Computational Statistics and Data Analysis 84 (2015) 54–67

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A penalized simulated maximum likelihood approach in parameter estimation for stochastic differential equations

Libo Sun, Chihoon Lee *, Jennifer A. Hoeting

Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA

ARTICLE INFO

Article history:
Received 12 May 2014
Received in revised form 3 October 2014
Accepted 12 November 2014
Available online 20 November 2014

Keywords:
Chronic wasting disease
Euler–Maruyama scheme
Penalized simulated maximum likelihood estimation
Partially observed discrete sparse data
Auxiliary importance sampling
Stochastic differential equations

ABSTRACT

We consider the problem of estimating parameters of stochastic differential equations (SDEs) with discrete-time observations that are either completely or partially observed. The transition density between two observations is generally unknown. We propose an importance sampling approach with an auxiliary parameter when the transition density is unknown. We embed the auxiliary importance sampler in a penalized maximum likelihood framework which produces more accurate and computationally efficient parameter estimates. Simulation studies in three different models illustrate promising improvements of the new penalized simulated maximum likelihood method. The new procedure is designed for the challenging case when some state variables are unobserved and moreover, observed states are sparse over time, which commonly arises in ecological studies. We apply this new approach to two epidemics of chronic wasting disease in mule deer.

1. Introduction

It is very important for ecologists and wildlife managers to understand the dynamics of infectious diseases, such as chronic wasting disease (CWD) which is a fatal contagious disease in cervid populations (Miller et al., 2006). Several ordinary differential equation models have been proposed by Miller et al. (2006) to describe the transmission mechanism of CWD. Stochastic epidemic models allow more realistic description of the transmission of disease as compared to deterministic epidemic models (Becker, 1979; Andersson and Britton, 2000). However, parameter estimation is challenging for discretely observed data for stochastic models (Sørensen, 2004; Jimenez et al., 2005). Stochastic differential equation (SDE) models are a natural extension of ordinary differential equation models and they may be simpler to derive and apply than Markov chain models. For example, the transition matrix in Markov chain models can be very complicated when there are several interacting populations (Allen and Allen, 2003; Allen et al., 2005). Moreover, SDEs have broader application areas, which include not only ecology and biology but also economics, finance, bioinformatics, and engineering.

Various methods for inferential problems for SDEs have been developed. The Hermite polynomial expansion approach proposed by Aït-Sahalia (2002, 2008) may perform poorly if the data are sparsely sampled (Stramer and Yan, 2007b). Moreover, this approach has some restrictions which could limit its application, especially for multivariate models (Lindström, 2012). Särkkä and Sottinen (2008) proposed an approach which uses an alternative SDE as an importance process and the Girsanov theorem to help evaluate the likelihood ratios of two SDEs. However, the diffusion coefficient of their model is state-independent, whereas general SDE models allow for a state-dependent diffusion coefficient. Recent developments have mainly been focused on Bayesian approaches (Eraker, 2001; Golightly and Wilkinson, 2005, 2006, 2011; Donnet et al., 2010), which can suffer a very slow rate of convergence as the dimension of the model increases and the data are sparsely sampled. We propose a penalized simulated maximum likelihood (PSML) approach which is computationally feasible.

* Corresponding author. Tel.: +1 970 491 7321.
E-mail addresses: sun@stat.colostate.edu (L. Sun), chihoon@stat.colostate.edu (C. Lee), jah@stat.colostate.edu (J.A. Hoeting).

http://dx.doi.org/10.1016/j.csda.2014.11.007
0167-9473/© 2014 Elsevier B.V. All rights reserved.

For a SDE model the transition density between two observations is known in only a few univariate cases. Pedersen (1995) firstly proposed a simulated maximum likelihood (SML) approach which integrates out the unobserved states using Monte Carlo integration with importance sampling. We refer to the basic sampler in this approach as the Pedersen sampler. Although the Pedersen sampler may provide estimates that are arbitrarily close to the true transition density, it is computationally expensive in practice. Durham and Gallant (2002) proposed several different importance samplers in a SML framework to improve the efficiency of the Pedersen sampler. They concluded their modified Brownian bridge (MBB) sampler has the best performance in terms of accuracy in root mean square error and efficiency in time. Richard and Zhang (2007) proposed an efficient importance sampling technique which converts the problem of minimizing the variance of an approximate likelihood to a recursive sequence of auxiliary least squares optimization problems. Pastorello and Rossi (2010) applied Richard and Zhang's approach to estimate the parameters of some univariate SDE models. However, the extension to multivariate SDEs with partially observed data is not trivial. Lindström (2012) introduced a regularized bridge sampler, which is a weighted combination of the Pedersen sampler and the MBB sampler, for sparsely sampled data.

The methods of Pedersen (1995) and Durham and Gallant (2002) have mainly been applied in the area of econometrics. Here we propose a methodology to improve the MBB sampler and the regularized sampler and extend them to the area of ecology. From an inferential viewpoint, practitioners must contend with two major challenges: (a) in the multivariate state space, some state variables are completely unobserved; (b) observed data are quite sparse over time. These are common features of ecological data. For example, the number of deaths for CWD in a wild animal population can be observed or estimated, but the numbers of infected and susceptible animals may be impossible or costly to obtain. Moreover, the time interval between two consecutive observations could be very long, usually weeks or even months. With such partially observed sparse data, the MBB approach no longer has the same promising results as in the univariate case. Although the regularized sampler in Lindström (2012) is designed for sparsely sampled data, the optimal choice of the weight parameter ρ (which is denoted as α in the cited paper) needs to be determined. We propose an importance sampling approach with an auxiliary parameter which provides more accurate estimates of the parameters of an SDE when the transition density is unknown. We embed the auxiliary importance sampler in a penalized maximum likelihood framework. The penalty term we add to the log likelihood is a constraint on selecting the importance sampler. We show via simulation studies that our approach improves the accuracy of parameter estimates and computational efficiency compared to the MBB sampler and the regularized sampler.

The remainder of the paper is organized as follows. In Section 2, we present the general multivariate SDE model. Section 3 provides brief descriptions of the Pedersen, MBB and regularized samplers. Section 4 describes our methodology in detail. Section 5 presents simulation studies for different models. Section 6 illustrates our method on a CWD dataset as a real world example. Section 7 concludes with a discussion.

2. Background

We begin with the general multivariate SDE model where some state variables are unobserved. Let $\boldsymbol{X}(t) = \{X_1(t), \ldots, X_k(t)\}^T$ denote a k-dimensional state variable vector at time $t \geq 0$. Consider a multivariate SDE model,

$$d\boldsymbol{X}(t) = f(\boldsymbol{X}(t), \boldsymbol{\theta})dt + g(\boldsymbol{X}(t), \boldsymbol{\theta})d\boldsymbol{W}(t) \tag{1}$$

with known initial condition $\boldsymbol{X}(t_0) = \boldsymbol{x}_0$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ is an unknown p-dimensional parameter vector, $\boldsymbol{W}$ is a k-dimensional standard Wiener process, and both functions $f : \mathbb{R}^k \times \Theta \to \mathbb{R}^k$ and $g : \mathbb{R}^k \times \Theta \to \mathbb{R}^{k\times k}$ are known. Note that the derivation below still holds for the case with unknown initial condition $\boldsymbol{X}(t_0)$, which can be treated as another unknown parameter. We assume that the SDE (1) has a unique weak solution. See Øksendal (2010, Chapter 5) for conditions that ensure this.

We assume that only a subset of the state process $\{\boldsymbol{X}_{\text{obs}}(t)\}_{t\geq 0}$ can be observed at discrete time points. It is natural to suppose only $\boldsymbol{X}_{\text{obs}}(t_i) = \{X_j(t_i), \ldots, X_k(t_i)\}$ is observed at t_i, for $1 < j \leq k$ and $i = 1, \ldots, n$, and all other state variables $\boldsymbol{X}_{-\text{obs}}(t_i) = \{X_1(t_i), \ldots, X_{j-1}(t_i)\}$, are unobserved. In the case of complete observation, that is when $j = 1$, a similar derivation as below can be obtained. Note that time intervals do not have to be equidistant.

The discrete-time likelihood of model (1) is given by

$$L(\boldsymbol{\theta}) = p(\boldsymbol{X}_{\text{obs}}(t_1)|\boldsymbol{X}(t_0), \boldsymbol{\theta}) \prod_{i=2}^{n} p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}), \boldsymbol{\theta}) \tag{2}$$

where $\boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})$ denotes all observations of $\boldsymbol{X}_{\text{obs}}$ from time t_1 to t_{i-1}. We omit the parameter $\boldsymbol{\theta}$ for brevity from now on. Notice that the term $p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))$ is not available in closed form except for simple cases. However, factoring the likelihood as in (2) allows us to evaluate the likelihood given by

$$p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})) = \int p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_{i-1}))p(\boldsymbol{X}_{-\text{obs}}(t_{i-1})|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))d\boldsymbol{X}_{-\text{obs}}(t_{i-1}).$$

A feasible approach to evaluate this integral is via Monte Carlo integration. That requires a method to draw samples from the distribution of $\boldsymbol{X}_{-\text{obs}}(t_{i-1})|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})$. It can be shown that (cf. Durham and Gallant, 2002)

$$p(\boldsymbol{X}_{-\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_i)) \propto \int p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))p(\boldsymbol{X}_{-\text{obs}}(t_{i-1})|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))d\boldsymbol{X}_{-\text{obs}}(t_{i-1}), \tag{3}$$