Chapter 13

Parameter estimation in discretely observed SDEs

13.1 Introduction

This chapter describes methods for estimating parameters in stochastic differential equations (SDEs). A brief introduction to the GMM method is given, but a major part of the presentation is devoted to a class of maximum likelihood methods which can be used for estimation parameters in both linear and non-linear SDEs.

It is clear that a method for estimating parameters of non-linear stochastic differential equations can also be used for estimating the parameters of a linear stochastic differential equation. If a linear model is considered, it is, however, advantageous to take the linearity into account at the estimation procedure. Likewise it is beneficial to simplify the method if it is known that the model is time invariant.

In this section we shall briefly introduce the various types of models which will be considered in the subsequent sections.

There are several reasons why it is advantageous to consider the continuous-discrete time state space approach:

- If any physical knowledge is available it is easily included in the model, since the system equation is described in continuous time.

- The estimated parameters are readily interpreted by experts in the field.

- Multivariate models are easily considered.

- Time varying models can be handled.

- Missing observations, as well as time varying sampling times, can be handled.

- It is often the case that less parameters are needed for the continuous time formulation than for the traditional discrete time formulation.

- The solution of the Itō equation is a Markov process.

The main limitation is the difficulty to derive estimators, or, rather, deriving the Maximum Likelihood estimator in closed form. Instead, we have to rely on approximations of the Maximum Likelihood estimator.

There are some perhaps unexpected results[1] regarding asymptotic properties of the estimators that are useful to know.

Example 13.1 (Estimating parameters in the drift and diffusion). *Consider the arithmetic Brownian motion*

$$dX(t) = \mu dt + \sigma dW(t). \tag{13.1}$$

The solution is well known, $X(t_{n+1}) - X(t_n) = \mu(t_{n+1} - t_n) + \sigma(W(t_{n+1}) - W(t_n))$, implying that each difference of observations is an independent Gaussian random variable. All that is needed is to compute the mean and covariance! The drift is estimated by computing the mean, and compensating for the sampling $\delta = t_{n+1} - t_n$

$$\hat{\mu} = \frac{1}{\delta N} \sum_{n=0}^{N-1} X(t_{n+1}) - X(t_n). \tag{13.2}$$

Expanding this expression reveals that the MLE for μ is given by

$$\hat{\mu} = \frac{X(t_N) - X(t_0)}{t_N - t_0}. \tag{13.3}$$

The only thing that matters is the lengths of the observation window; recording measurements inside the sample is of no use whatsoever for the drift parameter.

The situation is different for the diffusion (σ) parameter, as the MLE is given by

$$\hat{\sigma}^2 = \frac{1}{\delta(N-1)} \sum_{n=0}^{N-1} (X(t_{n+1}) - X(t_n) - \hat{\mu}\delta)^2 \xrightarrow{d} \sigma^2 \frac{\chi^2(N-1)}{N-1} \tag{13.4}$$

which converges to the correct quantity. The σ parameter can be well estimated by either having a short sampling interval and sampling frequently, or by having a fixed sampling frequency and sampling for a very long time.

13.2 High frequency methods

It should be clear from Example 13.1 that the variance term can be estimated using high frequency data. In fact, that variance estimator is closely linked to the convergence of the stochastic integral and the Itō formula. Here we assume that the process we study is a compound Poission Jump Diffusion

$$dX(t) = \mu(t, X(t))dt + \sigma(t, X(t))dW(t) + dZ(t) \tag{13.5}$$

with $Z(t) = \sum_{n=1}^{N(t)} J_n$ where J are the stochastic jumps and $N(t)$ is the Poisson process.

---

[1] Similar results exist in the theory on discrete time series analysis.

Let $\pi_N = [0 = \tau_0 < \tau_1 < \ldots \tau_N = T]$ be a partition of the interval $[0, T]$, where $\sup |\tau_{m+1} - \tau_m| \to 0$, and define

$$RV = \sum_{n=1}^{N} \left(X(\tau_n) - X(\tau_{n-1}) \right)^2. \qquad (13.6)$$

Then it follows the statistic RV, called the *Realized Variance* estimator, converges to the Quadratic Variation (QV) of the process, cf. Section 13.7.

$$RV \xrightarrow{p} \int \sigma(s, X(s))^2 \mathrm{d}s + \sum_{n=1}^{N(t)} J_n^2 = QV \qquad (13.7)$$

while a small modification called the *bipower* estimator defined as

$$BPV = \frac{\pi}{2} \sum_{n=1}^{N} |X(\tau_{n+1}) - X(\tau_n)| \, |X(\tau_n) - X(\tau_{n-1})| \qquad (13.8)$$

converges to the continuous component

$$BPV \xrightarrow{p} \int \sigma(s, X(s))^2 \mathrm{d}s. \qquad (13.9)$$

Shifting the computations a single step will eliminate jumps as the jumps only contribute to one of the terms, while the other term will be very small when the partitioning gets finer and finer.

The implication of these methods (Barndorff-Nielsen [2002], Barndorff-Nielsen and Shephard [2004]) is that high frequency data can be used to estimate the variance and jump parts with very good accuracy. However, market microstructure (Hansen and Lunde [2006], Zhang et al. [2005]), generally prevents us from going to the limit, and there seems to be a consensus that sampling too often will contaminate the data more than the additional variance reduction obtained due to the larger data set.

We illustrate the realized volatility by computing it on simulated data from a geometric Brownian motion; see Figure 13.1. The constant volatility of the log returns implies that the integrated squared volatility grows linearly.

Computing the realized variance and bipower variation on index returns on the Swedish OMXS30, as shown in top graph in Figure 13.2, reveals that the variance is not constant; cf. Section 5.5.2 and 5.5.3.

These statistics can also be used to investigate whether there are jumps as the difference $RV - BPV$ should equal to the sum of the squared jumps

$$RV - BPV \xrightarrow{p} \sum_{n=1}^{N(t)} J_n. \qquad (13.10)$$

This difference is computed in the lower graph in Figure 13.2, where some
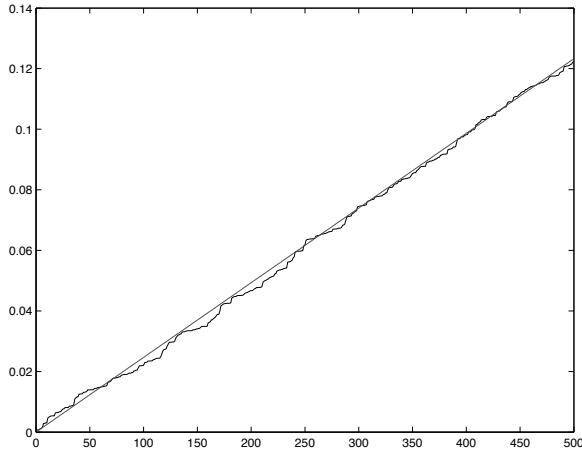
Figure 13.1: Realized variance computed on simulated log returns from a geometric Brownian motion
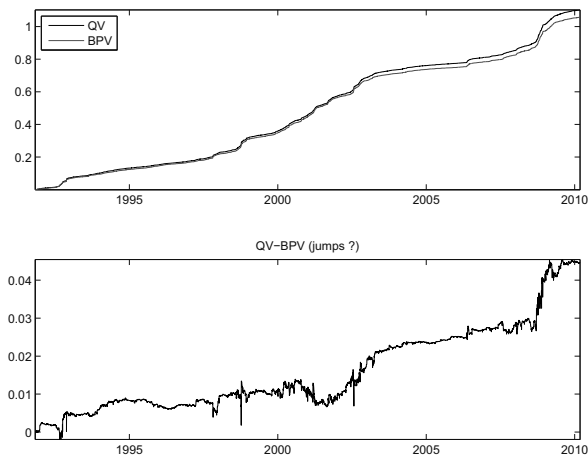

Figure 13.2: Realized variance and bipower variation computed on log returns on the OMXS30 index (top graph) and the difference between the bipower variation and realized variance (bottom graph).

evidence for jumps is presented. Notice that most jumps are rather modest in size, and also that the jumps often seem to cluster to periods of high variance.

Realized volatility was used in Phillips and Yu [2009] to derive Maximum Likelihood estimators for some diffusion processes by first estimating the integrated volatility and then using methods like those in Section 8.4.4.

13.3 Approximate methods for linear and non-linear models

The complexity estimating parameters in discretely observed diffusions, i.e., estimating parameters when the full state vector is observed as discrete time points, depends on a number of factors, most important being the dynamics of the process and the sampling frequency.

Many data sets are sampled at high frequency, compared to the dominant dynamics of the stochastic system, making the bias due to discretization of the SDEs using any of the schemes in Chapter 12 acceptable.

The simplest discretization, the explicit Euler method, would for the stochastic differential equation

$$\mathrm{d}X(t) = \mu(t,X(t))\mathrm{d}t + \sigma(t,X(t))\mathrm{d}W(t) \tag{13.11}$$

correspond to the *Discretized Maximum Likelihood* (DML) estimator given by

$$\hat{\theta}_{DML} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{n=1}^{N-1} \log \phi\left(X(t_{n+1}), X(t_n) + \mu(t_n,X(t_n))\Delta, \Sigma(t_n,X(t_n))\Delta\right) \tag{13.12}$$

where $\phi(x,m,P)$ is the density for a multivariate normal distribution with argument x, mean m and covariance P and

$$\Sigma(t,X(t)) = \sigma(t,X(t))\sigma(t,X(t))^T, \tag{13.13}$$

and $\Delta = t_{n+1} - t_n$ is the time between two consecutive observations.

Note that the Euler scheme and the Milstein scheme coincide when the diffusion term is independent of the state vector. Similarly, it was shown in Chapter 12 that higher order method can readily be derived when the diffusion term is independent of the state vector, which indicates that transformations to get rid of state dependence are generally a good idea.

13.4 State dependent diffusion term

If the SDE contains a state dependent diffusion term then the methods described in the previous section cannot be used directly. Frequently it is, however, possible to transform the SDE into an equivalent SDE where the diffusion term is independent of the state vector. The equivalent SDE contains the same parameters and describes a relation between the same input and output variables.

13.4.1 A transformation approach

We shall only consider the scalar case here. The transformation approach is based on the Itō formula which we restate here for convenience.

Theorem 13.1 (The Itō formula). *Let $X(t)$ be a solution to*

$$dX(t) = \mu(t, X(t))dt + \sigma(t, X(t))dW(t) \tag{13.14}$$

and $\varphi \colon \mathbb{R}^2 \mapsto \mathbb{R}$ be a $C^{1,2}(\mathbb{R}^2)$-function applied to $X(t)$

$$Y(t) = \varphi(t, X(t)). \tag{13.15}$$

Then it holds that

$$dY(t) = \left[\frac{\partial \varphi}{\partial t} + \mu \frac{\partial \varphi}{\partial X(t)} + \frac{1}{2}\sigma^2 \frac{\partial^2 \varphi}{\partial X(t)^2} \right] dt + g \frac{\partial \varphi}{\partial X(t)} dW(t). \tag{13.16}$$

Proof. See page 147. □

Notice that the diffusion term in the new Itō process $dY(t)$ is equal to the product of g and $\partial \varphi / \partial x$. This leads us to the following lemma.

Lemma 13.1. *Let $X(t)$ and φ be defined as above, then by choosing the transformation*

$$\varphi(x_t, t) = \int \frac{1}{\sigma(t, X(t))} \, dX(t) \tag{13.17}$$

the new process $Z(t) = \varphi(X(t), t)$ will be an Itō process with a constant diffusion term.

The procedure is easily generalized to the multivariate case where the matrix $\mathbf{g}(\mathbf{X}(t), t)$ only contains non-zero values on the diagonal, but is far more complicated in the general case (cf. Aït-Sahalia [2008]).

Example 13.2 (Transformation of the CKLS-model). *Consider the CKLS-model (Chan et al. [1992]):*

$$dX(t) = \alpha(\theta - X(t))dt + \sigma X(t)^\gamma dW(t) \tag{13.18}$$

where $X(t)$ is the short term interest rate, and α and θ are parameters related to the drift term and δ and γ are related to the diffusion term.

The drift term represents a tendency to pull the process back towards its long-term mean θ with the rate of adjustment determined by α. The diffusion term describes that the variability (the volatility) is increasing with x if $\gamma > 0$.

Some important models used in the field of finance are obtained as a special case of the CKLS-model:

$\gamma \;\; = \;\; 0$ *we get the Vasicek model (Vasicek [1977])*

$\gamma \;\; = \;\; \dfrac{1}{2}$ *we get the Cox–Ingersoll–Ross model (Cox et al. [1985])*

By applying the lemma, the following transformation is suggested in order

*to obtain a constant diffusion term in a new process which contains the same
parameters and relates the same input and output.*

$$Z(t) = \varphi(x_t, t) = \int \frac{1}{X(t)^\gamma} \, dX(t) = \frac{1}{1-\gamma} X(t)^{1-\gamma}. \tag{13.19}$$

The SDE for $Z(t)$ is found by using the Itō formula. Hence we need

$$\frac{\partial \varphi}{\partial t} = 0 \, , \; \frac{\partial \varphi}{\partial x} = X(t)^{-\gamma} \, , \; \frac{\partial^2 \varphi}{\partial x^2} = -\gamma X(t)^{-\gamma-1}. \tag{13.20}$$

Insertion into the Itō formula leads to the following SDE for $Z(t)$

$$dZ(t) = \left[\alpha\theta \left((1-\gamma) Z(t) \right)^{\frac{\gamma}{\gamma-1}} - \alpha (1-\gamma) Z(t) - \frac{\gamma\sigma^2}{2(1-\gamma) Z(t)} \right] dt \quad (13.21)$$

$$+ \sigma dW(t). \tag{13.22}$$

13.5 MLE for non-linear diffusions

It is rarely possible to do full maximum likelihood estimation for non-linear
diffusions, but the likelihood function can often be approximated arbitrarily
well. Some technical difficulties are avoided if the class of models is restricted
to discretely observable models.

There are at least three types of competing algorithms that are compu-
tationally efficient. The most general, and also the computationally least effi-
cient class of algorithms, are the Monte Carlo simulation based estimators. The
Fokker-Planck based estimators are computationally more efficient (at least in
low dimensions). Finally, the series expansion approach is the computationally
most efficient algorithm but it is also the most restrictive estimator. Common
for all estimators is that their computational efficiency is usually improved if
the diffusion term is independent of the states of the process; cf. Section 13.4.

13.5.1 Simulation-based estimators

The class of simulation based estimators can be applied to virtually any con-
tinuous time Markov process, for which a numerical scheme has been de-
rived. The basic idea is to calculate the unknown transition probability density
$p(x_t|x_s)$ by successive approximations. Assume that $s < \tau_1 < \tau_i < \tau_{N-1} < t$.
The transition probability density is then (due to the law of total probability
and the Markov property) given by

$$p(x_t|x_s) = \int p(x_t|x_{\tau_i}) p(x_{\tau_i}|x_s) dx_{\tau_i}.$$

This can be iterated N times, and the following expression is derived

$$p(x_t|x_s) = \int p(x_t|x_{\tau_N}) \int \dots \int p(x_{\tau_1}|x_s) dx_{\tau_1} \dots dx_{\tau_N}.$$

The $N \times d_X$ dimensional integral ($d_X = \dim(X(t))$) is easily calculated using Monte Carlo methods (e.g. Pedersen [1995]). An implementation of this algorithm is actually quite easy as the numerical representation of the integral

$$p(x_t|x_s) = \int p(x_t|x_\tau)p(x_\tau|x_s)\mathrm{d}x_{\tau_i}$$

is given by

$$p_K(x_t|x_s) = \frac{1}{K}\sum_{k=1}^{K}p(x_t|x_\tau^k)$$

where x_τ^k was generated from $p(x_\tau|x_s)$, see Chapter 12 for numerical schemes. However, the simple Monte Carlo approximation is usually too inaccurate to be useful (e.g. Durham and Gallant [2002]), and variance reduction is regularly applied. An efficient version was introduced in Durham and Gallant [2002], and uses a combination of antithetic variables, moment matching and importance sampling (so-called Brownian bridge sampler, see Problem 8.9). Lindström [2012a] improves that Brownian bridge sampler further by introducing a regularization term that improves the performance for sparsely sampled models.

The simulation based estimators converge as $1/\sqrt{K}$ regardless of dimension. This leads to relatively slow convergence in low dimensions but is very competitive in higher dimensions.

13.5.1.1 Jump diffusions

Similar methods have been developed for jump diffusions and Lévy driven SDEs (e.g. Hellquist et al. [2010], Lindström [2012b]). A computationally attractive alternative to direct maximization of the likelihood function is to use the EM algorithm (e.g. Sundberg [1974], Dempster et al. [1977]).

The EM algorithm finds the Maximum Likelihood estimate by iteratively alternating between computing the Expectation of the *intermediate quantity*

$$Q(\theta,\theta') = \mathbf{E}\left[\log p_\theta(X,Y)|Y,\theta'\right] \tag{13.23}$$

where X are some latent variable and Y the observations, and maximizing that quantity

$$\theta_{m+1} = \arg\max Q(\theta,\theta_m). \tag{13.24}$$

Computing the posterior distribution is similar to the simulated maximum likelihood methods. Define Y as the observations ($Y = Y_{t_1},\ldots,Y_{t_N}$) and $X = (X_{\tau_1},\ldots,X_{\tau_{N-1}})$ as latent values of the process in between the observation ($t_1 < \tau_1 < \ldots < \tau_{N-1} < t_N$). The posterior is then given by

$$p(X_{\tau_n}|Y_{t_n},Y_{t_{n+1}}) \propto p(Y_{t_{n+1}}|X_{\tau_n})p(X_{\tau_n}|Y_{t_n}). \tag{13.25}$$

It is therefore possible (this is very similar to what is done in the bootstrap

filter, see Section 14.10) to use resampling of particles generated from the naive dynamics to compute a sample from the posterior distribution (Lindström [2012b]).

The intermediate quantity will typically be easy to maximize as the transition density for a short time interval is well approximate using even simple discretization schemes (cf. Chapter 12), typically leading to a sum of logarithms of Gaussian densities.

13.5.2   Numerical methods for the Fokker–Planck equation

The Fokker-Planck based estimators apply deterministic numerical schemes to the Fokker-Planck equation, a parabolic partial differential equation governing the evoluation of the transition probability density.

The Crank-Nicholson scheme was applied in Lo [1988] while Lindström [2007] used higher order schemes as well. It is usually computationally efficient to use higher order finite difference schemes. The Fokker-Planck equation is typically solved by defining a grid in the domain of the process and approximating the derivatives in the domain with central differences on the grid. This reduces the partial differential equation

$$\frac{\partial p}{\partial t}(x,t) = -\frac{\partial}{\partial x}\left(\mu(x_t)p(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2(x_t)p(x,t)\right)$$

to a system of linear differential equations

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t}(t) = \mathbf{A}\mathbf{p}(t) + \mathbf{b}$$

which has a closed form solution. It was shown in Lindström [2007] that the rate of convergence of the Fokker-Planck based estimators is significantly faster than, i.e., simulation based estimators, but it is known the rate deteriorates as the dimension increases.

The partial differential equation approach can also be used to compute moments or eigenfunctions, which could lead to substantial computational savings as fewer PDEs need to be solved numerically. These can be used to derive GMM or Quasi-likelihood estimators (see Höök and Lindström [2014] for details).

13.5.3   Series expansion

The Aït-Sahilia estimator (Aït-Sahalia [2002, 2003]) derives a closed form expression of the transition probability density. This is achieved by approximating the true transition probability density by a standard Gaussian random variable and correcting for the deviations using Hermite polynomials (Hermite polynomials are orthogonal polynomials associated to the standard Gaussian

density). Unfortunately the correction only does converge if the true density is similar to the standard Gaussian density.

The solution in Aït-Sahalia [2002] is to transform the process $dX(t) = \mu_X(X(t))dt + \sigma_X(X(t))dW(t)$ to a new representation where the diffusion term is independent of $X(t)$; cf. Section 13.4. This can be achieved by introducing

$$Y(t) = \gamma(X(t)) = \int^{X(t)} \frac{du}{\sigma_X(u)}$$

resulting in

$$dY(t) = \left(\frac{\mu_X(\gamma^{-1}(Y(t)))}{\sigma_X(\gamma^{-1}(Y(t)))} - \frac{1}{2}\frac{\partial \sigma_X}{\partial x}(\gamma^{-1}(Y(t))) \right) dt + dW(t).$$

The second step is to normalize $Y(t)$ by its approximative standard deviation

$$Z(t) = \frac{Y(t) - Y(s)}{\sqrt{t - s}}.$$

The density for $Z(t)$ is then expanded in Hermite polynomials (orthogonal polynomials with respect to the Gaussian density) as

$$p_Z(z|y_s) = \phi(z) \left(1 + \sum_{j=1}^{\infty} \eta_Z^{(j)}(t - s, y_s; \theta) H_j(z) \right),$$

where the formula to calculate $\eta_Z^{(j)} = 1/(j!)E_Z[H_j(Z)]$ is given in Aït-Sahalia [2003], $H_j(z)$ are Hermite polynomials and $\phi(z)$ is the density of a standard Gaussian random variable. The closed form expression for $p_Z(z|y_s)$ is then used to calculate $p_X(x_t|x_s)$. It is not guaranteed that the Hermite expansion will generate a pdf (i.e., a density that $\geq$ for all function values and integrates to one). Another approach is then to work with log densities ($\log p$) rather than densities (p) (Aït-Sahalia [2002]). That would eliminate the risk for negative densities, but the approximation will still not integrate to one, unless infinitely many terms are included in the expansion.

The estimator is very efficient from a computational point of view when $t - s$ is small. A multivariate version of the algorithm was presented in Aït-Sahalia [2008], but the generalization is not as smooth as one could hope; the expansion is only valid for so-called *reducible diffusions*, which excludes many models in financial economics such as stochastic volatility models.

13.6   Generalized method of moments

The *Generalized Method of Moments* is a method attributable to Hansen [1982] which can be used for estimating parameters in stochastic differential equations. Similar ideas (called estimation function) can be found in Bibby and Sørensen [1995]; see Sørensen [2012] for a recent overview.

For the GMM method no explicit assumption about the distribution of the observations is made, but the method can include such assumptions. In fact, it is possible to derive the ML estimator as a special case of the GMM estimator.

The standard version of the GMM method requires that all the state variables are observed directly. This is clearly a drawback of the method, since it dramatically limits the complexity of the models which can be considered. In the modeling of spot interest rates, this is a severe restriction, due to the fact that the spot interest rate is not observed directly, but only implicitly via the bond prices. Furthermore, the estimation does not take place directly in the continuous-time model; the stochastic differential equations must be discretized (see Chapter 12), before using the GMM method, and the sampling time must be constant. The best reference to discretization techniques so far is Kloeden and Platen [1995] and Platen and Bruti-Liberati [2010]; see also Chapter 12.

13.6.1   GMM and moment restrictions

We shall introduce the moment restrictions by first considering an example which is of great interest in statistical finance.

Example 13.3 (GMM estimates of the CKLS model). *Consider the following stochastic differential equation*

$$dr(t) = (\theta + \eta r(t))dt + \sigma r(t)^{\gamma}dW(t) \tag{13.26}$$

where $W(t)$ is a standard Wiener process. This model, which is called the CKLS model (Chan et al. [1992]), is often used to model the short term risk-less interest rate. The structure of the CKLS model implies that the conditional mean and variance of the short term rate depend on the level of $r(t)$.

The observations are in general given in discrete time. Hence the model (13.26) has to be formulated in discrete time before setting up any moment restrictions which can be evaluated using the discrete time data.

In financial and economical applications the Euler approximation method (see Chapter 12) is most often used to formulate the discrete time equivalent of the continuous time model. Using the Euler approximation of model (13.26) we obtain the following discrete time model

$$r_{k+1} = r_k + (\theta + \eta r_k)\Delta + \varepsilon_{k+1} \tag{13.27}$$

where

$$\mathbf{E}[\varepsilon_{k+1}|B_k] = 0 \tag{13.28}$$

$$\mathbf{E}[\varepsilon_{k+1}^2|\mathscr{F}_k] = \sigma^2 r_k^{2\gamma}\Delta \tag{13.29}$$

and where $\mathscr{F}_k$ denotes the information set at time t_k.

Under the assumption that the restrictions implied by the model, i.e., by (13.27), (13.28) and (13.29) are true, the following moment restrictions follow

$$\mathbf{E}[\varepsilon_{t+1} - \mathbf{E}[\varepsilon_{t+1}|\mathscr{F}_t]] = 0, \tag{13.30}$$

$$\mathbf{E}[\varepsilon_{t+1}^2 - \mathbf{E}[\varepsilon_{t+1}^2|\mathscr{F}_t]] = 0. \tag{13.31}$$

The GMM procedure consists of replacing the expectation $\mathbf{E}[\cdot]$ with its sample counterpart, using the N observations available $\{r(k) \; ; \; k = 1, \cdots, N\}$.

Now we are ready to formulate the GMM method more formally. Assume that the observations are given as the following sequence of random vectors:

$$\{\mathbf{x}_t \; ; \; t = 1, \cdots, N\}$$

and let θ denote the unknown parameters ($\dim(\theta) = p$).

Let $\mathbf{f}(\mathbf{x}_t, \theta)$ be a $q \geq p$-dimensional zero mean function, which is chosen as some *moment restrictions* implied by the discretized model of $\mathbf{x}_t$.

The GMM estimates are found by minimizing

$$J_N(\theta) = \left(\frac{1}{N}\sum_{t=1}^N \mathbf{f}(\mathbf{x}_t, \theta)\right)^T \mathbf{W}_N \left(\frac{1}{N}\sum_{t=1}^N \mathbf{f}(\mathbf{x}_t, \theta)\right) \tag{13.32}$$

where $\mathbf{W}_N \in \mathbb{R}^{q \times q}$ is a positive semidefinite *weight matrix*, which defines a metric subject to which the quadratic form has to be minimized. The key observation is that any remaining bias in the moment conditions (e.g., (13.30) or (13.31)) is going to increase the J_N function.

Example 13.4 (GMM estimates of the CKLS model (cont.)). *Consider again the CKLS model given in Eq. (13.26), and the Euler approximative discrete time equivalent model given by (13.28)–(13.29). The unknown parameters are $\theta = (\theta, \eta, \gamma, \sigma)^T$.*

In order to use the GMM method we may use the moment restrictions given by

$$\begin{aligned}
\mathbf{f}(r_k, \theta) &= \begin{pmatrix} \varepsilon_{k+1} - \mathbf{E}[\varepsilon_{k+1}|\mathscr{F}_k] \\ (\varepsilon_{k+1} - \mathbf{E}[\varepsilon_{k+1}|\mathscr{F}_k])\, r_k \\ \varepsilon_{k+1}^2 - \mathbf{E}[\varepsilon_{k+1}^2|\mathscr{F}_k] \\ (\varepsilon_{k+1}^2 - \mathbf{E}[\varepsilon_{k+1}^2|\mathscr{F}_k])\, r_k \end{pmatrix} \\
&= \begin{pmatrix} \varepsilon_{k+1} \\ \varepsilon_{k+1} r_k \\ \varepsilon_{k+1}^2 - \sigma^2 r_k^{2\gamma}\Delta \\ (\varepsilon_{k+1}^2 - \sigma^2 r_k^{2\gamma}\Delta) r_k \end{pmatrix}. \tag{13.33}
\end{aligned}$$

Note that in the second and fourth equation the variable r_k is used as an instrumental variable. Those equations are justified by the fact that r_k is independent of the entities $(\varepsilon_{k+1} - \mathbf{E}[\varepsilon_{k+1}|\mathscr{F}_k])$ and $(\varepsilon_{t+1}^2 - \mathbf{E}[\varepsilon_{k+1}^2|\mathscr{F}_k])$.

Some extensions of the GMM for diffusions include *Simulated GMM* often called Simulated Method of Moments (SMM) (McFadden [1989]). Monte Carlo simulations can be used to compute arbitrarily accurate approximations of moments, thereby eliminating the bias caused by the discretization scheme. Solving PDEs is a computationally attractive alternative in many cases (Höök and Lindström [2014]).

The characteristic function is known for a fairly large class of processes, as presented in Section 7.5. These can be used to construct suitable moments without having to use any discretization scheme (Singleton [2001] and Chacko and Viceira [2003]). It can be useful to add many moment conditions (Carrasco et al. [2007]), who showed that Maximum Likelihood efficiency is achieved when a continuum of moments is employed in the estimation, although there is a practical upper limit when employing the method on real data (for example, the number of conditions has to be smaller than the number of observations).

Another approach suitable for diffusion (or Markov processes in general) is to use eigenfunctions of the generator as moments (Kessler and Sørensen [1999]). The downside is that these can be tricky to compute, and other methods may be preferred (i.e., maximum likelihood methods) for simple models where eigenfunctions are computable in closed form.

13.7   Model validation for discretely observed SDEs

Model validation for linear, discrete time models is a well known subject. Not only do we know how to estimate the parameters in most models, but there is also a large variety of suggestions on how we can validate the models.

Continuous-time models based on stochastic differential equations are harder to validate except for linear models, where validation techniques developed for discrete time models can be applied. Instead, we are often restricted to techniques telling us whether one model is better than another using e.g. LR and Wald tests, which is based on general theory on likelihood functions, not the models per se.

13.7.1   Generalized Gaussian residuals

Discrete time residuals are usually defined as standardized martingale increments, i.e.,

$$\varepsilon_k = \frac{X_k - \mathbf{E}[X_k|\mathscr{F}_{k-1}]}{\sigma(X_k|\mathscr{F}_{k-1})}$$

where $\{\varepsilon_k\}_{k\geq 0}$ is assumed to be a sequence of *iid* random variables. This definition is only valid for discrete time models, and using this definition on data generated by a diffusion process will not (in general) generate an *iid* sequence.

The difficulties encountered when working with validation of diffusion processes can be attributed to the (unknown) transition probability density, $p(x_s,s;x_t,t)$, implying that $\{\varepsilon_k\}_{k\geq 0}$ is not an *iid* sequence.

Let us, however, assume that we know (or can calculate) the transition probability density. We could then transform all innovations to a sequence of identically distributed random variables with a known distribution, making the validation problem easier.

This is done by calculating the conditional distribution function for all pairs of observation, $F_{X_k|X_{k-1}}(x) = \mathbb{P}(X_k \leq x_k|X_{k-1}) = \int_{-\infty}^{x_k} p(x_{t_{k-1}}, t_{k-1}; x, t_k) \mathrm{d}x$. Having calculated the distribution function, we use the fact that the distribution function applied to the stochastic variable is a uniformly distributed random variable

$$U_k = F_{X_{t_k}|X_{t_{k-1}}}(X_{t_k}; \boldsymbol{\theta}).$$

It can also be shown that the sequence $\{U_k\}_{k \geq 0}$ is pairwise independent (e.g. Lindström [2004, 2003]). Note that we have not assumed stationarity of the process, i.e., time variations, non-linearities and non-equidistantly sampled data are allowed. Stationarity will, however, often simplify the calculations of the cumulative transition probability density. The sequence can be used to obtain generalized Gaussian residuals by transforming the sequence once more:

$$Y_k = \Phi^{-1}(U_k).$$

The sequence $\{Y_k\}_{k \geq 0}$ will inherit the nice properties of $\{U_k\}_{k \geq 0}$ but the Gaussian residuals offer many advantages to uniform residuals. Most important is the property that uncorrelated Gaussian random variables are independent random variables, but it may also be argued that there are other advantages of using Gaussian residuals for validation (e.g. distributional tests and outlier detection).

The Gaussian residuals may be difficult to calculate as the inversion of the standard Gaussian distribution function is numerically sensitive. It is thus important that the transition probability density can be calculated accurately for all pairs of observations. It is recommended to approach this problem by solving the Fokker-Planck equation (see Section 13.5.2) numerically to calculate the transition probability density. Simulation could also be used but importance sampling would be needed to preserve the accuracy in the tails.

13.7.1.1   Case study

The first set of data used is simulated data from the Cox–Ingersoll–Ross model (Cox et al. [1985]). The model is widely used in the finance community as a model for short interest rates, but it has also been used as a model for volatility. The model is specified by the stochastic differential equation

$$\mathrm{d}X_t = \alpha(\beta - X(t))\mathrm{d}t + \sigma\sqrt{X(t)}\mathrm{d}W(t).$$

The data were simulated using a Milstein scheme using equidistant observations. The distance between the observations is $\Delta t = 1$ and the parameters used were chosen as $\alpha = 0.17$, $\beta = 0.05$ and $\sigma = 0.07$. The simulation produced
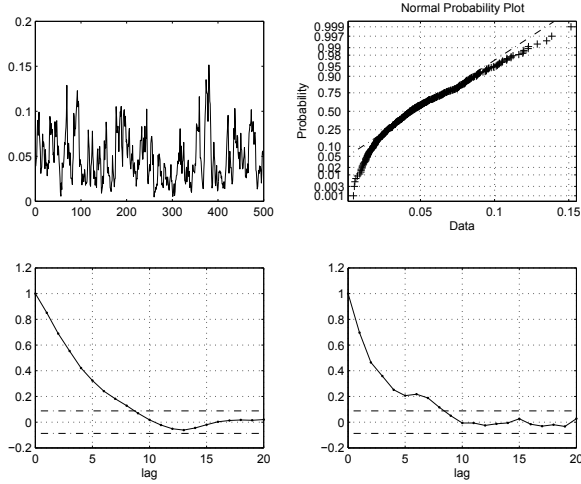
Figure 13.3: Simulated Cox–Ingersoll–Ross process (left) and corresponding normal probability plot.

500 observations, where 100 intermediate steps between each observation have been simulated to decrease the bias (Figure 13.3).

The sample autocorrelation and QQ plot for the generalized Gaussian residuals are presented in Figure 13.4. These are uncorrelated and Gaussian (and thus independent) which is not surprising since the data were simulated using the correct model.

The benchmark for stocks is the geometric Brownian motion, $dS(t) = \mu S(t)dt + \sigma S(t)dW(t)$. Let $\Delta_k = t_k - t_{k-1}$ be the sampling interval and let $\Delta W_i = W_{t_k} - W_{t_{k-1}}, \tilde{\mu} = (\mu - \sigma^2/2)$. The solution to the stochastic differential equation is given by $S_{t_k} = S_{t_{k-1}} e^{\tilde{\mu}\Delta_k + \sigma\Delta W_k}$. We obtain U_k by applying the distribution function. The conditional distribution of S_{t_k} is given by

$$\mathbb{P}(S_{t_{k-1}} e^{\tilde{\mu}\Delta_k + \sigma\Delta W_k} \leq s_k | S_{t_{k-1}}) = \Phi\left(\frac{\log\left(s_k/S_{t_{k-1}}\right) - \tilde{\mu}\Delta_k}{\sigma\sqrt{\Delta_k}}\right).$$

By applying the inverse of the distribution function of the standard Gaussian random variable, we find that

$$Y_k = \Phi^{-1}(U_k) = \frac{\log\left(s_k/S_{t_{k-1}}\right) - \tilde{\mu}\Delta_k}{\sigma\sqrt{\Delta_k}}.$$

The generalized Gaussian residuals are consistent with the ordinary standardized residuals (for linear diffusions) but the technique is also able to handle non-linear models. We have calculated the generalized Gaussian residuals
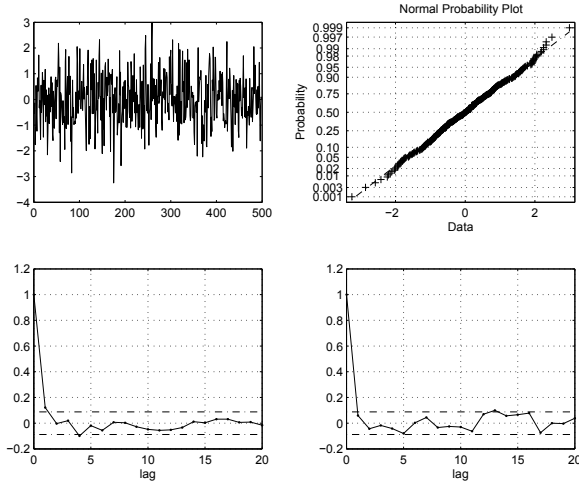
Figure 13.4: Sample autocorrelation and QQ plot for the generalized Gaussian residuals computed for the Cox–Ingersoll–Ross process.
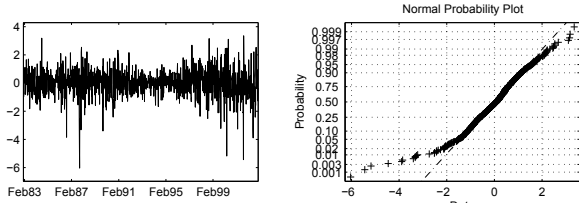

Figure 13.5: Generalized Gaussian residuals (left) and QQ plot (right) for continuously compounded returns on the S & P 500.

for the S & P 500 using weekly data from 1983–2002. The residuals are presented in Figure 13.5. It can be seen that the residuals are non-Gaussian and heteroscedastic, reaffirming the well-known fact that the S & P 500 is not generated by a geometric Brownian motion.

13.8   Problems

Problem 13.1
Consider the stochastic variable $X \in N(\mu, \sigma^2)$.

1. Specify two moment conditions that can be used for estimating the parameters μ and σ^2.

2. Specify an additional moment condition.

3. Simulate 100 observations of X. Estimate the parameters and their covariances based on these observations using the two moment conditions. Estimate the parameters and their covariances based on these observations using the three moment conditions. Comment on the results.

Problem 13.2 Consider the *GARCH* model (5.67) for $p = 1$ and $q = 1$, i.e., the *GARCH(1,1)* model.
1. Write down suitable moment conditions for estimating the parameters. Illustrate the choice of moment conditions!
 Hint: Use (5.69).

Problem 13.3 Consider the CKLS model

$$\mathrm{d}X(t) = \kappa(\theta - X(t))\mathrm{d}t + \sigma X(t)^{\gamma}\mathrm{d}W(t).$$

1. Calculate the simple estimation functions obtained from choosing $h(x) = x,\ x^2, x^3,\ x^4$.
2. Do an Euler approximation of the CKLS model with timestep Δ. Use this to derive suitable moment conditions for a GMM estimation of the parameters.