A Multiresolution Method for Parameter Estimation of Diffusion Processes

S. C. KOU, Benjamin P. OLDING, Martin LYSY, and Jun S. LIU

Diffusion process models are widely used in science, engineering, and finance. Most diffusion processes are described by stochastic differential equations in continuous time. In practice, however, data are typically observed only at discrete time points. Except for a few very special cases, no analytic form exists for the likelihood of such discretely observed data. For this reason, parametric inference is often achieved by using discrete-time approximations, with accuracy controlled through the introduction of missing data. We present a new multiresolution Bayesian framework to address the inference difficulty. The methodology relies on the use of multiple approximations and extrapolation and is significantly faster and more accurate than known strategies based on Gibbs sampling. We apply the multiresolution approach to three data-driven inference problems, one of which features a multivariate diffusion model with an entirely unobserved component.

KEY WORDS: Autocorrelation; Data augmentation; Euler discretization; Extrapolation; Missing data; Stochastic differential equation.

1. INTRODUCTION

Diffusion processes are commonly used in many applications and disciplines. For example, they have served to model price fluctuations in financial markets (Heston 1993), particle movement in physics (McCann, Dykman, and Golding 1999), and the dynamics of biomolecules in cell biology and chemistry (Golightly and Wilkinson 2008). Most diffusion processes are specified in terms of stochastic differential equations (SDEs). The general form of a one-dimensional SDE is

$$\mathrm{d}Y_t = \mu\,(Y_t, \boldsymbol{\theta})\,\mathrm{d}t + \sigma\,(Y_t, \boldsymbol{\theta})\,\mathrm{d}B_t,$$

where t is continuous time; Y_t is the underlying stochastic process; $\mu(\cdot)$ is the drift function, a function of both Y_t and a set of parameters $\boldsymbol{\theta}$; $\sigma(\cdot)$ is the diffusion function; and B_t is Brownian motion.

While an SDE model is specified in *continuous* time, in most applications, data can only be observed at *discrete* time points. For example, measurements of physical phenomena are recorded at discrete intervals—in chemistry and biology, molecular dynamics are often inferred from the successive images of camera frames. The price information in many financial markets is recorded at intervals of days, weeks, or even months. Inferring the parameters $\boldsymbol{\theta}$ of an SDE model from discretely observed data is often challenging because it is almost never possible to analytically specify the likelihood of these data (the list of special cases of SDEs that do admit an analytic solution is surprisingly brief). Inferring the parameters of a discretely observed SDE model is the focus of this article.

One intuitive approach to the problem is to replace the continuous-time model with a discrete-time approximation. To have the desirable accuracy, one often has to use a highly dense discretization. Dense discretization, however, leads to two challenging issues: (1) accurate discrete-time approximations often require the discretization time length to be shorter than the time lag between real observations, creating a missing data problem; (2) highly dense discretization often imposes an unbearable computation burden. In this article, we propose a new multiresolution Monte Carlo inference framework, which operates on different resolution (discretization) levels simultaneously. In letting the different resolutions communicate with each other, the multiresolution framework allows us to significantly increase both computational efficiency and accuracy of estimation.

1.1 Background

With direct inference of SDE parameters typically being infeasible, researchers have experimented with a wide number of approximation schemes. The methods range from using analytic approximations (Aït-Sahalia 2002) to using approaches that rely heavily on simulation (see Sørensen 2004 for a survey of various techniques). An alternate strategy to approximating the likelihood directly is to first approximate the equation itself and subsequently find the likelihood of the approximated equation. Among possible discretizations of SDEs (see Pardoux and Talay 1985 for a review), the Euler–Maruyama approach (Maruyama 1955; Pedersen 1995) is perhaps the simplest. It replaces the SDE with a stochastic *difference* equation:

$$\Delta Y_t = \mu\,(Y_{t-1}, \boldsymbol{\theta})\,\Delta t + \sigma\,(Y_{t-1}, \boldsymbol{\theta})\,\sqrt{\Delta t}\,Z_t,$$

where $\Delta Y_t = Y_t - Y_{t-1}$, Δt is the time lag between observations Y_{t-1} and Y_t, and Z_t are iid normal $\mathcal{N}(0, 1)$ random variables. In most cases, one cannot choose the rate at which data are generated—observation rate is typically dictated by equipment limitations or by historical convention—and applying the discretization scheme directly to the observed data may yield very inaccurate estimates.

More accurate inference is made possible, however, by incorporating the idea of missing data into the approximation approach. In this framework, the Δt of the discretization scheme can be reduced below the rate at which data are actually gathered. The complete data Y_t of the specified model then becomes either

S. C. Kou is Professor of Statistics, Harvard University, Cambridge, MA, 02138 (E-mail: *kou@stat.harvard.edu*). Benjamin P. Olding is Co-Founder and CTO of Jana, Boston, MA (E-mail: *ben@jana.com*). Martin Lysy is Assistant Professor of Statistics, University of Waterloo, Waterloo, ON, Canada (E-mail: *mlysy@uwaterloo.ca*). Jun S. Liu is Professor of Statistics, Harvard University, Cambridge, MA, 02138 (E-mail: *jliu@stat.harvard.edu*). This research is supported in part by the NIH (National Institutes of Health)/NIGMS (National Institute of General Medical Sciences) grant R01GM090202 and the NSF (National Science Foundation) grants DMS-0449204, DMS-0706989, and DMS-1007762.

Journal of the American Statistical Association
December 2012, Vol. 107, No. 500, Theory and Methods
DOI: 10.1080/01621459.2012.720899

missing or observed. Simulation could be used to integrate out the missing data and compute maximum likelihood estimates of the parameters (Pedersen 1995). The difficulty of this simulated maximum likelihood estimation method lies in the difficulty of finding an efficient simulation method. See Durham and Gallant (2002) for an overview.

The same methodology—combining the Euler–Maruyama approximation with the concept of missing data—can also be used to estimate posterior distributions in the context of Bayesian inference. For example, one can use the Markov chain Monte Carlo (MCMC) strategy of a Gibbs sampler to conditionally draw samples of parameters and missing data and form posterior estimates from these samples (Jones 1998; Elerian, Chib, and Shephard 2001; Eraker 2001). While the approximation can be made more accurate by reducing the discretization step size Δt, this will generally cause the Gibbs sampler to converge at a very slow rate. Not only does the reduction in discretization step size lead to more missing data—requiring more simulation time per iteration—but also adjacent missing data values become much more correlated, leading to substantially slower convergence.

For more efficient computation, Elerian, Chib, and Shephard (2001) suggested conditionally drawing missing data using random block sizes. Along similar lines but from a general perspective, Liu and Sabatti (2000) adapted group Monte Carlo methodology to this problem: changing the block size and using group Monte Carlo to update the blocks. Another possible approach to drawing missing data is to attempt to update all values in a single step. Roberts and Stramer (2001) proposed first transforming the missing data so that the variance is fixed and constant; then a proposal for all transformed missing data between two observations is drawn from either a Brownian bridge or an Ornstein–Uhlenbeck (OU) process and accepted using the Metropolis algorithm. Chib, Pitt, and Shephard (2004) proposed a different transformation method, avoiding the use of variance-stabilizing transformations. Golightly and Wilkinson (2008) extended this approach, proposing a global Gibbs sampling scheme that can be applied to a large class of diffusions (where reducibility is no longer required). Stuart, Voss, and Wilberg (2004) also investigated conditional path sampling of SDEs but employed a stochastic PDE-based approach instead. Beskos et al. (2006) proposed a method that not only draws all the missing data at once, as these other researchers have suggested, but does so using the actual SDE, rather than an Euler–Maruyama discretization. This is accomplished using exact retrospective sampling of the actual diffusion paths. For further details on this inference approach, see Beskos and Roberts (2005) and Beskos, Papaspiliopoulos, and Roberts (2009).

1.2 The Multiresolution Approach

While there has been much investigation on how to update missing data values in a Euler–Maruyama approximation scheme, all such schemes rely on a single discretization level for approximating the true likelihood. This leads to a delicate balance: on one hand, low-resolution (large Δt) approximations require less computation effort, but the results are inaccurate; on the other hand, high-resolution (small Δt) approximations are more accurate, but they require very intense computation. We propose a multiresolution framework, which simultaneously considers a collection of discrete approximations to estimate the posterior distributions of the parameters, such that different levels of approximations are allowed to communicate with one another. There are three critical advantages to this approach over using only one approximation level. First, the convergence rate of the MCMC simulation can be substantially improved: coarser approximations help finer approximations converge more quickly. Second, a more accurate approximation to the diffusion model can be constructed using multiple discretization schemes: each level's estimates of the posterior distribution can be combined and improved through extrapolation. Third, the overall accuracy of the posterior estimates can be augmented incrementally. If a smaller value of Δt is later determined necessary, the computational burden is considerably lower relative to starting a brand new sampler at the new value of Δt. This last feature allows the multiresolution framework to be most useful in practice, as the appropriate value of Δt is typically unknown at the outset of analysis. Allowing its value to be decreased incrementally over the course of analysis can be of great practical service.

Taken in combination, these three features of the multiresolution method allow for more computationally efficient, more accurate, and more convenient inference of the parameters. The remainder of this article is organized as follows: Section 2 introduces the general notation used in this article. Section 3 introduces the multiresolution sampler, a cross-chain MCMC algorithm between Euler–Maruyama approximations at different resolution levels. Section 4 describes how samples from these levels can be combined through extrapolation to form more accurate estimates of the true posterior distribution. Practical implementations of the multiresolution approach—combining multiresolution sampling with extrapolation—are presented in Section 5. The performance of the proposed method is illustrated with three different SDE applications (one in biophysics and two in finance) where no analytic form of the likelihood is presently known. The article concludes with a discussion in Section 6.

2. NOTATION AND TWO ILLUSTRATIVE EXAMPLES

It is instructive to examine simple examples of diffusions to better understand the details of different inference strategies. One of the simplest SDEs is the OU process:

$$\mathrm{d}Y_t = \gamma\,(\mu - Y_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t.$$

It is fortunate that the exact solution to this equation is known, thus allowing us to directly examine the error introduced by approximate inference strategies.

Let $\boldsymbol{Y} = (Y_0, Y_1, \ldots, Y_n)$ denote the $n+1$ values of observed data, beginning with an initial value Y_0. For simplicity, it is assumed that the observations $\boldsymbol{Y}$ have been made at regular time intervals of ΔT. The exact likelihood of $\boldsymbol{Y}$ under the OU process is:

$$f_{\text{exact}}(\boldsymbol{Y} \mid \mu, \gamma, \sigma) = \prod_{t=1}^{n} \frac{1}{\sqrt{\pi g}\sigma} \exp\left\{ -\frac{1}{g\sigma^2}\big((\mu - Y_t) - \sqrt{1-\gamma g}\,(\mu - Y_{t-1})\big)^2 \right\},$$

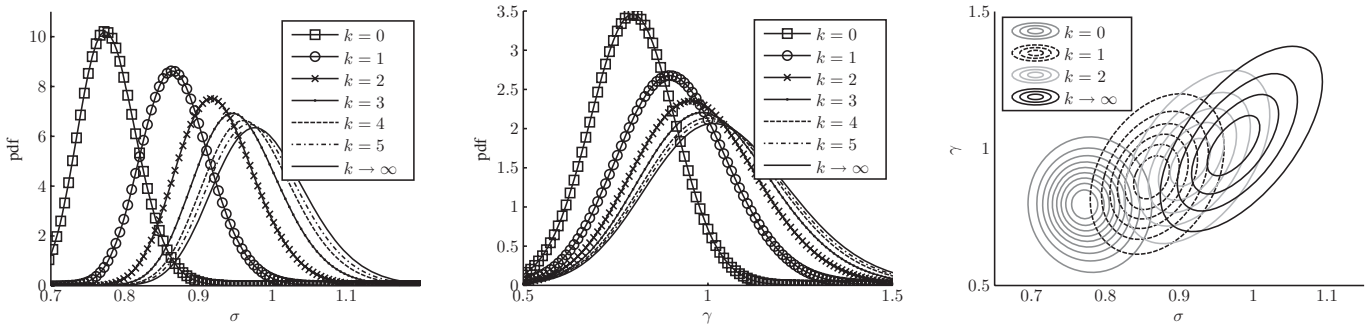


Figure 1. Euler–Maruyama approximation of the posterior of σ and γ in the OU process. Posteriors are based on 200 points of simulated data with $\Delta T = 0.5$, $\mu = 0$, and $\sigma = \gamma = 1$. The prior is $p(\mu, \gamma, \sigma) \propto \gamma/\sigma$. The third panel is a contour plot showing the joint parameter space. (a) $f_k(\sigma \mid \boldsymbol{Y})$, (b) $f_k(\gamma \mid \boldsymbol{Y})$, (c) $f_k(\sigma, \gamma \mid \boldsymbol{Y})$.

where $g = (1 - \exp(-2\gamma\Delta T))/\gamma$, and for simplicity, we ignore the initial distribution of Y_0 and treat it as a fixed value. To contrast this exact likelihood with Euler–Maruyama approximations to the likelihood, we introduce notation to describe the complete data—the union of the observations $\boldsymbol{Y}$ with the intermediate values of missing data. Let $\boldsymbol{Y}^{(k)}$ be the vector of complete data, where we put $2^k - 1$ regularly spaced values of missing data between two successive observations, such that the complete data interobservation time in $\boldsymbol{Y}^{(k)}$ is $\Delta t = \Delta T/2^k$. For example, $\boldsymbol{Y}^{(0)} = \boldsymbol{Y}$ and $\boldsymbol{Y}^{(1)} = (Y_0^{(1)}, Y_1^{(1)}, \ldots, Y_{2n-1}^{(1)}, Y_{2n}^{(1)})$. In this example with $k = 1$, the even indices correspond to observed values and the odd indices to missing values. Generally, the elements of the vector $\boldsymbol{Y}^{(k)}$ are labeled from 0 to $2^k n$, with every 2^kth element corresponding to an actual observation. The likelihood of the complete data under the Euler–Maruyama approximation is

$$f_k\big(\boldsymbol{Y}^{(k)} \mid \mu, \gamma, \sigma\big) = \prod_{j=1}^{2^k n} \frac{1}{\sqrt{2\pi\,\Delta t}\sigma} \exp\left\{ -\frac{1}{2\Delta t \sigma^2} \times \left(Y_j^{(k)} - Y_{j-1}^{(k)} - \gamma\,\Delta t\big(\mu - Y_{j-1}^{(k)}\big)\right)^2 \right\}.$$

Note that two different choices of k correspond to two different Euler–Maruyama approximations. The observed data will be the same, yet correspond to differently indexed elements. For instance, if $Y_j^{(k)}$ is an observed value of the process, then $Y_{2j}^{(k+1)}$ will be the identical value. For convenience, we use $\boldsymbol{Y}^{\{k\}}$ to denote all the *missing data* in the kth approximation scheme, $\boldsymbol{Y}^{\{k\}} = \boldsymbol{Y}^{(k)} \setminus \boldsymbol{Y}$.

The exact posterior distribution of the parameters in the OU process can be found by specifying a prior $p(\mu, \gamma, \sigma)$: $f_{\text{exact}}(\mu, \gamma, \sigma \mid \boldsymbol{Y}) \propto p(\mu, \gamma, \sigma) f_{\text{exact}}(\boldsymbol{Y} \mid \mu, \gamma, \sigma, Y_0)$. The Euler–Maruyama approximation is found by integrating out the missing data:

$$f_k(\mu, \gamma, \sigma \mid \boldsymbol{Y}) \propto \int_{\boldsymbol{Y}^{\{k\}}} p(\mu, \gamma, \sigma)\, f_k(\boldsymbol{Y}^{(k)} \mid \mu, \gamma, \sigma) \mathrm{d}\boldsymbol{Y}^{\{k\}}.$$

For the OU process, the posterior density $f_k(\mu, \gamma, \sigma \mid \boldsymbol{Y})$ can be calculated analytically. As $k \to \infty$, $f_k(\mu, \gamma, \sigma \mid \boldsymbol{Y})$ will approach the true posterior $f_{\text{exact}}(\mu, \gamma, \sigma \mid \boldsymbol{Y})$. This is illustrated in Figure 1, which plots the posteriors of $f_k(\sigma \mid \boldsymbol{Y})$ and $f_k(\gamma \mid \boldsymbol{Y})$ for several values of k, along with the respective true posteriors. These posteriors are based on 200 observations of a simulated OU process with $\Delta T = 0.5$, $\mu = 0$, $\gamma = 1$, and $\sigma = 1$. The noninformative (improper) prior $p(\mu, \gamma, \sigma) \propto \gamma/\sigma$ was used, following the example of Liu and Sabatti (2000).

As described in the introduction, the difficulty with this approximation scheme lies in the integration of the missing data. Unlike the OU process, most SDE applications require sampling of both the parameters and the missing data, and these are all strongly dependent on one another. Consider the common solution of using a Gibbs sampler to integrate out the missing data: the joint posterior of both parameters and missing data is sampled conditionally, one parameter or missing data value at a time. As k increases, not only does it take longer to iterate the sampler—as there is more missing data—but also each sequential draw is increasingly correlated. With all other values held constant, the conditional draws are almost deterministic: the sampler becomes nearly trapped. To illustrate this difficulty, a Gibbs sampler was run to generate samples from the posterior distributions of the parameters, using the same set of simulated data of the OU process as in Figure 1. The autocorrelations of sampled σ and γ are shown in Figure 2, both increasing substantially with k. This highlights the trade-off in using the Euler–Maruyama approximation approach. While it allows for numerical tractability, it can be very computationally expensive to achieve a high degree of accuracy relative to the true posterior specified by the original diffusion.

With its constant diffusion function, the OU process is a very special example of an SDE. A more complex SDE can help demonstrate some of the practical difficulties in working with these types of models. A good example of this is the Feller process—frequently referred to as the CIR model in the economics literature (Cox, Ingersoll, and Ross 1985)—where the diffusion function is not constant. The Feller process is

$$\mathrm{d}Y_t = \gamma\,(\mu - Y_t)\,\mathrm{d}t + \sigma\sqrt{Y_t}\mathrm{d}B_t. \tag{2.1}$$

The support of Y_t is 0 to ∞, and the parameters γ, μ, and σ are also constrained to be nonnegative. A closed-form solution to the joint posterior of parameters of the Feller process can be written using the special function $I_a(\cdot)$, the modified Bessel

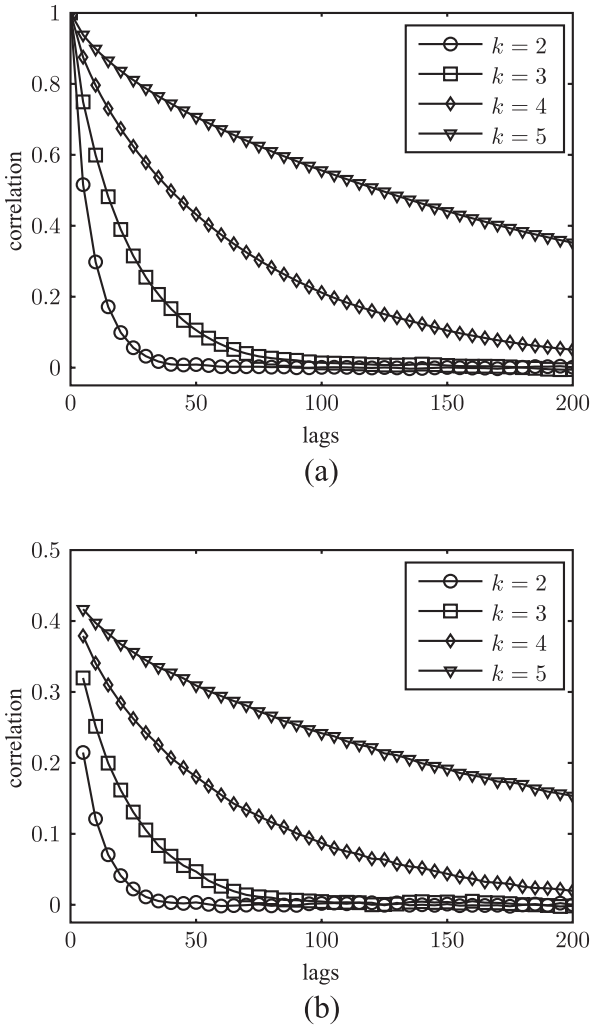


Figure 2. Autocorrelation of the posterior samples of σ and γ of the OU process from a Gibbs sampler output. Convergence slows as k increases. (a) Autocorrelation of σ, (b) autocorrelation of γ.

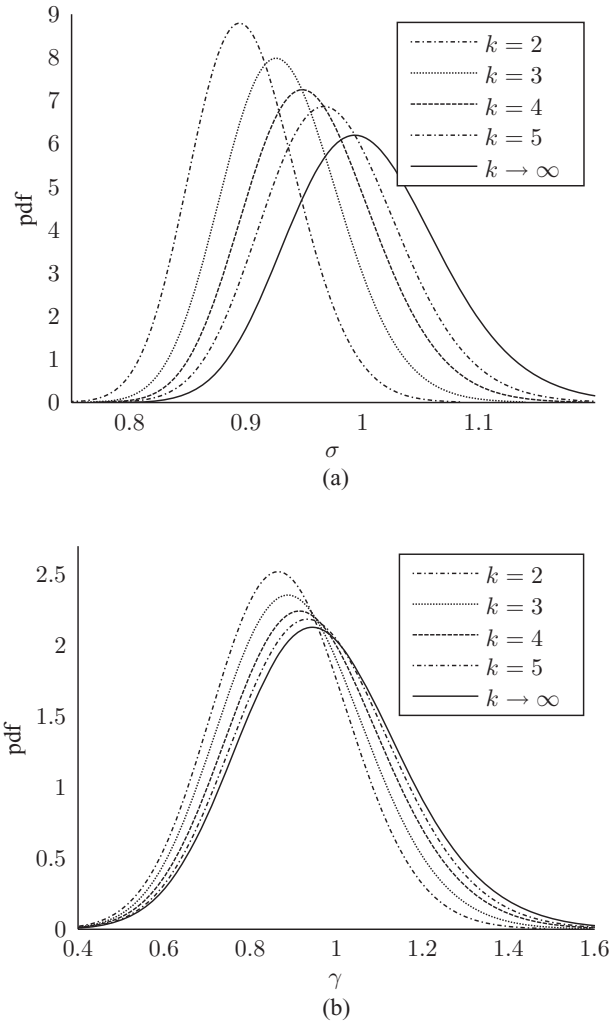


Figure 3. Posterior distributions of σ and γ in the Feller process based on Euler–Maruyama approximations. Posteriors are based on 200 points of simulated data with $\Delta T = 0.5$, and $\mu = \sigma = \gamma = 1$. $p(\mu, \gamma, \sigma) \propto \gamma/\sigma$. (a) $f_k(\sigma \mid Y^{(0)}, Y_0)$, (b) $f_k(\gamma \mid Y^{(0)}, Y_0)$.

function of order a (Kou and Kou 2004):

$$f_{\text{exact}}(\mu, \gamma, \sigma \mid \boldsymbol{Y}) \propto p(\mu, \gamma, \sigma) \left(\frac{2\gamma e^{\gamma \Delta T (\frac{\mu\gamma}{\sigma^2} - \frac{1}{2})}}{\sigma^2 (1 - e^{-\gamma \Delta T})} \right)^{n-1} \times \prod_{i=1}^{n-1} (Y_i / Y_{i-1})^{\frac{\mu\gamma}{\sigma^2} - \frac{1}{2}} \exp\left[- \frac{2\gamma (Y_i + e^{-\gamma \Delta T} Y_{i-1})}{\sigma^2 (1 - e^{-\gamma \Delta T})} \right] I_{\frac{2\mu\gamma}{\sigma^2} - 1} \left(\frac{4\gamma \sqrt{Y_{i-1} Y_i e^{-\gamma \Delta T}}}{\sigma^2 (1 - e^{-\gamma \Delta T})} \right). \tag{2.2}$$

This expression allows the error resulting from the Euler–Maruyama approximation to be examined directly. Figure 3 shows an example of different approximate posteriors using a simulated dataset from the Feller process. A total of 200 data points were drawn using $\Delta T = 0.5$, and μ, γ, and σ all equal to 1. We use the same prior $p(\mu, \gamma, \sigma) \propto \gamma/\sigma$ as before.

Here, the approximate Euler–Maruyama parameter posterior $f_k(\mu, \gamma, \sigma \mid \boldsymbol{Y})$ cannot be obtained analytically: a Gibbs sampler is used to integrate out the missing data instead. Using the prior above, the conditional distributions of each parameter γ, $\kappa = \gamma\mu$, and σ^2 are standard distributions: either a (truncated) normal or an inverse gamma. The conditional distribution of each value of missing data, however, is not a traditional one:

$$f_k\big(Y_j^{(k)} \big| \mu, \sigma, \gamma, Y_{j-1}^{(k)}, Y_{j+1}^{(k)}\big) \propto \big(Y_j^{(k)}\big)^{-1/2} \exp\left[- \frac{1}{2\sigma^2 \Delta t} \left(\frac{\big(Y_j^{(k)}\big)^2}{Y_{j-1}^{(k)}} - \left(1 - \gamma^2 \Delta t^2 + \frac{2\gamma\mu\Delta t}{Y_{j-1}^{(k)}} \right) Y_j^{(k)} + \frac{1}{Y_j^{(k)}} \big(Y_{j+1}^{(k)} - \gamma\mu\Delta t \big)^2 \right) \right]. \tag{2.3}$$

For most SDEs, the conditional distribution of missing data will not be a familiar one that can be easily sampled from. One possibility is to use a Metropolized-Gibbs step: first draw a new value of the missing data from a proposal distribution, then accept or reject the proposed draw according to the Metropolis–Hastings rule. Among many possible proposal distributions, a convenient one is

$$\pi_k\big(Y_j^{(k)} \mid \boldsymbol{\theta}, Y_{j-1}^{(k)}, Y_{j+1}^{(k)}\big) \sim \mathcal{N}\big(\big(Y_{j+1}^{(k)} + Y_{j-1}^{(k)}\big)/2, \sigma^2\big(Y_{j-1}^{(k)}, \boldsymbol{\theta}\big) \Delta t/2 \big).$$

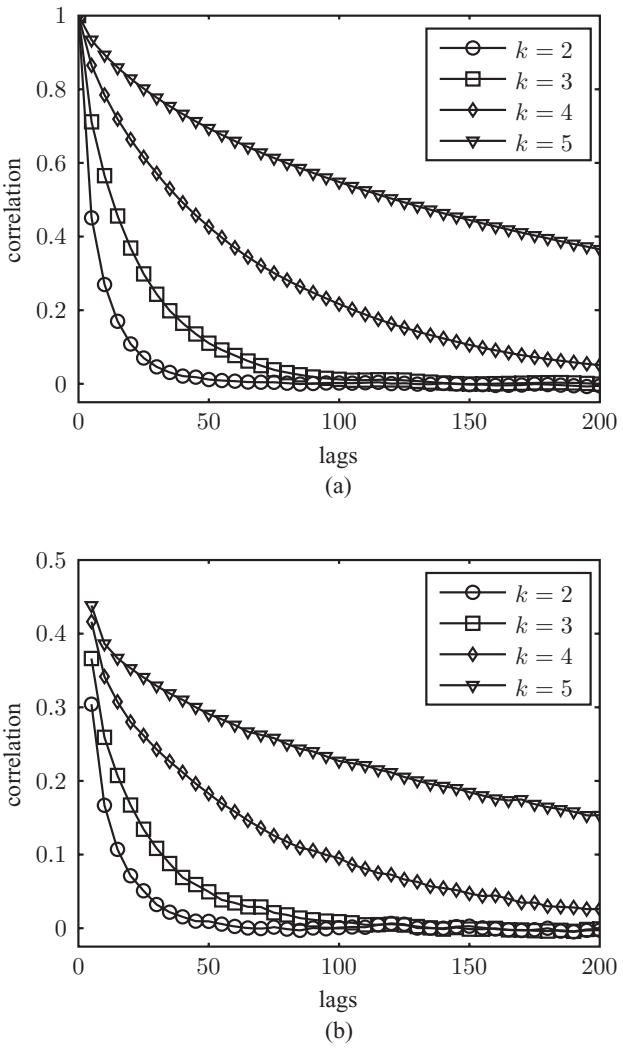


Figure 4. Autocorrelation of Feller process posterior samples σ and γ from the output of a Gibbs sampler. Convergence slows as k increases. (a) Autocorrelation of σ, (b) autocorrelation of γ.

This normal proposal has the advantage of being readily drawn from and asymptotically correct: as $\Delta t \to 0$, the acceptance rate approaches 1 (Eraker 2001). Note that when the support of the process is strictly positive, we can simply use a truncated normal distribution. Using this proposal, we applied the (Metropolized) Gibbs sampler to the Feller process. The results serve as a second illustration of the difficulty of using the Gibbs approach to integrate out the missing data as k becomes large. Figure 4 shows how the autocorrelations of σ and γ substantially increase with k.

The OU and Feller processes highlight the problems associated with applying a Gibbs sampler to computing posteriors under Euler–Maruyama approximations. While it may be theoretically possible to achieve arbitrary accuracy by selecting the appropriate value of k, it may not be practically feasible to wait for the Gibbs sampler to converge. Furthermore, the OU process and the Feller process are the rare cases where the difference between the approximate and true posteriors can be observed. In practice, the accuracy of a selected Euler–Maruyama approximation is unknown. One only knows that it converges to the correct distribution as $k \to \infty$.

3. MULTIRESOLUTION SAMPLING

3.1 The Sampler

Traditionally, the use of an Euler–Maruyama approximation requires a single resolution choice (corresponding to a single choice of Δt). The selection of a low resolution (large Δt) will result in a quickly converging sampling chain, which is, unfortunately, inaccurate. A high-resolution choice (small Δt) can result in a highly accurate estimate, yet will be slow—many samples will be required both for convergence and to build up an estimate of the posterior distribution.

In contrast, our proposed multiresolution sampler employs a *collection* of Euler–Maruyama discretization schemes at different resolutions. "Rough" approximations are used to locate the important regions of the parameter space, while "fine" approximations fill in and explore the local details. Low-resolution approximations quickly explore the global (parameter) space without getting stuck in one particular region; high-resolution approximations use the information obtained at the low-resolution explorations to yield accurate estimates in a relatively short time. By combining the strength of low and high resolutions (and mixing global and local explorations), this approach provides an inference method that is both fast and accurate. The key ingredient of the multiresolution sampler is to link different resolution approximations, using the empirical distribution of the samples collected at low resolutions to leap between states during high-resolution exploration.

In the multiresolution sampler, Euler–Maruyama approximations at m consecutive resolutions $k, k+1, \ldots, k+m-1$ are considered together. A sampling chain associated with each resolution is constructed. The multiresolution sampler starts from the lowest resolution chain k. This initial chain is sampled using any combination of local updates. For example, one may use the simple Gibbs algorithm to update the missing data $\boldsymbol{Y}^{\{k\}}$ and the parameters $\boldsymbol{\theta}$. Alternatively, one could combine the Gibbs algorithm with the block-update strategy of Elerian, Chib, and Shephard (2001) or the group-update algorithm of Liu and Sabatti (2000) to evolve $(\boldsymbol{Y}^{\{k\}}, \boldsymbol{\theta})$.

After an initial burn-in period, an empirical distribution of $(\boldsymbol{Y}^{\{k\}}, \boldsymbol{\theta})$ is constructed from the Monte Carlo samples. The multiresolution sampler then moves to the second lowest resolution chain, at level $k+1$. At each step of the multiresolution sampler, the state of $(\boldsymbol{Y}^{\{k+1\}}, \boldsymbol{\theta})$ is updated using one of two operations. With probability $1-p$, say 70%, the previous sample $(\boldsymbol{Y}_{\text{old}}^{\{k+1\}}, \boldsymbol{\theta}_{\text{old}})$ undergoes a local update step to yield the next sample. For example, in the case of Gibbs, this involves conditionally updating each element of $\boldsymbol{\theta}_{\text{old}}$ and each missing data value in $\boldsymbol{Y}_{\text{old}}^{\{k+1\}}$. With probability p, say 30%, a global, cross-resolution move is performed to leap $(\boldsymbol{Y}_{\text{old}}^{\{k+1\}}, \boldsymbol{\theta}_{\text{old}})$ to a new state.

The cross-resolution move is accomplished in three stages. First, a state $(\boldsymbol{Y}_{\text{trial}}^{\{k\}}, \boldsymbol{\theta}_{\text{trial}})$ is drawn uniformly from the empirical distribution formed by the earlier chain at resolution k. Second, $(\boldsymbol{Y}_{\text{trial}}^{\{k\}}, \boldsymbol{\theta}_{\text{trial}})$ is augmented to $(\boldsymbol{Y}_{\text{trial}}^{\{k+1\}}, \boldsymbol{\theta}_{\text{trial}})$ by generating the necessary additional missing data values (as missing data in the Euler approximations at levels k and $(k+1)$ have different dimensions). Third, $(\boldsymbol{Y}_{\text{trial}}^{\{k+1\}}, \boldsymbol{\theta}_{\text{trial}})$ is accepted to be the new sample with a Metropolis–Hastings type probability. As this cross-resolution step plays a pivotal role in the multiresolution

sampler's effectiveness, we shall describe it in full detail in Section 3.2.

After running the $(k+1)$-resolution chain for a burn-in period, an empirical distribution of $(\boldsymbol{Y}^{\{k+1\}}, \boldsymbol{\theta})$ is constructed from the posterior samples; this empirical distribution will in turn help the $(k+2)$-resolution chain to move. The multiresolution sampler on the $(k+2)$-resolution chain is then started and updated by the local move and the global cross-resolution move with probabilities $1-p$ and p. In the cross-resolution move, the old sample $(\boldsymbol{Y}_{\text{trial}}^{\{k+2\}}, \boldsymbol{\theta}_{\text{trial}})$ leaps to a new state with the help of the empirical distribution constructed by the $(k+1)$-resolution chain. In this way, the multiresolution sampler successively increases the resolution level until the Euler–Maruyama approximation with the finest resolution $k+m-1$ is reached. Each sampling chain (other than the one at the lowest resolution) is updated by two operations: the local move and the cross-resolution move. The basic structure of the multiresolution sampler is summarized in Algorithm 1.

Algorithm 1 The Multiresolution Sampler

1. Let $i=0$. Start from the k-resolution chain. Collect samples from $f_k(\boldsymbol{\theta}, \boldsymbol{Y}^{\{k\}} \mid \boldsymbol{Y})$ using any combination of local updating algorithms.
2. Discard some initial samples as burn-in, and retain the remaining samples as the empirical distribution of $(\boldsymbol{Y}^{\{k\}}, \boldsymbol{\theta})$ from $f_k(\boldsymbol{\theta}, \boldsymbol{Y}^{\{k\}} \mid \boldsymbol{Y})$.
3. Let $i \leftarrow i+1$. Start the $(k+i)$-resolution chain. Initialize the chain to a state $(\boldsymbol{Y}_{\text{old}}^{\{k+i\}}, \boldsymbol{\theta}_{\text{old}})$.
4. With probability $1-p$, perform a local update step to generate a new sample from $f_{k+i}(\boldsymbol{\theta}, \boldsymbol{Y}^{\{k+i\}} \mid \boldsymbol{Y})$, using any combination of local updates.
5. With probability p, perform a cross-resolution move:
   a. Randomly select a state $(\boldsymbol{Y}_{\text{trial}}^{\{k+i-1\}}, \boldsymbol{\theta}_{\text{trial}})$ from the empirical distribution of the $(k+i-1)$-chain.
   b. Augment $(\boldsymbol{Y}_{\text{trial}}^{\{k+i-1\}}, \boldsymbol{\theta}_{\text{trial}})$ to $(\boldsymbol{Y}_{\text{trial}}^{\{k+i\}}, \boldsymbol{\theta}_{\text{trial}})$ by generating additional missing data values.
   c. With a Metropolis–Hasting type probability r, accept $(\boldsymbol{Y}_{\text{trial}}^{\{k+i\}}, \boldsymbol{\theta}_{\text{trial}})$ as the next sample in the chain; with probability $1-r$, keep the previous values of $(\boldsymbol{Y}_{\text{old}}^{\{k+i\}}, \boldsymbol{\theta}_{\text{old}})$ as the next sample in the chain.
6. Rename the most recent draw as $(\boldsymbol{Y}_{\text{old}}^{\{k+i\}}, \boldsymbol{\theta}_{\text{old}})$, and repeat from Step 4 until a desired number of samples are achieved (typically determined in part by monitoring the chain for sufficient evidence of convergence).
7. Discard some initial samples of the chain as burn-in, and retain the remaining samples to form an empirical distribution of $(\boldsymbol{Y}^{\{k+i\}}, \boldsymbol{\theta})$ from $f_{k+i}(\boldsymbol{\theta}, \boldsymbol{Y}^{\{k+i\}} \mid \boldsymbol{Y})$. If a finer approximation to the SDE is desired, repeat from Step 3.

3.2 The Cross-Resolution Move

The cross-resolution move provides the means for successive resolution approximations to communicate with each other, allowing a rapidly mixing low-resolution approximation to speed up the convergence of a higher-resolution chain. There are two key insights behind the move. (1) As the amount of missing data increases, the posterior distributions of the parameters under different resolutions become closer; an example of this can be seen in Figures 1 and 3, which illustrate how the posterior distributions of $\boldsymbol{\theta}$ overlap to an increasing degree as k, the resolution level, increases. Notably, the high-resolution cases are where help is most needed because of the slow convergence of the local update. This suggests that in the sampling of a high-resolution chain (say $k=5$), generating proposals (independently) from a lower-resolution chain ($k=4$) will have a high chance of being accepted and will significantly speed up the high-resolution chain's convergence. (2) Although it is not feasible to directly draw from an Euler–Maruyama distribution, we can employ the empirical distribution to resolve this difficulty. With a sufficient number of samples, the empirical distribution built on them will be nearly identical to the analytic one. Furthermore, it is trivial to draw from an empirical distribution: simply select uniformly from the existing samples.

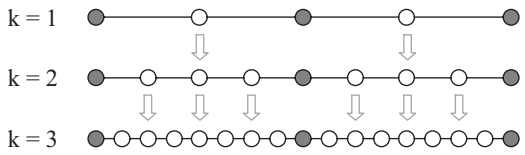


Figure 5. Graphic depicting three Euler–Maruyama approximations. Shaded circles represent observed data, while empty circles represent missing data. The arrows show how a draw from one approximation can be partially used as a proposal in the next.

Based on these two insights, the cross-resolution move is implemented in the multiresolution sampler by using the empirical distribution of a low-resolution chain to generate a new sample for the high-resolution chain. To carry this move out, it is important to note that different resolution levels do not share the same dimensionality. Thus, once a sample is drawn from the empirical distribution of a lower-resolution scheme, we must augment it with additional missing data values. A natural way of doing this is to divide the missing data at resolution $(k+1)$ into two groups, $\boldsymbol{Y}^{\{k+1\}} = \boldsymbol{Y}^{\{k\}} \cup \boldsymbol{Z}^{\{k+1\}}$, where $\boldsymbol{Z}^{\{k+1\}}$ are the additional missing data at resolution $(k+1)$. Figure 5 illustrates how such successive approximations line up relative to one another. Thus, the lower-resolution chain k generates the missing $\boldsymbol{Y}^{\{k\}}$, and we are free to propose the remaining $\boldsymbol{Z}^{\{k+1\}}$ from any distribution

$$T_{k+1}(\boldsymbol{Z}^{\{k+1\}} \mid \boldsymbol{\theta}, \boldsymbol{Y}^{\{k\}}, \boldsymbol{Y}).$$

Typically, the dimensionality of $\boldsymbol{Z}^{\{k+1\}}$ is high, but each of its components is *independent* of each other, conditioned on $\boldsymbol{\theta}$, $\boldsymbol{Y}^{\{k\}}$, and $\boldsymbol{Y}$, such that T_{k+1} boils down to independent draws from univariate distributions (or d-dimensional distributions for a d-dimensional SDE), which are much easier to construct.

Algorithm 2 summarizes the cross-resolution move from the kth approximation to the $(k+1)$th approximation. A reader familiar with the equi-energy sampler (Kou, Zhou, and Wong 2006) might note that the idea of letting different resolutions communicate with each other echoes the main operation of the equi-energy sampler, in which a sequence of distributions indexed by a temperature ladder is simultaneously studied: the flattened distributions help the rough ones to be sampled faster. Indeed, it was the equi-energy sampler's noted efficiency that motivated our idea of the cross-resolution move. We conclude this section by giving practical guidelines for how to choose the

Algorithm 2 Cross-Resolution Move of Multiresolution Sampler

1. Let $(\boldsymbol{\theta}_{\text{old}}, \boldsymbol{Y}_{\text{old}}^{\{k+1\}})$ be the current set of parameters and missing data. Draw $(\boldsymbol{\theta}_{\text{trial}}, \boldsymbol{Y}_{\text{trial}}^{\{k\}})$ from the empirical distribution of $f_k(\boldsymbol{\theta}, \boldsymbol{Y}^{\{k\}} \mid \boldsymbol{Y})$. Let $\pi_k^{\text{trial}} = f_k(\boldsymbol{\theta}_{\text{trial}}, \boldsymbol{Y}_{\text{trial}}^{\{k\}} \mid \boldsymbol{Y})$.
2. Draw $\boldsymbol{Z}_{\text{trial}}^{\{k+1\}}$ from a distribution $T_{k+1}(\boldsymbol{Z}^{\{k+1\}} \mid \boldsymbol{\theta}_{\text{trial}}, \boldsymbol{Y}_{\text{trial}}^{\{k\}}, \boldsymbol{Y})$. Let $\tau_{k+1}^{\text{trial}} = T_{k+1}(\boldsymbol{Z}_{\text{trial}}^{\{k+1\}} \mid \boldsymbol{\theta}_{\text{trial}}, \boldsymbol{Y}_{\text{trial}}^{\{k\}}, \boldsymbol{Y})$. Recall that $\boldsymbol{Y}_{\text{trial}}^{\{k+1\}} = \boldsymbol{Y}_{\text{trial}}^{\{k\}} \cup \boldsymbol{Z}_{\text{trial}}^{\{k+1\}}$. Let $\pi_{k+1}^{\text{trial}} = f_{k+1}(\boldsymbol{\theta}_{\text{trial}}, \boldsymbol{Y}_{\text{trial}}^{\{k+1\}} \mid \boldsymbol{Y})$.
3. Similarly, let $\pi_k^{\text{old}} = f_k(\boldsymbol{\theta}_{\text{old}}, \boldsymbol{Y}_{\text{old}}^{\{k\}} \mid \boldsymbol{Y})$, $\tau_{k+1}^{\text{old}} = T_{k+1}(\boldsymbol{Z}_{\text{old}}^{\{k+1\}} \mid \boldsymbol{\theta}_{\text{old}}, \boldsymbol{Y}_{\text{old}}^{\{k\}}, \boldsymbol{Y})$, and $\pi_{k+1}^{\text{old}} = f_{k+1}(\boldsymbol{\theta}_{\text{old}}, \boldsymbol{Y}_{\text{old}}^{\{k+1\}} \mid \boldsymbol{Y})$. Accept $(\boldsymbol{\theta}_{\text{trial}}, \boldsymbol{Y}_{\text{trial}}^{\{k+1\}})$ as the next sample from $f_{k+1}(\boldsymbol{\theta}, \boldsymbol{Y}^{\{k+1\}} \mid \boldsymbol{Y})$ with probability

$$r = \min\left\{1, \frac{\pi_{k+1}^{\text{trial}} / \left(\pi_k^{\text{trial}} \tau_{k+1}^{\text{trial}}\right)}{\pi_{k+1}^{\text{old}} / \left(\pi_k^{\text{old}} \tau_{k+1}^{\text{old}}\right)}\right\}.$$

Otherwise, with probability $1 - r$, keep $(\boldsymbol{\theta}_{\text{old}}, \boldsymbol{Y}_{\text{old}}^{\{k+1\}})$ as the next sample.

proposal distribution T_{k+1}, and how to determine the appropriate probability p of a cross-resolution move.

3.2.1 Choosing T_{k+1}. We are free to choose the distribution T_{k+1} to conditionally augment the additional missing data (Step 2 of Algorithm 2). A good choice, however, will make the acceptance rate of the independence move approach 1 as k increases. A simple proposal is

$$T_{k+1}\left(Y_j^{(k+1)} \middle| \boldsymbol{\theta}, Y_{j-1}^{(k+1)}, Y_{j+1}^{(k+1)}\right) \sim \mathcal{N}\left(\left(Y_{j+1}^{(k+1)} + Y_{j-1}^{(k+1)}\right)/2, \sigma^2\left(Y_{j-1}^{(k+1)}, \boldsymbol{\theta}\right)\Delta t/2\right)$$

independently for each $Y_j^{(k+1)} \in \boldsymbol{Z}^{\{k+1\}}$, where $\Delta t = \Delta T / 2^{k+1}$. This is the proposal used to update the missing data in the Gibbs sampler of Section 2. To see how the cross-resolution move improves the Monte Carlo convergence, let us turn to the OU example process introduced in Section 2. The autocorrelations of the OU process parameters σ and γ under the cross-resolution move are shown in Figure 6. These can be directly contrasted with the Gibbs sampler autocorrelations shown in Figure 2, as the identical dataset was used in both samplers. In addition to the evident improvement of the autocorrelation, we note that in the cross-resolution move—in contrast to the local update move—the autocorrelation *decreases* as k increases. This reflects the fact that the acceptance rate is increasing as the successive Euler–Maruyama approximations increasingly overlap with one another.

A good choice of T_{k+1} can make the multiresolution sampler very efficient. On the other hand, a poor choice of T_{k+1} can result in a low acceptance rate of the cross-resolution proposal. There does not appear to be, however, a foolproof recipe that guarantees a good distribution T_{k+1} for any arbitrary SDE. One useful technique that can make T_{k+1} easier to choose is to transform some aspect of the SDE to stabilize the variance (Roberts and Stramer 2001). For instance, if Y_t is a Feller process (2.1) and we let $Z_t = f(Y_t) = 2\sqrt{Y_t}$, then by Itō's formula $\mathrm{d}Z_t = (\frac{2\mu}{Z_t} - \frac{Z_t}{2} - \frac{\sigma^2}{2Z_t\gamma})\gamma \mathrm{d}t + \sigma \mathrm{d}B_t$. The distribution of missing data under Z_t, with its constant variance function, is much

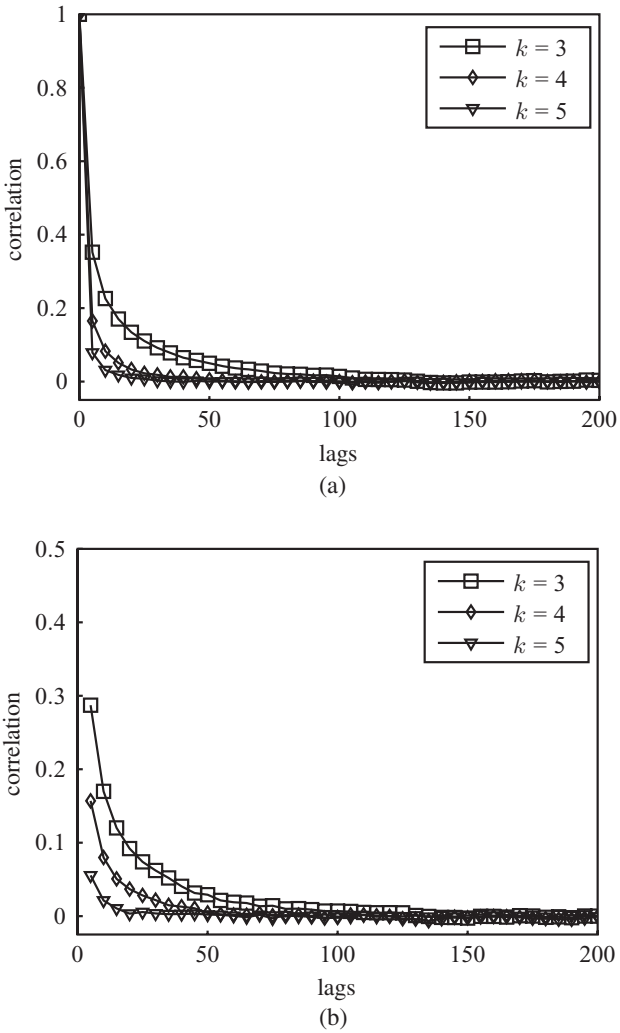


Figure 6. Autocorrelation of OU process parameters σ and γ from the output of a multiresolution sampler. Convergence improves as k increases. (a) Autocorrelation of σ, (b) autocorrelation of γ.

closer to a normal than the original Y_t. Figure 7 shows the autocorrelation of σ and γ from the output of the multiresolution sampler on Z_t. As k increases, the convergence rate of the multiresolution sampler improves. This stands in contrast to Figure 4 of the Gibbs sampler.

3.2.2 Choosing p. The probability p of making a cross-chain move in the multiresolution sampler (or the fraction of moves on a deterministic schedule) can be chosen as follows. Consider a local-update MCMC algorithm (e.g., the Gibbs sampler or the block update algorithm). For a given quantity of interest $\tau = h(\boldsymbol{\theta})$, we may approximate the effective sample size E_G of these local updates up to first order by

$$E_G \approx N \frac{1 - \eta}{1 + \eta},$$

where N is the number of MCMC iterations and η is the lag-1 autocorrelation of τ: $\eta = \text{cor}(\tau^{(t)}, \tau^{(t+1)})$ (see, for instance, Liu 2001, sec. 5.8). Now suppose that at each cycle of the local updates, a cross-resolution move targeting $p(\boldsymbol{\theta}, \boldsymbol{Y}^{\{k+1\}} \mid \boldsymbol{Y})$ with acceptance rate a is made with probability p. Then $\tau^{(t)}$ and $\tau^{(t+1)}$ are independent with probability ap and have correlation η with probability $1 - ap$, such that the lag-1 autocorrelation of τ using these cross-resolution moves decreases to $(1 - ap)\eta$.

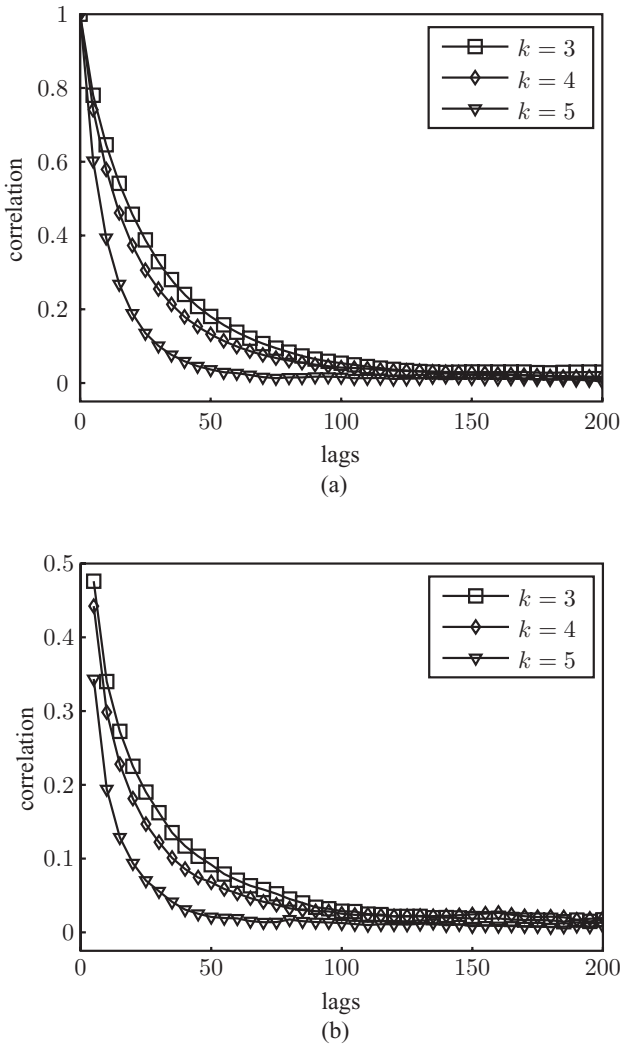


Figure 7. Autocorrelation of variance-stabilized Feller process parameters σ and γ from the output of a multiresolution sampler. Convergence improves as k increases. (a) Autocorrelation of σ, (b) autocorrelation of γ.

If E_M denotes the effective sample size of the multiresolution sampler combining local updates with cross-resolution proposals, the efficiency of this algorithm relative to the local updates alone can be measured as

$$\frac{E_M}{E_G} = \frac{(1-\eta+ap\eta)(1+\eta)}{(1+\eta-ap\eta)(1-\eta)}. \tag{3.1}$$

The value of p can then be adjusted if a and η are known, or estimated after an initial pilot run. For instance, if the basic Gibbs sampler has lag-1 autocorrelation $\eta = 0.75$ for a parameter of interest, it takes $ap = 0.25$ to double the effective sample size. For $\eta = 0.9$, we only need $ap = 0.1$, which helps quantify the great potential of multiresolution sampling when the autocorrelations of the local updates are high.

4. MULTIRESOLUTION INFERENCE

The multiresolution sampler uses the rapid convergence of low-resolution chains to in turn speed up the high-resolution chains. At the completion of sampling, the multiresolution sampler has samples from several approximation levels. For the subsequent statistical inference, a naive approach might be to simply focus on the highest resolution approximation—since it is the most accurate—and ignore the low-resolution samples, treating them as merely a computational by-product of the procedure. This approach, however, does not use all the samples effectively, wasting a great deal of both information and computation. In fact, the different approximations can be combined by extrapolation to significantly reduce the estimation error.

4.1 Multiresolution Extrapolation

Extrapolation is a technique often used in numerical analysis. It is a series acceleration method that combines successive approximations to reduce error. Richardson extrapolation (Richardson 1927) is a general statement of the approach, which can be applied whenever a function $F(h)$ converges to a value $F_0 = \lim_{h\to 0} F(h)$. Consider the expansion of such a limit:

$$F_0 = F(h) + a_m h^m + O(h^{m'}), \tag{4.1}$$

where $m' > m$ and $a_m \neq 0$. Taking the view that $F(h)$ is an approximation to the limit F_0, two successive approximations $F(h)$ and $F(\frac{h}{s})$ can be combined to form a more accurate estimate of F_0 by eliminating the $a_m h^m$ term in the expansion:

$$R(h) = \frac{s^m F(\frac{h}{s}) - F(h)}{s^m - 1} = F_0 + O(h^{m'}).$$

Compared with $F(h)$, the error in $R(h)$ is at least an order smaller. Additional extrapolation can be applied recursively to $R(h)$ to eliminate even higher-order terms in the expansion. The Romberg method of integration is an example of Richardson extrapolation applied to numerical integration (Romberg 1955). Richardson extrapolation has also been applied to simulating and numerically solving SDEs (Talay and Tubaro 1990; Kloeden, Platen, and Hofmann 1995; Durham and Gallant 2002).

In our Bayesian inference of diffusions, the multiresolution sampler gives us samples from several Euler–Maruyama approximations of the posterior distribution. Our goal is to combine them to have a more accurate estimate of the true posterior. To do so, we perform extrapolation. This multiresolution extrapolation allows us to reduce the discretization error by an order or more. For example, suppose a function $g(\boldsymbol{\theta})$ of the parameters is of scientific interest. An extrapolated point estimate can be obtained by first calculating the posterior mean or median of $g(\boldsymbol{\theta})$ based on the samples from each Euler–Maruyama approximation and then performing an extrapolation. Similarly, a $1-\alpha$ credible interval of $g(\boldsymbol{\theta})$ can be obtained by calculating its $\alpha/2$ and $1-\alpha/2$ quantiles from each Euler–Maruyama approximation and then performing an extrapolation on these quantiles. For most inference problems, point and interval estimation suffices. Occasionally, one might want to look at the marginal posterior density of a particular parameter θ_j. In this case, we can perform extrapolation on a kernel density estimate $\hat{f}(\theta_j)$ at each value of θ_j on a grid. By piecing together these extrapolated values, we obtain an extrapolated estimate for the marginal posterior density of θ_j.

A key ingredient of successful extrapolation is establishing the exponent m in Equation (4.1). We will show in the Appendix that the Euler–Maruyama approximation for the posterior distribution has the exponent $m = 1$ for the posterior mean, quantiles, and kernel density estimates.

As an example of the method, consider combining the $k = 2$ and $k = 3$ approximations of a given quantile α of θ_j. Let

us designate this extrapolated quantile estimate as $R_\alpha^{-1}(2 \cup 3)$. With $m = 1$, and with the $k = 3$ approximation having twice the discretization rate as the $k = 2$ approximation, we have the formula

$$R_\alpha^{-1}(2 \cup 3) = 2F_\alpha^{-1}(k = 3) - F_\alpha^{-1}(k = 2).$$

Combining $k = 3$ and $k = 4$ is similar:

$$R_\alpha^{-1}(3 \cup 4) = 2F_\alpha^{-1}(k = 4) - F_\alpha^{-1}(k = 3).$$

Combining $k = 2$, $k = 3$, and $k = 4$, however, is different. Rather than combine the quantiles directly, we (recursively) combine the extrapolated estimates $R_\alpha^{-1}(2 \cup 3)$ and $R_\alpha^{-1}(3 \cup 4)$ together:

$$R_\alpha^{-1}(2 \cup 3 \cup 4) = \frac{1}{3}\left(4R_\alpha^{-1}(3 \cup 4) - R_\alpha^{-1}(2 \cup 3)\right).$$

Note that here this combination is to eliminate the next higher-order term; thus, in this formula, $m = 2$.

4.2 Illustration of Multiresolution Extrapolation

To provide an illustrative example, extrapolated density approximations for the OU, Feller, and variance-stabilized Feller processes are displayed in Figure 8. Several observations immediately follow from this figure:

1. Combining two posterior estimates through extrapolation significantly reduces the error. Combining the approximations by using only 3 and 7 interpolated missing data points between observations ($k = 2$ and $k = 3$), for example, generally produces an estimate that is as accurate or even *more* accurate than the corresponding estimate based on a single approximation using 31 values of missing data ($k = 5$). This illustrates a major advantage of the multiresolution approach: using the combined strength of the multiresolution sampler and extrapolation, one does not always require a highly dense discretization rate for an accurate result; proper combination of low-resolution approximations can often lead to a better result than a single high-resolution approximation.
2. A comparison between the Feller and variance-stabilized Feller results again highlights the advantage of using a variance-stabilizing transformation wherever possible.
3. Combining three Euler–Maruyama approximation schemes (in this example, $k = 2$, 3, and 4) can be effective at reducing the overall error, as this eliminates both the first- and second-order errors. Thus, even in cases where the discretization error is largely in higher-order terms, the benefit derived from using extrapolation has the potential to be quite significant.

These observations suggest that whenever the computational challenge of sampling from a high-dimensional Euler–Maruyama approximation is substantial, it can be more efficient to sample from several lower-dimensional approximations and combine the resulting estimates with a final extrapolation step.

5. MULTIRESOLUTION METHOD IN PRACTICE

In this section, we shall apply the multiresolution approach to three realistic SDE models, one in biophysics and two in finance. Comparisons were made to chains that used only the simple Gibbs-type local updates. However, it is worth emphasizing that any strategy that increases the efficiency of the Gibbs sampler can be incorporated into the multiresolution sampler's local updates. This includes the block-update strategies of Elerian, Chib, and Shephard (2001) or the group moves of Liu and Sabatti (2000). The metric we use for comparison is the relative mean squared error (MSE) $\widehat{R}$, the ratio of the MSE of the Gibbs approach to the MSE of the multiresolution approach. Both MSEs are taken relative to the true posterior parameter distribution in each example. Since the true posterior in these nontrivial examples cannot be obtained analytically, we performed an exhaustive search. Higher and higher resolution chains were run to full convergence (many millions of iterations), until the last chain matched the extrapolated estimate of the two chains directly below it to within 0.1 standard deviations on 50 equally spaced quantiles of each parameter's marginal density. This last chain was then retained as a proxy for the ground truth.

5.1 Double-Well Potential Model for Optical Trap

The following general potential model is used to model a wide number of natural phenomena:

$$\mathrm{d}Y_t = -U'(Y_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t,$$

where $U(x)$ is a potential function and $U'(x)$ is the first derivative of $U(x)$ with respect to x. In a variety of circumstances, such as enzymatic reactions and electron transfer, the potential function is characterized as having a double well. In such cases, the following potential is often used as a model:

$$U(x) = \gamma(x^2 - \beta^2)^2 + \gamma c(x^3/3 - \beta^2 x).$$

The SDE model corresponding to data Y_t observed in this potential is thus

$$\mathrm{d}Y_t = -\left(4Y_t^3 + cY_t^2 - 4\beta^2 Y_t - c\beta^2\right)\gamma\,\mathrm{d}t + \sigma\,\mathrm{d}B_t.$$

Note that $U(x)$ has local minima at $\pm\beta$ and a local maximum at $-c/4$, provided $c < 4|\beta|$. Figure 9(a) plots the double-well potential $U(x)$.

We apply this model to an example from biophysics. In this case, Y_t describes the location of a particle when placed in an optical trap. McCann, Dykman, and Golding (1999) studied the behavior of a submicrometer-sized dielectric particle in a double-well optical trap. They acquired the location of the particle in time using a high-speed camera. While McCann, Dykman, and Golding have not made their data publicly available, they have published their estimates of the double-well potential itself, as well as some of the inferred particle positions over time. We fit the double-well potential model to these results and found values of $\beta = 0.1725$, $c = 0.0259$, $\gamma = 5000$, and $\sigma = 3$. Using these parameters, we simulated this process and sampled observations at a rate of $\Delta T = 1$ ms to record a total of 500 data points. An example of simulated observations from the process are plotted in Figure 9(b).

Using an exhaustive numerical search, we determined that resolution level $k = 5$ was indistinguishable from our proxy for the ground truth. We compare the ratio of the MSE of the Gibbs approach to that of the multiresolution method as follows. After a burn-in period of 10,000 iterations, we ran the Gibbs sampler for 1000 iterations at resolution $k = 5$, that is,

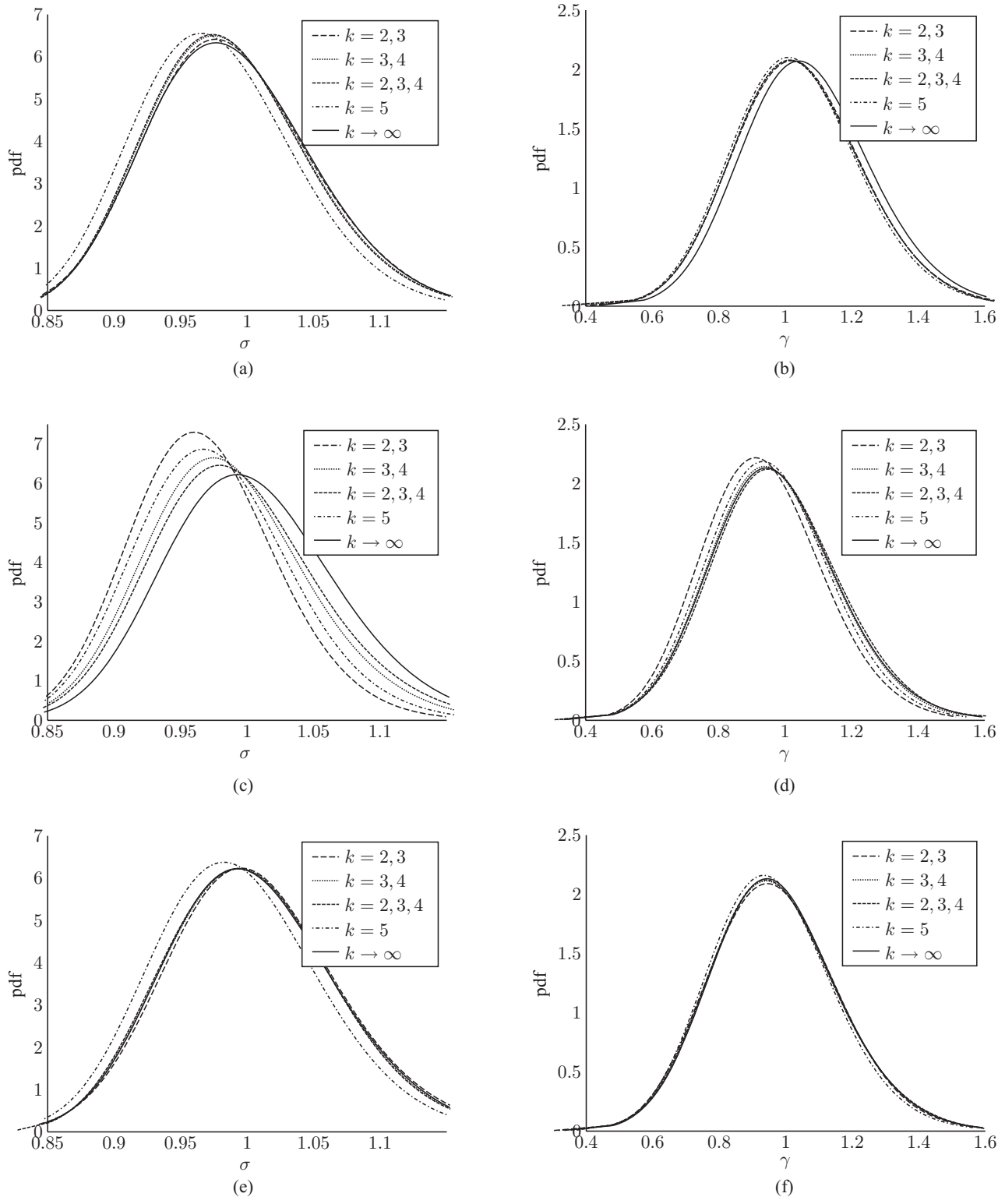
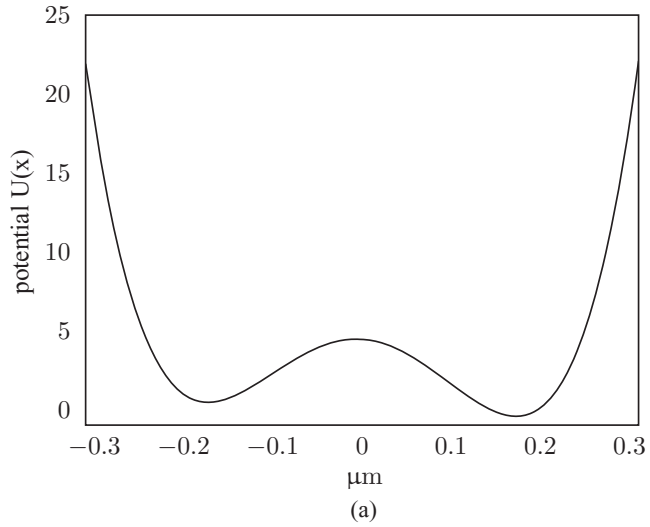


Figure 8. Posterior distribution estimates of σ and γ for different diffusions. Posterior estimates are created by combining two or more Euler–Maruyama estimates of the posterior quantiles and reconstructing the estimate of the distribution. (a) $f_k(\sigma \mid \boldsymbol{Y})$, OU process; (b) $f_k(\gamma \mid \boldsymbol{Y})$, OU process; (c) $f_k(\sigma \mid \boldsymbol{Y})$, Feller process; (d) $f_k(\gamma \mid \boldsymbol{Y})$, Feller process; (e) $f_k(\sigma \mid \boldsymbol{Y})$, variance-stabilized Feller; (f) $f_k(\gamma \mid \boldsymbol{Y})$, variance-stabilized Feller.

with 31 values of missing data between observations. A prior $p(\gamma, \beta^2, c, \sigma) \propto \gamma/\sigma \cdot 1\{c < 4|\beta|\}$ is used to obtain the parameter posteriors, where $1\{\cdot\}$ denotes the indicator function. With this prior the conditional parameter draws of γ, $\kappa = \gamma c$, β^2, and σ^2 are truncated normals or inverse gamma. We recorded the time it took to draw these 1000 samples, then gave the same time budget to the multiresolution sampler on levels $k = 3$ and $k = 4$, that is, with 7 and 15 values of missing data between observations. At level $k = 4$, the lag-1 parameter autocorrelations were around 0.85 and the cross-resolution proposals from $k = 3$



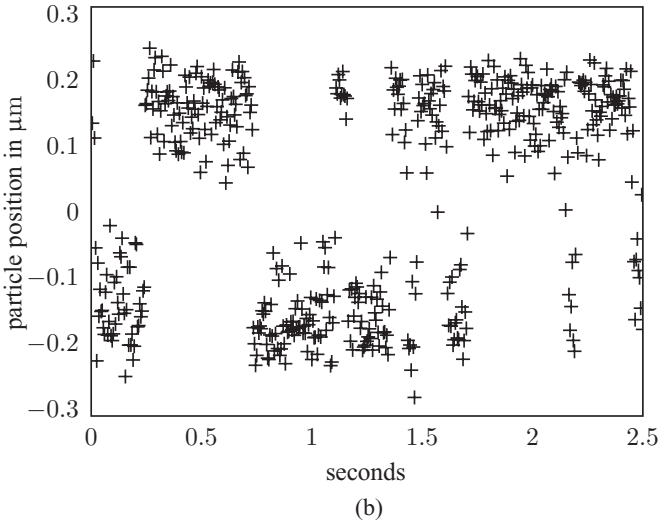


Figure 9. Example of simulated data from a double-well potential model. $\beta = 0.1725$, $c = 0.0259$, $\gamma = 5000$, and $\sigma = 3$. (a) Double-well potential, (b) simulated data.

to $k = 4$ had a 30% acceptance rate. We set the cross-resolution proposal rate to $p = 0.5$, such that the multiresolution sampler at $k = 4$ is expected to have twice the effective sample size of the Gibbs sampler at $k = 4$ according to the rule-of-thumb in (3.1).

Each sampler (Gibbs and multiresolution) was run many times starting from different initial values, to produce the ratio of MSE between the Gibbs and multiresolution estimates displayed in Table 1. Here, the multiresolution sampler is roughly two to three times as efficient as a single Gibbs sampler. This is roughly the value we expect, assuming that (3.1) holds and that the computation time for Gibbs samplers doubles with each k.

5.2 Generalized CIR Model for U.S. Treasury Bill Rate

Diffusions are often used as models for short-term interest rates in the field of mathematical finance. Chan et al. (1992) had suggested using the generalized Cox, Ingersoll, and Ross (gCIR) model:

$$dY_t = \gamma(\mu - Y_t)\,dt + \sigma Y_t^{\psi}\,dB_t,$$

Table 1. Ratios of MSE. Estimates of posterior quantiles from a Gibbs sampler versus those from the multiresolution method for the double-well potential model over the same amount of computer time

γ MSE ratio		c MSE ratio	
	$\widehat{R} \pm \widehat{\mathrm{sd}}(\widehat{R})$		$\widehat{R} \pm \widehat{\mathrm{sd}}(\widehat{R})$
$Q_{0.05}$	2.4 ± 0.65	$Q_{0.05}$	1.6 ± 0.34
$Q_{0.25}$	2.3 ± 0.64	$Q_{0.25}$	2.2 ± 0.45
$Q_{0.5}$	2.2 ± 0.59	$Q_{0.5}$	3.1 ± 0.66
$Q_{0.75}$	2.1 ± 0.54	$Q_{0.75}$	3.0 ± 0.59
$Q_{0.95}$	1.8 ± 0.45	$Q_{0.95}$	2.8 ± 0.59

where γ, μ, σ, ψ, and Y_t are all nonnegative. Both the OU and Feller processes are special cases of this generalized process: $\psi = 0$ is the OU process and $\psi = 1/2$ is the Feller process.

We apply the gCIR model to interest rate data consisting of 16 years of monthly records, from August 1982 to November 1998, of the 3-month U.S. Treasury Bill rate, as compiled by the Federal Reserve Board. This data, shown in Figure 10, is available for download at *http://research.stlouisfed.org/fred2/series/TB3MA/downloaddata?cid=116*. The data has been converted into a fraction by dividing by 100 (thus 0.1 is a rate of 10%). There are 196 observations in total.

The prior used in our investigations is $p(\gamma, \mu, \sigma, \psi) \propto \gamma/\sigma \cdot 1\{0 \le \psi \le 1\}$. This is the same prior on ψ used by Roberts and Stramer (2001). We used $\Delta T = 1/12$ to reflect that the data were recorded monthly. Our exhaustive numerical evaluation of the ground truth yielded posterior means of μ, γ, σ, and ψ equal to 0.0471, 0.1923, 0.0628, and 0.6851. respectively.

Following burn-in (10,000 iterations), we ran the Gibbs sampler for 10,000 iterations at the appropriate level $k = 5$ (as determined by the exhaustive numerical search). We ran the multiresolution sampler on $k = 2$ and $k = 3$ for the same amount of time allocated to the Gibbs sampler. In this case, the lag-1 autocorrelations for $k = 3$ were around 0.95 while the multiresolution acceptance rate was again around 30%. Setting the cross-resolution move probability to $p = 0.5$ was expected to increase efficiency by a factor of 4. The resulting posteriors of the two chains $k = 2$ and $k = 3$ were combined using multiresolution

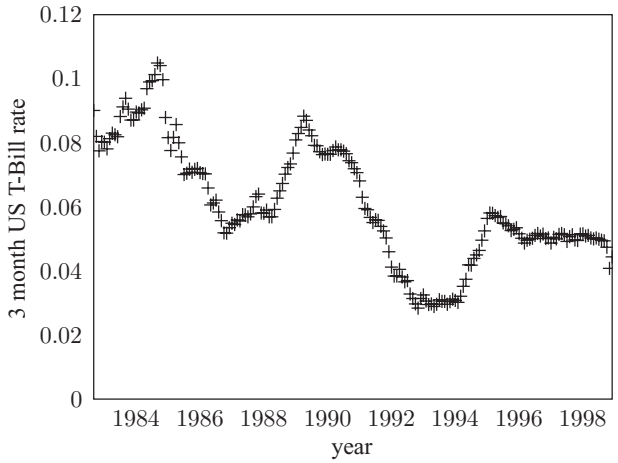


Figure 10. Sixteen years of monthly 3-month U.S. Treasury Bill rate data, as compiled by the Federal Reserve Board.

Table 2. Ratios of MSE. Estimates of posterior quantiles from a Gibbs sampler versus those from the multiresolution method for the gCIR process over the same amount of computer time

σ MSE ratio		ψ MSE ratio	
	$\widehat{R} \pm \widehat{\mathrm{sd}}(\widehat{R})$		$\widehat{R} \pm \widehat{\mathrm{sd}}(\widehat{R})$
$Q_{0.05}$	32 ± 6.2	$Q_{0.05}$	13 ± 2.0
$Q_{0.25}$	17 ± 3.8	$Q_{0.25}$	10 ± 1.7
$Q_{0.5}$	11 ± 2.0	$Q_{0.5}$	10 ± 1.5
$Q_{0.75}$	11 ± 1.8	$Q_{0.75}$	16 ± 2.7
$Q_{0.95}$	18 ± 2.4	$Q_{0.95}$	38 ± 6.0

extrapolation into final estimates of posterior quantiles. The simulation was independently repeated multiple times for both the Gibbs sampler and the multiresolution method.

Table 2 shows the ratio of the MSE of the Gibbs estimate to the MSE of the multiresolution approach, for a range of posterior quantiles of σ and ψ, two parameters of particular interest to researchers studying short-term interest rate. For this particular model and dataset, extrapolation allows us to skip two resolution levels $k = 4$ and $k = 5$, such that the multiresolution approach is seen to be 10 to 30 times more efficient than a standard Gibbs sampler.

5.3 Stochastic Volatility Model

So far, we have benchmarked the multiresolution approach against a single Gibbs sampler of an Euler–Maruyama approximation. The added cost of obtaining multiresolution samples is well offset by the increasing autocorrelation as the resolution k increases. It should be pointed out, however, that for univariate SDEs, there exists an alternative data augmentation scheme that does not use Euler–Maruyama discretization, or any direct discretization of the complete diffusion path Y_t itself. Instead, it is based on a factorization of Y_t with respect to a parameter-free Brownian measure, made possible by the Girsanov change-of-measure theorem. This approach was first considered by Roberts and Stramer (2001) and has been developed, for instance, in Beskos et al. (2006).

Borrowing from the terminology employed by these authors, we have implemented one such "exact-path" scheme on the double-well and gCIR models presented earlier. Although the conditional parameter draws are more difficult than with the Euler–Maruyama approximation, the autocorrelations of $\boldsymbol{\theta}$ were much lower, both discretization schemes having the same level of accuracy for a given resolution k. While it is possible to implement a multiresolution sampler on the exact-path scheme, the benefit of reducing small parameter autocorrelations even further is rather modest and generally does not make up for the cost of obtaining multiresolution samples in the first place.

An important step of the exact-path scheme above is to transform the given diffusion process Y_t to a different diffusion process $Z_t = \eta(Y_t, \boldsymbol{\theta})$ with unit diffusion

$$\mathrm{d}Z_t = \alpha(Z_t, \boldsymbol{\theta})\mathrm{d}t + \mathrm{d}B_t.$$

It is easy to show that $\eta(y, \boldsymbol{\theta}) = \int \sigma^{-1}(y, \boldsymbol{\theta})\mathrm{d}y$ satisfies this requirement in the univariate case. However, for multidimensional diffusion processes, such a transformation generally does not exist. A simple example is Heston's (1993)) stochastic volatility model for a financial asset S_t,

$$\begin{aligned} \mathrm{d}S_t &= \alpha S_t \mathrm{d}t + V_t^{1/2} S_t \, \mathrm{d}B_{St} \\ \mathrm{d}V_t &= -\gamma(V_t - \mu)\mathrm{d}t + \sigma V_t^{1/2} \, \mathrm{d}B_{Vt}, \end{aligned} \tag{5.1}$$

where the two Brownian motions B_{St} and B_{Vt} have correlation $\mathrm{cor}(B_{St}, B_{Vt}) = \rho$. In typical applications, only discrete observations $\boldsymbol{S} = (S_0, \ldots, S_n)$ of the financial asset are recorded. The "instantaneous variance" or *volatility* process V_t is completely unobserved.

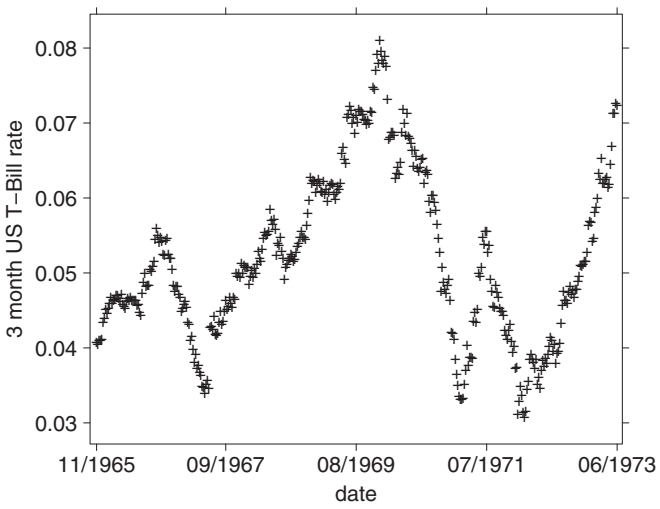


Figure 11. Weekly observations of 3-month U.S. Treasury Bill rates.

Implementation of the exact-path scheme for Heston's model is not as simple as in the univariate case, but can be achieved by using simultaneous time-scale transformations $t \mapsto \vartheta_V(t)$ and $t \mapsto \vartheta_S(t)$ (Kalogeropoulos, Roberts, and Dellaportas 2010). Even then, the transformations are only possible because the volatility V_t is itself marginally a diffusion process. While extending the exact-path approach to the more general multivariate setting appears to pose a considerable technical challenge, the Euler–Maruyama Gibbs-type scheme can easily be adapted to multiple dimensions. This simple scheme does, however, suffer from a heavy computational burden, which stands to be greatly alleviated by the multiresolution approach.

We have fit Heston's stochastic volatility model to 400 weekly 3-month U.S. Treasury Bill rates from November 5, 1965, to June 29, 1973, displayed in Figure 11. Inference was performed using Euler–Maruyama posterior approximations on the transformed process $X_t = \log(S_t)$ and $Z_t = 2V_t^{1/2}$. Since there are 252 trading days in a year, the financial convention for weekly data is to set $\Delta T = 5/252$. We used the prior $p(\alpha, \gamma, \mu, \sigma, \rho) \propto \gamma\sigma^2$: a variety of noninformative priors were found to give very similar results.

Posterior densities and autocorrelations for σ and ρ are displayed in Figure 12, for Gibbs samplers at resolution levels $k = 0$ to $k = 4$. Since the volatility process V_t is unobserved, the $n + 1 = 400$ volatility points $\boldsymbol{V} = (V_0, \ldots, V_n)$ corresponding to the observed data $\boldsymbol{S}$ must also be integrated out, which has a considerable impact on the mixing time of the Gibbs samplers. Even at the lowest level $k = 0$, the lag-1 autocorrelation of σ is 0.98, the highest of any autocorrelation encountered in

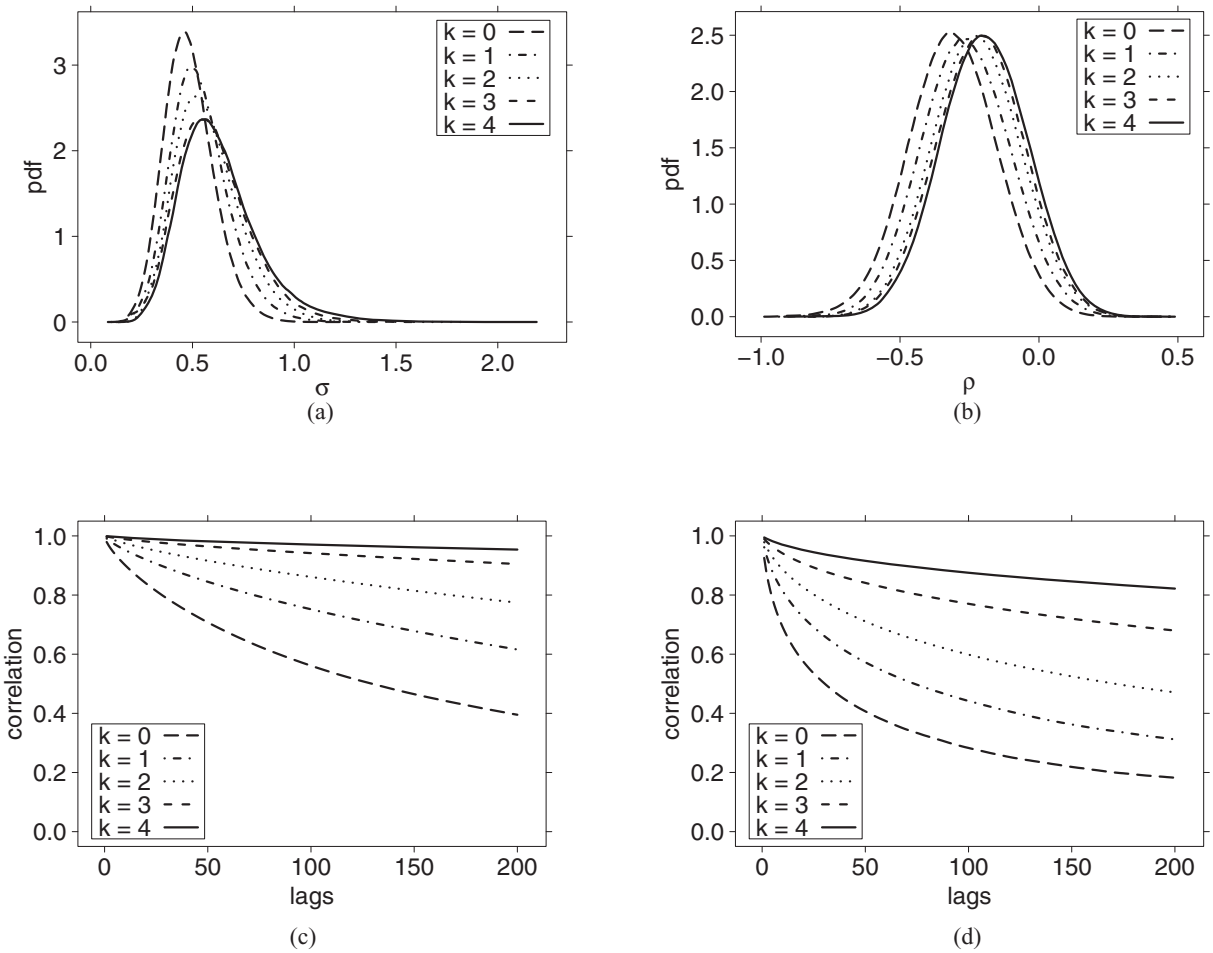


Figure 12. Densities and autocorrelations for Heston's model parameters σ and ρ. (a) $f_k(\sigma \mid \boldsymbol{Y})$, (b) $f_k(\rho \mid \boldsymbol{Y})$, (c) autocorrelation of σ, (d) autocorrelation of ρ.

the previous examples. At level $k = 4$, over 20 million Gibbs samples were required to give the posterior densities their full convergence shape.

In the following evaluation, we compare the multiresolution approach not only to a single $k = 4$ Gibbs sampler, but also to *parallel* Gibbs samplers running at $k = 2$ and $k = 3$. This accounts for the widespread availability of simultaneous computing resources, allowing researchers to run several Euler–Maruyama approximations at once and later combine them to produce estimates by extrapolation.

In total, three Gibbs samplers were run for 200,000 iterations each, at $k = 2$, 3, and 4. The first two Gibbs samplers $k = 2$ and $k = 3$ were combined to form extrapolated parameter estimates. To benefit from available technology, a parallelized version of the multiresolution sampler was implemented as follows. First, a Gibbs sampler is started at $k = 0$ and run for some burn-in period. Then, another Gibbs sampler is started at $k = 1$, and both samplers are run *simultaneously*; the cross-resolution proposals linking these samplers can now be drawn uniformly from an ever-increasing pool of samples. After another burn-in period, a third Gibbs sampler is started at $k = 2$ and run alongside the two others. It is linked to the $k = 1$ sampler by cross-resolution proposals, which continue to link $k = 1$ to $k = 0$. Finally, the $k = 3$ Gibbs sampler is added to the ensemble, with cross-resolution proposals connecting all four samplers. Multiresolution extrapolation is then performed using the last two levels $k = 2$ and $k = 3$.

A direct time comparison between the Gibbs samplers and the multiresolution sampler is difficult and perhaps uninformative in this setting. Instead, we assume that the computation time for Gibbs samplers at different resolutions scales as $O(2^k)$ for the same number N of posterior iterations. We also assume that the cost of computing one cross-resolution proposal and acceptance rate, when correctly implemented, is negligible compared with the cost of computing one full cycle of missing data *and* parameter updates in the Gibbs sampler. In our experience, this tends to be the case when the complete data themselves are the parameters' sufficient statistics. Thus, each step of the multiresolution sampler consists of both a local update cycle and a cross-resolution move. Now, suppose that the multiresolution sampler is given N iterations at $k = 0$, then spends N iterations running at $k = 0$ and $k = 1$ together; N iterations at $k = 0, 1, 2$; and N iterations at $k = 0, 1, 2, 3$. This is equivalent to $N(1 + 1/2 + 1/4 + 1/8) \approx 2N$ iterations of the Gibbs sampler at $k = 3$ and N iterations of the Gibbs sampler at $k = 4$.

Extrapolated quantiles using multiresolution samplers with $N = 100{,}000$ iterations are compared with the extrapolated quantiles of the Gibbs samplers at $k = 2$ and $k = 3$ in Table 3. Even though the cross-resolution acceptance rate is only around 15%, the MSE of the extrapolated Gibbs samplers is generally

Table 3. Ratio of MSEs for extrapolated Gibbs samplers ($k = 2, 3$) to multiresolution sampler

α MSE ratio		γ MSE ratio		β MSE ratio	
	$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$		$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$		$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$
$Q_{0.05}$	9.6 ± 2.3	$Q_{0.05}$	10 ± 2.8	$Q_{0.05}$	12 ± 3.4
$Q_{0.25}$	1.4 ± 0.39	$Q_{0.25}$	7.4 ± 1.8	$Q_{0.25}$	6.6 ± 1.6
$Q_{0.5}$	1.1 ± 0.34	$Q_{0.5}$	5.1 ± 1.1	$Q_{0.5}$	4.6 ± 1
$Q_{0.75}$	1.6 ± 0.5	$Q_{0.75}$	4.2 ± 0.85	$Q_{0.75}$	3.9 ± 0.85
$Q_{0.95}$	9.6 ± 2.9	$Q_{0.95}$	3.3 ± 0.4	$Q_{0.95}$	2.6 ± 0.4

σ MSE ratio		ρ MSE ratio	
	$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$		$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$
$Q_{0.05}$	11 ± 2.8	$Q_{0.05}$	18 ± 2.5
$Q_{0.25}$	5.6 ± 1.2	$Q_{0.25}$	14 ± 2.2
$Q_{0.5}$	3.9 ± 0.75	$Q_{0.5}$	10 ± 1.5
$Q_{0.75}$	3.3 ± 0.6	$Q_{0.75}$	8.3 ± 1.3
$Q_{0.95}$	2.2 ± 0.3	$Q_{0.95}$	7.9 ± 1.2

three to ten times higher than for the multiresolution sampler. Moreover, this assumes that the user running the Gibbs samplers either knows that extrapolation between $k = 2$ and $k = 3$ is sufficient or happens to run them in parallel at the first step of the analysis. With the multiresolution sampler, it is not as crucial to know or guess the "correct" resolution (or combination of resolution levels) in advance, as higher-resolution levels can be sampled incrementally at a substantially lower cost.

We next give $N = 200{,}000$ iterations to each step of the parallelized multiresolution sampler—N iterations for $k = \{0\}, \{0, 1\}, \{0, 1, 2\}, \{0, 1, 2, 3\}$—to compare with the 200,000 iterations of the single Gibbs sampler at $k = 4$. Both samplers require about the same amount of computation as discussed in the previous paragraph. Ratios of MSEs comparing the single Gibbs sampler to the multiresolution sampler with extrapolation are computed in Table 4. In this case, the multiresolution approach is 5 to 20 times more efficient than a single Gibbs sampler.

6. CONCLUSION

We have proposed a multiresolution Bayesian inference approach for estimating the parameter posterior of diffusion models. The method calls for samples to be drawn not just from one but multiple Euler–Maruyama approximations that communicate with each other. The fast but rough approximations help speed up the fine ones using cross-resolution moves. Moreover, combining the samples using multiresolution extrapolation can improve accuracy by an order or more, allowing the overall discretization level to be much lower than if a single chain had been used.

In our illustrations of the multiresolution sampler, we used the Gibbs-type move for local updating. In practice, any strategy that increases the sampling efficiency at a fixed resolution can be incorporated into the multiresolution sampler as well. This includes, for example, the block-update strategy of Elerian, Chib, and Shephard (2001) or the group-update strategy of Liu and Sabatti (2000). Our multiresolution approach thus complements

Table 4. Ratio of MSEs for single Gibbs sampler ($k = 4$) to multiresolution sampler

α MSE ratio		γ MSE ratio		β MSE ratio	
	$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$		$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$		$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$
$Q_{0.05}$	120 ± 26	$Q_{0.05}$	17 ± 4.4	$Q_{0.05}$	17 ± 4.6
$Q_{0.25}$	22 ± 5	$Q_{0.25}$	5.8 ± 1.4	$Q_{0.25}$	4.2 ± 1.2
$Q_{0.5}$	8.8 ± 2.2	$Q_{0.5}$	5.4 ± 1.2	$Q_{0.5}$	4.8 ± 1.2
$Q_{0.75}$	9.1 ± 2.3	$Q_{0.75}$	6 ± 1.2	$Q_{0.75}$	7.2 ± 1.6
$Q_{0.95}$	87 ± 19	$Q_{0.95}$	8.5 ± 1.6	$Q_{0.95}$	27 ± 4

σ MSE ratio		ρ MSE ratio	
	$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$		$\widehat{R} \pm \widehat{\text{sd}}(\widehat{R})$
$Q_{0.05}$	26 ± 7	$Q_{0.05}$	60 ± 10
$Q_{0.25}$	6.3 ± 1.3	$Q_{0.25}$	41 ± 8
$Q_{0.5}$	3.2 ± 0.65	$Q_{0.5}$	21 ± 5
$Q_{0.75}$	2.9 ± 0.55	$Q_{0.75}$	16 ± 3.8
$Q_{0.95}$	3.3 ± 0.6	$Q_{0.95}$	14 ± 2.8

these existing methods by allowing them to be accelerated by cross-resolution moves.

Another practical advantage of the multiresolution method is how the precision of its estimates can be improved incrementally. Rarely does one know ahead of time what the correct value of $(2^k - 1)$—the number of missing data values between observations—will actually be. The idea of running a computationaly intensive sampler at some level k only to find out that an even higher-level approximation must be started from scratch is certainly unappealing. In contrast, the additional computation time for each level of the multiresolution sampler is considerably smaller. Proceeding incrementally allows the appropriate level k to be naturally determined over the course of the analysis.

We have implemented the multiresolution approach in one- and two-dimensional settings. The same methodology can be applied to general multidimensional diffusions, and even to jump diffusions (e.g., Kou 2002), and infinite-activity processes such as the variance-gamma process (Madan, Carr, and Chang 1998) as well. It is likely that in these more complicated settings, a fully parallel version of the multiresolution sampler as in Section 5.3 will be most desirable. This version of the sampler is referred to as an *interacting* MCMC algorithm by Fort, Moulines, and Priouret (2011), evoking the one-way relation between the "target" chain at level $k+1$ and the "auxiliary" chain at level k. In their article, Fort, Moulines, and Priouret established several convergence results for a similar implementation of the equi-energy sampler, contingent on (1) the choice of local updates and (2) a regularity condition on the target and auxiliary densities, which can be routinely verified in practice. It would be very interesting to see whether such results can be established for the multiresolution sampler as well. More complicated inferential settings may also call for more creative cross-resolution missing data proposals T_{k+1}. With the absence of a variance-stabilizing transformation in multiple dimensions, the multiresolution sampler could potentially be combined with dynamic importance weighting (Wong and Liang 1997) to achieve higher cross-resolution acceptance rates. Further investigation of these ideas is currently under way.

APPENDIX: THE EXPANSION ORDER OF POSTERIOR ESTIMATES

In this section, we show that the posterior mean, quantiles, and kernel density estimates of parameters under the Euler–Maruyama discretization scheme have exponent $m = 1$ in the expansion (4.1). Let us restate the general form of the SDE as

$$\mathrm{d}Y_t = \mu(Y_t, t, \boldsymbol{\theta})\mathrm{d}t + \sigma(Y_t, t, \boldsymbol{\theta})\mathrm{d}B_t.$$

We assume the discrete observations $\boldsymbol{Y} = \boldsymbol{Y}^{(0)}$ occur at times $\boldsymbol{t} = \{t_0, \ldots, t_n\}$. For notational ease, we rewrite $\boldsymbol{Y}$ as $\boldsymbol{y} = (y_0, \ldots, y_n)$ and denote $Y(\boldsymbol{t}) = \{Y(t_1), \ldots, Y(t_n)\}$.

Without loss of generality, let us assume $t_0 = 0$. Then the Euler-Maruyama approximation $Y^{(k)}(t)$, with time discretization $\Delta t = \Delta T/2^k$, is given by

$$\begin{aligned} Y^{(k)}((j+1)\Delta t) = {} & Y^{(k)}(j\Delta t) + \mu(Y^{(k)}(j\Delta t), j\Delta t, \boldsymbol{\theta})\Delta t \\ & + \sigma\left(Y^{(k)}(j\Delta t), j\Delta t, \boldsymbol{\theta}\right)\left(B_{(j+1)\Delta t} - B_{j\Delta t}\right), \end{aligned}$$

where $j = 0, 1, 2, \ldots$. Using the notation established in Section 2, $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, f is the density function of $Y(\boldsymbol{t})$, and f_k is the density function of the Euler–Maruyama approximation $Y^{(k)}(\boldsymbol{t})$. We assume that $Y^{(k)}(0)$ and $Y(0)$ are drawn from the same distribution.

In examining weak convergence, we are interested in determining how the posterior expectation $E(g(\boldsymbol{\theta})|Y^{(k)}(\boldsymbol{t}) = \boldsymbol{y})$ under the Euler–Maruyama discretization approximates the true posterior expectation $E(g(\boldsymbol{\theta})|Y(\boldsymbol{t}) = \boldsymbol{y})$ as a function of k. In real applications, however, owing to measurement, equipment, and rounding errors, the realistic posterior expectation accessible to us is best stated as $E(g(\boldsymbol{\theta})|Y(\boldsymbol{t}) \in (\boldsymbol{y} - \varepsilon, \boldsymbol{y} + \varepsilon))$, where ε is a small number corresponding to the level of numerical precision. This posterior expectation

$$\begin{aligned} & E\left(g(\boldsymbol{\theta})\,|Y(\boldsymbol{t}) \in (\boldsymbol{y} - \varepsilon, \boldsymbol{y} + \varepsilon)\right) \\ & \quad = \frac{\int \mathrm{d}\boldsymbol{\theta} \int_{y_0-\varepsilon}^{y_0+\varepsilon} \cdots \int_{y_n-\varepsilon}^{y_n+\varepsilon} g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) f(Y(\boldsymbol{t})|\boldsymbol{\theta}) \mathrm{d}Y(\boldsymbol{t})}{\int \mathrm{d}\boldsymbol{\theta} \int_{y_0-\varepsilon}^{y_0+\varepsilon} \cdots \int_{y_n-\varepsilon}^{y_n+\varepsilon} p(\boldsymbol{\theta}) f(Y(\boldsymbol{t})|\boldsymbol{\theta}) \mathrm{d}Y(\boldsymbol{t})} \end{aligned}$$

involves many step functions $1_{[y_i-\varepsilon, y_i+\varepsilon]}(z)$, which are not mathematically convenient. Thus, we replace the step function by a smooth (four times continuously differentiable) kernel density function w and focus instead on how $E_{\varepsilon,w}(g(\boldsymbol{\theta})|Y^{(k)}(\boldsymbol{t}) \simeq \boldsymbol{y})$, our shorthand notation for

$$\begin{aligned} & E_{\varepsilon,w}(g(\boldsymbol{\theta})|Y^{(k)}(\boldsymbol{t}) \simeq \boldsymbol{y}) \\ & \quad = \frac{\int \mathrm{d}\boldsymbol{\theta} \cdots \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) f_k(Y^{(k)}(\boldsymbol{t})|\boldsymbol{\theta}) \prod_i \left[\frac{1}{\varepsilon} w\left(\frac{Y^{(k)}(t_i) - y_i}{\varepsilon}\right)\right] \mathrm{d}Y^{(k)}(\boldsymbol{t})}{\int \mathrm{d}\boldsymbol{\theta} \cdots \int p(\boldsymbol{\theta}) f_k(Y^{(k)}(\boldsymbol{t})|\boldsymbol{\theta}) \prod_i \left[\frac{1}{\varepsilon} w\left(\frac{Y^{(k)}(t_i) - y_i}{\varepsilon}\right)\right] \mathrm{d}Y^{(k)}(\boldsymbol{t})}, \end{aligned}$$

approximates

$$\begin{aligned} & E_{\varepsilon,w}(g(\boldsymbol{\theta})|Y(\boldsymbol{t}) \simeq \boldsymbol{y}) \\ & \quad = \frac{\int \mathrm{d}\boldsymbol{\theta} \cdots \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) f(Y(\boldsymbol{t})|\boldsymbol{\theta}) \prod_i \left[\frac{1}{\varepsilon} w\left(\frac{Y(t_i) - y_i}{\varepsilon}\right)\right] \mathrm{d}Y(\boldsymbol{t})}{\int \mathrm{d}\boldsymbol{\theta} \cdots \int p(\boldsymbol{\theta}) f(Y(\boldsymbol{t})|\boldsymbol{\theta}) \prod_i \left[\frac{1}{\varepsilon} w\left(\frac{Y(t_i) - y_i}{\varepsilon}\right)\right] \mathrm{d}Y(\boldsymbol{t})}. \end{aligned}$$

Theorem 1. Suppose the following three conditions hold for an SDE:

1. $\mu(x, t, \boldsymbol{\theta})$ and $\sigma^2(x, t, \boldsymbol{\theta})$ have linear growth, that is, $\mu^2(x, t, \boldsymbol{\theta}) + \sigma^2(x, t, \boldsymbol{\theta}) \le K(\boldsymbol{\theta})(1 + x^2)$ for every $\boldsymbol{\theta}$;
2. $\mu(x, t, \boldsymbol{\theta})$ and $\sigma^2(x, t, \boldsymbol{\theta})$ are twice continuously differentiable with bounded derivatives for every $\boldsymbol{\theta}$; that is, $|\frac{\partial}{\partial t}\mu(x, t, \boldsymbol{\theta})|$, $|\frac{\partial}{\partial x}\mu(x, t, \boldsymbol{\theta})|$, $|\frac{\partial^2}{\partial x^2}\mu(x, t, \boldsymbol{\theta})|$, $|\frac{\partial}{\partial t}\sigma^2(x, t, \boldsymbol{\theta})|$, $|\frac{\partial}{\partial x}\sigma^2(x, t, \boldsymbol{\theta})|$, and $|\frac{\partial^2}{\partial x^2}\sigma^2(x, t, \boldsymbol{\theta})|$ are all bounded by $N(\boldsymbol{\theta})$;
3. $\sigma^2(x, t, \boldsymbol{\theta})$ is bounded from below for every $\boldsymbol{\theta}$, that is, $\sigma^2(x, t, \boldsymbol{\theta}) \ge \lambda(\boldsymbol{\theta}) > 0$.

Then, for any integrable function g,

$$E_{\varepsilon,w}(g(\boldsymbol{\theta})|Y^{(k)}(\boldsymbol{t}) \simeq \boldsymbol{y}) - E_{\varepsilon,w}(g(\boldsymbol{\theta})|Y(\boldsymbol{t}) \simeq \boldsymbol{y}) = \frac{C_g}{2^k} + o(2^{-k}),$$

where C_g is a constant that does not depend on k.

Proof. We note

$$\begin{aligned} & E_{\varepsilon,w}(g(\boldsymbol{\theta})|Y^{(k)}(\boldsymbol{t}) \simeq \boldsymbol{y}) \\ & \quad = \frac{\int p(\boldsymbol{\theta}) g(\boldsymbol{\theta}) E\left\{\prod_{i=1}^n \left[\frac{1}{\varepsilon} w\left(\frac{Y^{(k)}(t_i) - y_i}{\varepsilon}\right)\right] \middle| \boldsymbol{\theta}\right\} \mathrm{d}\boldsymbol{\theta}}{\int p(\boldsymbol{\theta}) E\left\{\prod_{i=1}^n \left[\frac{1}{\varepsilon} w\left(\frac{Y^{(k)}(t_i) - y_i}{\varepsilon}\right)\right] \middle| \boldsymbol{\theta}\right\} \mathrm{d}\boldsymbol{\theta}}. \end{aligned} \tag{A.1}$$

Denote $v_j(Y^{(k)}(t_j), \boldsymbol{\theta}) = E\{\prod_{i=j+1}^n [\frac{1}{\varepsilon} w(\frac{Y^{(k)}(t_i) - y_i}{\varepsilon})] | Y^{(k)}(t_j), \boldsymbol{\theta}\}$, and $u_j(Y(t_j), \boldsymbol{\theta}) = E\{\prod_{i=j+1}^n [\frac{1}{\varepsilon} w(\frac{Y(t_i) - y_i}{\varepsilon})] | Y(t_j), \boldsymbol{\theta}\}$. Then, we have the recursion

$$\begin{aligned} & v_l(Y^{(k)}(t_l), \boldsymbol{\theta}) \\ & \quad = E\left\{\frac{1}{\varepsilon} w\left(\frac{Y^{(k)}(t_{l+1}) - y_{l+1}}{\varepsilon}\right) v_{l+1}(Y^{(k)}(t_{l+1}), \boldsymbol{\theta}) \middle| Y^{(k)}(t_l), \boldsymbol{\theta}\right\} \\ & u_l(Y(t_l), \boldsymbol{\theta}) = E\left\{\frac{1}{\varepsilon} w\left(\frac{Y(t_{l+1}) - y_{l+1}}{\varepsilon}\right) u_{l+1}(Y(t_{l+1}), \boldsymbol{\theta}) \middle| Y(t_l), \boldsymbol{\theta}\right\} \end{aligned}$$

from the Markov property. By theorem 14.1.5 of Kloeden and Platen (1992), for any smooth (fourth continuously differentiable) function q,

$$E\{q(Y^{(k)}(t_{l+1}),\theta)|Y^{(k)}(t_l)=y,\boldsymbol{\theta}\}-E\{q(Y(t_{l+1}),\boldsymbol{\theta})|Y(t_l)=y,\boldsymbol{\theta}\}=\frac{A_q}{2^k}+o(2^{-k}),$$

where the constant A_q does not depend on k, y, or l. It follows that if we assume $v_{l+1}(y,\boldsymbol{\theta})-u_{l+1}(y,\boldsymbol{\theta})=B_{l+1}/2^k+o(2^{-k})$, then

$$\begin{aligned}
&v_l(y,\boldsymbol{\theta})\\
&=E\left\{\frac{1}{\varepsilon}w\left(\frac{Y^{(k)}(t_{l+1})-y_{l+1}}{\varepsilon}\right)v_{l+1}(Y^{(k)}(t_{l+1}),\boldsymbol{\theta})\middle|Y^{(k)}(t_l)=y,\boldsymbol{\theta}\right\}\\
&=E\left\{\frac{1}{\varepsilon}w\left(\frac{Y^{(k)}(t_{l+1})-y_{l+1}}{\varepsilon}\right)\left[u_{l+1}(Y^{(k)}(t_{l+1}),\boldsymbol{\theta})\right.\right.\\
&\quad\left.\left.+\frac{B_{l+1}}{2^k}+o(2^{-k})\right]\middle|Y^{(k)}(t_l)=y,\boldsymbol{\theta}\right\}\\
&=E\left\{\frac{1}{\varepsilon}w\left(\frac{Y^{(k)}(t_{l+1})-y_{l+1}}{\varepsilon}\right)u_{l+1}(Y^{(k)}(t_{l+1}),\boldsymbol{\theta})\middle|Y^{(k)}(t_l)=y,\boldsymbol{\theta}\right\}\\
&\quad+\frac{B_{l+1}}{2^k}+o(2^{-k})\\
&=E\left\{\frac{1}{\varepsilon}w\left(\frac{Y(t_{l+1})-y_{l+1}}{\varepsilon}\right)u_{l+1}(Y(t_{l+1}),\boldsymbol{\theta})\middle|Y(t_l)=y,\boldsymbol{\theta}\right\}\\
&\quad+\frac{A}{2^k}+\frac{B_{l+1}}{2^k}+o(2^{-k})\\
&=u_l(y,\boldsymbol{\theta})+\frac{A+B_{l+1}}{2^k}+o(2^{-k}).
\end{aligned}$$

Therefore, using backward induction, we obtain that $v_0(x,\boldsymbol{\theta})-u_0(x,\boldsymbol{\theta})=B_0/2^k+o(2^{-k})$, which, combined with the assumption that $Y^{(k)}(t_0)$ and $Y(t_0)$ have the same distribution, implies that

$$E\left\{\prod_{i=1}^n\left[\frac{1}{\varepsilon}w\left(\frac{Y^{(k)}(t_i)-y_i}{\varepsilon}\right)\right]\middle|\boldsymbol{\theta}\right\}-E\left\{\prod_{i=1}^n\left[\frac{1}{\varepsilon}w\left(\frac{Y(t_i)-y_i}{\varepsilon}\right)\right]\middle|\boldsymbol{\theta}\right\}=\frac{C(\boldsymbol{\theta})}{2^k}+o(2^{-k}),$$

for some constant $C(\boldsymbol{\theta})$ depending on $\boldsymbol{\theta}$. Taking this result back to (A.1), we obtain

$$\begin{aligned}
&E_{\varepsilon,w}(g(\boldsymbol{\theta})|Y^{(k)}(\boldsymbol{t})\simeq\boldsymbol{y})\\
&=\left\{\int p(\boldsymbol{\theta})g(\boldsymbol{\theta})E\left\{\prod_{i=1}^n\left[\frac{1}{\varepsilon}w\left(\frac{Y(t_i)-y_i}{\varepsilon}\right)\right]\middle|\boldsymbol{\theta}\right\}\mathrm{d}\boldsymbol{\theta}\right.\\
&\quad\left.+\frac{\Delta T}{2^k}\int p(\boldsymbol{\theta})g(\boldsymbol{\theta})C(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}+o(2^{-k})\right\}\Bigg/\left\{\int p(\boldsymbol{\theta})\right.\\
&\quad\left.\times E\left\{\prod_{i=1}^n\left[\frac{1}{\varepsilon}w\left(\frac{Y(t_i)-y_i}{\varepsilon}\right)\right]\middle|\boldsymbol{\theta}\right\}\mathrm{d}\boldsymbol{\theta}+\frac{\Delta T}{2^k}\int p(\boldsymbol{\theta})C(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}+o(2^{-k})\right\}\\
&=\frac{\int p(\boldsymbol{\theta})g(\boldsymbol{\theta})E\left\{\prod_{i=1}^n\left[\frac{1}{\varepsilon}w\left(\frac{Y(t_i)-y_i}{\varepsilon}\right)\right]\middle|\boldsymbol{\theta}\right\}\mathrm{d}\boldsymbol{\theta}}{\int p(\boldsymbol{\theta})E\left\{\prod_{i=1}^n\left[\frac{1}{\varepsilon}w\left(\frac{Y(t_i)-y_i}{\varepsilon}\right)\right]\middle|\boldsymbol{\theta}\right\}\mathrm{d}\boldsymbol{\theta}}+\frac{C_g}{2^k}+o(2^{-k})\\
&=E_{\varepsilon,w}(g(\boldsymbol{\theta})|Y(\boldsymbol{t})\simeq\boldsymbol{y})+\frac{C_g}{2^k}+o(2^{-k}).
\end{aligned}$$

□

We make explicit use of this theorem by noting the following corollary on the posterior cdf and quantiles:

Corollary 1. The posterior cdf $F^j_{\varepsilon,w}$ of the jth parameter θ_j satisfies

$$F^j_{\varepsilon,w}(z|Y^{(k)}(\boldsymbol{t})\simeq\boldsymbol{y}):=E_{\varepsilon,w}(1(\theta_j\le z)|Y^{(k)}(\boldsymbol{t})\simeq\boldsymbol{y})=F^j_{\varepsilon,w}(z|Y(\boldsymbol{t})\simeq\boldsymbol{y})+\frac{C}{2^k}+o(2^{-k}).$$

If the posterior cdf $F^j_{\varepsilon,w}$ has nonzero derivative, then the quantile $F^{-1}_{\varepsilon,w,j}$ of θ_j satisfies

$$F^{-1}_{\varepsilon,w,j}(\alpha|Y^{(k)}(\boldsymbol{t})\simeq\boldsymbol{y})=F^{-1}_{\varepsilon,w,j}(\alpha|Y(\boldsymbol{t})\simeq\boldsymbol{y})+\frac{C'}{2^k}+o(2^{-k}),$$

for fixed $0<\alpha<1$.

Proof. Taking $g(\boldsymbol{\theta})$ of Theorem 1 to be the indicator function $1(\theta_j\le z)$ immediately yields the first equation. The assumption that $F^j_{\varepsilon,w}$ has nonzero derivative enables us to invert it to obtain the second equation. □

We can make the connection between Corollary 1 and Equation (4.1) explicit by noting $h=\Delta T/2^k$. Therefore, to apply extrapolation to the quantiles of a parameter posterior, we should use the exponent $m=1$.

Similarly, suppose that we wish to estimate the density $f(\theta_j)$ of parameter j at a specific value $\theta_j=x$. Suppose that a kernel density estimate $\hat{f}(x)$ is of the form

$$\hat{f}(x)=\frac{1}{Nh}\sum_{i=1}^N K\left(\frac{x-\theta_j^{(i)}}{h}\right),$$

where K is a (symmetric) kernel, h is a bandwidth parameter, and $\theta_j^{(1)},\ldots,\theta_j^{(N)}$ is a collection of N samples from $f(\theta_j)$. In this case, for fixed h, $\hat{f}(x)$ can be seen as a sample estimate of

$$E\left\{\frac{1}{h}K\left(\frac{x-\theta_j}{h}\right)\middle|\boldsymbol{Y}(\boldsymbol{t})\simeq\boldsymbol{y}\right\},$$

such that $g(\boldsymbol{\theta})=K((x-\theta_j)/h)/h$. As long as the kernel K is integrable, Theorem 1 also applies. Moreover, if the kernel density estimate $\hat{f}(x)$ at each resolution level is normalized, then so is the density estimate obtained by extrapolation.

[Received June 2011. Revised July 2012.]

REFERENCES

Aït-Sahalia, Y. (2002), "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach," *Econometrica*, 70, 223–262. [1558]

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2009), "Monte Carlo Maximum Likelihood Estimation for Discretely Observed Diffusion Processes," *The Annals of Statistics*, 37, 223–245. [1559]

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006), "Exact and Computationally Efficient Likelihood-Based Estimation for Discretely Observed Diffusion Processes" (with discussion), *Journal of the Royal Statistical Society,* Series B, 68, 333–382. [1559,1569]

Beskos, A., and Roberts, G. O. (2005), "Exact Simulation of Diffusions," *The Annals of Applied Probability*, 15, 2422–2444. [1559]

Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992), "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate," *Journal of Finance*, 47, 1209–1227. [1568]

Chib, S., Pitt, M. K., and Shephard, N. (2004), "Likelihood Based Inference for Diffusion Driven Models," Technical Report [online], Nuffield College, University of Oxford. Available at *http://www.nuff.ox.ac.uk/Economics/papers/2004/w20/chibpittshephard.pdf* [1559]

Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1985), "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–408. [1560]

Durham, G. B., and Gallant, A. R. (2002), "Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes," *Journal of Business and Economic Statistics*, 20, 297–338. [1559,1565]

Elerian, O., Chib, S., and Shephard, N. (2001), "Likelihood Inference for Discretely Observed Nonlinear Diffusions," *Econometrica*, 69, 959–993. [1559,1562,1566,1571]

Eraker, B. (2001), "MCMC Analysis of Diffusion Models With Application to Finance," *Journal of Business and Economic Statistics*, 19, 177–191. [1559,1562]

Fort, G., Moulines, E., and Priouret, P. (2011), "Convergence of Adaptive and Interacting Markov Chain Monte Carlo Algorithms," *The Annals of Statistics*, 39, 3262–3289. [1572]

Golightly, A., and Wilkinson, D. J. (2008), "Bayesian Inference for Nonlinear Multivariate Diffusion Models With Error," *Computational Statistics and Data Analysis*, 52, 1674–1693. [1558,1559]

Heston, S. L. (1993), "A Closed-Form Solution for Options With Stochastic Volatility With Applications to Bond and Currency Options," *Review of Financial Studies*, 6, 327–343. [1558,1569]

Jones, C. S. (1998), "A Simple Bayesian Method for the Analysis of Diffusion Processes," *SSRN eLibrary* [online]. Available at *http://ssrn.com/paper=111488*. [1559]

Kalogeropoulos, K., Roberts, G. O., and Dellaporta, P. (2010), "Inference for Stochastic Volatility Models Using Time Change Transformations," *The Annals of Statistics*, 38, 784–807. [1569]

Kloeden, P. E., and Platen, E. (1992), *Numerical Solution of Stochastic Differential Equations*, Berlin: Springer-Verlag. [1573]

Kloeden, P. E., Platen, E., and Hofmann, N. (1995), "Extrapolation Methods for the Weak Approximation of Itō Diffusions," *SIAM Journal on Numerical Analysis*, 32, 1519–1534. [1565]

Kou, S. G. (2002), "A Jump-Diffusion Model for Option Pricing," *Management Science*, 48, 1086–1101. [1572]

Kou, S. C., and Kou, S. G. (2004), "A Diffusion Model for Growth Stocks," *Mathematics of Operations Research*, 29, 191–212. [1561]

Kou, S. C., Zhou, Q., and Wong, W. H. (2006), "Equi-Energy Sampler With Applications in Statistical Inference and Statistical Mechanics" (with discussion), *The Annals of Statistics*, 34, 1581–1652. [1563]

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer-Verlag. [1564]

Liu, J. S., and Sabatti, C. (2000), "Generalised Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation," *Biometrika*, 87, 353–369. [1559,1560,1562,1566,1571]

Madan, D. B., Carr, P. P., and Chang, E. C. (1998), "The Variance Gamma Process and Option Pricing," *European Finance Review*, 2, 79–105. [1572]

Maruyama, G. (1955), "Continuous Markov Processes and Stochastic Equations," *Rendiconti del Circolo Matematico di Palermo*, 4, 48–90. [1558]

McCann, L. I., Dykman, M., and Golding, B. (1999), "Thermally Activated Transitions in a Bistable Three-Dimensional Optical Trap," *Nature*, 402, 785–787. [1558,1566]

Pardoux, E., and Talay, D. (1985), "Discretization and Simulation of Stochastic Differential Equations," *Acta Applicandae Mathematicae*, 3, 23–47. [1558]

Pedersen, A. R. (1995), "A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations," *Scandinavian Journal of Statistics*, 22, 55–71. [1558]

Richardson, L. F. (1927), "The Deferred Approach to the Limit. Part I. Single Lattice," *Philosophical Transactions of the Royal Society of London,* Series A, 226, 299–361. [1565]

Roberts, G. O., and Stramer, O. (2001), "On Inference for Partially Observed Nonlinear Diffusion Models Using the Metropolis-Hastings Algorithm," *Biometrika*, 88, 603–621. [1559,1564,1568,1569]

Romberg, W. (1955), "Vereinfachte Numerische Integration," *Der Kongelige Norske Videnskaber Selskab Forhandlinger*, 28, 30–36. [1565]

Sørensen, H. (2004), "Parametric Inference for Diffusion Processes Observed at Discrete Points in Time: A Survey," *International Statistical Review*, 72, 337–354. [1558]

Stuart, A. M., Voss, J., and Wilberg, P. (2004), "Conditional Path Sampling of SDEs and the Langevin MCMC Method," *Communications in Mathematical Sciences*, 2, 685–697. [1559]

Talay, D., and Tubaro, L. (1990), "Expansion of the Global Error for Numerical Schemes Solving Stochastic Differential Equations," *Stochastic Analysis and Applications*, 8, 483–509. [1565]

Wong, W. H., and Liang, F. (1997), "Dynamic Weighting in Monte Carlo and Optimization," *Proceedings of the National Academy of Sciences*, 94, 14220–14224. [1572]