Penalized importance sampling for parameter estimation
in stochastic differential equations

Libo Sun∗, Chihoon Lee†, and Jennifer A. Hoeting‡
Department of Statistics, Colorado State University, Fort Collins, Colorado

Abstract

We consider the problem of estimating parameters of stochastic differential equations with discrete-time observations that are either completely or partially observed. The transition density between two observations is generally unknown. We propose a penalized importance sampling approach to approximate the transition density. Simulation studies in three different models illustrate promising improvements of the new penalized importance sampling 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.

Keywords: Chronic wasting disease; Euler-Maruyama scheme; Maximum likelihood estimation; Partially observed discrete sparse data; Penalized importance sampling; Stochastic differential equations.

1 Introduction

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, such as stochastic differential equations (SDEs) (Sørensen, 2004; Jimenez et al., 2005). Recent developments have mainly been focused on Bayesian approaches (Eraker, 2001; Goliak, 2001; Goj, 2005; 2006; Toni et al., 2009; Donnet et al., 2010; Drovandi and Pettitt, 2011). Here we propose a frequentist approach which is computationally feasible. The transition density between two observations is known in only a few univariate cases. Pedersen (1995) proposed to integrate out the unobserved states using Monte Carlo simulation with importance sampling, which is called simulated maximum likelihood estimation (SMLE). This approach is promising from a theoretical point of view, but it is computationally expensive. Durham and Gallant (2002) proposed several different importance samplers
to improve the efficiency of SMLE. They conclude their modified Brownian bridge (MBB) has the best performance in terms of accuracy in root mean square error and efficiency in time.

The methods of Pedersen (1995) and Durham and Gallant (2002) have mainly been applied in the area of econometrics. Here we propose methodology to improve Durham and Gallant’s MBB approach and extend it 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 chronic wasting disease (CWD), a fatal contagious disease in cervid populations (Miller et al., 2006), in a wild animal population can be observed, but the numbers of infected and susceptible animals are 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. We propose a penalized importance sampling approach to approximate the transition density. The penalty term we add to the log likelihood is a constraint on selecting the importance sampler. We show via a simulation study that our approach significantly improves the accuracy of parameter estimates and efficiency of the method compared to the MBB.

In Section 2, we present the general multivariate SDE epidemic model. Section 3 gives a brief description of SMLE and MBB. Section 4 describes our methodology in detail. Section 5 presents some simulation studies for different models. Section 6 illustrates our method on a CWD dataset as an real world example. Section 7 concludes with discussion.

2 Background

Let \( X(t) = \{X_1(t), \ldots, X_k(t)\}^T \) denote a \( k \)-dimensional state variable vector at time \( t \). Consider a general multivariate SDE model,

\[
dX(t) = f(X(t), \theta)dt + g(X(t), \theta)dW(t)
\]

with known initial condition \( X(0) = x_0 \), and \( \theta \in \Theta \subseteq \mathbb{R}^p \) is an unknown \( p \)-dimensional parameter vector, \( W \) is a \( k \)-dimensional standard Wiener process, and both functions \( f : \mathbb{R}^k \times \Theta \rightarrow \mathbb{R}^k \) and \( g : \mathbb{R}^k \times \Theta \rightarrow \mathbb{R}^{k \times k} \) are known. We also 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 partial one-dimensional state process can be observed at discrete time points. It is natural to suppose only \( X_k(t_i) \) is observed at \( t_i, \) for \( i = 1, \ldots, n \), and all other \( X_j(t_i), \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, k-1 \), are unobserved. Note that time intervals do not have to be equidistant.

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

\[
L(\theta) = p(X_k(t_1)|X(t_0)) \prod_{i=2}^{n} p(X_k(t_i)|X(t_0), X_k(t_1:t_{i-1}))
\]
where $X_k(t_1 : t_{i-1})$ denotes all observations of $X_k$ from time $t_1$ to $t_{i-1}$, and the likelihood of $X_k(t_i)$ given all previous observations is given by

$$p(X_k(t_i) | X(t_0), X_k(t_1 : t_{i-1})) = \int p(X_k(t_i) | X(t_{i-1})) p(X_{-k}(t_{i-1}) | X(t_0), X_k(t_1 : t_{i-1})) dX_{-k}(t_{i-1}),$$

where $X_{-k}$ denotes $\{X_1, \cdots, X_{k-1}\}$. A feasible approach to evaluate this integral is via Monte Carlo integration. That requires a method to draw samples from the distribution of $X_{-k}(t_{i-1}) | X(t_0), X_k(t_1 : t_{i-1})$. It can be shown that

$$p(X_{-k}(t_i) | X(t_0), X_k(t_1 : t_i)) = \int p(X_{-k}(t_i) | X(t_{i-1}), X_k(t_i)) p(X_{-k}(t_{i-1}) | X(t_0), X_k(t_1 : t_i)) dX_{-k}(t_{i-1}),$$

Plugging (3) into (2), then

$$p(X_{-k}(t_i) | X(t_0), X_k(t_1 : t_i)) \propto \int p(X_{-k}(t_i) | X(t_{i-1}), X_k(t_i)) p(X_{-k}(t_{i-1}) | X(t_0), X_k(t_1 : t_i)) dX_{-k}(t_{i-1}).$$

Plugging (3) into (2), then

$$p(X_{-k}(t_i) | X(t_0), X_k(t_1 : t_i)) \propto \int p(X(t_i) | X(t_{i-1})) p(X_{-k}(t_{i-1}) | X(t_0), X_k(t_1 : t_{i-1})) dX_{-k}(t_{i-1}).$$

Hence, given the distribution of $X_{-k}(t_{i-1}) | X(t_0), X_k(t_1 : t_{i-1})$, we can approximate the distribution of $X_{-k}(t_i) | X(t_0), X_k(t_1 : t_i)$ by Monte Carlo integration on (4).

Therefore it is left to find the transition probability density $p(X(t_i) | X(t_{i-1}))$ which has no closed form in most cases. The Euler-Maruyama scheme (Kloeden and Platen, 1992) is a common approach to approximate this density which is given by

$$X(t + \delta) - X(t) \approx f(X(t), \theta)\delta + g(X(t), \theta)(W(t + \delta) - W(t)),$$

where $\delta$ is called step size, $W(t + \delta) - W(t)$ follows a multivariate normal distribution with variance matrix $\delta I_{k \times k}$, where $I$ is the identity matrix. This Euler-Maruyama scheme works well if the step size is small. Hence, if the time interval between two observations is small enough, we can approximate $p(X(t_i) | X(t_{i-1}))$ using a multivariate normal density.

If the time interval between observations is large, the above approximation will introduce bias. We can partition the interval $t_{i-1}$ to $t_i$ to $M$ subintervals such that $\delta = (t_i - t_{i-1}) / M$ is small enough for the Euler-Maruyama scheme. By the Markov property, $p(X(t_i) | X(t_{i-1}))$ can be estimated by

$$\int \prod_{m=1}^{M} p(X(t_{i-1} + m\delta) | X(t_{i-1} + (m-1)\delta)) dX((t_{i-1} + \delta) : (t_i - \delta)),$$
which was proven in Pedersen (1995).

Then, our goal is to approximate (6). Using importance sampling, we draw \( J \) samples, \( \{X^{(j)}((t_{i-1} + \delta) : (t_i - \delta)) : j = 1, \ldots, J \} \), from an importance sampler \( q \), then (6) can be approximated by

\[
\frac{1}{J} \sum_{j=1}^{J} \prod_{m=1}^{M} \frac{p(X^{(j)}((t_{i-1} + m\delta) : (t_i - \delta)))}{q(X^{(j)}((t_{i-1} + \delta) : (t_i - \delta)))}.
\] (7)

The convergence of the importance sampling estimator (7) as \( J \to \infty \) is shown by Geweke (1989). The expectation of (7) is \( p(X(t_i)|X(t_{i-1})) \), regardless of the choice of the importance sampler \( q \). The variance of (7) becomes

\[
\frac{1}{J} \left[ \int \prod_{m=1}^{M} \frac{p^2(X^{(j)}((t_{i-1} + m\delta) : (t_i - \delta)))}{q(X^{(j)}((t_{i-1} + \delta) : (t_i - \delta)))} dX((t_{i-1} + \delta) : (t_i - \delta)) - p^2(X(t_i)|X(t_{i-1})) \right],
\] (8)

which attains its minimum of 0 when

\[
q(X^{(j)}((t_{i-1} + \delta) : (t_i - \delta))) = \frac{\prod_{m=1}^{M} p(X^{(j)}((t_{i-1} + m\delta) : (t_i - \delta)))}{p(X(t_i)|X(t_{i-1}))}.
\] (9)

Thus in theory a single sample is sufficient to approximate \( p(X(t_i)|X(t_{i-1})) \). However, in practice this is infeasible because \( p(X(t_i)|X(t_{i-1})) \) is unknown.

In order to decrease (8) and reduce the sample size \( J \), we want to choose an importance sampler \( q(X^{(j)}((t_{i-1} + \delta) : (t_i - \delta))) \) that is as close as possible to \( \prod_{m=1}^{M} p(X^{(j)}((t_{i-1} + m\delta) : (t_i - \delta))) \), which is the principle of choosing the proposal density in importance sampling.

3 Importance sampler

3.1 Simulated maximum likelihood estimation

The simulated maximum likelihood estimation (SMLE) method was the first simulation-based approach proposed to approximate a transition density (Pedersen 1995, Santa-Clara 1997). SMLE constructs the importance sampler \( q \) by simulating \( J \) paths on each subinterval just using the Euler-Maruyama scheme conditional on \( X(t_{i-1}) \), so the first \( M - 1 \) terms in (7) are canceled. Hence, (7) reduces to

\[
\frac{1}{J} \sum_{j=1}^{J} p(X(t_i)|X^{(j)}(t_i - \delta)).
\]

One can simulate \( J \) trajectories of all \( k \)-dimensional state process \( X \) from time \( t_{i-1} \) to time \( t_i - \delta \) by using the Euler-Maruyama scheme with the step size \( \delta \). Although the
SMLE has a very simple form, it is well known that it is computationally intense in practice (Durham and Gallant, 2002), especially for a multivariate SDE model. SMLE can introduce excessive variance in the simulation of all possible transition probabilities even with a very large number of simulated trajectories.

3.2 Modified Brownian bridge

A more efficient importance sampler is called MBB, which was originally proposed by Durham and Gallant (2002) for the univariate case and modified by Golightly and Wilkinson (2006) for multivariate case. Instead of simulating a path on each subinterval using the Euler approximation based on \(X(t_{i-1})\) as in Pederson’s SMLE, this method draws \(X((t_{i-1} + \delta) : (t_i - \delta))\) conditional on \(X(t_{i-1})\) and \(X_k(t_i)\). Here, we outline the procedure. See Golightly and Wilkinson (2006) for more details.

Let \(X^m\) denote \(X(t_{i-1} + m\delta)\), and partition the drift and diffusion functions in (11) as

\[
\mathbf{f}(\mathbf{X}) = \begin{pmatrix} f_{-k}(\mathbf{X}) \\ f_k(\mathbf{X}) \end{pmatrix},
\]

and

\[
\mathbf{g}^T(\mathbf{X}) \mathbf{g}(\mathbf{X}) = \begin{bmatrix} G_{-k,-k}(\mathbf{X}) & G_{-k,k}(\mathbf{X}) \\ G_{k,-k}(\mathbf{X}) & G_{k,k}(\mathbf{X}) \end{bmatrix}.
\]

Then the MBB draws \(X^{m+1}\) from the density

\[
q(X^{m+1} | X^m, X_k(t_i)) = \phi(X^{m+1}; X^m + \eta_m\delta, \Sigma_m\delta),
\]

(10)

where \(\phi(x; \mu, \Sigma)\) is a multivariate normal density with mean vector \(\mu\) and covariance matrix \(\Sigma\). Here

\[
\eta_m = \left( f_{-k}(X^m) + \frac{G_{-k,k}(X^m)}{\delta(M-m)G_{k,k}(X^m)}[X_k(t_i) - (X_k(t_{i-1} + m\delta) + f_k(X^m)(M-m-1)\delta)] \\ (X_k(t_i) - X_k(t_{i-1} + m\delta))/[\delta(M-m)] \right),
\]

(11)

and

\[
\Sigma_m = \begin{bmatrix} G_{-k,-k}(X^m) - \frac{G_{-k,k}(X^m)G_{k,-k}(X^m)}{M-m} & \frac{M-m-1}{M-m}G_{-k,k}(X^m) \\ \frac{M-m-1}{M-m}G_{k,-k}(X^m) & \frac{M-m-1}{M-m}G_{k,k}(X^m) \end{bmatrix},
\]

(12)

for \(m = 0, 1, \cdots, M-2\). For \(m = M-1\), we draw \(X_{-k}(t_i)\) conditional on \(X^{M-1} = X(t_i - \delta)\) and \(X_k(t_i)\), which is the conditional multivariate normal by the Euler-Maruyama scheme. By recursively drawing from (10) one can get a Brownian bridge, \(X((t_{i-1} + \delta) : (t_i - \delta))\) conditioned on starting at \(X(t_{i-1})\) and finishing at \(X_k(t_i)\).

4 Penalized importance sampling

To find an even more efficient importance sampler, we need to minimize (8), the variance of the approximation of the transition density. 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. Here we consider a similar idea but combine the auxiliary parameter with model parameters by maximizing the log likelihood with a constraint on the coefficient of variation of the importance sampler.

We maximize the log likelihood subject to the sum of the coefficient of variation of the Monte Carlo approximation of the transition density being less than a prespecified level. Suppose a family of auxiliary importance samplers \( \{q_\rho\} \) has been selected, where \( \rho \) is the auxiliary parameter. Hence, our goal is to find \( \hat{\rho} \) that minimizes the sum of the coefficient of variation of the Monte Carlo approximation of the transition density. The penalized importance sampling estimator \( (\hat{\theta}, \hat{\rho}) \) is defined by

\[
(\hat{\theta}, \hat{\rho}) = \arg \max \sum_{i=1}^{n} \log(p_i^*) \text{ subject to } \sum_{i=1}^{n} \hat{c}v(p_i^*) \leq s,
\]

where \( s \geq 0 \) is a tuning parameter, \( p_i^* \) is the approximate transition density approximated via importance sampling and \( \hat{c}v(p_i^*) \) is the sample coefficient of variation of \( p_i^* \). Notice that this is reminiscent of LASSO (Tibshirani, 1996). This is equivalent to maximizing a penalized log likelihood,

\[
l^*(\theta, \rho) = \sum_{i=1}^{n} \log(p_i^*) - \lambda \sum_{i=1}^{n} \hat{c}v(p_i^*),
\]

where \( \lambda \) in (14) has a one-to-one mapping to \( s \) in (13).

The constraint, \( \sum_{i=1}^{n} \hat{c}v(p_i^*) \leq s \), is equivalent to a constraint on the effective sample size,

\[
\hat{N}(q, p) = \frac{J}{1 + \frac{1}{n} \sum_{i=1}^{n} \hat{c}v(p_i^*)} \geq \frac{J}{1 + \frac{s}{n}}.
\]

The effective sample size measures how much the importance sampler density \( q \) differs from the target density \( p \), and it can be interpreted as \( J \) weighted samples are worth \( \hat{N}(q, p) \) unweighted i.i.d. samples drawn exactly from target density \( p \) (Givens and Hoeting, 2012, Chapter 6). Effective sample size can be used as a measure of computational efficiency.

The tuning parameter \( s \) controls how close the importance sampler density \( q \) is to the transition probability density. Let \( s^0 \) be the sum of the coefficient of variation for the approximation of the transition density by SMLE. When \( s < s^0 \) the resulting penalized importance sampler will have smaller variance (8) than SMLE. When \( s = 0 \), the constraint in (13) requires that the importance sampler \( q \) attains its ideal case (9). However, as \( s \to 0 \), \( \lambda \to \infty \) so the log likelihood plays no role in estimating \( \theta \).

We can estimate the tuning parameter \( s \) or \( \lambda \) using various techniques. We choose the value that minimizes the estimated prediction error,

\[
\frac{1}{L} \sum_{\ell=1}^{L} \sum_{i=1}^{n} [\hat{X}_k^{(\ell)}(t_i) - X_k(t_i)]^2,
\]

where \( \hat{X}_k^{(\ell)}(t_i) \) is the \( \ell \)th simulated \( X_k \) at observation time \( t_i \) by the Euler-Maruyama scheme (5) with \( \theta = \hat{\theta} \). The number of simulations \( L \) is chosen arbitrarily and is set to 1000 here.
The class of auxiliary importance samplers that we consider here is given by
\[
\{N(X^{m+1}; X^m + \eta_m \delta, \rho \Sigma_m \delta), \rho \in (0, 1)\},
\]
(15)
where \(\eta_m, \Sigma_m\) are defined in (11) and (12), and \(\rho\) is the shrinkage coefficient, which will be estimated as an auxiliary parameter in the penalized log likelihood (14). Adding the shrinkage coefficient was a key improvement over MBB but the penalty term in (13) allows estimation of the auxiliary parameter \(\rho\) and leads to improved performance over MBB. One can also choose other families of auxiliary importance samplers, but (15) is a good starting point for illustration of the method. Other distributions, such as the Student’s \(t\) distribution, might also be a suitable choice.

We can use the parametric bootstrap (Efron, 1982, Chapter 5) to obtain confidence intervals for the estimator, which proceeds as follows. First, based on the parameter estimates from the real dataset, we can generate a large number of new datasets by using the Euler-Maruyama scheme for the SDE model. For each new simulated dataset, we obtain estimates of parameters using the penalized importance sampling method (13). Then we compute the confidence interval from those estimates using the corresponding quantiles.

5 Simulation studies

Here, we compare the performance of our penalized importance sampler (IS) with MBB on simulated datasets for three different models. For all the optimization algorithms in this paper, we use an implementation of the Nelder-Mead algorithm for derivative-free optimization (Varadhan and Borchers, 2011) in R (R Development Core Team, 2011). The initial values for the parameters are chosen arbitrarily.

5.1 Ornstein-Uhlenbeck process

We first consider a univariate SDE, the Ornstein-Uhlenbeck process
\[
dX = (\theta_1 - \theta_2 X)dt + \theta_3 dW,
\]
with known initial condition \(X(0)\), and the parameter \(\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+\). The parameter \(\theta_2\) is the speed of reversion, \(\theta_1/\theta_2\) is the long-run equilibrium value of the process, and \(\theta_3\) is interpreted as the volatility. We generate 100 datasets, each including 100 observations, with parameter \(\theta_0 = (0.0187, 0.2610, 0.0224)\) as reported in Aït-Sahalia (2002).

The transition density between two observations is given by
\[
X(t_{i+1})|X(t_i) \sim N\left(\frac{\theta_1}{\theta_2} + \left(X(t_i) - \frac{\theta_1}{\theta_2}\right) e^{-\theta_2 \Delta}, \frac{\theta_3^2 (1 - e^{-2\theta_2 \Delta})}{2 \theta_2}\right),
\]
where \(\Delta = t_{i+1} - t_i\) for \(i = 1, \ldots, N - 1\). Hence, the exact likelihood is known for this case and we can obtain the exact maximum likelihood estimator of the parameters \(\theta = (\theta_1, \theta_2, \theta_3)\). We compute the bias and the root mean square error (RMSE) of the simulated maximum likelihood estimators \(\hat{\theta}_r\) with respect to the exact maximum likelihood estimators \(\hat{\theta}_{MLE}\),
Table 1: The bias and RMSE of the simulated maximum likelihood estimates with respect to the exact maximum likelihood estimates for the Ornstein-Uhlenbeck process.

| Method       | $J = 8$                        | $J = 16$                        |
|--------------|--------------------------------|--------------------------------|
|              | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_1$ | $\theta_2$ | $\theta_3$ |
| Bias         |            |            |            |            |            |            |
| MBB          | 0.0046     | 0.0382     | 0.0035     | 0.0017     | 0.0121     | 0.0012     |
| Penalized IS | 0.0010     | -0.0102    | 0.0001     | -0.0012    | -0.0115    | -0.0003    |
| RMSE         |            |            |            |            |            |            |
| MBB          | 0.0115     | 0.0982     | 0.0093     | 0.0074     | 0.0585     | 0.0057     |
| Penalized IS | 0.0020     | 0.0235     | 0.0021     | 0.0017     | 0.0209     | 0.0013     |

defined by $\frac{1}{100} \sum_{r=1}^{100} (\hat{\theta}_r - \hat{\theta}_{MLE})$ and $\sqrt{\frac{1}{100} \sum_{r=1}^{100} (\hat{\theta}_r - \hat{\theta}_{MLE})^2}$, respectively. For both MBB and the penalized IS methods, we consider $M = 8$ subintervals but with two levels of the number of simulated sample paths, $J = 8$ and $J = 16$. The penalized IS with $J = 8$ takes approximately the same computational time as the MBB with $J = 16$. That is less than 3 minutes to get the simulated maximum likelihood estimator on a Quad Core Intel Xeon W3565 3.2 GHz with CentOS 6 computer. The computational time grows rapidly in $J$ for both algorithms. The results are shown in Table 1.

The penalized IS has much better performance than MBB in terms of reducing bias and RMSE, especially when the number of sample paths is small ($J = 8$). However, for the same computation time, the penalized IS with $J = 8$ dramatically improves performance over MBB with $J = 16$. To obtain a similar level of accuracy, MBB requires much larger $J$, which means much longer computation time. Generally speaking, as $J$ increases, the difference between MBB and the penalized IS decreases (results not shown). A similar phenomenon has also been observed for MBB and SMLE; with a sufficiently large $J$ the approximated transition density by either SMLE or MBB converges to the true transition density under mild regularity conditions (Durham and Gallant, 2002).

Figure 1 shows that some estimates are far away from the exact maximum likelihood estimates for MBB. This typically happens when the MBB algorithm gets stuck at the initial value for optimization, which is also an indication of the poor approximation of the likelihood. The proposed penalized importance sampler has better performance in this regard.

5.2 Stochastic volatility model

Next, we consider the two-dimensional stochastic volatility model in Durham and Gallant (2002), which is given by

\[
\begin{align*}
    dX &= (\theta_1 + \theta_2 X)dt + e^{H/2}X^{1/2}dW_1, \\
    dH &= (\theta_3 + \theta_4 H)dt + \theta_5 dW_2,
\end{align*}
\]

(16)

with known initial condition $(X(0), H(0))$, where $W_1$ and $W_2$ are two independent Wiener processes. Only the process $X$ is observed at $t_i$ for $i = 1, \ldots, n$ and the volatility factor $H$ is assumed to be unobserved.
We again generate 100 datasets, each including 500 observations, with parameter $\theta_0 = (\theta_1 = 0.0002, \theta_2 = -0.002, \theta_3 = -0.3, \theta_4 = -0.03, \theta_5 = 0.3)$. In this case, the exact transition density is no longer available. We can only compute the bias and the RMSE of the simulated maximum likelihood estimators $\hat{\theta}_r$ with respect to the true parameters $\theta_0$, defined by $\frac{1}{100} \sum_{r=1}^{100} (\hat{\theta}_r - \theta_0)$ and $\sqrt{\frac{1}{100} \sum_{r=1}^{100} (\hat{\theta}_r - \theta_0)^2}$, respectively. Estimates are obtained using $J = 128$ sample paths and $M = 8$ subintervals as suggested in Durham and Gallant (2002). Again, the penalized IS performs better than MBB for all parameters except $\theta_5$ (Table 2).

Note that researchers have found that estimation improves when the stochastic volatility model (16) is reparameterized using $\theta_1/\theta_2$ and $\theta_3/\theta_4$ instead of $\theta_1$ and $\theta_3$, respectively (Durham and Gallant, 2002), that is

\[
\begin{align*}
\frac{dX}{dt} &= \theta_2 (X - \mu) dt + \sigma e^{H \frac{X}{2}} dW_1, \\
\frac{dH}{dt} &= \theta_4 H dt + \theta_5 dW_2,
\end{align*}
\]

where $\mu = -\theta_1/\theta_2$ and $\sigma = e^{-\theta_3/\theta_4}$. We can see the improvement of the penalized IS over MBB becomes greater when using $\theta_3/\theta_4$ instead of $\theta_3$.

### 5.3 CWD direct model

The specific model and background are introduced in Section 6. Again, we generate 100 datasets, each including 21 annual observations from two distinct CWD epidemics similar to the real dataset in Section 6, by using the CWD direct model (17) with parameter ($\beta_0 = \ldots$)
Table 2: The bias and RMSE of the simulated maximum likelihood estimates with respect to the true parameters for the stochastic volatility model.

| Method          | $\theta_1$  | $\theta_2$  | $\theta_1/\theta_2$ | $\theta_3$  | $\theta_4$  | $\theta_3/\theta_4$ | $\theta_5$  |
|-----------------|-------------|-------------|----------------------|-------------|-------------|----------------------|-------------|
| Bias            | MBB         | $-0.0001$   | $0.00003$            | $0.0312$    | $0.1788$    | $0.0186$             | $1.1654$    | $0.0014$            |
|                 | Penalized IS| $-0.0001$   | $0.00002$            | $0.0388$    | $0.1673$    | $0.0162$             | $-0.3782$   | $-0.1181$           |
| RMSE            | MBB         | $0.0002$    | $0.0002$             | $0.0787$    | $0.1917$    | $0.0199$             | $2.6436$    | $0.0583$            |
|                 | Penalized IS| $0.0001$    | $0.0002$             | $0.0620$    | $0.1733$    | $0.0168$             | $1.0604$    | $0.1240$            |

Table 3: The bias and RMSE of the simulated maximum likelihood estimates with respect to the true parameters for CWD direct model.

| Method       | $\beta$ | $\mu$ |
|--------------|---------|-------|
| Bias         | MBB     | 0.03  | 0.06  |
|              | Penalized IS | 0.02  | 0.02  |
| RMSE         | MBB     | 0.07  | 0.11  |
|              | Penalized IS | 0.03  | 0.06  |

0.03, $\mu_0 = 0.20$). The step size of the Euler-Maruyama scheme is $1/12$ of the time interval between each pair of observations, which is one month in this case. Estimates are obtained using $J = 48$ sample paths. The initial condition $X(0) = (S(0), I(0), C(0))^T$ is set to be the same as the real dataset. The exact transition density is not available for this case. The bias and RMSE of the simulated maximum likelihood estimates with respect to the true parameters are shown in Table 3, which indicate similar promising improvements over MBB.

Two key differences in this simulation that are of particular interest are that the states $S$ and $I$ are unobserved and the time between observations is long (yearly). The fact that penalized importance sampling does well in this context is promising for this and other applications in ecology. From Tables 1–3, we note that the performance of the penalized IS over MBB tends to be more significant as the dimension of the SDE model increases and observed data become more sparse over time.

6 Chronic wasting disease example

Deer populations and ecosystems can be severely disrupted by the contagious prion disease, chronic wasting disease (CWD) [Miller et al., 2006]. In order to reduce the potential damages caused by CWD, it is important to understand the transmission mechanisms of CWD. Several deterministic epidemic models were proposed by Miller et al. (2006) in order to portray the transmission of CWD. Here, based on one of those deterministic models, we firstly derive a CWD SDE model using the technique described in Allen (2003, Chapter 8). Then, we implement the proposed penalized simulated maximum likelihood method to the dataset studied in Miller et al. (2006), which consisted of annual observations of cumulative mortality from two distinct CWD epidemics in captive mule deer held at the Colorado Division of Wildlife Foothills Wildlife Research Facility in Fort Collins, Colorado. The first epidemic
occurred from 1974 to 1985. The second epidemic occurred in a new deer herd from 1992 to 2001. The dataset also includes the annual number of new deer added to the herd and the per capita losses due to natural deaths and removals.

6.1 CWD direct model

CWD may be transmitted to susceptible animals directly from infected animals (Miller et al., 2006). We portray this direct transmission using an SDE model. Let \( X(t) = (S(t), I(t), C(t))^T \), where \( S \) is the number of susceptible animals, \( I \) is the number of infected animals, \( C \) is the total number of accumulate deaths from CWD over time. Only \( C \) is observed at \( t_i \), for \( i = 1, \ldots, n \), and the other two state variables, \( S \) and \( I \), are unobserved. The unknown parameters to be estimated in the statistical model are denoted by \( \theta = (\beta, \mu) \), where \( \beta \) is the direct transmission coefficient, \( \mu \) is the per capita CWD mortality rate, and \( W = (W_1, W_2, W_3)^T \) is a 3-dimensional standard Wiener process. Then the direct transmission SDE model is given by

\[
\begin{align*}
    dS &= [a - S(\beta I + m)]dt + B_{11}dW_1 + B_{12}dW_2 + B_{13}dW_3, \\
    dI &= [\beta SI - I(\mu + m)]dt + B_{21}dW_1 + B_{22}dW_2 + B_{23}dW_3, \\
    dC &= \mu Idt + B_{31}dW_1 + B_{32}dW_2 + B_{33}dW_3,
\end{align*}
\]

with known initial condition \( X(0) = (S(0), I(0), C(0))^T \), where \( a \) is the known number of susceptible animals annually added to the population via births or importation, \( m \) is the known per capita natural mortality rate, and \( B = (B_{ij}) \) is the positive definite square root of the covariance matrix, \( B = \sqrt{\Sigma} \) with

\[
\Sigma = \begin{bmatrix}
    a + S(\beta I + m) & -\beta SI & 0 \\
    -\beta SI & \beta SI + I(\mu + m) & -\mu I \\
    0 & -\mu I & \mu I
\end{bmatrix}.
\]

Here, we briefly explain how the above SDE model is derived. See Allen (2003, Chapter 8) for more details. Let \( X_\delta = X(t + \delta) - X(t) \) be the increment during the time interval \( \delta \). If \( \delta \) is sufficiently small, we can assume at most one animal is infected or died during the time interval \( \delta \). The probability of an event that more than one infection or death has occurred during time \( \delta \) is of order \( \delta^2 \), which can be neglected. Then we can approximate the mean of \( X_\delta \) for \( \delta \) sufficiently small to order \( \delta \) by

\[
E[X_\delta] \approx f_\delta = \begin{pmatrix} a - S(\beta I + m) \\ \beta SI - I(\mu + m) \\ \mu I \end{pmatrix} \delta.
\]

Furthermore, we can also approximate the covariance of \( X_\delta \) for \( \delta \) sufficiently small by

\[
V[X_\delta] = E[(X_\delta)(X_\delta)^T] - E(X_\delta)E(X_\delta)^T \approx E[(X_\delta)(X_\delta)^T] = \Sigma. \tag{20}
\]

The matrix \( \Sigma \) in (18) is positive definite and hence has a positive definite square root \( B = \sqrt{\Sigma} \). It can be shown that (19) and (20) are estimates of order \( \delta \). We also assume \( X_\delta \) follows normal distribution with mean vector \( f_\delta \) and covariance matrix \( B^2\delta = \Sigma \). Thus,

\[
X(t + \delta) \approx X(t) + f_\delta + B\sqrt{\delta\eta}, \tag{21}
\]
where $\eta \sim N(0, I_{3 \times 3})$ and $I$ is the identity matrix. This is exactly one iteration of the Euler-Maruyama scheme for a system of SDE (17). As a result, the dynamical system (21) converges in the mean square sense to the system of SDEs (17) as $\delta \to 0$.

6.2 Results

We set the number of simulations (sample paths) $J$ as 48, and set the step size $\delta$ of the Euler-Maruyama scheme as $1/12$ of the time interval between each pair of observations. The simulated maximum likelihood estimates based on the penalized importance sampler are $\hat{\theta} = (\hat{\beta}, \hat{\mu}) = (0.03, 0.21)$ with 90% confidence intervals $[0.03, 0.10]$ and $[0.16, 0.34]$, respectively. To measure the goodness of fit, 100 simulated trajectories of cumulative number of deaths for CWD using $\theta$ and the real CWD data are overlaid in Figure 2. In general, the penalized IS successfully captures the pattern of the CWD death data over time, especially for such a small sample size. Miller et al. (2006) also proposed a more complex deterministic model, which we could also extend to a corresponding stochastic model, however the model quickly becomes over-parameterized due to the limited sample size and complexity of the model. Instead, we only consider this direct model as an application example.

The basic reproductive number $R_0$, which is the average number of secondary cases generated by one infected individual over the course of its infectious period when the entire population is susceptible, is important in biology and epidemiology (Anderson and May, 1992). Usually people consider the situation in which the majority of a closed population is susceptible, that is $S(0)/N \approx 1$. For deterministic models, if $R_0 > 1$ then the infection will spread in a population, and if $R_0 \leq 1$, the infection will die out monotonically. For stochastic models, the probability that there is no epidemic equals $1$ if $R_0 \leq 1$ and $\left( \frac{1}{R_0} \right)^{I(0)}$ if $R_0 > 1$ (Allen and Burgin, 2000), where $I(0)$ is the initial number of infected animals. The traditional interpretation of $R_0$ is not available here because the population is not closed; $a$ in (17) is the known number of susceptible annually added to the population. However, we want to point out that our method can be used to estimate $R_0$ for cases when the population is closed and the other assumptions of $R_0$ hold. For example, assuming a natural mortality rate of $m = 0.15$ (Miller et al., 2006), the corresponding estimate for the basic reproductive number $R_0$ equals $\frac{\hat{\beta} N_0}{\hat{\mu} + m} \approx 0.09 N_0$ with 90% confidence interval $[0.07 N_0, 0.32 N_0]$, where $N_0$ is the initial population or susceptible size. Hence, we would expect that CWD will spread if a few infected animals, like one or two, are introduced to a closed susceptible population with size at least $1/0.09 \approx 11$.

7 Conclusion and discussion

The dynamics of many ecological problems can be well described by a multivariate stochastic differential equation system. However, the transition densities of discrete-time observations are not known for most interesting models. Pedersen (1995) and Santa-Clara (1997) proposed the SMLE, which is computationally intensive to achieve a reasonable accuracy. Durham and Gallant (2002) suggested the MBB, which significantly improves SMLE. The efficient importance sampling approach introduced by Richard and Zhang (2007) can further improve
Figure 2: The 100 simulated trajectories of the cumulative number of deaths for CWD (dotted line) are obtained by using CWD direct model (17) with estimated parameters. The circled points are the observed CWD data.

the accuracy, however it requires another stage of optimization to obtain the best importance sampler which would be even more computationally intensive.

We propose the penalized importance sampling approach, which provides a balanced approach to achieve accurate parameter estimates with efficient computation times for these complex stochastic models. The key idea is the introduction of a penalty term to select a better importance sampler in order to reduce the number of simulated sample paths. We compare the new method to MBB for three different models in a simulation study, and also show an application for a real dataset. From those results, we conclude that the penalized importance sampling is a significant improvement over MBB while keeping the computational cost low.

Stramer and Yan (2007) concluded the optimal choice for the number of Monte Carlo simulations $J$ in (7) is of the order $O(M^2)$ for the MBB approach, where $M$ is the number of subintervals between two observations. One can choose a number smaller than this as a starting point for the proposed penalized importance sampling method in practice. More formal guidance is under investigation. Other theoretical work, such as the convergence of the penalized simulated MLE and the asymptotic distribution of the penalized simulated MLE, will also be considered as future work.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. EF-0914489 (Sun and Hoeting). The research work of Chihoon Lee is supported in part by the Army Research Office (W911NF-12-1-0494) and the National Security Agency
This research also utilized the CSU ISTeC Cray HPS System, which is supported by NSF Grant CSN-0923386. We are grateful to N. Thompson Hobbs for introducing us to this problem, and Michael W. Miller and the Colorado Division of Parks and Wildlife for sharing the data. We also would like to thank John Geweke and Garland Durham for stimulating conversations.

References

Aït-Sahalia, Y. (2002). Transition densities for interest rate and other nonlinear diffusions. *The Journal of Finance*, 54(4):1361–1395.

Allen, L. (2003). *An Introduction to Stochastic Processes with Applications to Biology*. Pearson/Prentice Hall Upper Saddle River (New Jersey).

Allen, L. and Burgin, A. (2000). Comparison of deterministic and stochastic SIS and SIR models in discrete time. *Mathematical Biosciences*, 163(1):1–33.

Anderson, R. and May, R. (1992). *Infectious Diseases of Humans: Dynamics and Control*. Wiley Online Library.

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*. Springer Verlag.

Becker, N. (1979). The uses of epidemic models. *Biometrics*, 35(1):295–305.

Donnet, S., Foulley, J.-L., and Samson, A. (2010). Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. *Biometrics*, 66(3):733–741.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233.

Durham, G. and Gallant, A. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–338.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics*, 19(2):177–191.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339.

Givens, G. and Hoeting, J. (2012). *Computational Statistics*. Wiley Series in Probability and Statistics, second edition.
Golightly, A. and Wilkinson, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788.

Golightly, A. and Wilkinson, D. J. (2006). Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing*, 16(4):323–338.

Jimenez, J., Biscay, R., and Ozaki, T. (2005). Inference methods for discretely observed continuous-time stochastic volatility models: a commented overview. *Asia-Pacific Financial Markets*, 12(2):109–141.

Kloeden, P. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag.

Miller, M., Hobbs, N., and Tavener, S. (2006). Dynamics of prion disease transmission in mule deer. *Ecological Applications*, 16(6):2208–2214.

Øksendal, B. (2010). *Stochastic Differential Equations: An Introduction with Applications*. Springer.

Pedersen, A. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 22(1):55–71.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. ISBN 3-900051-07-0.

Richard, J. and Zhang, W. (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385–1411.

Santa-Clara, P. (1997). Simulated likelihood estimation of diffusions with an application to the short term interest rate. *University of California at Los Angeles, Anderson Graduate School of Management*.

Sørensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3):337–354.

Stramer, O. and Yan, J. (2007). Asymptotics of an efficient Monte Carlo estimation for the transition density of diffusion processes. *Methodology and Computing in Applied Probability*, 9(4):483–496.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Varadhan, R. and Borchers, H. W. (2011). *dfoptim: Derivative-free Optimization*. R package version 2011.8-1.