ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 8, 2004

Estimating parameters of stochastic differential equations using a criterion function based on the Kolmogorov–Smirnov statistics

Daria Filatova, Marek Grzwaczewski and David McDonald

Abstract. We introduce a method for the estimation of stochastic differential equation (SDE) coefficients from panel data. The method involves matching the distribution of the experimental/field data with a panel of simulated data generated by a Monte Carlo experiment. The fit between the two distributions is assessed by means of Kolmogorov–Smirnov goodness-of-fit statistic leading to a confidence function computed from an incomplete gamma function. A numerical optimization algorithm then optimizes the choice of parameters to maximize this function.

1. Introduction

Deterministic mathematical model can be used as an explanation of some physical phenomenon as, for example, dynamic characteristics of a frame structure, technological process dynamics, resource management, population dynamics, or economical systems behavior. This approach provides good results only if input signal is free of noise or noise influence is insignificant. However, in most of these cases noise influence is considerable and such a description brings along the following problems: presence of stochastic parameters leads stationary system into unstable state with respect to moments of a phase vector; decrease of stability region; essential differences in the trajectory's behavior of "noise free" system and "noise" system in spite of coincidence in the sense of mean square error. In order to avoid these problems it is recommended to use stochastic mathematical models and somehow

Received October 23, 2003.

2000 *Mathematics Subject Classification.* 65C20, 60H10.

Key words and phrases. Stochastic differential equations, parameter estimation, Kolmogorov–Smirnov statistic, simulation experiments.

to identify parameters of these models. One possibility is to use stochastic differential equation (SDE). The problem of SDE parameters identification has been studied extensively for both continuous and discrete data. We focus our research to provide a numerical method, based on maximum likelihood principle, for the estimation of SDE parameters using discrete panel data.

2. Stochastic differential equations

Let us consider the SDE with an initial condition of the form $Y(0) = y_0$ $(t \geq 0)$ of the type

$$dY(t) = a(t, Y(t))\,dt + b(t, Y(t))\,dW(t), \tag{1}$$

where $Y(t)$ is the solution to be determined, $a(t, Y(t))$ and $b(t, Y(t))$ are the given functions, called the drift and diffusion, respectively, and $dW(t)$ is the increment of a standard Wiener process. In integral notation, solution $Y(t)$ can be written as

$$Y(t) = Y(s) + \int_s^t a(r, Y(r))\,dr + \int_s^t b(r, Y(r))\,dW(r), \tag{2}$$

where on the right-hand side of (2) the first integral is Riemann and the second integral is stochastic one. Moreover, the solution of (1) can be difficult to determine explicitly because it appears on both, the left- and right-hand sides of equation (2). As it is possible to see, the main difference between ordinary differential equations and SDE is determined by the Wiener process, $W(t)$, which can be defined as a continuous Gaussian process with independent increments and the following properties:

$$W(0) = 0,\ E[W(t)] = 0,\ Var[W(t) - W(s)] = t - s,\ 0 \leq s \leq t. \tag{3}$$

Suppose that interval $[s, t]$ is divided into subintervals (not nessesarily uniformly) by points t_i: $s = t_0 < t_1 < ... < t_n = t$. Then the value of an integral is defined conventionally to be the limit of a sequence of sums over the subintervals $[t_r, t_{r+1}]$, $0 \leq r \leq n-1$, as n tends to infinity. Specifically

$$\int_s^t b(r, Y(t))\,dW(r) = \lim_{n\to\infty} \sum_{r=0}^{n-1} b(\tau_r, Y(\tau_r))\{W(t_{r+1}) - W(t_r)\}, \tag{4}$$

where $\tau_r \in [t_r, t_{r+1}]$ and the location of τ_r is unimportant. We will use Ito definition of stochastic integral, which has the martingale property, but unfortunately does not obey the rules of ordinary calculus. Let $U(t, Y(t))$

be a continuous deterministic scalar function, then Ito's formula connecting $U(t, Y(t))$ and $U(s, Y(s))$ $(t \geq s)$ gives us

$$U(t, Y(t)) = U(s, Y(s)) + \int\limits_s^t \left(\frac{\partial U}{\partial t} + a \frac{\partial U}{\partial Y} + \frac{1}{2} \frac{\partial^2 U}{\partial Y^2} \right) dr + \int\limits_s^t b \frac{\partial U}{\partial Y} dW(r), \tag{5}$$

where $Y(t)$ is the solution of the SDE (2) and $U(t, Y(t))$ is at least twice differentiable so that all the integrals in (5) exist. This result can be presented in Ito's lemma

$$dU = \frac{\partial U}{\partial t} dt + \frac{\partial U}{\partial Y} dY(t) + \frac{1}{2} b^2 \frac{\partial^2 U}{\partial Y^2} dt, \tag{6}$$

where U can be sometimes interpreted as a probability density function.

One of the simplest cases of SDE is the linear Ito stochastic differential equation

$$dY(t) = aY(t) + bY(t)\, dW(t), \tag{7}$$

where a and b are constants. If we take $U = \log(Y)$, it may be verified from Ito's formula (5) that (7) has a stochastic solution

$$Y(t) = Y(0) \exp\left[\left(a - \frac{1}{2} b^2 \right) t + bW(t) \right]. \tag{8}$$

So, we have sketched the basic ideas of solving SDE. Notice that very few specific SDEs have known explicit solutions and hence the task of finding $Y(t)$ is much more complicated as it seems. One of the possible solutions of this problem is to identity the process parameters on a basis of observations and their numerical computation. We wish to develop a method for estimating the parameters of SDE of type (1) based on the main properties of SDE and to get a numerical solution.

3. The SDE models examined

We examine SDE of type (1) in the general form, where a and b are constant parameters to be estimated. Using subscript t to denote time, the specific equations are

(A) $dY_t = aY_t dt + bY_t dW_t;$

(B) $dY_t = Y_t (1 - Y_t)\, dt + bY_t^2 dW_t;$

(C) $dY_t = aY_t (1 - Y_t)\, dt + bY_t dW_t;$

(D) $dY_t = aY_t (1 - Y_t)\, dt + \frac{1}{5} bY_t (1 - Y_t)\, dW_t;$

(E) $dY_t = Y_t (1 - Y_t)^a\, dt + \frac{1}{2} Y_t^{2b} dW_t.$

These equations were chosen because they are of the form used frequently for modeling renewable resource systems. Obtaining point and interval estimates of the drift and diffusion parameters of such models is important in applied work because these estimates provide a tool for testing hypotheses about the state of the system and the relative importance of stochastic influences on it. Observed data for stochastic processes are recorded in discrete time, regardless of whether the system is described better by a continuous or a discrete model. One advantage of using a continuous model is that, in principle, its solution can be used for any time interval, without altering the meaning or interpretation of the model parameters. Estimation of the SDE parameters requires the solution or an approximation to it. We concentrate on two methods for finding discrete-time approximations to the solution of Equations (A)–(E); namely the strong Euler scheme that attains convergence of order 0.5 (Kloeden et al. (1994), pp. 140–142) and the strong Taylor scheme that attains convergence of order 1.5 (Kloeden et al. (1994), pp. 162–163).

4. Parameter estimation using a criterion function based on the Kolmogorov–Smirnov statistic

The Kolmogorov–Smirnov statistic adapted to a two-sample problem provides the basis for goodness-of-fit method for SDE parameter estimation. The two-sample Kolmogorov–Smirnov goodness-of-fit test is used to compare the empirical distribution functions of two samples. In the present paper one of these samples is generated as if observed from a fully-specified SDE and the other one is generated from the same SDE but under the assumption that the coefficients are unknown. Parameter estimation in practice requires one of the samples Y to be observed and the other $\widetilde{Y}$ to be generated by the SDE that is used to model the data-generating process. Let us denote by $F_{Y,n}$ and $F_{\widetilde{Y},m}$ the empirical distribution functions of Y and $\widetilde{Y}$ with n and m sample paths respectively. In this case the Kolmogorov–Smirnov two-sample test statistic

$$D_{n,m} = \max_{Y} \left| F_{Y,n}(y) - F_{\widetilde{Y},m}(y) \right| \tag{9}$$

is the maximum absolute difference between the two empirical distributions. This statistic can be used to test the (null) hypothesis that the population distributions are identical and, therefore, both samples have been drawn from the same population. The two-sample Kolmogorov–Smirnov statistic has an asymptotic null distribution given by

$$\lim P\left(\sqrt{\frac{nm}{n+m}}D_{n,m} \leq D\right) = KS\,(D) \tag{10}$$

where

$$KS\,(D) = 1 - 2\sum_{j=1}^{+\infty}(-1)^{j-1}\exp\left(-2j^2D^2\right).$$

A large value of D, and therefore a small value of $KS\,(D)$, indicates that the null hypothesis is unlikely to be true, whereas small values D support the null hypothesis. In the present paper we are concerned with replicated time series data. This provides the opportunity to evaluate the Kolmogorov–Smirnov statistic for each time period, D_t. Because we are dealing with a stochastic process that we assume to be modeled adequately by an equation of the general form (1), we must estimate the drift and diffusion parameters so that the entire time series is taken into account. We do this by analogy with maximum-likelihood estimation, taking the product of Kolmogorov–Smirnov statistics computed at each time point as our criterion function. Using the asymptotic null distribution, this yields

$$\Phi = \prod_{t=t_0}^{T} KS\,(D_t) \to \max. \tag{11}$$

Given a set of observations or simulated observations giving rise to $F_{Y,n}\,(y)$, this criterion function is maximized with respect to the drift and diffusion parameters of an SDE that is used in the evaluation of $F_{\widetilde{Y},m}\,(y)$.

5. Empirical results

Estimation of the drift and diffusion parameters of equations (A)–(E) was conducted using the above criterion functions. Fifty realizations of eleven "observed" data points (i.e. $n = 50$, $T = 11$) were generated for each model using the initial value $y_0 = 0.5$ and true parameters $a = 1.0$ and $b = 0.5$. The derivative-free simplex method of Nelder and Mead (1965) was then used to estimate a and b with starting values of $\widehat{a} = 1.3$ and $\widehat{b} = 0.4$, initial value $\widehat{y}_0 = 0.1$, number of simulated replications $m = 50$, and time-series length $T = 11$. Both, the Euler and Taylor SDE solver schemes, referred to above, were used. Parameter estimation was carried out 500 times for each equation and solver scheme. The mean and standard deviation of the parameter estimates are reported in Table 1. It should be noted that, because of the heavy computational burden associated with enumerating $KS(D)$ in (10), we used the approximation

$$KS\left(D\right) \approx 1 - \exp\left(-2D^2\right) \tag{12}$$

which is the asymptotic distribution of the one-sided two-sample test outlined by Gibbons (1985), pp. 130–131.

Table 1. Mean and standard deviation of parameter estimates

Model	Scheme	Mean value	Standard deviation
A	Taylor	$\overline{a} = 1.0189$ $\overline{b} = 0.5224$	$s_a = 0.0335$ $s_b = 0.0324$
A	Euler	$\overline{a} = 1.0152$ $\overline{b} = 0.5192$	$s_a = 0.0313$ $s_b = 0.0299$
B	Taylor	$\overline{b} = 0.5189$	$s_b = 0.0211$
C	Taylor	$\overline{a} = 1.0552$ $\overline{b} = 0.5353$	$s_a = 0.0447$ $s_b = 0.0271$
D	Taylor	$\overline{a} = 0.9906$ $\overline{b} = 0.5155$	$s_a = 0.0103$ $s_b = 0.0310$
E	Taylor	$\overline{a} = 1.0418$ $\overline{b} = 0.5009$	$s_a = 0.0455$ $s_b = 0.0145$
E	Euler	$\overline{a} = 1.0298$ $\overline{b} = 0.5069$	$s_a = 0.0455$ $s_b = 0.0125$

The results presented in Table 1 reveal a small bias in the estimates of a and b over the 500 replications of parameter estimation. In all cases, however, the mean of the point estimates is well within 1.5 standard deviations of the true value. Interestingly the lower order of convergence of the Euler scheme is associated with smaller bias and standard deviation estimates for Equations (A) and (E). This is likely to be a result of the use of intermediate time steps with the equation solver to improve the SDE simulations.

In practice we can use only approximation of statistics $KS\left(D\right)$.In one of the experiments the maximum value $j = 7$ were applied when estimating the parameters of Equation (E). This gave mean and standard deviation of 0.998 and 0.038 for $\widehat{a}$ and of 0.499 and 0.013 for $\widehat{b}$, what means an apparent reduction in bias, as well as an improvement in precision. On the basis of these results it seems that there is room for further investigation of the form of criterion function used for parameter estimation.

6. Conclusion

Estimation of the parameters of five linear and nonlinear SDE using a criterion function based on Kolmogorov–Smirnov statistic has been examined. Although the method used is demanding computationally, the results are satisfactory with respect to both point and interval estimation.

References

Gibbons, J. D. (1985). *Nonparametric Statistical Inference (2nd Ed.)*. Marcel Dekker, New York.

Kloeden, P. E., Platen, E. and Schurz, H. (1994). *Numerical Solution of SDE through Computer Experiments*. Springer-Verlag, Berlin.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7**, 308–313.

University of Kielce, Krakowska 11, 25029 Kielce, Poland
E-mail address: daria_filatova@interia.pl

Politechnical University of Radom, Malczewskiego 20A, 26600 Radom, Poland
E-mail address: mgrzyw@interia.pl

CSIRO Marine Research, GPO Box 1538, Hobart Tasmania 7001, Australia