this respect, a primary division arises from the times at which observations are recorded. This leads to distinguish between continuous-time observations and discrete-time observations. In what follows we will deal only with the latter, which is the common situation faced in practice. In turns, a further classification comes from whether or not all the components of the modelled process are observed and some additional noise is involved in the recorded data. On this basis the following kinds of inference problems can be specified:

**D1)** Estimation of diffusion models from complete observations

**D2)** Estimation of diffusion models from partial observations
D3) Estimation of diffusion models from partial and noisy observations.

The first inference problem (D1) can be defined as follows: given a scalar model for an asset price $S$ (e.g., (1)) and a time series $\{S_{tk}\}$ of observations of $S$, estimate the unknown parameters of the model (i.e., $\alpha$ and $\sigma$). Here, complete observations means that all the components of the diffusion process are observed, which is typically a rare case in multidimensional stochastic volatility models. This is the reason why many of the well known inference methods for diffusion processes find application inside the finance field in relatively few cases such as the estimation of one-dimensional models or the estimation of a set of one-dimensional models with mutual interactions.

The second inference problem (D2) above can be described as follows: given a multidimensional model for an asset price $S$ (e.g., (2)) and a time series $\{S_{tk}\}$ of observations of $S$, estimate the unknown parameters of the model (i.e., $\alpha_i$, $\beta_i$ and $\gamma_i$) and the unobserved components of the diffusion process (i.e., $\sigma_{tk}$).

In both problems (D1) and (D2) it has been assumed than the data $\{S_{tk}\}$ are free from observation noise, which is frequently a non-realistic simplification. For instance, a model that describes the daily evolution of an asset price ignores the well known fluctuations of the price within each business day (Fama 1965, French 1980, Shiryaev 1999). Another limitation of (D1)-(D2) is that they do not allow for nonlinear observations of $S$ (see Brennan & Xia 2005, for an example). These two restrictions can be overcome, in a natural way, into the framework of the so-called continuous-discrete state space models. For this, in addition to equation (4) it is introduced a (possibly nonlinear) discrete-time observation equation, whose noise term reflects the variability of the price at each business day. This poses the third inference problem (D3) which can be defined as follows: given a multidimensional model for an asset price $S$ (e.g., (3)), an observation equation like

$$y_{tk} = h(S_{tk}) + e_{tk}$$

and data $\{y_{tk}\}$, estimate the unknown parameters (i.e., $\alpha_i$, $\beta_i$ and $\gamma_i$) and the unobserved components (i.e., $\phi_{tk}$, $\lambda_{tk}$) of the diffusion process. Here, $h$ is in general a nonlinear function, and $\{e_{tk} : e_{tk} \sim N(0, \Sigma)\}$ is a sequence of independent Gaussian random vectors which are also independent of $w$. Note that problem (D2) can be thought of as included in (D3) just by taking as observation equation the identity map.

A second classification of inference methods is according to the nature of the parameter space. In this sense it is important to distinguish between parametric and nonparametric volatility models. In the former the parameter space has finite dimension while in the latter it is infinite-dimensional, typically a functional space. Although the problems (D1)-(D3) above where stated in terms of parametric models, they can be straightforwardly extended to nonparametric models as well. For instance, estimation from discrete-time complete observations of a scalar nonparametric volatility model can be defined as follows: estimate the unknown drift and diffusion coefficients $f$ and $g$ of the diffusion process $dS_t = f(S_t) + g(S_t)dw$, given a time series $\{S_{tk}\}$ of observations of $S$.

A third classification is according to the way in which prior information about unknown aspects of the model is expressed and used for making inferences. In this respect, two major
classes are bayesian and non-bayesian methods.

A fourth criterion of classification is according to the specific statistical techniques used for obtaining the estimates. It is worthy of note that some of them require complete information about the distribution of the observed processes (e.g., bayesian and maximum likelihood methods) while others need only partial information (e.g., the two first conditional moments or several unconditional moments in some M and Z methods).

Estimation methods of all these classes constitute a rather well-studied topic of statistical inference in case of diffusion processes observed continuously in time (see, e.g., Prakasa Rao 1999). On the contrary, estimation from discrete-time observations is still a challenging subject with many aspects open to future research. A basic difficulty is that, except for a few simple examples, it is not possible to know in closed-form the joint distribution of discrete-time observations. To overcome this, a number of analytical and simulated approximations have been developed in the last years.

In this paper a review of the variety of existing methods for estimating diffusion type stochastic volatility models from discrete-time observations is given. The classification criteria mentioned above are used as guidelines to organize the presentation under a unique conceptual framework. The exposition is focused on three major aspects: 1) main ideas underlying the methods, 2) theoretical properties and practical performance of estimators, and 3) applications in finance. From the theoretical point of view, special emphasis is done on bias, consistency, asymptotic normality and efficiency of the estimators. Whereas preservation of the underlying theoretical properties by the computed estimators as well as availability of feasible and efficient algorithms for actual inference will be the center of attention from the practical point of view. Advantages and drawbacks of alternative methods are discussed, and a number of comments are given with the purpose of aiding applied researchers in the complex task of selecting suitable methods for the problems at hand.

2 Inference methods for diffusion processes from complete observations

For simplicity in the further presentation, it is convenient to use the following more compact notation for a general diffusion volatility model:

\[ dx(t) = a(x(t); \theta)dt + \sum_{i=1}^{m} \sigma_i(x(t); \theta)d\mathbf{w}^i(t) = a(x(t); \theta)dt + \mathbf{\sigma}(x(t); \theta)d\mathbf{w}(t), \text{ for } t \geq t_0 \in \mathbb{R}, \]

where \( x(t) \in \mathbb{R}^d \) is the state vector at the time instant \( t \), \( \theta \) is a set of unknown parameters to be estimated, \( \mathbf{w} = (\mathbf{w}^1,...,\mathbf{w}^m) \) is an \( m \)-dimensional standard Wiener process, \( \mathbf{\sigma} = (\sigma_1,...,\sigma_m) \), and \( a, \sigma_i(\cdot; \theta) : \mathbb{R}^d \rightarrow \mathbb{R}^d \) are nonlinear vector functions for each \( \theta \). We will also denote \( \mathbf{b}(x; \theta) = \mathbf{\sigma}(x; \theta)\mathbf{\sigma}^T(x; \theta) \). Notice that the model is assumed autonomous, though most of the methods to be discussed can be straightforwardly extended to non-autonomous cases.

It will be assumed that the data is a vector \( \mathbf{x}^{obs} = (x_1, ..., x_n) \) of \( n \) consecutive obser-
vations \( x_i \) of \( x \) at discrete times \( t = t_i, i = 1, \ldots, n \). The initial value of \( x \) at time \( t_0 \) is denoted by \( x_0 \), and \( h_i = t_i - t_{i-1} \). For notational simplicity, equidistant observation times are assumed, i.e., \( h_i = h \) for \( i = 1, \ldots, n \).

In this Section all components of the vectors \( x_i \) are supposed to be observed.

### 2.1 Approximate maximum likelihood methods

Due to the Markov property of diffusion processes, (conditional on the initial value \( x_0 \)) the likelihood function associated with the discrete-time observation \( x^{obs} \) for the model (5) can be written as

\[
L_n(\theta) = \prod_{i=1}^{n} p_\theta(h_i, x_{i-1}, x_i),
\]

where \( p_\theta(h_i, x_{i-1}, x_i) \) denotes the transition density function between two consecutive observations, \( x_{i-1} \) and \( x_i \), i.e., the conditional density of \( x_i = x(t_i) \) given \( x_{i-1} = x(t_{i-1}) \).

The most common efficient non-bayesian method of estimation in statistical theory is based on the maximum likelihood approach. The Maximum Likelihood (ML) estimator is defined as

\[
\hat{\theta}_{ML} = \arg \max_{\theta} L_n(\theta).
\]

However, in case of discrete-time observations of a diffusion process this approach has a major drawback: explicit expressions for the transition density \( p \), and so for the likelihood, are available only in a few simple cases (see for instance Lo, 1988). To overcome this, a number of approximation procedures have been proposed, as is described below.

#### 2.1.1 Approximation through the numerical solution of the Kolmogorov Forward Equation

For any given \( x \in \mathbb{R}^d \), the transition density of the diffusion process (5) satisfies the so-called Kolmogorov Forward Equation

\[
\frac{\partial}{\partial t} p_\theta(t, x, y) = -\sum_{i=1}^{d} \frac{\partial}{\partial y_i}(a_i(x, y) p_\theta(t, x, y)) + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2}{\partial y_j \partial y_i} (b_{ij}(y; \theta) p_\theta(t, x, y))
\]

with initial condition \( p_\theta(0, x, y) = \delta(x - y) \), where \( \delta \) is the Dirac delta function.

A natural idea is to approximate the solution of (6) by some numerical method. Poulsen (1999) proposes to use the Crank-Nicholson finite difference method to solve each of the \( n \) Kolmogorov Forward Equations corresponding to \( x = x_i, i = 1, \ldots, n \), in order to obtain approximations to the densities \( p_\theta(h_i, x_{i-1}, \cdot) \) involved in the likelihood function. In this way, an approximation \( \ln L_n^\Delta(\theta) \) to \( \ln L_n(\theta) \) is provided with error \( O(\Delta^2) \), where \( \Delta \) is the size of the \((t, y)\)-grid of the finite difference method. Then, an approximate maximum likelihood estimator is defined as

\[
\hat{\theta}_\Delta = \arg \max_{\theta} \ln L_n^\Delta(\theta).
\]
It is proved that if $\Delta = \Delta(n)$ tends to zero faster than $n^{-1/4}$ as $n \to \infty$ then $\hat{\theta}_{\Delta(n)} \to \hat{\theta}_{ML}$ as $n \to \infty$. The method has been used to fit the so called CKLS short-term interest rate model (Chan et al. 1992) from actual data.

The main drawback of this method is the large computer time required by the numerical solution of $n$ systems of $d$-dimensional partial differential equations for each tentative value of $\theta$ in the process of maximizing $L^\Delta_n(\theta)$ (Jensen & Poulsen, 2002).

### 2.1.2 Analytical approximation based on Hermite-Taylor expansions

Consider first the case of a scalar diffusion model. The main idea behind the method is to construct explicit, non-Gaussian approximations to $p_{\theta}(h, x_{i-1}, x_i)$ by means of a truncated Hermite series expansions, whose coefficients are in turns approximated by truncated Taylor series in $h$. The resulting approximation to $\ln p_{\theta}(h, x_{i-1}, x_i)$ is a (deformed) Gaussian approximation term plus a polynomial in $h$.

For this, first an invertible transformation of $x$ is carried out to a process $z$ with distribution closer to the Gaussian one. More explicitly, for fixed $x_{i-1}$ and $\theta$, the approximation is based on the following steps (Ait-Sahalia 1999, 2002):

1. Transform $x$ to a new process $z$ such that, conditional on $x_{i-1}$, $z(h)$ is approximately $N(0, 1)$ distributed and $z(0) = 0$. This is achieved by setting

   $$ z = g_{x_{i-1}, \theta}(x) = \frac{1}{\sqrt{h}} \int_{x_{i-1}}^{x} \frac{du}{\sigma(u, \theta)}. \quad (7) $$

   Then, the process $z$ satisfies the equation

   $$ dz = c(z, \theta)dt + \frac{1}{\sqrt{h}}dw, $$

   where

   $$ c(z, \theta) = \frac{a(g_{x_{i-1}}^{-1}(z); \theta)}{\sigma(g_{x_{i-1}}^{-1}(z); \theta)} \frac{1}{\sqrt{h}} - \frac{1}{2\sqrt{h}} \frac{\partial}{\partial x} \sigma(g_{x_{i-1}}^{-1}(z); \theta). $$

2. Approximate the transition density $p_{\theta}^z(h, 0, .)$ of the process $z$ starting at $z_0 = 0$ by an Hermite series expansion truncated after $J$ terms, say $p_{\theta}^{z,J}(h, 0, .)$.

3. In turns, approximate the coefficients of the expansion in step (2) by means of a truncated Taylor expansion in powers of $h$.

4. Use the inverse transformation $z \to x$ to approximate $p_{\theta}(h, x_{i-1}, x_i)$ by

   $$ p_{\theta}^{z,J}(h, x_{i-1}, x_i) = \frac{1}{\sqrt{h}\sigma(u, \theta)} p_{\theta}^{z,J}(h, 0, g_{x_{i-1}}(x_i)). $$
Then, the likelihood $L_n(\theta)$ is approximated by $L'_n(\theta) = \prod_{i=1}^{n} p_\theta^i(h, x_{i-1}, x_i)$, and the approximate maximum likelihood estimator is defined as

$$\hat{\theta}_J = \arg\max_\theta L'_n(\theta).$$

Explicit formulae are given for carrying out the computations resulting from these steps leading to $p_\theta^i(h, x_{i-1}, x_i)$ (Ait-Sahalia 1999, 2002). Moreover, it has been proved that if $J = J(n)$ tends to infinite as $n \to \infty$ then $\hat{\theta}_{J(n)} \to \hat{\theta}_{ML}$.

Recently, the method has been extended to time-inhomogeneous diffusion processes (Egorov et al. 2003) and to multidimensional processes (Ait-Sahalia, 2004). In the latter, $\ln p_\theta^i(h, x_{i-1}, x_i)$ takes the form of a Gaussian approximation term plus a polynomial expansion in terms of both $h$ and $x_i - x_{i-1}$.

This method has been used to fit both time-homogeneous and time-inhomogeneous models from simulated data, for instance, the Black-Scholes, CIR and Hull-White models. Theoretical and simulation results have shown that this approach provides accurate estimates with low computer time. However, a major drawback is that, though explicit, the required expressions are very cumbersome to derive. Moreover, enough terms must be included in the Taylor expansions in the step (3) above for moderate values of $h$. Otherwise the accuracy and efficiency of the estimate deteriorate.

### 2.1.3 Approximations via simulation

For any $0 < \Delta < h$, the Chapman-Kolmogorov equation states

$$p_\theta(h, x_{i-1}, x_i) = \int p_\theta(h - \Delta, x_{i-1}, y)p_\theta(\Delta, y, x_i)dy = E(p_\theta(\Delta, X(t_i - \Delta), x_i) / x_{i-1}),$$

where $E(. / x_{i-1})$ is the expected value with respect to the random variable $X(t_i - \Delta)$ conditional to $X(t_{i-1}) = x_{i-1}$. This suggests the following algorithm (Pedersen 1995a): for $i = 1, \ldots, n$,

1. Split each interval $[t_{i-1}, t_i]$ between two consecutive observations into $N + 1$ subintervals of length $\Delta_N = h / (N + 1)$, where $N$ is a given integer;

2. Approximate $p_\theta(h, x_{i-1}, x_i)$ by

$$p_\theta^N(h, x_{i-1}, x_i) = E(\tilde{p}(\Delta_N, \tilde{x}(t_i - \Delta_N), x_i) / x_{i-1}),$$

where $\tilde{x}(t_i - \Delta_N)$ is the approximation of $x(t_i - \Delta_N)$ obtained by means of the Euler-Maruyama integrator with step-size $\Delta_N$ and initial point $\tilde{x}(t_{i-1}) = x_{i-1}$. Here, $\tilde{p}(\Delta_N, \tilde{x}(t_i - \Delta_N), x_i)$ is the Gaussian transition density induced by such numerical method.
3. Approximate the expected value in the last expression by Monte Carlo, which results in

\[ p_{\theta}^{N,M}(h, x_{i-1}, x_i) = \frac{1}{M} \sum_{m=1}^{M} \tilde{p}(\Delta_N, \tilde{x}^m(t_i - \Delta_N), x_i), \]

where \( \tilde{x}^m(t_i - \Delta_N) \) are independent simulations of \( x(t_i - \Delta_N) \) by the Euler-Maruyama method with step-size \( \Delta_N \) and initial point \( \tilde{x}^m(t_{i-1}) = x_{i-1} \).

Then, by using \( p_{\theta}^{N}(h, x_{i-1}, x_i) \) as transition density one obtains the approximate likelihood

\[ L_n^N(\theta) = \prod_{i=1}^{n} p_{\theta}^{N}(h, x_{i-1}, x_i), \]

and so the approximate maximum likelihood estimator

\[ \hat{\theta}^N = \arg \max_{\theta} L_n^N(\theta). \]

In a similar way, \( p_{\theta}^{N,M}(h, x_{i-1}, x_i) \) defines an approximate likelihood \( L_n^{N,M}(\theta) \) and a corresponding estimator \( \hat{\theta}^{N,M} \).

It has been proved that there exists a sequence \( N(n) \to \infty \) such that \( \hat{\theta}^{N(n)} \to \theta_{ML} \) in probability as \( n \to \infty \), i.e., such that the approximate maximum likelihood estimator \( \hat{\theta}^{N(n)} \) is asymptotically equivalent to the ML estimator (Pedersen 1995b). On the other hand, the additional Monte Carlo step 3) above provides an approximation \( p_{\theta}^{N,M}(h, x_{i-1}, x_i) \) to \( p_{\theta}^{N}(h, x_{i-1}, x_i) \) with error of size \( O\left(\frac{1}{N \sqrt{M}}\right) \).

However, in the step 3 above the simulated value \( \tilde{x}^m(t_i - \Delta_N) \) at time \( t_i - \Delta_N \) can be far from the observed value \( x_i \) at time \( t_i \), which causes an unlikely large jump in the simulated process and so the need for many simulations to obtain acceptable convergence. To overcome this, a modification of this step has been proposed on the basis of an importance sampling algorithm conditioning on both \( x_{i-1} \) and \( x_i \) (Elerian et al. 2001). This variant allows for the reduction of the number \( M \) of simulations but at the expense of computing the mode of \( \log p_{\theta}(x(t_i - \Delta_N), x_i / x_{i-1}) \) as a function of \( x(t_i - \Delta_N) \) and its Hessian evaluated at the mode.

In general, large numbers of subintervals and of simulated trajectories, \( N \) and \( M \), are necessary in order to make these approximations accurate enough. Nevertheless, the method has been successful applied to the estimation of the CKLS model (Honore, 1997).

In Hurn et al. (2003) a related approach for approximating the likelihood via simulation is presented, which is based on kernel density estimation.

In Jansen & Poulsen 2002 a comparative study between Maximum Likelihood Methods have been carried out. Their conclusions are: 1) Analytical Approximation Method and Kolmogorov Forward Equation Method are much faster than Approximation via Simulation; and 2) Kolmogorov Forward Equation Method is less accurate than the other two approximate maximum likelihood methods.
2.2 Bayesian methods

According to the general Bayesian inference approach, a Bayesian estimator of the parameter $\theta$ can be obtained by a simple average of values of $\theta$ sampled from the posterior probability distribution

$$p(\theta / x^{\text{obs}}) = \frac{L_n(\theta) p(\theta)}{\int L_n(\theta) p(\theta) d\theta},$$

where $L_n(\theta)$ is the likelihood function and $p(\theta)$ is a specified prior distribution for $\theta$.

Several estimators of this kind were proposed early in the literature, but their use have been limited to certain particular kinds of diffusion processes for which $L_n(\theta)$ is explicitly given or easy to approximate (see Prakasa Rao, 1999 for details). For general diffusion models, more sophisticated approximate methods must be applied in order to sample from said posterior distribution. Below we shall briefly present the promising methodology that has been recently proposed for this purpose on the basis of Markov Chain Monte Carlo algorithms (Eraker 2001, Elerian 2001, Robert & Stramer 2001).

2.2.1 Markov Chain Monte Carlo (MCMC) methods

The main idea behind this method is to use MCMC algorithms for sampling $\theta$ from a convenient approximation to the posterior distribution $p(\theta / x^{\text{obs}})$.

For this purpose, introduce the following intermediate (unobserved) values of the process $x$ between two consecutive observations $x_{i-1}$ and $x_i : x_{i,k} = x(t_{i-1} + k\Delta_N)$, $k = 1, ..., N$, where $N \in \mathbb{N}$ is a given integer and $\Delta_N = h/((N + 1)$. For all $i$, denote also $x_{i,0} = x(t_{i-1}) = x_{i-1}$, $x_{i,N+1} = x(t_i) = x_i$ and $\bar{x} = \{\bar{x}_1, ..., \bar{x}_n\}$, where $\bar{x}_i = (x_{i,1}, ..., x_{i,N})$ is the vector of unobserved values of $x$ between $x_{i-1}$ and $x_i$. For $\Delta_N$ small enough, the transition density $p(\Delta_N, x_{i,k}, x_{i,k+1} / \theta)$ between two consecutive points $x_{i,k}, x_{i,k+1}$ is approximated by the Gaussian density $\tilde{p}(\Delta_N, x_{i,k}, x_{i,k+1} / \theta)$ induced by the Euler-Maruyama approximation with step-size $\Delta_N$.

The intermediate points $\bar{x}$ are regarded as missing values in the Bayesian approach. The marginal distribution of the distribution $p(\theta, \bar{x} / x^{\text{obs}})$ provides the desired posterior distribution $p(\theta / x^{\text{obs}})$, so the problem is reduced to sample from $p(\theta, \bar{x} / x^{\text{obs}})$. This is carried out by means of MCMC methods, i.e., by constructing a Markov chain $(\bar{x}^j, \theta^j) (j = 1, 2, ...)$ whose limiting density is $p(\theta, \bar{x} / x^{\text{obs}})$. Specifically, the Gibbs sampling proceeds as follows. The $j$-th iteration is conducted in two steps:

S1) Sample $\bar{x}^j$ from the density $p(\bar{x} / x^{\text{obs}}, \theta^{j-1})$

S2) Sample $\theta^j$ from the density $p(\theta / x^{\text{obs}}, \bar{x}^j)$.

To start the Markov chain, $\theta^0$ can be sampled from the prior distribution $p(\theta)$ and $\bar{x}^0$ can be defined by linear interpolation between the observations $x_i$. Finally, the Bayesian estimator is computed as

$$\hat{\theta} = \frac{1}{J - J_0} \sum_{j=J_0}^J \theta^j$$

after some burn-in time $J_0$ of the Markov chain.
Steps $S1$ and $S2$) can not be carried out exactly because the sampling distributions are unknown explicitly. This is overcome by applying the Metropolis-Hasting algorithm in combination with the Gaussian approximation. Specifically, notice that the Markov property of the diffusion implies that

$$p(\bar{x}/x^{obs}, \theta) = \prod_{i=1}^{n} p(\bar{x}_i/x_{i-1}, x_i, \theta).$$

In turns, the factors in this expression are approximated as

$$p(\bar{x}_i/x_{i-1}, x_i, \theta) \approx p^N(\bar{x}_i/x_{i-1}, x_i, \theta),$$

where

$$p^N(\bar{x}_i/x_{i-1}, x_i, \theta) \propto p^N(\bar{x}_i, x_{i-1}, x_i/\theta) = \prod_{k=0}^{N} \tilde{p}(\Delta N, x_{i,k}, x_{i,k+1}/\theta).$$

Here $\propto$ means 'is proportional to'. Since in general the normalizing constant in $p^N(\bar{x}_i/x_{i-1}, x_i, \theta)$ can not be found, the Metropolis-Hasting algorithm is used to sample from $p^N(\bar{x}_i/x_{i-1}, x_i, \theta)$ on the basis of $p^N(\bar{x}_i, x_{i-1}, x_i/\theta)$. That is, given some proposal distribution $q_1(x_i/\bar{x}_{i-1})$ (e.g., some Gaussian proposal), step $S1$ is substituted by a Metropolis-Hasting step. In a similar fashion, given some proposal $q_2(. / \theta_{i-1})$ for $\theta$ (e.g., the prior distribution $p(\theta)$), step $S2$ above is replaced by a corresponding Metropolis-Hasting step.

Several variants for implementing this method are discussed in Eraker 2001, Elerian 2001, and Robert & Stramer 2001. See also these papers for applications of the method to the estimation of various interest rates models from actual and simulated data.

The main difficulty with this bayesian approach is that it is very computing time demanding. Furthermore, convergence issues of the involved Markov chains have not been theoretically studied so they are only empirically checked in each particular case.

### 2.3 M and Z methods

M estimators are obtained by maximizing a suitable function of the data and parameters, while Z estimators are obtained by computing the zeros of a certain function of the data and parameters (van der Vaart, 1998). Under mild regularity conditions both approaches are equivalent. A simple example is the maximum likelihood estimator under standard assumptions. This section deals with other types of M and Z estimators that do not require complete information on the transition distribution of the diffusion process $x$.

#### 2.3.1 Pseudo-likelihood methods

This kind of estimators are derived by maximizing the likelihood $L_n^h(\theta)$ of a discrete process $\tilde{x}$ that approximates the diffusion process $x$. That is,

$$\hat{\theta} = \arg \max_{\theta} L_n^h(\theta),$$
where \( L^h_n(\theta) = \prod_{i=1}^{n} p^h_\theta(h, x_{i-1}, x_i) \) and \( p^h_\theta(h, x_{i-1}, x_i) \) is the transition density function from \( \tilde{x}_{i-1} \) to \( \tilde{x}_i \) evaluated at the observations \( x_{i-1} \) and \( x_i \).

Typically \( \tilde{x} \) is taken as the discretization of (5) obtained through a numerical integrator with step-size \( h \). The numerical integrators more commonly used are the Euler-Maruyama scheme (Prakasa Rao 1983, Yoshida 1992, Florens-Zmirou 1989) and the Local Linearization (LL) schemes (Ozaki 1992, Shoji-Ozaki 1997, 1998). In both cases, the mean \( \mu^h_\theta(x_{i-1}) \) and variance \( \sigma^h_\theta(x_{i-1}) \) of the Gaussian density \( p^h_\theta(h, x_{i-1}, x_i) \) are known functions induced by the conditional moments of the discretization.

For moderate values of \( h \), a comparative study carried out by simulations (Shoji & Ozaki 1997) shows that the pseudo-likelihood estimators based on LL schemes provide better results than those based on the Euler-Maruyama method and a Generalized Method of Moments (see next subsection).

In particular, if the diffusion term does not depend on \( \theta \) and the Euler-Maruyama discretization is used, the variance \( \sigma^h_\theta(x_{i-1}) \) does not depend on \( \theta \), and the resulting M estimator is called the least squares estimator.

For fixed \( h \), pseudo-likelihood estimators based on numerical integrators are inconsistent (Florens-Zmirou 1989). However, this does not mean that they have no usefulness. In practical situations in which the bias is small enough (usually for small \( h \)) they are a convenient choice due to their simplicity and computational efficiency (Durham & Gallant 2002, Singer 2002). Even more, in the worst situation, when the bias introduced by the discretization can not be disregarded, such estimators can be used as initial values in the iterative computation of less biased M or Z estimators by means of more sophisticated methods.

For the pseudo-likelihood method based on the Euler-Maruyama integrator, Clement 1995 introduced a correction for the resulting bias \( \hat{\theta} - \theta \) by means of simulations. Alternatively, on the basis of analytical expansions in powers of \( h \), Kessler 1997 proposed corrections to \( \mu^h_\theta(x_{i-1}) \) and \( \sigma^h_\theta(x_{i-1}) \) that improve their approximations to \( E_\theta(x_i/x_{i-1}) \) and \( V_\theta(x_i/x_{i-1}) \), respectively.

Pseudo-likelihood methods based on LL schemes have been applied to the estimation of some models for short-term interest rate from actual data (Shoji & Ozaki, 1996).

### 2.3.2 Methods of the Moments

The generalized method of the moments defines an estimators of \( \theta \) as

\[
\hat{\theta} = \arg \min_\theta R_n(\theta),
\]

where \( R_n(\theta) = G_n(\theta)\Omega G_n(\theta) \) is quadratic form, \( G_n(\theta) \) is a function of the data and \( \Omega \) is a (definite non-negative) weight matrix. Typically,

\[
G_n(\theta) = m_n - M(\theta),
\]
which matches the empirical moments

\[ m_n = \frac{1}{n} \sum_{i=1}^{n} g(x_{i-1}, x_i) \]

corresponding to some vector function \( g \) with their associated population moments

\[ M(\theta) = E_\theta (g(x_{i-1}, x_i)) . \]

In most models the population moments \( M(\theta) \) are not known exactly and are difficult to approximate analytically. The simulated method of the moments (Clement 1997) overcomes this by replacing \( M(\theta) \) by its approximation through simulation, i.e.,

\[ M_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} g(\tilde{x}_{i-1}(\theta), \tilde{x}_i(\theta)). \]

Here \( \tilde{x}_i(\theta) \) are simulated values of the process \( x \) at a large number \( N > n \) of times \( t = t_i, i = 1, ..., N \), obtained by the numerical integration of the model with parameter value \( \theta \).

Other well known estimators of this class that also involve simulated data are the indirect inference method (Gourieroux et. al 1993) and the efficient method of the moments (Gallant & Tauchen 1996). The main idea behind these methods is to ‘project’ the original model onto some more simple auxiliary model described by an auxiliary vector parameter \( \beta \). In these approaches the contrast function \( G_n(\theta) \) is defined on the basis of a function \( S \) that is usually taken as the log-likelihood function under the auxiliary model. For instance, in case of the indirect inference method

\[ G_n(\theta) = \hat{\beta} - \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}^i(\theta), \]

where

\[ \hat{\beta} = \arg \max_\beta S(\mathbf{x}^{obs}, \beta), \]

\[ \hat{\beta}^i(\theta) = \arg \max_\beta S(\tilde{x}_i(\theta), \beta), \]

\( \tilde{x}_i(\theta) \) is the \( i \)-th simulated trajectory of \( x \) computed by means of a numerical integrator (i.e., the Euler-Maruyama scheme), and \( N \) is the number of simulations. In the efficient method of the moments,

\[ G_n(\theta) = \sum_{i=1}^{N} \frac{\partial}{\partial \beta} S(\tilde{x}_i(\theta), \beta). \]

These estimators are in general consistent and asymptotically normal. However, for finite samples, the quality of the estimators strongly depends on an appropriate selection
of the contrast function \( G_n(\theta) \). Moreover, the indirect inference method and the efficient method of the moments require a suitable selection of the auxiliary function \( S \) and the parameter \( \beta \) (Tauchen 1986, Ferson & Foerster 94, Hansen et. al. 1996). Unfortunately there is no practical or theoretical guidelines to carried out this task. Closeness of \( S \) to the, in general unknown, true likelihood is needed to yield efficiency. In addition, such methods that involve simulations are very computing time demanding because a large number of simulations is needed to achieve negligible bias.

Methods of the Moments are quite popular in financial studies mainly because they can also be applied to the case of partial observations, which is common in stochastic volatility models. Examples are the estimation of interest rates and asset price models (Duffie & Singleton 1993, Hansen et al. 1995; Gallant & Tauchen 1997, 1998; Gourieroux & Monfort 1996, etc.).

### 2.3.3 Quasi-likelihood methods

The central idea behind quasi-likelihood methods is to obtain an estimator \( \hat{\theta} \) that combines the advantages of maximum likelihood estimation and weighted least squares estimation with minimum loss of efficiency. The quasi-likelihood estimator \( \hat{\theta} \) is obtained by solving the equation

\[
F_n(x^{obs}, \theta) = 0,
\]

where \( F_n \) is a so called optimal estimating function (within some given class of estimating functions). Specifically, \( F_n \) is required to satisfy the following two main conditions: 1) \( E_{\theta_0}(F_n(x^{obs}, \theta)) = 0 \) if and only if \( \theta = \theta_0 \), and 2) \( F_n(x^{obs}, \theta) \) maximize the Estimating Information (EF) criterion within the given class of estimating functions. See Heidy 97 for details.

The main classes of optimal estimating functions for diffusion processes are: martingales (Bibby & Sorensen 1995, 1996, 1997), simple estimating functions (Kessler 2000, Jacobsen 2001, Sorensen 2001, Bibby & Sorensen 2001) and prediction-based estimating functions (Sorensen 2000). All them have the general form

\[
F_n(x^{obs}, \theta) = \sum_{i=1}^{n} f(x_{i-1}, x_i, \theta).
\]

In martingale optimal estimating functions, \( f \) has the form

\[
f(x_{i-1}, x_i, \theta) = \sum_{j=1}^{J} \alpha_j(x_{i-1}, \theta) h_j(x_{i-1}, x_i, \theta),
\]

where \( h_j(x_{i-1}, x_i, \theta) \) is a scalar martingale with respect the filtration generated by the data, i.e., \( E_{\theta}(h_j(x_{i-1}, x_i, \theta)/x_{i-1}) = 0 \) for all \( \theta \), and the vectors \( \alpha_j(x_{i-1}, \theta) \) are the associated optimal weight functions (obtained by maximizing the EF criteria). Most martingale estimating functions used in practice are obtained by setting

\[
h_j(x_{i-1}, x_i, \theta) = g_j(x_i, \theta) - E_{\theta}(g_j(x_i, \theta)/x_{i-1}),
\]
where \( g_j(x_i, \theta) \) are given scalar functions with finite expectation. For instance, \( g_j(x_i, \theta) \)
may be polynomials in \( x_i \). In particular, they include martingale estimating functions that are linear (Bibby & Sorensen 1995), quadratic (Bibby & Sorensen 1996, 1997) and those based on eigen-functions of the differential operator associated with the diffusion process (Kessler & Sorensen 1999). The latter are available in explicit form only for some simple models. An important class of linear estimating functions is obtained from discretization of the continuous-time score function and adjustment by its compensator (Bibby & Sorensen 1995).

Consistency and asymptotic normality of estimators based on martingale estimating functions are directly derived from the asymptotic theory of martingales under mild conditions.

In case of prediction-based estimating functions both \( h_j \) and \( \alpha_j \) can depend also on \( x_k, k \leq i - 1 \). Specifically, for \( 1 \leq i \leq n \), \( 1 \leq j \leq N \), let \( P_{i,j} \) be a closed linear subspace of the \( L^2(P_\theta) \)-space of square integrable (scalar) random variables that are measurable with respect to the \( \sigma \)-algebra of events generated \( x_1, ..., x_i \). Then, the prediction-based estimating function corresponding to these subspaces are defined by setting

\[
h_j(x_1, ..., x_i, \theta) = g_j(x_i) - \pi_{i-1}^j(\theta),
\]

where \( g_j \) are given functions and \( \pi_{i-1}^j(\theta) \) is the minimum mean square predictor of \( g_j(x_i) \) in \( P_{i-1,j} \). In practice, \( P_{i-1,j} \) is taken as the finite-dimensional linear space generated by a given set of functions \( r_{ijk}(x_1, ..., x_{i-1}), k = 0, ..., q_{ij} \), where \( h_{ij0} \equiv 1 \). That is, \( \pi_{i-1}^j(\theta) \) is a linear predictor in the variables \( r_{ijk}(x_1, ..., x_{i-1}) \) and 1. In this approach, in contrast with previous ones, the optimal estimating function depends only on unconditional moments which makes the computation of the estimators more simple.

In the method of simple estimating functions, \( f(x_{i-1}, x_i, \theta) \) is defined as a function of only \( x_i \), say

\[
f(x_{i-1}, x_i, \theta) = f(x_i, \theta).
\]

This function is usually constructed ad hoc by trying to mimic the score function of the invariant distribution of \( x \). The resulting estimator does not allow for the estimation of parameters that are not identifiable by the (marginal) stationary distribution of the process. For this reason the method is applicable only to situations in which the data are stationary and the relevant parameters can be determined by the stationary distribution.

In general, by construction, quasi-likelihood estimators are consistent, asymptotically normal and efficient within a given class of estimating functions. But the specification of such a class is in general made ad hoc, and so its efficiency may be quite poor in comparison with maximum likelihood estimation. Furthermore, in general the optimal estimating function \( F_n \) must be approximated because there are no closed expressions for \( \alpha_j \) and/or \( h_j \). In this respect few works have been made. Approximation through simulation for some particular classes of estimating functions has been proposed (Bibby & Sorensen 1995, Kessler & Paredes 2002). Therefore, up to now, it is not clear in general how to compute approximate quasi-likelihood estimators that preserve the properties of
their theoretical counterparts and how much loss of efficiency could be involved in such approximations.

Applications in finance have been limited to theoretical studies of some models.

2.4 Nonparametric methods

In nonparametric modeling, no parametric forms are given for the unknown coefficients of the Stochastic Differential Equations (SDEs) but only some mild smoothness conditions are assumed.

Consider an scalar SDE \( dx = a(x,t)dt + \sigma(x,t)dw \). In case that \( a \) [resp. \( \sigma \)] is a known function, this equation can be transformed into a pure additive SDE \( dy = f(y,t)dt + dw \) (whitening) [resp. a pure multiplicative SDE \( dy = g(y,t)dw \)], where \( f \) [resp. \( g \)] is an unknown function to be estimated. A different situation is faced when both the drift and diffusion coefficients are unknown. All these cases will be considered below. For simplicity the presentation is restricted to scalar diffusion models but most of the results are also valid for multidimensional diffusions.

Notice that, in contrast with parametric estimators, in order to achieve consistency all nonparametric estimators require that the time interval \( h = h(n) \) between observations tends to zero as the sample size \( n \) increases.

2.4.1 Estimation of the drift coefficient

For diffusion models with known diffusion coefficient \( \sigma \) and stationary data, the nonparametric estimator

\[
\hat{a}(x) = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\lambda_i} K_2^{(1)}((x - x_i)/\lambda_i) \\
\varepsilon + \sum_{i=1}^{n} K_1((x - x_i)/\lambda_i)
\]

was proposed in Banon (1977). Here, \( K_1 \) and \( K_2 \) are kernels functions, \( K_2^{(1)} \) is the derivative of \( K_2 \), \( \varepsilon \) is a fixed positive number and \( \lambda_i \) is a decreasing sequence of positive numbers. This estimator is based on the fact that

\[
a(x) = \frac{1}{2\pi(x)} \frac{d}{dx} (b(x) \pi(x)),
\]

where \( \pi \) denotes the stationary distribution of \( x \) and \( b = \sigma^2 \).

A similar kernel estimator can be derived from the well known property \( a(x) = \lim_{t \to 0} E((x_t - x)/x) \) (Nguyen & Tuan 1982).

Computational details and properties of these two estimators are reviewed in Prakasa Rao 1999.

Alternatively, by using the identity

\[
da^{(i)}(x) = a^{(i+1)}(x)dx + \frac{b}{2} a^{(i+2)}(x)dt
\]
for the derivatives $a^{(i)}$ of $a$, the nonparametric estimator

$$
\hat{a}(x) = \sum_{i=0}^{n} \hat{a}^{(i)}(x_i) \frac{(x - x_i)^i}{k!}
$$

(9)

was proposed in Shoji (2000). Here, the coefficients $\hat{a}^{(i)}(x_i)$ are obtained by means of the Kalman filter algorithm. Asymptotic properties of this estimator have not been studied.

2.4.2 Estimation of the diffusion coefficient

For additive SDE with diffusion coefficient $\sigma$, the estimator

$$
\hat{b}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - x_{i-1})(x_i - x_{i-1})^T
$$

is directly derived from the well known property $b(x) = \lim_{t \to 0} E((x_t - x)(x_t - x)^T/x)$.

For diffusion models with known drift coefficient $a$ and stationary data, the identity

$$
b(x) = \frac{2}{\pi(x)} \int_{0}^{x} a(u)\pi(u)du
$$

allows for the approximation of the diffusion coefficient $b$ as

$$
\hat{b}(x) = \frac{2}{\hat{\pi}(x)} \int_{0}^{x} a(u)\hat{\pi}(u)du,
$$

where

$$
\hat{\pi}(x) = \frac{1}{n} \sum_{i=1}^{n} K((x - x_i)/\lambda)
$$

is the the kernel estimator of the stationary distribution $\pi$ (Prakasa Rao, 1999).

See also Florens-Zmirou (1993) and Jacod (2000) for other kinds of kernel estimators that can be obtained from approximations to the local time of $x$ in $x_0$ during $[0, t]$,

$$
L_t(x_0) = \lim_{\delta \to 0} \frac{1}{\delta} \int_{0}^{t} 1_{\{|x(s) - x_0| < \delta\}} ds,
$$

(10)

Consistency and asymptotic normality of such estimators have been proved.

In case of known drift coefficient $a$ and additive noise, i.e., $dx = a(x)dt + b(t)dw$, estimators of $b$ based on wavelet expansions have been proposed in Genon-Catalot et al., 1992, and Hoffmann, 1999.
2.4.3 Estimation of both the drift and diffusion coefficients

Eplett 1987 derived kernel estimators for \( a \) and \( b \) by approximating the continuous time likelihood function of \( x \) through a kernel function \( K \) (see also Prakasa Rao 1999). In Jiang & Knight 1997 other kinds of kernel estimators for \( a \) and \( b \) were obtained by taking into account the properties (8), (10), and the kernel approximation for \( \frac{1}{\pi(x)} \frac{d}{dx} \pi(x) \) introduced by Banon & Nguyen 1981. It has been proved that these kernel estimators are consistent and asymptotically normal.

For SDEs with additive noise, in addition to the estimator (9) for \( a \), a maximum likelihood estimator for \( b \) was proposed in Shoji 2002. Statistical properties of these estimators remain to be studied.

Finally, we refer to Fan and Zhang 2003 and references therein for further nonparametric estimators of (scalar) diffusion models.

Applications in finance of nonparametric estimators include analysis of the prices and interest rates of derivative securities (Ait-Sahalia 1996, Jiang 1998, Jiang & Knight 1997).

3 Inference methods for diffusion processes from partial observations

As it was early mentioned in the Introduction, only one component of multidimensional stochastic volatility models is usually observed. This, obviously, complicates the inference problem to be solved. In this kind of situation the methods of the moments, as well as the bayesian and the quasi-likelihood methods have been used. However, it is not clear whether the other inference methods described in the previous Section can be applied.

Alternatively, in case of simple two-dimensional models some particular methods have been proposed. We just refer to recent works in Genon-Catalot et. al. 1999, 2000 and in Sorensen 2003.

On the other hand, the case of partial observations without measurement noise can be thought of as a limit case of partial and noisy observations which shall be dealt with in the next section.

4 Inference methods for diffusion processes from partial and noisy observations

This section deals with situations in which observations \( x_i \) of the diffusion process \( x \) described by (5) are not available but only data \( y_i \) obtained according to an observation equation

\[
y_i = h(x_i) + e_i.
\]

Here, \( h \) is a (possibly nonlinear) known function, and \( e_i, \ i = 1,\ldots,n \), are independent random vectors with Gaussian distribution \( N(0, \Sigma) \). The dimension of \( y_i \) can be less than that of \( x \). That is, \( y_i \) are noisy and possibly partial observations of the process \( x \). Equations (5)-(11) describe what is called a continuous-discrete state space model.
4.1 Maximum likelihood methods

In particular, for linear continuous-discrete state space models (with additive noise) the equations (5)-(11) have the form

\[ dx = Ax \, dt + Bdw \]
\[ y_k = Cx(t_k) + e_{tk} \]

where \(A\), \(B\) and \(C\) are constant matrices.

For this model the maximum likelihood estimator of \(\vartheta = (A, B)\) is

\[ \widehat{\vartheta}_{ML} = \arg \min_{\vartheta} \{ N \ln(2\pi) + \sum_{k=1}^{N} \ln(\det(\Sigma_{t_k}/t_{k-1})) + \nu_{t_k}^\top (\Sigma_{t_k}/t_{k-1})^{-1} \nu_{t_k} \}, \]

where \(\nu_{t_k}^\top = y_k - Cx_{t_k}/t_{k-1}\) is the discrete-time innovation with variance \(\Sigma_{t_k}/t_{k-1} = CP_{t_k}/t_{k-1}C^\top + \Sigma\) (Schweppve 65). Here, \(x_{t_k}/t_{k-1} = E_{\vartheta}(x(t_k)/Y_{t_k-1})\) and \(P_{t_k}/t_{k-1} = E_{\vartheta}((x(t_k) - x_{t_k}/t_{k-1})(x(t_k) - x_{t_k}/t_{k-1})^\top/Y_{t_k-1})\), where \(E_{\vartheta}(\cdot/.)\) denotes conditional expectation and \(Y_{t_k} = \{y_{t_j} : t_j \leq t_k\}\) are the observations up to \(t_k\).

From a practical point of view these estimators can be efficiently computed by using the Kalman filter (Schweppve 1965) or EM algorithms (Singer 1993). Consistency and asymptotic normality of maximum likelihood estimators in general state space models has been recently proved in (Jensen & Pedersen 1999).

Unfortunately, this explicit form for \(\widehat{\vartheta}_{ML}\) is limited to linear state space models with additive noise or to models that can be reduced to the former one (by using the transformation (7), for instance).

In Duffee & Stanton 2004 the method has been applied to the estimation of some one-factor and two factors term structure models from simulated data. A comparative study carried out in that paper shown that the maximum likelihood estimator based on Kalman filter performs much better than the efficient method of the moments and the approximate maximum likelihood via simulations method.

4.2 M and Z methods

4.2.1 Prediction-based estimating functions

For general continuous-discrete state space models (5)-(11), the prediction-based estimating function \(F_n\) is defined as in section 2.3.3 but now the optimal weights \(\alpha_j(x_1, ..., x_{i-1}, \theta)\) are obtained by considering that the each observation \(y_i\) satisfies the observation equation (11) (Nolsoe et al., 2000). The resulting quasi-likelihood estimator has the same statistical and computational properties as the prediction-based estimators based on noise-free observations.

Applications in finance have been limited to simulations with the CIR model (Nolsoe et al., 2000).
4.2.2 Innovation method

For general state space models (5)-(11), the innovation estimator $\hat{\vartheta}$ is defined by

$$\hat{\vartheta} = \arg\min_{\vartheta} Q(\vartheta),$$

where

$$Q(\vartheta) = n \ln(2\pi) + \sum_{i=1}^{n} \ln(\det(\Sigma_{t_i/t_{i-1}})) + \nu_{t_i}^\top(\Sigma_{t_i/t_{i-1}})^{-1}\nu_{t_i};$$

$\nu_{t_i} = y_{t_i} - E(h(x(t_i))/Y_{t_{i-1}})$ is called the discrete-time innovation, and $\Sigma_{t_k/t_{k-1}}$ is the variance of $\nu_{t_k}$ (Ozaki 1994, Jimenez & Ozaki 2002).

This approach is motivated by the following appealing properties of the innovations:

1. $\nu_{t_1}, \ldots, \nu_{t_n}$ are uncorrelated random vectors no matter the step size $h$.

2. In case of a continuous-time observation equation $y(t) = h(x(t)) + \xi(t)$, where $\xi(t)$ is a white noise, the continuous-time innovation $\{\nu(t) : t \geq 0\}$ defined by

$$\nu(t) = y(t) - E(h(x(t))/\{y(s) : s \leq t\})$$

is a white noise.

3. The discrete-time innovation $\nu_{t_i}$ converges to the continuous-time innovation $\nu(t_i)$ as $h \to 0$.

Therefore, $\nu_{t_1}, \ldots, \nu_{t_n}$ are approximately Gaussian and independent random vectors for $h$ small enough.

For all $h$, consistency and asymptotic normality of innovation estimators are directly obtained from the asymptotic theory of prediction error methods (Ozaki 1994, Nolsoe et al. 2000).

From a practical point of view, recursive filters like local linearization filters (Ozaki 1994, Shoji 1998, Jimenez & Ozaki 2003), extended Kalman filter (Nielsen et al. 2000, Singer 2002), and second order filters (Nielsen et al. 1998, 2000, Singer 2002) have been used for approximating the values of $\nu_{t_k}^\top$ and $\Sigma_{t_k/t_{k-1}}$. Alternative, simulated filters like particle filters (del Moral, 2001) might be used, but with a substantive increase of the computational cost. For fixed $h$, the approximate innovation estimates obtained in such a way are biased but the bias is, in general, negligible in a number of practical situations (see Jimenez & Ozaki 2002, and references therein).

A remarkable characteristic of all these approximate innovation methods is that the fitness of the model to actual data can be easily evaluated by testing whether the distribution of the fitted-innovation deviates from a Gaussian distribution (Ozaki et al. 2000, Jimenez & Ozaki 2002).
In a simulation study with the CIR model, approximate innovation methods provided similar or better results than those obtained by prediction-based estimating functions but with much lower computational cost (Nielsen et. al, 2000). Similar results have been reported in a comparative study with the approximate likelihood via simulation method (Singer, 2002).

The innovation method has been applied for statistical analyses of the CIR model (Brigo & Hanzon 1998, Geyer & Pichler 1999, Nielsen et al. 2000), the Black-Scholes-Coutadon model for the US stock market (Nielsen et. al, 2000), the volatility structure of LIBOR markets (Chiarella et al. 2005), as well as the micro-market model and a time-deformed model for the currencies rates market (Ozaki & Iino 2001, Ozaki et al. 2001, Ozaki & Jimenez 2002).

5 Discussion

As it have been shown in the previous sections there is a wide variety of inference methods for continuous-time stochastic volatility models observed at discrete times.

From the theoretical point of view, all of them provide consistent and asymptotically normal estimators except pseudo-likelihood methods based on numerical integrators for fixed step-size $h$. Although it is desirable to use estimation techniques that have good asymptotic properties, their finite-sample properties are more important in actual applications (Duffee & Stanton 2004).

Table I summarizes several relevant features of the methods that have been reviewed in this paper. It shows that, in realistic situations, no method is the best with respect to all criteria but there is a clash between statistical precision and cost to compute the estimate.

In practice, many factors affect the quality of the estimators: the number of observations $n$, the time interval $h$ between observations, the number of simulations $M$ required by some methods, the truncation errors induced by the finite precision of numerical computations, the convergence of the optimization (resp. sampling or continuation) algorithm used to compute $M$ (resp. bayesian or Z) estimators, etc.

Thus, the selection of the convenient inference method to estimate a particular model from actual data is not an easy task. For this purpose an intuitive and reasonable criterion is the following: the inference method should provide suitably accurate estimates from finite samples with lowest computational cost (Singer 2002). According to this the following strategy seems to be a reasonable proposal for a large number of practical situations: 1) Generate data that resemble the specific recording conditions of the actual data (number of observations, time interval between observations, etc). This can be done by using numerical schemes to integrate the model with a given set of parameters $\theta_0$. 2) According to the kind of observations (complete, partial or partial plus noise) and the kind of model (parametric or not), choose a simple and low-cost inference method. 3) Apply the chosen method to estimate the parameters of the model from the generated data. 4) If the bias of the estimator of $\theta_0$ is negligible and its variance is small enough according to desired precision, then use the method to estimate the model from the actual
data, else return to step 2) to select a new inference method (more costly but providing less biased estimators).

If the inference method has been selected according this strategy, then it is guaranteed that the estimator has the desired precision for the model under consideration. Therefore, there is confidence that the deviation of the estimate from the actual value of the parameter is within an acceptable range if the model fits the data. Assessment of fitness is also a difficult task. For this, the distribution of the fitted-innovations can be used as a useful tool (Ozaki et al. 2000, Jimenez & Ozaki 2002).

6 Conclusion

Statistical inference on stochastic volatility models is an exciting research field with many theoretical and practical open problems still to be solved. In this paper, a number of inference methods have been reviewed making emphasis on underlying ideas, theoretical properties of the estimators and the viability of their implementations to solve actual problems. Advantages and drawbacks of the methods have been discussed with the hope to encourage further theoretical studies and aid applied researchers to select a suitable method for the problem at hand.

Acknowledgments

The authors are grateful to Prof. Miura (Hitotsubashi University) for his invitation to give a talk on the subject of this paper in the 5th International Conference on Statistical Finance and Financial Engineering at Hitotsubashi University, Tokyo and for exhorting us to write this paper. We thank also to Prof. Honda (Hitotsubashi University) for calling our attention to the kind of inference problems with nonlinear observation.
References


Econometric Theory 13, 615—645.


<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>Efficiency</th>
<th>Available algorithm</th>
<th>Implementation Cost</th>
<th>Computational Cost</th>
<th>Partial Observ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood: Kolmogorov Equation</td>
<td>small</td>
<td>high</td>
<td>yes</td>
<td>low</td>
<td>high</td>
<td>-</td>
</tr>
<tr>
<td>Likelihood: Analytical Approximation</td>
<td>small</td>
<td>high</td>
<td>yes</td>
<td>high</td>
<td>low</td>
<td>-</td>
</tr>
<tr>
<td>Likelihood: Approximation via Simulation</td>
<td>small</td>
<td>high</td>
<td>yes</td>
<td>low</td>
<td>high</td>
<td>-</td>
</tr>
<tr>
<td>Bayesian</td>
<td>small</td>
<td>high</td>
<td>yes</td>
<td>low</td>
<td>high</td>
<td>yes</td>
</tr>
<tr>
<td>Pseudo likelihood</td>
<td>large</td>
<td>low</td>
<td>yes</td>
<td>low</td>
<td>low</td>
<td>-</td>
</tr>
<tr>
<td>Methods of the moments</td>
<td>small</td>
<td>low to medium$^2$</td>
<td>yes</td>
<td>medium</td>
<td>high</td>
<td>yes</td>
</tr>
<tr>
<td>Quasi-likelihood</td>
<td>small</td>
<td>low to medium$^3$</td>
<td>not always</td>
<td>medium</td>
<td>medium to high$^4$</td>
<td>yes+noise</td>
</tr>
<tr>
<td>Non-parametric</td>
<td>large</td>
<td>low</td>
<td>yes</td>
<td>low</td>
<td>low</td>
<td>-</td>
</tr>
<tr>
<td>Innovation</td>
<td>medium or small$^1$</td>
<td>medium</td>
<td>yes</td>
<td>low</td>
<td>low or high$^4$</td>
<td>yes+noise</td>
</tr>
</tbody>
</table>

Table I. Summary of the main properties of the inference methods. Here, it has been assumed the worst case in which $h$ is moderate and $n$ is relatively small, otherwise the bias and efficiency of all the methods are similar (small and high, repetitively) whereas the other four aspect remain same. The implementation cost includes the work to obtain the needed formulae of the algorithms for a given model. In box (1) the bias is medium or small depending on the filter algorithm used to compute the innovation. In boxes (2) and (3) the efficiency increases when, respectively, $R_n^{(1)}(\theta)$ and $F_n(x_{\text{obs}}, \theta)$ are closer to the score function. In boxes (4) the computational cost is high when simulations are used to compute the estimates.