Commun Nonlinear Sci Numer Simulat 62 (2018) 445–453

Contents lists available at ScienceDirect

Commun Nonlinear Sci Numer Simulat

journal homepage: www.elsevier.com/locate/cnsns

Research paper

# Detecting the sampling rate through observations

Isao Shoji

*Tokyo University of Science, Fujimi, Chiyoda, Tokyo 102-0071, Japan*

ARTICLE INFO

*Article history:*
Received 1 September 2017
Revised 19 January 2018
Accepted 23 February 2018
Available online 26 February 2018

*Keywords:*
Sampling rate
Stochastic differential equation
Maximum likelihood estimation
Kullback–Leibler divergence

ABSTRACT

This paper proposes a method to detect the sampling rate of discrete time series of diffusion processes. Using the maximum likelihood estimates of the parameters of a diffusion process, we establish a criterion based on the Kullback–Leibler divergence and thereby estimate the sampling rate. Simulation studies are conducted to check whether the method can detect the sampling rates from data and their results show a good performance in the detection. In addition, the method is applied to a financial time series sampled on daily basis and shows the detected sampling rate is different from the conventional rates.



## 1. Introduction

The time evolution of stochastic dynamic phenomena is formulated by stochastic differential equations (SDE). For practical use of these equations, we have to estimate their parameters from data. Here, we should pay our attention to the qualitative difference between the model formulated by an SDE and data used for estimation because the model is expressed in the continuous time framework while the data are observed at discrete times. To bridge the difference, we have to develop schemes for discretization to make a model fit to data rather than obtain continuous time data because the latter is basically difficult due to accuracy of observation by experimental devices. Actually, there are various studies to pursue this idea; see, for example, [4], [5,6,8,9] and [11].

It is particularly important for the discretization to determine the sampling interval of data $\Delta t$, or the sampling rate $1/\Delta t$. When using small $\Delta t$, the discretized model of $\Delta t$ would show the short term behavior, whereas the model would show the longer time behavior when using larger $\Delta t$. Conventionally, we often determine how long $\Delta t$ is, depending on the sampling frequency such as daily, monthly, or annually. This convention, however, is useless to determine the sampling interval. For example, suppose time series data observed every day. When using 1 day as the unit time scale, $\Delta t$ is equal to 1, where as it is equal to 1/365 when using 1 year as the unit time scale. The discretized model would show the somewhat longer time behavior in the former case but the short term behavior in the latter case although the data themselves are just identical. Moreover, from a relativistic point of view, observers in different inertial frames would observe different $\Delta t$ because of their different elapsed time even if we take no account of the unit time scale; one observer at a rest frame $O$ observes data every $\Delta t$, while another observer moving at velocity $v$ relative to $O$ observes the same data every $\Delta t' = \Delta t\sqrt{1 - v^2/c^2}$ for the speed of light $c$.

Additionally, $\Delta t$ has a significant influence on the performance of discretization. Deriving a discretized model from an SDE is basically equivalent to solving the SDE. Since it is generally difficult to obtain the exact solution to such an SDE, especially as nonlinear stochastic differential equations, approximation methods are often used for the derivation. The per-

*E-mail address:* shoji@rs.tus.ac.jp

https://doi.org/10.1016/j.cnsns.2018.02.031
1007-5704/

formance of the approximation is measured by the rate of convergence in $\Delta t$. For example, when applying the well-known Euler method to SDEs, $O(\Delta t)$ is known as its rate of convergence; see [7] for example. Thus the better the performance of approximation the shorter $\Delta t$.

Unless running simulations, however, we are unable to take $\Delta t$ as short as we want. Rather, in applications we have to identify how long it is. Given time series, we first assume a statistical model, or a data generating process, which is thought to generate the time series with specific values of its parameters and $\Delta t$. In parameter estimation we have main interest in what values those parameters have, but often do not care about what value $\Delta t$ is. Since the observed time series are assumed to follow the data generating process, the value of $\Delta t$ as well as the values of the parameters must be estimated from the time series.

At a first glance, the estimation looks easy. However, we cannot estimate the parameters and $\Delta t$ simultaneously. Take a Brownian motion with drift for example, which are formulated by $dX_t = \mu dt + \sigma dB_t$, where $B_t$ is the standard Brownian motion. Here, we want to estimate $\mu$, $\sigma$, and $\Delta t$ from discrete time series $\{X_{t_k}\}_{1\le k\le n}$ of its time interval, $t_k - t_{k-1} = \Delta t$. To this end, the maximum likelihood estimation can be applied because $X_{t_k} - X_{t_{k-1}}$ follows the normal distribution with mean $\mu\Delta t$ and variance $\sigma^2\Delta t$. It can be easily seen that we are unable to estimate the parameters and the sampling interval simultaneously. Alternatively, we could estimate the parameters and the sampling interval separately as follows. Let $\Delta t$ be fixed at some value and then maximize the log-likelihood function with respect to parameters. Repeat this procedure for one possible sampling interval after another. Among those obtained maximized log-likelihoods, it seems that we have only to choose the sampling interval at which the maximized log-likelihood attains the maximum. This straightforward approach, however, is shown to fail to work.

On the basis of the Kullback–Leibler divergence, this paper proposes the mean log-likelihood as criterion for the detection. In this approach, we first estimate the parameters of an SDE via the maximum likelihood method from one discrete sample path given a possible sampling rate. And then, we construct a mean log-likelihood from another sample paths independent of each other. Repeating this procedure by changing one possible sampling rate after another, we choose the sampling rate at which the mean log-likelihood attains the maximum.

To check the performance of the proposed method numerically, simulation studies are conducted by using well-known SDEs with several combinations of sampling rates and the number of observations. In addition, we apply the method to financial time series sampled on daily basis and detect its sampling rate implied by the data. The analysis shows that the detected sampling rate is different from such rates as implied by conventionally taking 1 week, month, or year as the unit time scale.

The paper is organized as follows. In Section 2 we show the maximized log-likelihood fails to work as criterion for comparing sampling rates, but instead the mean log-likelihood is proposed. In Section 3 the numerical experiments are carried out to check the numerical performance of the proposed method. In Section 4 an empirical application to the Japanese stock price index is provided and the concluding remarks are given in the final section.

## 2. Methods of detecting the sampling rate

Suppose a diffusion process $X_t$ starting at $x_0$ satisfies the following SDE,

$$\begin{aligned} dX_t &= \mu(X_t;\theta)dt + \sigma(X_t;\theta)dB_t \\ X_0 &= x_0, \end{aligned} \tag{1}$$

where $B_t$ is the standard Brownian motion and $\mu(X_t;\theta)$ and $\sigma(X_t;\theta)$ represent functions of $X_t$ with an unknown parameter vector $\theta$. Here, we have time series with an equidistant sampling interval, or a discrete sample path of the process $\{X_{t_k}\}_{1\le k\le n}$ $(t_k - t_{k-1} = \tau_0)$, thereby we estimate the unknown $\theta$. Unlike usual settings, however, we assume that we have no information on the physical sampling interval $\tau_0$ or rate $1/\tau_0$ although we may know those data are sampled on a daily, weekly, monthly, or annually basis. Those conventional sampling frequency means neither sampling interval nor rate because the sampling interval or rate is determined by the unit time scale which is usually unknown to experimenters. Therefore, we have to estimate not only the parameter vector but also the sampling interval or rate.

In the following discussion, we focus one-dimensional diffusion processes for simplicity. Those arguments, however, can be easily extended to multi-variate diffusion processes.

### 2.1. Maximized log-likelihood

Here, consider the maximum likelihood (ML) method for example. Let $l(\theta)$ be the logarithm of the likelihood function for given $\{X_{t_k}\}_{1\le k\le n}$. We get the ML estimate $\hat{\theta}$ by maximizing $l(\theta)$ with respect to $\theta$ when the sampling interval is known. In our settings, however, the true sampling interval $\tau_0$ is unknown. So, first we tentatively assign a possible value to the sampling interval, denoted by $\tau$, and then carry out the maximum likelihood estimation. The estimate obtained in such a way depends on $\tau$ used for estimation, and thus the estimate should be expressed as a function of $\tau$, or $\hat{\theta}(\tau)$.

In terms of the ML method, the greater $l(\hat{\theta}(\tau))$ is more desirable. Thus, it seems that we have only to choose $\tau$ which gives the largest $l(\hat{\theta}(\tau))$ among possible sampling intervals. But, this straightforward approach fails. Consider the differential