Monte Carlo State-Space Likelihoods by Weighted Posterior Kernel Density Estimation

Perry DE VALPINE

Maximum likelihood estimation and likelihood ratio tests for nonlinear, non-Gaussian state-space models require numerical integration for likelihood calculations. Several methods, including Monte Carlo (MC) expectation maximization, MC likelihood ratios, direct MC integration, and particle filter likelihoods, are inefficient for the motivating problem of stage-structured population dynamics models in experimental settings. An MC kernel likelihood (MCKL) method is presented that estimates classical likelihoods up to a constant by weighted kernel density estimates of Bayesian posteriors. MCKL is derived by using Bayesian posteriors as importance sampling densities for unnormalized kernel smoothing integrals. MC error and mode bias due to kernel smoothing are discussed and two methods for reducing mode bias are proposed: "zooming in" on the maximum likelihood parameters using a focused prior based on an initial estimate and using a posterior cumulant-based approximation of mode bias. A simulated example shows that MCKL can be much more efficient than previous approaches for the population dynamics problem. The zooming-in and cumulant-based corrections are illustrated with a multivariate variance estimation problem for which accurate results are obtained even in 20 parameter dimensions.

KEY WORDS: Importance sampling; Monte Carlo expectation maximization; Monte Carlo kernel likelihood; Monte Carlo likelihood ratio; Population dynamics; State-space model.

1. INTRODUCTION

A difficulty for frequentist likelihood-based analysis with mechanistic models is that the models often include unknown states or process errors, so that likelihood calculations require high-dimensional integrations. Approaches to maximum likelihood estimation using Monte Carlo (MC) integration in this setting include MC expectation maximization (MCEM) (Wei and Tanner 1990; Chan and Ledolter 1995; McCulloch 1997; Robert and Casella 1999; Booth and Hobert 1999; Hürzeler and Künsch 2001; Levine and Casella 2001); MC likelihood ratio (MCLR) estimation via importance sampling (Geyer and Thompson 1992; Geyer 1994, 1996; McCulloch 1997); basic MC integration; more advanced MC integration, such as importance sampling (Durbin and Koopman 1997, 2000); and particle filtering (PF) (Gordon, Salmond, and Smith 1993; Kitagawa 1996, 1998; Pitt and Shephard 1999; Doucet, de Freitas, and Gordon 2001b; Hürzeler and Künsch 2001). However, for my motivating model, a biologically stage-structured population dynamics model for a replicated experimental setting, none of these methods is very efficient.

I present a faster approach based on estimating likelihoods from weighted kernel density estimates of a Bayesian posterior, which is motivated by importance sampling the kernel smoothing integral. I consider convergence and accuracy of the method, including MC error and mode bias due to kernel smoothing. I give methods to gain accuracy by "zooming in" on the likelihood maximum using focused priors, which is equivalent to importance sampling near the maximum, and to estimate the mode bias due to kernel smoothing using posterior cumulants. I give a second example showing that even in 10–20 dimensions, these methods can provide high accuracy for chi-squared hypothesis test cutoffs. This approach may be widely and easily applicable, because it can take advantage of the boom in computational methods for Bayesian posterior sampling.

My investigation is motivated by the need to fit demographic models to time series from ecological population dynamics experiments. In population ecology there is a wide gap between plausible mechanistic models and models commonly used to analyze data. Population dynamics experiments typically produce replicated time series structured by species, life stages within species, and/or location. These common types of experiments (reviewed by Hairston 1989, Underwood 1997; Resetarits and Bernardo 1998) follow a long tradition that includes classic work by Huffaker (1958) on predatory and herbivorous spider mites and by Gause (1934) and Luckinbill (1973) on predator and prey protozoans. Population dynamics experiments are used to study direct and indirect species interactions (e.g., Wootton 1994; Rosenheim, Kaya, Ehler, Marois, and Jaffee 1995; Karban, English-Loeb, and Hougen-Eitzman 1997; Murdoch, Nisbet, McCauley, de Roos, and Gurney 1998; Ellner et al. 2001; Rosenheim 2001; Snyder and Ives 2001), population cycles (e.g., Nicholson and Bailey 1935; Gurney, Blythe, and Nisbet 1980; McCauley, Nisbet, Murdoch, de Roos, and Gurney 1999), species' roles in ecosystem functioning (e.g., Naeem, Thompson, Lawler, Lawton, and Woodfin 1994; Downing and Leibold 2002), trophic cascades (e.g., Carpenter and Kitchell 1993; Ives, Carpenter, and Dennis 1999; Strong, Whipple, Child, and Dennis 1999; Pace, Cole, Carpenter, and Kitchell 1999; Klug, Fischer, Ives, and Dennis 2000), and laboratory "model" systems (e.g., Costantino, Desharnais, Cushing, and Dennis 1997; Kaunzinger and Morin 1998; Holyoak 2000; Dennis, Desharnais, Cushing, Henson, and Costantino 2001).

I use agricultural insect ecology to illustrate the model-fitting problem for population dynamics experiments. Entomologists routinely conduct laboratory, greenhouse, and field experiments in which abundances of eggs, immatures, and adults of different species are estimated at several times under various plant conditions, predator communities, or other experimental treatments. Currently, such data are almost always analyzed with generalized linear models that do not reflect processes of reproduction, growth, mortality, and predation that produced the data. In contrast, the models used for theoretical studies of population

Perry de Valpine is Postdoctoral Researcher, Department of Integrative Biology, University of California, Berkeley, CA 94720-3140 (E-mail: *devalpin@nceas.ucsb.edu*). This work was supported by the National Center for Ecological Analysis and Synthesis, a center funded by the National Science Foundation (grant DEB-0072909), the University of California (UC), and UC Santa Barbara; and the U.S. Department of Agriculture National Research Initiative Competitive Grants Program.

Journal of the American Statistical Association
June 2004, Vol. 99, No. 466, Theory and Methods
DOI 10.1198/016214504000000476

dynamics describe these processes and often involve nonlinear predator–prey interactions, time lags arising from development, and other factors that can produce complex dynamics (e.g., Metz and Diekmann 1986; Tuljapurkar and Caswell 1997; Gurney and Nisbet 1998).

Because of their relative dimensionality of noises, states, observations, and replicates, experimental population dynamics problems have a different balance of efficiency issues than other state-space or hidden-variables problems. MC state-space research has typically considered methods for single long time series, such as for financial data (Carlin, Polson, and Stoffer 1992; Shephard and Pitt 1997; Tanizaki and Mariano 1998; Pitt and Shephard 1999; Durham and Gallant 2002) or fisheries catch and survey records (Bjørnstad, Fromentin, Stenseth, and Gjosaeter 1999; Meyer and Millar 1999; Millar and Meyer 2000). At the other extreme, MC methods for experimental data have been applied to generalized linear mixed models, where data are iid and the likelihood requires a MC integration over unknown random variables (Clayton 1996; McCulloch 1997; Booth and Hobert 1999). Experimental population dynamics data mix these features in the form of short, replicated time series. In addition, realistic population dynamics models are often continuous in time or have a short time step, so that the space of random disturbances entering the process may be of much higher dimension than the observation space. Also, the aim here is to calculate approximate likelihood ratio tests, whereas most MC state-space research has addressed nonexperimental settings with more focus on filtering and smoothing than on parameter estimation and testing.

I compare five MC methods for likelihood maximization. Two of the main approaches in the literature are MCEM (Wei and Tanner 1990; Chan and Ledolter 1995; McCulloch 1997; Robert and Casella 1999; Booth and Hobert 1999) and MCLR (Geyer and Thompson 1992, Geyer 1994, 1996; McCulloch 1997), which both use alternating steps of MC sampling and local maximization (but see Levine and Casella 2001 for potential improvements). Two other natural candidates for the problem are PF (Gordon et al. 1993; Kitagawa 1996; Pitt and Shephard 1999; Doucet et al. 2001b) and basic MC integration, possibly with importance sampling, which I call *MC direct* (MCD). The method developed here, *MC kernel likelihood* (MCKL), temporarily treats parameters as random variables for purposes of Markov Chain MC (MCMC) sampling and then uses a weighted kernel density estimator to approximate a classical likelihood surface. MCKL is related to kernel density estimation of Bayesian posterior distributions (West 1993; Chen 1994; Givens and Raftery 1996; Liu and West 2001), but the approach here of using kernel density estimates to recover the classical likelihood surface for a state-space model appears to be new.

Next I introduce the MC likelihood methods in detail. For MCKL, I discuss convergence, MC error and smoothing mode bias, zooming in by resampling near the mode to reduce smoothing bias, and estimating smoothing bias from posterior cumulants. I then introduce an example population model and hypothetical experiment, and compare maximum likelihood convergence of the different methods. Finally, I use a problem of multivariate standard deviation estimation to illustrate and evaluate the zooming and cumulant-based corrections.

2. MONTE CARLO LIKELIHOOD METHODS

Suppose that there are n experimental units with data vectors $\mathbf{Y}_i$, $i = 1, \dots, n$, which include both multiple times and multiple observation dimensions. To be more explicit, denote $\mathbf{Y}_i = (\mathbf{Y}_i(1), \dots, \mathbf{Y}_i(T))$, where $\mathbf{Y}_i(t)$ is a vector of observations at time t and there is a fixed set of observation times, $t = 1, \dots, T$. (For notational simplicity, I assume the same set of observation times for each replicate.) The methods here require only that the model be amenable to various MC algorithms: MCMC for the MCKL, MCEM, and MCLR methods and sequential particle filtering for the PF method. For each replicate i, let $\boldsymbol{\nu}_i$ be the unknown states or process noises, with probability density $\Pr(\boldsymbol{\nu}_i)$, and let $\Pr(\mathbf{Y}_i|\boldsymbol{\nu}_i)$ be the probability density of observations given states. Denote $\boldsymbol{\nu}_i = (\boldsymbol{\nu}_i(1), \dots, \boldsymbol{\nu}_i(T))$, where $\boldsymbol{\nu}_i(t)$ are the process noises from time $t - 1$ to t, which may include many model time steps (and noise values) between observation times.

Each method aims to maximize the likelihood integral for the d-dimensional parameter vector $\boldsymbol{\Theta}$,

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \int \Pr(\mathbf{Y}_i|\boldsymbol{\nu}_i) \Pr(\boldsymbol{\nu}_i)\, d\boldsymbol{\nu}_i, \tag{1}$$

or $l(\boldsymbol{\Theta}) = \log L(\boldsymbol{\Theta})$. Although (1) makes sense with $\boldsymbol{\nu}$ as either states or process noises, from here on I use it as process noises and write $\mathbf{X}(\boldsymbol{\nu})$ for the states. In other words, $\Pr(\mathbf{Y}_i|\boldsymbol{\nu}_i)$ will usually involve a calculation of state dynamics from the process noises, with the observation density related to the state values. In the example that follows, $\boldsymbol{\nu}$'s are random environmental variations, $\mathbf{X}$ is a time trajectory of dynamics for an age cohort model with a 1-day time step, and $\mathbf{Y}$ is a time series of estimates taken every 10 days of the abundance of several life stages, each of which is a summation of multiple age cohorts. For general introduction of each method, I do not need to specify dimensions for states, noises, or observations. Treatment of initial conditions is assumed to be included in the notation. If these conditions are fixed, then the state calculations start from them; if they are random, then they are included in the process noises. Dependence of probabilities on $\boldsymbol{\Theta}$ is suppressed in the notation. Finally, everything is written for continuous variables, but could be easily adapted to discrete variables.

Only the MCD and PF methods estimate the likelihood without introducing an unknown constant. For the other methods, after estimating the maximum likelihood estimator (MLE), it is necessary to estimate the likelihood at the MLE for purposes of likelihood ratio approximate hypothesis tests, and I assume that this is feasible. For my population dynamics example, I do this through an importance-sampled MC estimate of (1), with an estimate of each $\Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i)$ as the importance density for each $\Pr(\boldsymbol{\nu}_i)$ (Shephard and Pitt 1997).

2.1 Monte Carlo Direct

The most obvious approach, MCD draws a large sample $\{\boldsymbol{\nu}_i^{(j)}\}_{j=1}^{m}$ from each $\Pr(\boldsymbol{\nu}_i)$ and uses

$$l(\boldsymbol{\Theta}) \approx \sum_{i=1}^{n} \log\left[\frac{1}{m}\sum_{j=1}^{m} \Pr\bigl(\mathbf{Y}_i|\boldsymbol{\nu}_i^{(j)}\bigr)\right]. \tag{2}$$

This method suffers from the inefficiency of basic MC integration: a large variance of $\Pr(\mathbf{Y}_i|\boldsymbol{\nu}_i^{(j)})$ (Robert and Casella

1999, sec. 3.2). However, it has the potential efficiency that the same noise sample and state trajectories can be used for each replicate within a treatment group. If the same model applies to $i \in \mathbf{I}$, where $\mathbf{I}$ defines a treatment group, then a single sample $\{\boldsymbol{\nu}_{\mathbf{I}}^{(j)}\}_{j=1}^{m}$ and only one calculation of each $\mathbf{X}(\boldsymbol{\nu}_{\mathbf{I}}^{(j)})$ can be used in the inner sum of (2) for each $i \in \mathbf{I}$. This may be useful if $\mathbf{X}(\boldsymbol{\nu}_{i}^{(j)})$ is the most computationally intensive step for each j. For likelihood maximization, MCD provides a smooth surface in $\boldsymbol{\Theta}$ if the same MC sample is used for all $\boldsymbol{\Theta}$.

2.2 Particle Filter

A basic PF, or bootstrap filter (Doucet et al. 2001a), likelihood approximation uses a sampling importance-resampling (SIR)-type algorithm at each observation time to obtain a sample from $\Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i)$, as follows. The algorithm handles each replicate separately, and subscripts i are omitted in this section and replaced with time subscripts. Denote $\mathbf{Y}_t = \mathbf{Y}(t)$, $\boldsymbol{\nu}_t = \boldsymbol{\nu}(t)$, $\mathbf{Y}_{1:t} = (\mathbf{Y}_1, \dots, \mathbf{Y}_t)$, $\boldsymbol{\nu}_{1:t} = (\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_t)$, and $\boldsymbol{\nu}_{1:t|1:s} = (\boldsymbol{\nu}_{1:t}|\mathbf{Y}_{1:s})$. Also define dummy variables $\mathbf{Y}_0 = \varnothing$ and $\mathbf{Y}_{1:0} = \varnothing$. Factor the likelihood sequentially for each replicate, so that

$$\Pr(\mathbf{Y}|\boldsymbol{\Theta}) = \prod_{t=1}^{T} \Pr(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}). \tag{3}$$

PF uses the following steps, starting with $t = 1$ and a sample $\{\boldsymbol{\nu}_{1:t|1:t-1}^{(j)}\}_{j=1}^{m}$, from $\Pr(\boldsymbol{\nu}_{1:t|1:t-1})$:

1. Estimate $\Pr(\mathbf{Y}_t|\mathbf{Y}_{1:t-1})$ by direct MC integration as $\frac{1}{m}\sum_{j=1}^{m} \Pr(\mathbf{Y}_t|\boldsymbol{\nu}_{1:t|1:t-1}^{(j)})$.
2. Define normalized weights $w^{(j)} = \Pr(\mathbf{Y}_t|\boldsymbol{\nu}_{1:t|1:t-1}^{(j)}) / \sum_{j=1}^{m} \Pr(\mathbf{Y}_t|\boldsymbol{\nu}_{1:t|1:t-1}^{(j)})$.
3. Generate a sample from $\Pr(\boldsymbol{\nu}_{1:t|1:t})$ by drawing $(\boldsymbol{\nu}_{1:t|1:t-1}^{(j)})$ with probability $w^{(j)}$.
4. Generate a sample from $\Pr(\boldsymbol{\nu}_{1:t+1|1:t})$ by drawing $\boldsymbol{\nu}_{t+1|1:t}^{(j)}$ from $\Pr(\boldsymbol{\nu}_{t+1}|\boldsymbol{\nu}_{1:t|1:t}^{(j)})$ and setting $\boldsymbol{\nu}_{1:t+1|1:t}^{(j)} = (\boldsymbol{\nu}_{1:t|1:t}^{(j)}, \boldsymbol{\nu}_{t+1|1:t}^{(j)})$. Increment t and return to step 1.

Gordon et al. (1993) showed that this procedure works asymptotically (as $m \to \infty$), but the primary difficulty in practice is sample degradation as samples are lost in step 3 as t increases (Doucet et al. 2001b). For likelihood maximization, PF has the serious difficulty of not providing a smooth surface in $\boldsymbol{\Theta}$. Hürzeler and Künsch (2001) handled this by loess smoothing the PF estimate of $L(\boldsymbol{\Theta})$, a step that I do not try here because it would be computationally expensive in the dimensionality of my examples.

2.3 Monte Carlo Expectation Maximization

MCEM works by using MC samples for the expectations of the EM algorithm. The MCEM algorithm is to start with $\boldsymbol{\Theta}_0$, obtain samples $\{\boldsymbol{\nu}_{i}^{(j)}\}_{j=1}^{m}$ from $\Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i, \boldsymbol{\Theta}_0)$ for each replicate i, find $\boldsymbol{\Theta}_1$ by maximizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{1}{m} \log \Pr\bigl(\mathbf{Y}_i, \boldsymbol{\nu}_i^{(j)}|\boldsymbol{\Theta}_1\bigr) \approx \sum_{i=1}^{n} \int \log \Pr(\mathbf{Y}_i, \boldsymbol{\nu}_i|\boldsymbol{\Theta}_1) \Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i, \boldsymbol{\Theta}_0)\, d\boldsymbol{\nu}_i, \tag{4}$$

and finally set $\boldsymbol{\Theta}_0$ to the new value of $\boldsymbol{\Theta}_1$ and repeat the procedure. The samples from $\Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i, \boldsymbol{\Theta}_0)$ for each i could in general be obtained with an MCMC or other algorithm. This method requires a sample for each replicate for each iteration of the algorithm, although Levine and Casella (2001) suggested simply reweighting the previous sample according to importance sampling principles, if it covers the region of interest for the next maximization step. MCEM also has the advantages and drawbacks of the EM algorithm (Chan and Ledolter 1995; McCulloch 1997); it can be slow to converge and can converge to local maxima.

2.4 Monte Carlo Likelihood Ratio

MCLR works by using MC samples to approximate likelihood ratios. The approach is based on the importance sampling identity

$$l(\boldsymbol{\Theta}) - l(\boldsymbol{\Theta}_S) = \sum_{i=1}^{n} \log\left[\int \frac{\Pr(\mathbf{Y}_i, \boldsymbol{\nu}_i|\boldsymbol{\Theta})}{\Pr(\mathbf{Y}_i, \boldsymbol{\nu}_i|\boldsymbol{\Theta}_S)} \Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i, \boldsymbol{\Theta}_S)\, d\boldsymbol{\nu}_i\right]. \tag{5}$$

For MC approximation, one starts with a fixed initial guess $\boldsymbol{\Theta}_S$ and obtains samples (like MCEM) $\{\boldsymbol{\nu}_{i}^{(j)}\}_{j=1}^{m}$, from $\Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i, \boldsymbol{\Theta}_S)$ for each replicate i and maximizes over $\boldsymbol{\Theta}$

$$l(\boldsymbol{\Theta}) - l(\boldsymbol{\Theta}_S) \approx \sum_{i=1}^{n} \log\left[\sum_{j=1}^{m} \frac{\Pr(\mathbf{Y}_i, \boldsymbol{\nu}_i^{(j)}|\boldsymbol{\Theta})}{\Pr(\mathbf{Y}_i, \boldsymbol{\nu}_i^{(j)}|\boldsymbol{\Theta}_S)}\right]. \tag{6}$$

The efficiency of this method depends on how the importance sampling approximation breaks down as $\boldsymbol{\Theta}$ gets far from $\boldsymbol{\Theta}_S$, and MCLR has been observed to perform poorly in some cases (Geyer 1994; McCulloch 1997). For example, if $\boldsymbol{\Theta}$ is the variance of a normal distribution, then for $\boldsymbol{\Theta} > \boldsymbol{\Theta}_S$, the MC approximate integrals may not have finite variance (Robert and Casella 1999, sec. 5.3.2). Geyer (1996) suggested iterating the procedure—maximize $\boldsymbol{\Theta}$ in a trust region around $\boldsymbol{\Theta}_S$, set $\boldsymbol{\Theta}_S$ equal to $\boldsymbol{\Theta}$, and start over—but this does not necessarily solve the convergence issue. Geyer (1994) and Geyer (1996) suggested using importance sampling distributions that are mixtures of the conditional noise distributions from several different reference parameters. This seems to work well for a one-dimensional parameter problem of an Ising model (Geyer 1994), but would gain complexity for higher-dimensional parameter spaces. Like MCEM, MCLR requires a sample for each replicate for each iteration of the algorithm.

3. MONTE CARLO KERNEL LIKELIHOOD METHOD

The MCKL method appears to be a new approach to maximum likelihood parameter estimation for state-space models, but it is related to kernel density estimation of Bayesian posterior densities (West 1993; Chen 1994; Givens and Raftery 1996; Liu and West 2001). The approach involves temporarily treating parameters as having probability densities and sampling from a posterior density as in Bayesian methods. The likelihood can then be estimated up to a constant as a weighted kernel density estimate, with weights obtained by viewing the posterior as an importance sampling density. (For a wide prior, the likelihood can also be estimated up to a constant as an unweighted kernel density estimate divided by the prior; I briefly discuss this version later.) MCKL does not involve a Bayesian interpretation of parameters, because only likelihoods are recovered in the end, but I use terms like "prior" and "posterior" for consistency with Bayesian methods.

MCKL involves the following steps:

1. Choose a prior $\Pr(\boldsymbol{\Theta})$ and use an MCMC (or some other algorithm, such as a joint parameter-state particle filter; Liu and West 2001) to obtain a sample from

$$\Pr(\boldsymbol{\Theta}, \boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_n | \mathbf{Y}_1, \dots, \mathbf{Y}_n) \propto \Pr(\boldsymbol{\Theta}) \prod_{i=1}^{n} \Pr(\mathbf{Y}_i, \boldsymbol{\nu}_i | \boldsymbol{\Theta}). \tag{7}$$

The $\boldsymbol{\Theta}$ dimensions of this sample are a sample from the posterior,

$$\begin{aligned}\Pr_S(\boldsymbol{\Theta}) \equiv \Pr(\boldsymbol{\Theta}|\mathbf{Y}_1, \dots, \mathbf{Y}_n) &= \frac{L(\boldsymbol{\Theta}) \Pr(\boldsymbol{\Theta})}{C_S} \\ &= \int \cdots \int \Pr(\boldsymbol{\Theta}, \boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_n | \mathbf{Y}_1, \dots, \mathbf{Y}_n)\, d\boldsymbol{\nu}_1 \cdots d\boldsymbol{\nu}_n, \end{aligned} \tag{8}$$

where $C_S = \int L(\boldsymbol{\Theta}) \Pr(\boldsymbol{\Theta})\, d\boldsymbol{\Theta}$.

2. Maximize the following kernel density estimate of the likelihood up to the unknown constant C_S:

$$\hat{L}_{\mathbf{h}}(\boldsymbol{\Theta}) = \frac{1}{m} \sum_{j=1}^{m} K_{\mathbf{h}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(j)}) w^{(j)}, \tag{9}$$

$$w^{(j)} = \frac{1}{\Pr(\boldsymbol{\Theta}^{(j)})},$$

where $\{\boldsymbol{\Theta}^{(j)}\}_{j=1}^{m}$ are the sample points and $K_{\mathbf{h}}$ is a normalized kernel smoother function with multivariate bandwidth $\mathbf{h} = (h_1, \dots, h_d)$. For convenience, I assume throughout that $K_{\mathbf{h}}$ is orthogonal and oriented along the coordinate axes; that is, $K_{\mathbf{h}}(\boldsymbol{\Theta}, \boldsymbol{\eta}) = \prod_{l=1}^{d} K_{h_l}(\Theta_l, \eta_l)$ where h_l, Θ_l, and η_l are the l-axis components of $\mathbf{h}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\eta}$ in d dimensions and K_{h_l} is a one-dimensional kernel smoother. For example, for Gaussian $K_{\mathbf{h}}$, I use a diagonal covariance matrix with entries $(h_1^2, \dots, h_d^2)$. I also assume $K_{\mathbf{h}}$ symmetric in every dimension, with $K_{\mathbf{h}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(j)}) = K_{\mathbf{h}}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(j)})$.

To see that (9) is an unnormalized, importance-sampled, kernel estimate of L, consider a function $g(\boldsymbol{\Theta})$ such that $\int g(\boldsymbol{\Theta}) L(\boldsymbol{\Theta})\, d\boldsymbol{\Theta}$ exists. The usual importance sampling motivation is

$$\int g(\boldsymbol{\Theta}) L(\boldsymbol{\Theta})\, d\boldsymbol{\Theta} = \int g(\boldsymbol{\Theta}) \frac{L(\boldsymbol{\Theta})}{\Pr_S(\boldsymbol{\Theta})} \Pr_S(\boldsymbol{\Theta})\, d\boldsymbol{\Theta}, \tag{10}$$

where $\Pr_S$ is an importance density. MCKL uses the special choice $\Pr_S(\boldsymbol{\Theta}) = \Pr(\boldsymbol{\Theta}) L(\boldsymbol{\Theta}) / C_S$ because L cannot be calculated directly, giving the MC estimate,

$$\int g(\boldsymbol{\Theta}) L(\boldsymbol{\Theta})\, d\boldsymbol{\Theta} \approx \frac{C_S}{m} \sum_{j=1}^{m} \frac{g(\boldsymbol{\Theta}^{(j)})}{\Pr(\boldsymbol{\Theta}^{(j)})}, \tag{11}$$

$$\boldsymbol{\Theta}^{(j)} \sim \Pr_S(\boldsymbol{\Theta}), \qquad j = 1, \dots, m,$$

MCKL uses $g(\boldsymbol{\Theta}^{(j)}) = K_{\mathbf{h}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(j)})$ and drops the normalization constant C_S.

3.1 Convergence

Next I examine several aspects of MCKL convergence with a focus on understanding rather than technical proofs. Romano (1988) proved almost sure convergence of modes of kernel estimates as $m \to \infty$ and $h \to 0$ with $h \gg \log(m)/m$ (see also Grund and Hall 1995). These proofs have the useful feature of requiring regularity only in a neighborhood of the true mode, so the possibilities of nonintegrability or infinite "moments" of L are not problematic. Simply put, in reasonable cases kernel density estimates and their derivatives converge to their true values as $m \to \infty$ and $h \to 0$ in a coordinated way, so mode estimates converge to the true mode.

To transfer proofs of convergence of kernel mode estimates to the MCKL setting, compare unweighted, normalized kernel density estimates, for which the proofs were developed, to the weighted, unnormalized kernel density estimates of MCKL. The relation between the unweighted and weighted estimates is exactly the relation between simple MC integration and importance-sampled MC integration, and it is well known that for a well-chosen importance density, the latter converges at least as well as the former and is often more accurate. The unknown normalizing constant of MCKL is bounded over proper priors (assuming bounded L), and scaling by a bounded constant also leaves proofs of mode convergence intact. Thus mode convergence transfers to both the importance sampling and scaling aspects of MCKL. Moreover, MCKL can be more accurate for mode estimation in a region of high posterior density than unweighted kernel density estimates, essentially because it can use more sample points near the mode.

3.2 Accuracy in Practice

As in kernel estimation of entire densities, the key challenge in kernel mode estimation is more practical than theoretical—that is, to obtain useful choices of $h > 0$ and $m < \infty$. Because $h > 0$ in practice, it is useful to consider accuracy for fixed h as $m \to \infty$. Denote the true likelihood as L_0, with true mode $\boldsymbol{\Theta}_0$, and the smoothed likelihood as $L_{\mathbf{h}} = K_{\mathbf{h}} * L_0$, where "$*$" denotes convolution, with true mode $\boldsymbol{\Theta}_{\mathbf{h}} = \boldsymbol{\Theta}_0 + \Delta\boldsymbol{\Theta}_{\mathbf{h}}$ and MCKL mode estimate $\hat{\boldsymbol{\Theta}}_{\mathbf{h}}$. Here $\hat{\boldsymbol{\Theta}}_{\mathbf{h}} - \boldsymbol{\Theta}_{\mathbf{h}}$ is the MC error and $\Delta\boldsymbol{\Theta}_{\mathbf{h}}$ is the smoothing bias for fixed $\mathbf{h}$. As before, I use $K_{\mathbf{h}}$ orthogonal and oriented along the coordinate axes, with $\mathbf{h} = (h_1, \dots, h_d)$. I use the derivative notation $\partial_l = \partial/\partial\Theta_l$, $\partial_{lm} = \partial^2/\partial\Theta_l\partial\Theta_m$, and so on, and $\nabla L = (\partial_1 L, \dots, \partial_d L)$, $(\nabla^2 L)_{lm} = \partial_{lm} L$. As $m \to \infty$, $C_S \hat{L}_{\mathbf{h}} \to L_{\mathbf{h}}$ by the law of large numbers. In this setting, I consider MC variance of the MCKL mode estimate as well as the more challenging problem of estimating and reducing its smoothing bias.

3.3 Monte Carlo Variance

I use M-estimator theory (van der Vaart 1998) to describe the MC error. The MCKL mode estimator $\hat{\boldsymbol{\Theta}}_{\mathbf{h}}$ is a solution to the estimating equations

$$\nabla \hat{L}_{\mathbf{h}}(\boldsymbol{\Theta}) = \frac{1}{m} \sum_{j=1}^{m} \frac{\nabla K_{\mathbf{h}}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(j)})}{\Pr(\boldsymbol{\Theta}^{(j)})} = 0. \tag{12}$$

Define $\psi_j(\boldsymbol{\Theta}) = K_{\mathbf{h}}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(j)})/\Pr(\boldsymbol{\Theta}^{(j)})$. General M-estimator theory gives that $\sqrt{m}(\hat{\boldsymbol{\Theta}}_{\mathbf{h}} - \boldsymbol{\Theta}_{\mathbf{h}})$ is asymptotically normal with mean $\mathbf{0}$ and variance

$$E[\nabla^2\psi(\boldsymbol{\Theta}_{\mathbf{h}})]^{-1}E\big[(\nabla\psi(\boldsymbol{\Theta}_{\mathbf{h}}))(\nabla\psi(\boldsymbol{\Theta}_{\mathbf{h}}))^T\big] \times E[\nabla^2\psi(\boldsymbol{\Theta}_{\mathbf{h}})]^{-1}, \quad (13)$$

where expectations are with respect to $\Pr_S(\boldsymbol{\Theta})$ (van der Vaart 1998, sec. 5.3, using $\nabla\psi$ instead of ψ). In terms of $L_{\mathbf{h}}$,

$$E[\nabla^2\psi(\boldsymbol{\Theta}_{\mathbf{h}})]^{-1} = C_S[\nabla^2 L_{\mathbf{h}}(\boldsymbol{\Theta}_{\mathbf{h}})]^{-1}, \quad (14)$$

and the (l, m) matrix component of the middle expectation is

$$E\big[(\nabla\psi(\boldsymbol{\Theta}_{\mathbf{h}}))(\nabla\psi(\boldsymbol{\Theta}_{\mathbf{h}}))^T\big]_{lm} = \frac{1}{C_S}\int \frac{[\partial_l K_{\mathbf{h}}(\boldsymbol{\Theta}_{\mathbf{h}} - \boldsymbol{\Theta})][\partial_m K_{\mathbf{h}}(\boldsymbol{\Theta}_{\mathbf{h}} - \boldsymbol{\Theta})]}{\Pr(\boldsymbol{\Theta})} L(\boldsymbol{\Theta})\,d\boldsymbol{\Theta}. \quad (15)$$

These results require that $E[\nabla^2\psi(\boldsymbol{\Theta}_{\mathbf{h}})]$ be invertible and, nontrivially, that (15) exist for all $l, m = 1, \dots, d$. (Use of m as a dimension subscript is separate from its use as sample size throughout.)

When $\Pr(\boldsymbol{\Theta})$ and $K_{\mathbf{h}}$ are Gaussian with covariances $\boldsymbol{\Sigma}_{\Pr}$ and $\boldsymbol{\Sigma}_K$, (15) is finite when $2\boldsymbol{\Sigma}_K^{-1} - \boldsymbol{\Sigma}_{\Pr}^{-1}$ is positive definite, which is always true for sufficiently small $\boldsymbol{\Sigma}_K$. When $\boldsymbol{\Sigma}_{\Pr} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$ (i.e., $\boldsymbol{\Sigma}_{\Pr}$ and $\boldsymbol{\Sigma}_K$ have the same eigenvalues), this amounts to $h_l^2 < 2\sigma_l^2$ for $l = 1, \dots, d$. MC error of $\hat{\boldsymbol{\Theta}}_{\mathbf{h}}$ can be estimated by a plug-in estimator for (13) or by a bootstrap (for independent samples) or a moving-block bootstrap for sequentially-dependent samples such as from MCMC (Kunsch 1989; Efron and Tibshirani 1993; Mignani and Rosa 1995). I now focus on the more difficult problem of smoothing bias.

3.4 Smoothing Bias

The usual conundrum of kernel density estimation is that estimation of smoothing bias requires more accurate knowledge of the density than is provided from the kernel estimates. In the current context, this is revealed by approximating $\Delta\boldsymbol{\Theta}_{\mathbf{h}}$ for small $\mathbf{h}$ using the standard Taylor series for kernel estimates (Silverman 1986; Scott 1992, sec. 6.3). Using $\boldsymbol{\Theta} = (\Theta_1, \dots, \Theta_d)$, $\boldsymbol{\Theta}_{\mathbf{h}} = (\Theta_{\mathbf{h},1}, \dots, \Theta_{\mathbf{h},d})$, $K_{\mathbf{h}}(\boldsymbol{\Theta}) = \prod_{m=1}^d K_{h_m}(\Theta_m)$, $K_{h_m}(\Theta_m) = K(\Theta_m/h_m)/h_m$, $\partial_m K_{h_m}(\Theta_m) = K'(\Theta_m/h_m)/h_m^2$, and $\Delta\Theta_{\mathbf{h},m} = \Theta_{\mathbf{h},m} - \Theta_m = h_m t_m$, all for $m = 1, \dots, d$, leads to

$$\begin{aligned} 0 &= \partial_l L_{\mathbf{h}}(\boldsymbol{\Theta}_{\mathbf{h}}) \\ &= \int\Bigg(\prod_{m\neq l} K(t_m)\Bigg)\frac{K'(t_l)}{h_l} L(\Theta_{\mathbf{h},1} - h_1 t_1, \dots, \Theta_{\mathbf{h},d} - h_d t_d)\,d\mathbf{t}. \end{aligned} \quad (16)$$

Expanding L around $\boldsymbol{\Theta}_{\mathbf{h}}$ gives

$$\begin{aligned} 0 = \int\Bigg(\prod_{m\neq l} K(t_m)\Bigg)&\frac{K'(t_l)}{h_l} \\ \times\Bigg[L(\boldsymbol{\Theta}_{\mathbf{h}}) &- \sum_i h_i t_i\,\partial_i L(\boldsymbol{\Theta}_{\mathbf{h}}) + \frac{1}{2}\sum_{i,j} h_i h_j t_i t_j\,\partial_{ij} L(\boldsymbol{\Theta}_{\mathbf{h}}) \\ &- \frac{1}{6}\sum_{i,j,k} h_i h_j h_k t_i t_j t_k\,\partial_{ijk} L(\boldsymbol{\Theta}_{\mathbf{h}}) + O(h^4)\Bigg]d\mathbf{t}. \end{aligned} \quad (17)$$

Integrating and expanding L around $\boldsymbol{\Theta}_0$ relates $\Delta\boldsymbol{\Theta}_{\mathbf{h}}$ to $\mathbf{h}$. For Gaussian $K_{\mathbf{h}}$ this gives, for each l,

$$\begin{aligned} 0 = &\sum_{m=1}^d \Delta\Theta_{\mathbf{h},m}\,\partial_{lm} L(\boldsymbol{\Theta}_0) \\ &- \sum_{j=1}^d \frac{h_j^2}{2}\Bigg(\partial_{ljj} L(\boldsymbol{\Theta}_0) + \sum_{m=1}^d \Delta\Theta_{\mathbf{h},m}\,\partial_{ljjm} L(\boldsymbol{\Theta}_0)\Bigg) \\ &+ O(h^4). \end{aligned} \quad (18)$$

Estimating $\Delta\boldsymbol{\Theta}_{\mathbf{h}}$ from (18) is useful only if the derivatives of L are known with greater accuracy than the kernel estimates used to estimate $\hat{\boldsymbol{\Theta}}_{\mathbf{h}}$ in the first place. In the usual kernel density estimation context, with a fixed data sample, higher-order estimates are the primary option to reduce smoothing bias (Jones and Signorini 1997), but these are ultimately limited by sample size. The MCKL situation is quite different, because much greater improvements can be obtained by simulating additional points in the region of the MLE, which I consider next.

3.4.1 Reducing Smoothing Bias by Zooming in. Suppose that one has an initial estimate $\hat{\boldsymbol{\Theta}}_{\mathbf{h}}$, using $\mathbf{h}$, m, $\Pr(\boldsymbol{\Theta})$, and $\Pr_S(\boldsymbol{\Theta})$ as well as some rough estimate of a new prior, $\Pr'(\boldsymbol{\Theta})$, that will yield a new posterior, $\Pr'_S(\boldsymbol{\Theta})$ with more weight than $\Pr_S(\boldsymbol{\Theta})$ on $\boldsymbol{\Theta}_{\mathbf{h}}$. The following approximation of the MC error (13) shows that, for sufficiently small $\mathbf{h}$, if $\Pr'_S(\boldsymbol{\Theta})$ puts more weight than $\Pr_S(\boldsymbol{\Theta})$ on $\boldsymbol{\Theta}_{\mathbf{h}}$, then it reduces (13). This implies that with a sample of size m from $\Pr'_S$, an improved estimate $\hat{\boldsymbol{\Theta}}_{\mathbf{h}'}$, with $h'_l < h_l$, $l = 1, \dots, d$, can be obtained with smaller smoothing bias and no greater MC error than for $\hat{\boldsymbol{\Theta}}_{\mathbf{h}}$. A natural choice for $\Pr'$ is a parametric estimate of $\Pr_S$, which performs well in Example 2 of Section 5. Alternatively, $\Pr'$ might be based on an estimate of $\Delta\boldsymbol{\Theta}_{\mathbf{h}}$, which could come from a higher-order kernel estimate or, for a wide prior, from the cumulant-based estimate given below.

Define $\mathbf{A}^S$ (with dependence on $\mathbf{h}$ implicit) to be the matrix factor of (13) that depends on Pr (or $\Pr_S$), that is, $\mathbf{A}^S_{lm} = C_S^2 E[(\nabla\psi(\boldsymbol{\Theta}_{\mathbf{h}}))(\nabla\psi(\boldsymbol{\Theta}_{\mathbf{h}}))^T]_{lm}$, which, from (15), is

$$A^S_{lm} = \int [\partial_l K_{\mathbf{h}}(\boldsymbol{\Theta}_{\mathbf{h}} - \mathbf{y})][\partial_m K_{\mathbf{h}}(\boldsymbol{\Theta}_{\mathbf{h}} - \mathbf{y})]\frac{L(\mathbf{y})^2}{\Pr_S(\mathbf{y})}\,d\mathbf{y}. \quad (19)$$

Using the same type of expansion as (17) gives, for the lth diagonal element,

$$\begin{aligned} \mathbf{A}^S_{ll} = &\frac{1}{h_l^2\prod_{m=1}^d h_m} \\ &\times\Bigg[\frac{L(\boldsymbol{\Theta}_{\mathbf{h}})^2}{\Pr_S(\boldsymbol{\Theta}_{\mathbf{h}})}[K]^{d-1}[K'] \\ &\quad + [K]^{d-2}[K'][K_{(2)}]\sum_{i\neq l} h_i^2\,\partial_{ii}\Bigg(\frac{L(\boldsymbol{\Theta}_{\mathbf{h}})^2}{\Pr_S(\boldsymbol{\Theta}_{\mathbf{h}})}\Bigg) \\ &\quad + [K]^{d-1}[K'_{(2)}]h_l^2\,\partial_{ll}\Bigg(\frac{L(\boldsymbol{\Theta}_{\mathbf{h}})^2}{\Pr_S(\boldsymbol{\Theta}_{\mathbf{h}})}\Bigg) + O(h^4)\Bigg], \end{aligned} \quad (20)$$

where $[K] = \int K(t)^2\,dt$, $[K'] = \int K'(t)^2\,dt$, $[K_{(2)}] = \int t^2 K(t)^2\,dt$, and $[K'_{(2)}] = \int t^2 K'(t)^2\,dt$. Off diagonal-terms can be derived similarly, but have no $O(1)$ term inside the brackets. For sufficiently small $\mathbf{h}$, the leading term indicates that asymptotic variance decreases as $\Pr_S$ puts more weight at $\boldsymbol{\Theta}_{\mathbf{h}}$. This suggests the interpretation that, to first order, using $\Pr'$ with sample size m' is equivalent to using Pr with sample size $m'_e = m' \Pr'_S(\boldsymbol{\Theta}_{\mathbf{h}}) / \Pr_S(\boldsymbol{\Theta}_{\mathbf{h}})$, which I call the "effective m" for $\Pr'_S$ relative to $\Pr_S$.

Given more specific knowledge of L, one could derive more careful choices of $\Pr'$, but because L is unknown, it is useful that mismatches of $\Pr'$, $\mathbf{h}'$, and m' can be easily diagnosed. The primary danger is that if $\mathbf{h}'$ is too large, then spurious maxima can occur in the tails of $L_{\mathbf{h}}$ due to the $1/\Pr'$ weights, which is like the danger of infinite variance of an importance sampled integral with a light-tailed importance density. Fortunately, this can be easily diagnosed by examining the distribution of weights in (9) at $\hat{\boldsymbol{\Theta}}_{\mathbf{h}'}$ and by making independent calculations of $L(\hat{\boldsymbol{\Theta}}_{\mathbf{h}'})$. There is also the usual pitfall for kernel density situations that for $\mathbf{h}'$ too small, MC error can be large.

3.4.2 Estimating Smoothing Bias From Posterior Cumulants. A different approach, which shows good results in Example 2 (Sec. 5), uses the multivariate Edgeworth approximation (e.g., Severini 2000, sec. 2.3) to estimate smoothing mode bias. For a density $f(\mathbf{y})$ with mean zero,

$$f(\mathbf{y}) \approx \phi(\mathbf{y}, \boldsymbol{\Sigma}) \left[1 + \frac{1}{6} \sum_{ijk} \kappa_{ijk} H_{ijk}(\mathbf{y}, \boldsymbol{\Sigma}) \right], \qquad (21)$$

where ϕ is the multivariate normal density, κ_{ijk} are the third cumulants of f, and H_{ijk} are the Hermite polynomials

$$H_{ijk} = z_i z_j z_k - \lambda_{jk} z_i - \lambda_{ik} z_j - \lambda_{ij} z_k, \qquad (22)$$

where $\mathbf{z} = \boldsymbol{\Sigma}^{-1}\mathbf{y}$ and $\lambda_{ij} = (\boldsymbol{\Sigma}^{-1})_{ij}$. Following the approximation of the distance between the mean and mode of a univariate, unit-variance distribution as approximately $-\frac{1}{2}\kappa_3$ (Stuart and Ord 1994, exercise 6.20) from keeping $O(z)$ and larger terms in the derivative of f, the kth dimension of the mean-mode distance, denoted as $\mathbf{y}^*_f(k)$, is approximated by

$$\mathbf{y}^*_f(k) \approx -\frac{1}{2} \sum_{ij} \kappa_{ijk} (\boldsymbol{\Sigma}^{-1})_{ij}. \qquad (23)$$

Convolution with a Gaussian kernel adds covariance matrices but leaves the mean and third cumulants unchanged. Therefore, for the convolved density $g = K_{\mathbf{h}} * f$, where $K_{\mathbf{h}}$ has covariance matrix $\boldsymbol{\Sigma}_{\mathbf{h}}$, the difference between the modes of g and f is approximately

$$\mathbf{y}^*_g(k) - \mathbf{y}^*_f(k)$$
$$\approx -\frac{1}{2} \sum_{ij} \kappa_{ijk} \left[\left((\boldsymbol{\Sigma} + \boldsymbol{\Sigma}^2_{\mathbf{h}})^{-1} \right)_{ij} - (\boldsymbol{\Sigma}^{-1})_{ij} \right], \qquad (24)$$

$k = 1, \ldots, d$. This approximation can be used for MCKL with a wide prior and $f = \Pr_S$, $g = K_{\mathbf{h}} * \Pr_S$, to estimate the smoothing mode bias of the posterior as an approximation to that of L.

3.5 Choice of h

Complex methods for choosing $\mathbf{h}$ have been proposed for general kernel mode estimation problems (e.g., Romano 1988; Grund and Hall 1995), but here I consider a simpler choice motivated by the asymptotically Gaussian shape of L and by the potential for zooming in sufficiently so that L can be approximated as Gaussian near its mode. If L is unit Gaussian, and Pr and $K_{\mathbf{h}}$ are Gaussian with mean 0 and variances σ_p and h in each dimension, then the asymptotic variance (13) in each dimension l is

$$\operatorname{var}(\hat{\Theta}_{\mathbf{h},l}) \approx \frac{(\sigma_p^2)^{d+1}(1+h^2)^{d+2}}{m(1+\sigma_p^2)^{d/2} h^{d+2} (h^2(\sigma_p^2 - 1) + 2\sigma_p^2)^{d/2+1}}, \qquad (25)$$

which, for $\sigma_p \to \infty$, is

$$\operatorname{var}(\hat{\Theta}_{\mathbf{h},l}) \approx \frac{(1+h^2)^{d+2}}{m h^{d+2} (h^2 + 2)^{d/2+1}}. \qquad (26)$$

Given a target value of $\operatorname{var}(\hat{\Theta}_{\mathbf{h},l})$ and a feasible sample size m, (25) or (26) can be numerically solved for h. I consider choosing the target value by fixing $p = \Pr(L(\hat{\boldsymbol{\Theta}}_{\mathbf{h}})/L(\boldsymbol{\Theta})) > q$, where p and q must be chosen. Solutions of (25) and (26) give somewhat counterintuitive results, because increasing h *increases* accuracy in the Gaussian case (i.e., as $h \to \infty$, MCKL estimates the mean, which for symmetric L is the mode), but this approach at least translates h into a useful accuracy statement. A coordinate scaling still must be estimated, because (25) and (26) use unit variance for L. A simple choice is to use standardized principal components of the posterior (for a wide prior). Another possibility would be to estimate the Hessian of L around its mode and then scale axes so that it approximates a unit-Gaussian mode. Although further results on automated, optimal choices of prior and kernel bandwidth could improve MCKL, manual choices are nevertheless practicable.

3.6 Unweighted Monte Carlo Kernel Likelihood

For a wide, flat prior, one can also consider an unweighted density estimate divided by the prior,

$$\tilde{L}(\boldsymbol{\Theta}) = \frac{1}{m \Pr(\boldsymbol{\Theta})} \sum_{j=1}^{m} K_{\mathbf{h}}\left(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(j)}\right). \qquad (27)$$

I refer to (27) as the UMCKL estimate. In practice, UMCKL is useful only for wide priors because for fixed $\mathbf{h}$, $\tilde{L}(\boldsymbol{\Theta}) \to \frac{K_{\mathbf{h}} * \Pr_S(\boldsymbol{\Theta})}{\Pr(\boldsymbol{\Theta})}$ as $m \to \infty$. For narrow Pr, this function can have no maximum. For example, if L is $N(0,1)$ and Pr is $N(0, \sigma^2_{\Pr})$, then $K_{\mathbf{h}} * \Pr_S$ is $N(0, \frac{\sigma^2_{\Pr}}{1+\sigma^2_{\Pr}} + h^2)$, so that for $h^2 > \frac{\sigma^4_{\Pr}}{1+\sigma^2_{\Pr}}$, $\frac{K_{\mathbf{h}} * \Pr_S}{\Pr(\boldsymbol{\Theta})}$ has only a global minimum, and in practice one might not sample with m large enough to choose h below this criterion. For my population dynamics example with a nearly flat prior, UMCKL gives virtually the same results as MCKL, but MCKL is more general because of its potential for zooming in.

4. EXAMPLE 1: MAXIMUM LIKELIHOOD ESTIMATION FOR A POPULATION DYNAMICS EXPERIMENT

I compared convergence efficiency for the five MC maximum likelihood methods for a hypothetical age-structured population dynamics experiment. Simulated experiments were started with known conditions of 10 adult insects placed on each of 20 plants, 10 each grown in treatment and control conditions. Noisy observations were taken for eggs, juveniles, and adults every 10 days for 40 days (Fig. 1). This represents a simple, typical ecological experiment.

4.1 Stage-Structured Population Model

For the population model, $n(a,t)$ is the number of individuals of age a at time t, with a and t taking integer values and $1 \le a \le a_{\max}$, so the state at time t is the age vector, $\mathbf{n}(t) = [n(1,t), n(2,t), \ldots, n(a_{\max},t)]$. Using a state-space format, denote the state dynamics by

$$\mathbf{n}(t+1) = \mathbf{A}\big(\mathbf{n}(t), \boldsymbol{\nu}_{t+1}\big)\mathbf{n}(t) \tag{28}$$

and an observation at time t by

$$\mathbf{Y}_t = G\big(\mathbf{B}\mathbf{n}(t), \boldsymbol{\epsilon}_t\big), \tag{29}$$

where the matrix $\mathbf{A}$ determines day cohort survival and reproduction, matrix $\mathbf{B}$ sums day cohorts into stage totals (e.g., the first several day cohorts may be eggs, which are estimated together), G models estimation of stage totals, $\boldsymbol{\nu}_t$ is process noise, and $\boldsymbol{\epsilon}_t$ is observation noise.

The matrix $\mathbf{A}$ is called a Leslie matrix in population biology. In an age cohort model, it can have non-0 elements only in the top row, the subdiagonal, and the lower right corner. In the top row, $\mathbf{A}_{1,a}$ is the daily reproductive rate of age a individuals. In the subdiagonal, $\mathbf{A}_{a+1,a}$ is the survival rate of age a individuals who grow into age $a+1$. Cohort $a_{\max}$ serves to collect individuals with age $\ge a_{\max}$, with survival $\mathbf{A}_{a_{\max},a_{\max}}$. (See Caswell 2001 for a thorough treatment of matrix models in ecology.) Model (28) fits in the more general class of physiologically structured models with $n(a,t)$ the density of individuals age a at time t, with a and t taking continuous values (Metz and Diekmann 1986; Tuljapurkar and Caswell 1997; Gurney and Nisbet 1998). These models have a wide and complex range of potential dynamics, and it seems unlikely that fitting methods that perform inadequately for my example can handle more complex cases.

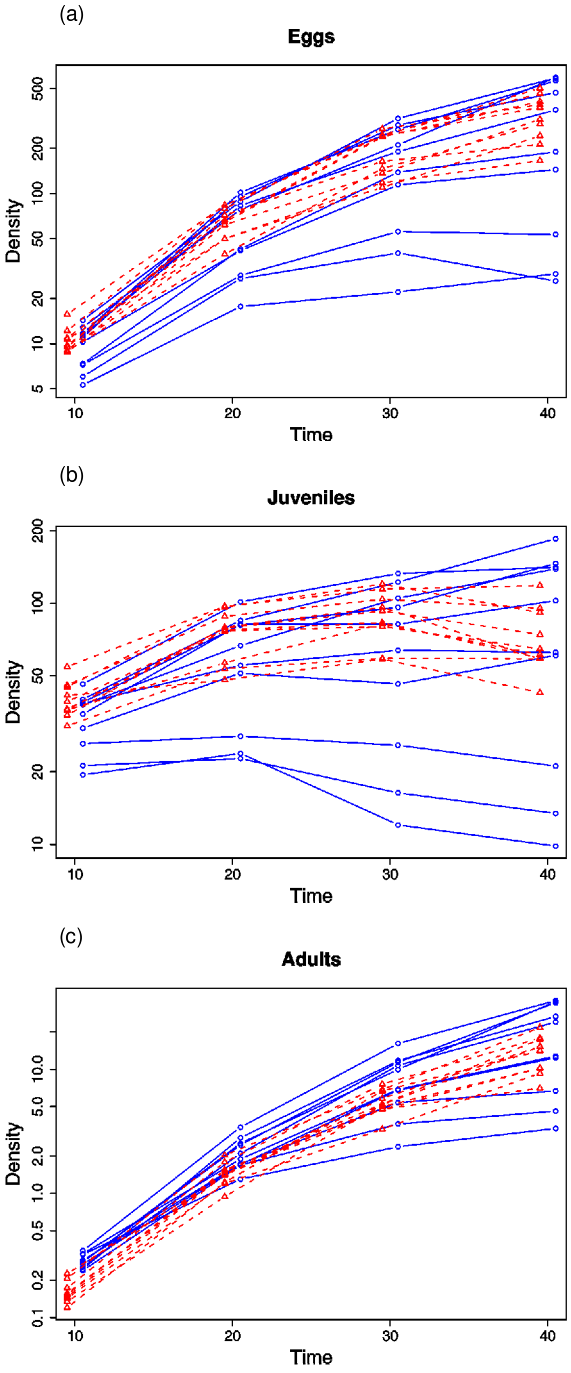

Figure 1. Simulated Data for Example 1: Control (—○—) and Treatment (-△-) Trajectories for (a) Eggs, (b) Juveniles, and (c) Adults for the Simulated Population Experiment for Which Likelihood Maximization Comparisons Were Conducted.

For my simulated examples, I assume that $\mathbf{A}$ does not depend on $\mathbf{n}$ and that day cohorts in the same life stage share the same demographic rates—a common, realistic assumption for arthropods with distinct stages (instars) separated by molting (Gurney, Nisbet, and Lawton 1983). Define individuals with $0 \le a < L_E$ as eggs, those with $L_E \le a < (L_E + L_J)$ as juveniles, and those with $a \ge (L_E + L_J)$ as adults. Let D_s be the maximum integer less than L_s, for $s \in \{E, J\}$, and let S_s be the daily survival rate for an individual in stage $s \in \{E, J, A\}$. For each stage transition there may be one day cohort with individuals in both stages, so let $h_{s_1 s_1}$ and $h_{s_1 s_2}$ be the fractions of individuals in the $s_1 \to s_2$ day cohort who experience the survival rates for stages s_1 and s_2, with the obvious choice of $h_{EE} = L_E - D_E$, $h_{EJ} = 1 - h_{EE}$, $h_{JJ} = L_J - D_J$, and $h_{JA} = 1 - h_{JJ}$. Now the non-0 elements of $\mathbf{A}$ are

$$\mathbf{A}_{a+1,a} = \begin{cases} S_E, & a \le D_E \\ h_{EE}S_E + h_{EJ}S_J, & a = D_E + 1 \\ S_J, & D_E + 1 < a \le D_J \\ h_{JJ}S_J + h_{JA}S_A, & a = D_J + 1 \end{cases} \tag{30}$$

and

$$A_{a,a} = S_A, \qquad a = a_{\max} = D_J + 2. \tag{31}$$

Only adults reproduce, so

$$\mathbf{A}_{1,a} = \begin{cases} \beta S_E(t), & a = a_{\max} \\ h'_{JA}\beta S_E(t), & a = D_J + 1, \end{cases} \tag{32}$$

where h'_{JA} is the fraction of the $J \to A$ cohort (age $D_J + 1$) in stage A at time $t + 1$, defined by $h'_{JA} = h_{JA}S_A/(h_{JJ}S_J + h_{JA}S_A)$.

For environmental randomness in the survival rates, I let S_E, S_J, and S_A change every 2 days as follows:

$$S_s(z_s(t)) = \frac{\exp[a_s + b_s z_s(t)]}{1 + \exp[a_s + b_s z_s(t)]} \tag{33}$$

$$z_s(t) \sim N(0, 1), \qquad t \text{ odd}, \tag{34}$$

and

$$z_s(t) = z_s(t-1), \qquad t \text{ even}, \tag{35}$$

where $s \in \{E, J, A\}$. For simplicity, a reasonable time scale of environmental variation has been assumed to be 2 days. One could let environmental variation be autocorrelated and/or have a different time scale. For variation in fecundity, let β for each experimental replicate be gamma distributed with mean μ_β and standard deviation σ_β.

I assume that observations are made every 10 days at times $t_1 = 10, \ldots, t_4 = 40$. To define the stage summation matrix $\mathbf{B}$, assign indices $E = 1$, $J = 2$, and $A = 3$ for the rows of $\mathbf{B}$, giving

$$\mathbf{B}_{s,a} = \begin{cases} 1, & D_{s-1} + 1 < a \le D_s \\ h'_{s-1,s}, & a = D_{s-1} + 1 \\ h'_{s,s+1}, & a = D_{s+1} + 1 \\ 0, & \text{otherwise}, \end{cases} \tag{36}$$

with $s \in \{E, J, A\}$, h'_{EJ} defined by similar logic as for h'_{JA} before, and $D_0 \equiv -1$. The observations are distributed as

$$\mathbf{Y}_s(t) \sim N(\mathbf{Bn}(t)_s, \sigma = .1\mathbf{Bn}(t)_s + .01), \tag{37}$$

where $s \in \{E, J, A\}$. I also assume that the investigators have conducted replicated observation trials independently of the experiment and know this model of the observation distributions.

For the experimental length of 40 days, with one β value and a three-dimensional $\mathbf{z}$ value for every 2 days, the dimension of $\boldsymbol{\nu}$ was 61 for each experimental unit. Control parameters were $\mu_\beta = 4$, $\sigma^2_\beta = 1$, $S_E(0) = .9$, $S_J(0) = .8$, $S_A(0) = .7$, $b_E = b_J = b_A = .1$, $L_E = 4.0$, and $L_J = 6.0$. Treatment parameters were: $\mu_\beta = 8.0$, $S_E(0) = .86$, $S_J(0) = .76$, $S_A(0) = .66$, and all others identical to control. The treatment here represents some change in plant growth conditions, such as water or nutrient regime, induction of secondary chemical defenses, or different plant type. The assumed effect of the treatment is a shift in life history strategy toward higher fecundity at the cost of lower survival; insects often have a high range of phenotypic plasticity. These parameters lead to very rapid population growth, with 10s to 1,000s of eggs and 10s of adults by the end of the experiment, which can happen with real organisms. I considered the idealized situation that the experimenters use the correct model structure for analysis, including knowledge of the observation model (37). I used the null hypothesis that all parameters are equal between treatment and control, and the alternative that μ_β, a_E, a_J, and a_A may vary between treatment and control.

4.2 Implementation Issues

For the MCEM and MCLR methods, I used a block sampling Metropolis–Hastings algorithm to sample from $\Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i)$ for fixed $\boldsymbol{\Theta}$ (Liu, Wong, and Kong 1994; Roberts and Sahu 1997; Shephard and Pitt 1997; Liu and Sabatti 2000) with blocks of adjacent $\mathbf{z}$ values. I sampled separately from β, with log-normal proposals iterated five times and blocks of 10, 5, and 1 adjacent $\mathbf{z}$'s, with normal proposals. This sampling is cumbersome because, unlike in many state-space models, there is no guarantee of the existence, or an easy solution for, the noise that moves one state to another arbitrary state. This means that the state trajectory must be recalculated for each proposal from the earliest proposed change to the end of the trajectory. Simpler situations could have been devised for the examples here, but the goal was to maintain generality, with the most general case being recalculation of the entire state process for any change in $\boldsymbol{\nu}$. The sampler produced a well-mixed sample, recording the sample after every fifth full iteration.

A second issue for the MCEM and MCLR methods was how large the MCMC sample should be for each experimental unit for each optimization iteration. After some experimentation, I show results that start with $m = 1{,}000$ and then, after fast initial progress toward the MLE, use either $m = 1{,}000$ or $m = 5{,}000$. The performance of these algorithms has not been fully optimized, but the results suggest that further optimization is unwarranted in this case.

For the MCD method, there was a choice of whether to importance sample for β and/or $\mathbf{z}$. Because there was only a single β for each trajectory, this value was very important, and I used an importance density that was $N(\mu_\beta, \sigma = 1.1\sigma_\beta)$. The mapping from $\boldsymbol{\nu}$ (61-dimensional) to $\mathbf{X}(\boldsymbol{\nu})$ (12-dimensional) is roughly degenerate in the sense that many different $\boldsymbol{\nu}$ values can produce similar $\mathbf{X}(\boldsymbol{\nu})$. This diminishes the role of extreme $\mathbf{z}$ values, which appears to be why importance sampling for the $\mathbf{z}$'s offered little or no improvement in preliminary trials and was not used.

For the MCKL method, sampling from $\Pr(\boldsymbol{\Theta}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n | \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ was considerably more complicated than sampling each $\Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i)$ for fixed $\boldsymbol{\Theta}$ because of strong correlations between $\boldsymbol{\Theta}$ and each $\boldsymbol{\nu}_i$. I reparameterized by replacing a_E, a_J, and a_A with $S_E = S_E(0)$, $S_J = S_J(0)$, and $S_A = S_A(0)$. Metropolis–Hastings steps for joint parameter-process noise directions were developed based on biological interpretations. For example, S_E was negatively correlated with β_i because higher fecundity can compensate for lower egg survival to produce similar population trajectories. To take advantage of this, I used a Metropolis–Hastings proposal in the parameterization $(S_E, k_1, k_2, \ldots, k_n) = g(S_E, \beta_1, \beta_2, \ldots, \beta_n)$, with g defined by $k_i = \beta_i S_E^{L_E}$, which is the number of eggs surviving to become juveniles in the mean environment ($z_E = 0$), for $i = 1, \ldots, n$.

The transformed density is

$$\Pr(S_E, k_1, \ldots, k_n) = \frac{\Pr(S_E, \beta_1, \ldots, \beta_n)}{S_E^{nL_E}}, \tag{38}$$

where the denominator is the determinant of the Jacobian of $g(\cdot)$. Then for a proposal density $q(S'_E|S_E)$, adjusting the β_i's to keep the k_i's constant, the Metropolis–Hastings ratio in the original coordinates is

$$\frac{[\prod_{i=1}^n \Pr(\mathbf{Y}_i|\boldsymbol{\nu}'_i, \boldsymbol{\Theta}')] \Pr(S'_E)[\prod_{i=1}^n \Pr(\beta'_i|\boldsymbol{\Theta}')] S_E^{nL_E} q(S_E|S'_E)}{[\prod_{i=1}^n \Pr(\mathbf{Y}_i|\boldsymbol{\nu}_i, \boldsymbol{\Theta})] \Pr(S_E)[\prod_{i=1}^n \Pr(\beta_i|\boldsymbol{\Theta})] S_E'^{nL_E} q(S'_E|S_E)}, \tag{39}$$

where prime indicates a proposal value and model terms that cancel [e.g., $\Pr(\mathbf{z})$] have been omitted. This approach is similar to the generalized Gibbs steps of Liu and Sabatti (2000). The proposal density $q(S'_E|S_E)$ was a reflected normal distribution centered on S_E, with reflection points at 0 and 1, so $q(S_E|S'_E) = q(S'_E|S_E)$.

Other Metropolis–Hastings steps included sampling from S_J holding each $\beta_i S_J^{L_J}$ constant; from (S_E, S_J) holding each $\beta_i S_E^{L_E} S_J^{L_J}$ constant; from S_A holding each $\beta_i/(1 - S_A)$ (lifetime reproductive output of an adult in the mean environment) constant; from $L_E - L_J$ holding $L_E + L_J$ constant; from $L_E - L_J$ holding $L_E + L_J$ and each $\beta_i S_E^{L_E} S_J^{L_J}$ constant; from each parameter separately, in some cases log-transformed; and from process noises using the sampler described for the MCEM and MCLR methods. Samples were recorded after every 15 iterations to ensure good mixing. For the convergence study, MCMC sample sizes were 2,000, (increments of 2,000), ..., 8,000, (increments of 4,000), ..., 40,000.

For MCKL, I used a Gaussian kernel on the standardized principal components of the $\boldsymbol{\Theta}$ sample, with h chosen by solving (26) so that $\Pr(L(\hat{\boldsymbol{\Theta}})/L(\boldsymbol{\Theta}) > .95) = .95$. (A Jacobian adjustment would be required for a coordinate transformation after sampling and before kernel estimation, but the constant Jacobian of a linear transformation just introduces another scaling constant.) For the constrained ($d = 10$) parameter space, this gave h from .86 ($m = 2{,}000$) to 0.55 ($m = 40{,}000$) and from 1.03 ($m = 2{,}000$) to .67 ($m = 40{,}000$) for the unconstrained ($d = 14$) parameter space. Figure 2 shows posterior profile contours from the unconstrained posterior; the MCMC algorithms mixed well, and the approximately normal assumption for (26) seems reasonable. The prior was proper but virtually flat throughout the region of the posterior.

For all methods, $L(\hat{\boldsymbol{\Theta}}_\mathbf{h})$ was estimated by importance sampling. Normal approximations $\Pr_i^S(\boldsymbol{\nu}_i)$ to each $\Pr_S(\boldsymbol{\nu}_i|\mathbf{Y}_i)$ were estimated from MCMC samples, using the same state sampler as in MCEM and MCLR, and were used as the importance density in

$$l(\boldsymbol{\Theta}) \approx \sum_{i=1}^n \log\left[\frac{1}{m}\sum_{j=1}^m \Pr\bigl(\mathbf{Y}_i|\boldsymbol{\nu}_i^{(j)}\bigr)\frac{\Pr(\boldsymbol{\nu}_i^{(j)})}{\Pr_i^S(\boldsymbol{\nu}_i^{(j)})}\right],$$

$$\boldsymbol{\nu}_i^{(j)} \sim \Pr_i^S(\boldsymbol{\nu}_i), \tag{40}$$

with $m = 10{,}000$. MCMC samples for estimating each $\Pr_i^S$ had $m = 2{,}000$.

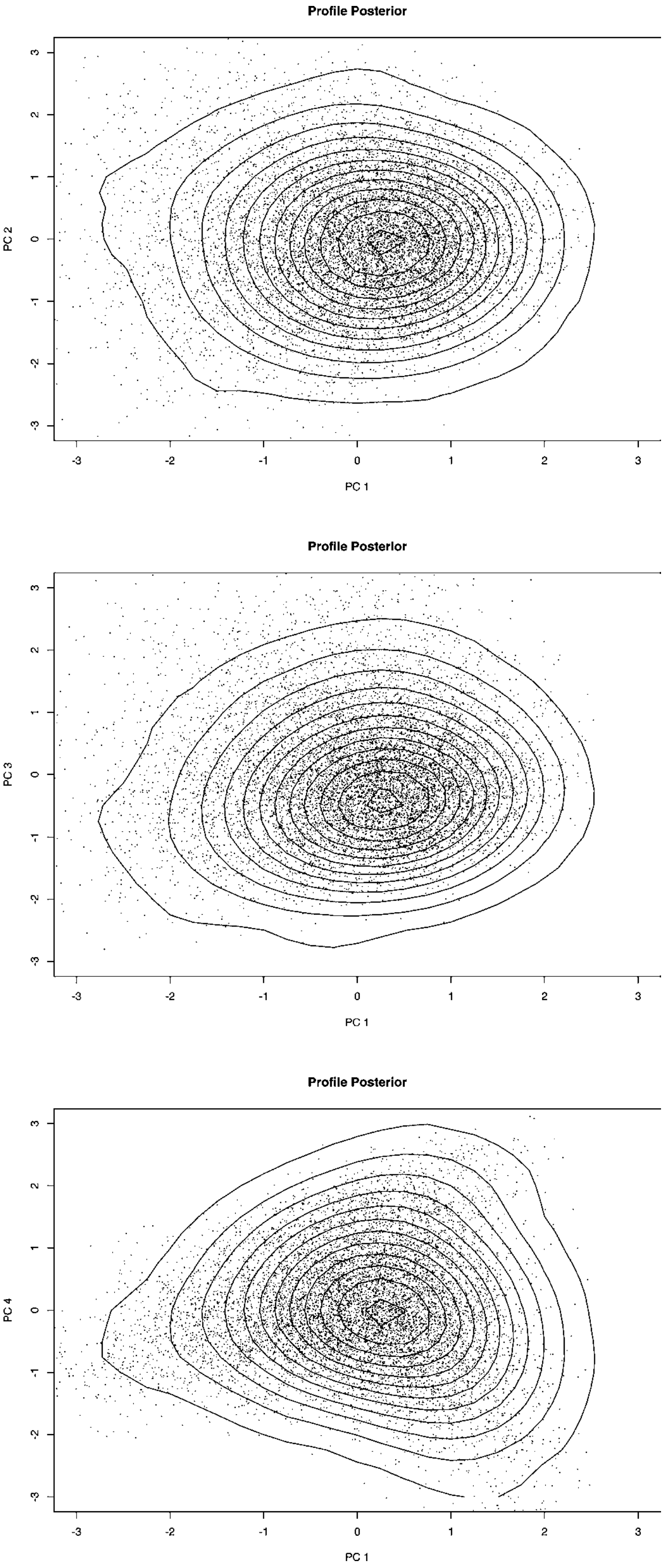

Figure 2. Posterior Profile Contours (maximized over all other dimensions) for Three Pairs of Standardized Principal Components of the Posterior Sample for the Unconstrained Parameter Space of the Simulated Population Experiment, Plotted Over a Thinned Marginal Sample.

For all methods, optimization was conducted using a standard Nelder–Mead simplex algorithm (Press, Teukolsky, Vetterling, and Flannery 1992). Restarting an optimization algorithm after first convergence is often a good safety step. I restarted the

MCKL optimization, which has almost no computational burden because the kernel likelihood calculations are fast once the sample is obtained. For the MCD method, I optimized with an initial sample of $m = 10{,}000$, followed by a restart with $m = 10{,}000, 20{,}000, 30{,}000$, or $40{,}000$. For PF, I used a sample size scheme such that for the unconstrained parameter space ($d = 10$), the total number of state trajectories calculated over all n replicates matched the totals for MCD for each m. For example, samples of (2,500, 1,250, 1,250, 1,250) for $[\mathbf{v}_i(10), \mathbf{v}_i(20), \mathbf{v}_i(30), \mathbf{v}_i(40)]$ give the same total trajectories as $m = 10{,}000$ in MCD. The larger samples for $\mathbf{v}_i(10)$ alleviate the PF difficulty of sample thinning and are computationally cheap, because the same sample can be used for all replicates up to the first observation.

4.3 Results

MCKL converged quickly to the maximum likelihood as MC sample size (and computation time) increased (Fig. 3). MCEM, MCLR, MCD, and PF all show quick initial movement toward the maximum likelihood, but then very slow subsequent convergence. The results shown in Figure 3 were produced by starting with (the same) parameter values moderately far away from the MLEs and are typical of other simulated datasets that were tried. For MCKL, the initial conditions have little effect because the MCMC sampler moves quickly around the parameter posterior. This appears to be a relatively unskewed likelihood surface (Fig. 2), so that even for the smallest sample size ($m = 2{,}000$) with relatively large kernel bandwidth, the MCKL estimate is quite good. For initial parameters even further away, MCEM and MCLR perform even worse. For some parameters close to the MLEs, MCEM and MCLR perform reasonably, but I have no results about how close is close enough or how one could assess in practice whether one of these methods has achieved a correct maximum. McCulloch (1997) suggested using MCEM first, followed by MCLR once parameters close to the MLE are known. My results do not rule out that MCLR may be useful near the MLE. Also note that for MCD and PF, the lines connect points in increasing order of MC sample size (10,000 to 40,000 for the restart sample) and illustrate that time to convergence can depend on the sample.

Comparison of MCD and PF shows that PF is more efficient for likelihood estimation at fixed $\mathbf{\Theta}$ but that MCD provides a smooth surface that allows optimization (Fig. 4). Hürzeler and Künsch (1998, 2001) used a loess smoother on a regular grid of PF likelihood evaluations in one and two dimensions, but this would be unwieldy for higher dimensions. Stavropoulos and Titterington (2001) smoothed the PF likelihood surface by resampling particles from a kernel density estimate of $\Pr(\mathbf{v}_{1:t|1:t-1})$. With this approach, the likelihood surface would be smooth with respect to a fixed sample of underlying random variables used for the simulations, but the filter densities at each time would be distorted by kernel smoothing.

I used MCKL (with $m = 10{,}000$) to fit 100 simulated experiments, and found that despite MC and smoothing errors, it offers dramatic improvement over ANOVA for detecting the treatment effect (de Valpine 2003). In all cases, -2 times the likelihood ratio was in the extreme tail of a χ^2_4 distribution [the minimum value was 115, compared with $P(\chi^2_4 > 33.4) = 10^{-6}$], strongly rejecting the null. In contrast, a t test with Welch correction for unequal variances on the log distributions of eggs, juveniles, or adults at the end of the experiment rejected the null in only 44, 17, and 2 out of 100 cases. ANOVA is the conventional analysis method in applied entomology, so the inaccuracies in MC state-space likelihoods are tiny relative to the dramatic improvement over common practice offered by using population models for hypothesis tests with population data. See de Valpine (2003) for more simulated power comparisons.

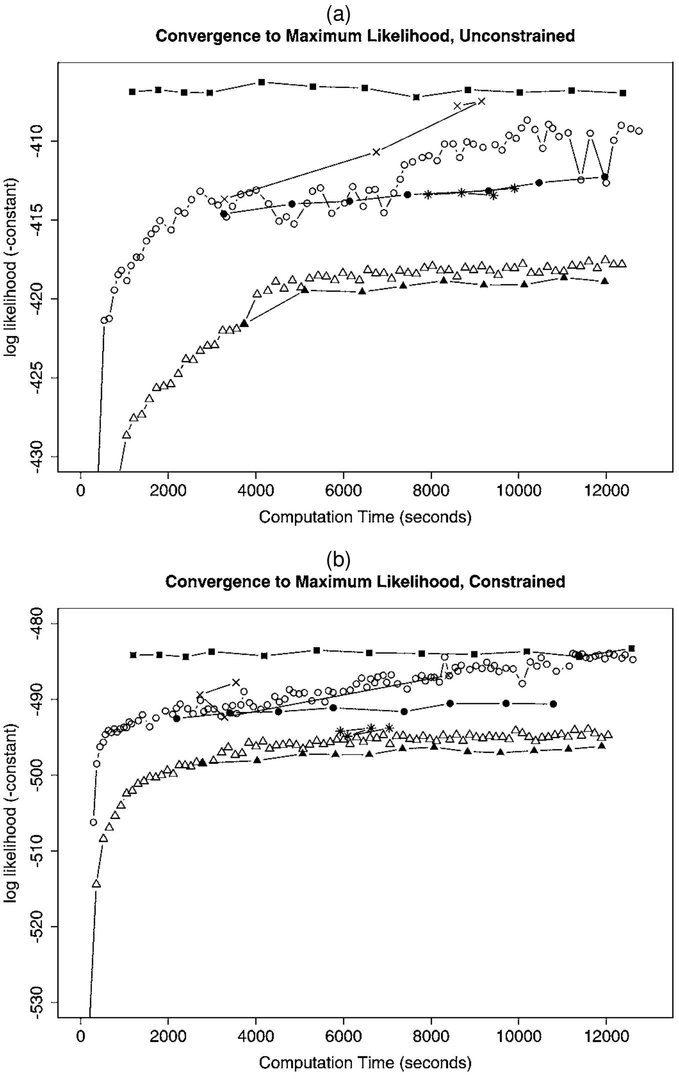

Figure 3. Convergence to Maximum Likelihood Estimates for Five MC Likelihood Approximations for (a) Unconstrained and (b) Constrained Parameter Spaces of the Population Model. Starting parameters for optimization were μ_β = true value + 5, a_E = true value − 1.0, σ^2_β = true value + 1.0, b_A = true value + .1, and the true value of all other parameters. For the constrained optimization, starting values were the average of control and treatment starting values. MCD and PF points are connected in order of MC sample size; maximization time depends on the sample, so a particular larger sample can happen to give faster maximization than a smaller one. For MCKL, the MCMC sampler starts with the same initial parameter values as the other methods but quickly samples throughout the posterior. Even for the smallest MCKL sample, m = 2,000, the MCKL provides a fairly good maximum likelihood estimate (■, MCKL; ×, MCD; ∗, PF; ○, MCLR 1K; •, MCLR 5K; △, MCEM 1K; ▲, MCEM 5K).

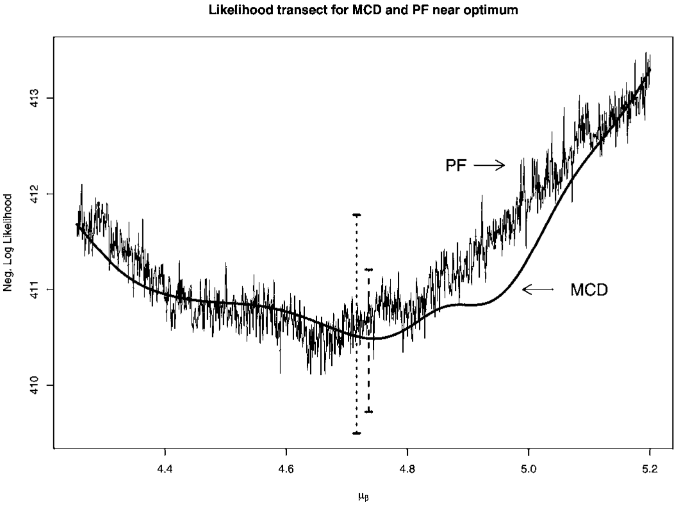

Figure 4. Profile of Log-Likelihood Estimates for MCD and PF Along a Discrete Transect of μ_β and a_E (values not shown) Near the MLE of the Population Model. A bootstrapped 95% confidence interval for a single parameter value is shown for each method, offset left (MCD) and right (PF) of the μ_β value.

5. EXAMPLE 2: ZOOMING AND CUMULANT CORRECTION

To investigate the accuracy of MCKL with zooming and cumulant-based corrections as dimensionality increases, I considered estimation of d-dimensional normal standard deviations from $n = 10$ d-dimensional unit normal data points. I assume that the data are independent but that the standard deviations must be estimated jointly. This an artificial but useful situation giving the conditions of interest: a d-dimensional likelihood surface that is skewed, so smoothing mode bias is substantial; easy posterior simulation; and easy calculation of true MLEs for comparison.

Define σ_l as the standard deviation and $\eta_l = 1/\sigma_l^2$ for each dimension l. Define each dimension of the prior as

$$\Pr(\sigma_l) \propto \left(\frac{1}{\sigma_l^2}\right)^{\alpha-1} e^{-r/\sigma_l^2}. \tag{41}$$

Because $\Pr(\sigma_l) = 2|\eta_l|^{1.5}\Pr(\eta_l)$, (41) can be simulated by $\eta_l \sim$ gamma with shape $\alpha - 1.5$ and rate r. Similarly, the posterior $\Pr_S(\sigma_l) = \Pr(\sigma_l|\mathbf{Y})$ can be simulated by $\eta_l|\mathbf{Y} \sim$ gamma with shape $\alpha|\mathbf{Y} = \alpha - 1.5 + .5n$, $r|\mathbf{Y} = r + .5\sum_{i=1}^{n}\mathbf{Y}_i^2$. I used $\alpha = 1$, $r = .1$ (mean = standard deviation = 10) for an uninformative prior.

I calculated the cumulant correction (24) from posterior cumulants. For a zooming correction, I used an estimate of $\Pr_S$, with $\alpha|\mathbf{Y}$ and $r|\mathbf{Y}$ estimated from the sample, as a zoomed prior $\Pr'$. For the unzoomed prior I calculated h_l, $l = 1, \ldots, d$, by solving (26) for $\Pr(L(\hat{\boldsymbol{\Theta}}_{\mathbf{h}})/L(\boldsymbol{\Theta}_0) > .99) = .9$ and scaling by the standard deviation of the sample. For the zoomed case, I estimated m_e' using the ratio of unzoomed to zoomed posterior densities, estimated by kernel estimates with bandwidth $.75 * \mathbf{h}$, with $\mathbf{h}$ from the unzoomed case. I then re-solved (26) with $m = m_e'$ to obtain a smaller $\mathbf{h}$ for the zoomed case. I also applied the cumulant correction from the unzoomed posterior moments to estimate $\Delta\boldsymbol{\Theta}_{\mathbf{h}}$ for the zoomed $\mathbf{h}$. I view these $\mathbf{h}$ choices as ad hoc but reasonable—the kind of practical decisions often made in other importance sampling and density estimation contexts, which invite more theoretical development but may be useful nonetheless.

Figure 5 shows distributions of the log-likelihood error, $\log(L(\hat{\boldsymbol{\Theta}}_{\mathbf{h}})/L(\boldsymbol{\Theta}_0))$, where $\boldsymbol{\Theta}_0$ is the true MLE, from 100 simulated datasets with $d = 5, 10, 15, 20$. For all d, I used $m =$ 40,000, which is excessive for small d but simplifies comparisons across d values. The right-hand axes show the p values that would result from the maximum likelihood errors if the true p value from a χ_d^2 test for $-2\log(L)$ is .05 and the MCKL estimate is for an unconstrained parameter space. The cumulant correction and zooming methods offer dramatic improvements, and in a real analysis zooming could be iterated and sample sizes increased for greater accuracy.

6. DISCUSSION

At least for my motivating class of problems, with replicated short state-space datasets, MCKL provides efficiency gains over other MC maximum likelihood methods that can facilitate practical use. MCKL's greatest strength may be in quickly locating approximate MLEs, and Example 2 shows that it can be very accurate. Nevertheless, further work is warranted on reducing smoothing mode bias or, if no smoothing bias is acceptable, perhaps switching to another method once MCKL has quickly located a neighborhood of the MLE. Further work on automated bandwidth selection and protocols for zooming would also facilitate the application of MCKL.

I have focused on relatively basic versions of each method, but all of the methods have the potential for improvements and combinations. For example, Levine and Casella (2001) considered an MCEM method where the sample approximating $\Pr(\boldsymbol{\nu}_i|\mathbf{Y}_i, \boldsymbol{\Theta}_0)$ can be obtained by reweighting the sample from the previous iteration, according to importance sampling principles, instead of running a new MCMC each time. A similar idea could be used for MCLR, although in both cases it improves only the repeated sampling aspect of their efficiency, not the repeated local maximization aspect.

Posterior sampling by means other than MCMC may improve MCKL implementation and efficiency. Gordon et al. (1993), Kitagawa (1998), and Kitagawa and Sato (2001) proposed sampling $\Pr(\boldsymbol{\Theta}, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n|\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ by particle filtering parameters and states jointly with artificial parameter "dynamics," such as $\boldsymbol{\Theta}(t+1) = \boldsymbol{\Theta} + \text{noise}$, to alleviate sample degradation. Liu and West (2001) addressed the sample degradation problem using West's (1993) approach of kernel smoothing the parameter dimensions (see also Liu and Chen 1998; and Berzuini, Best, Gilks, and Larizza 1997). Another approach by Gilks and Berzuini (2001) and Berzuini and Gilks (2001) is to use MCMC steps to mix particles after each filter step.

For the population model example, the correlations between parameter and process noise dimensions were problematic, and efficient samplers were obtained only by drawing on understanding of the model dynamics, which takes away from the generality of the approach. An alternative way to formulate the model would be with random variables for each transition between each day class. This would inflate the dimension of the process noise space, but would allow calculation of Metropolis–Hastings ratios without recalculating entire state trajectories

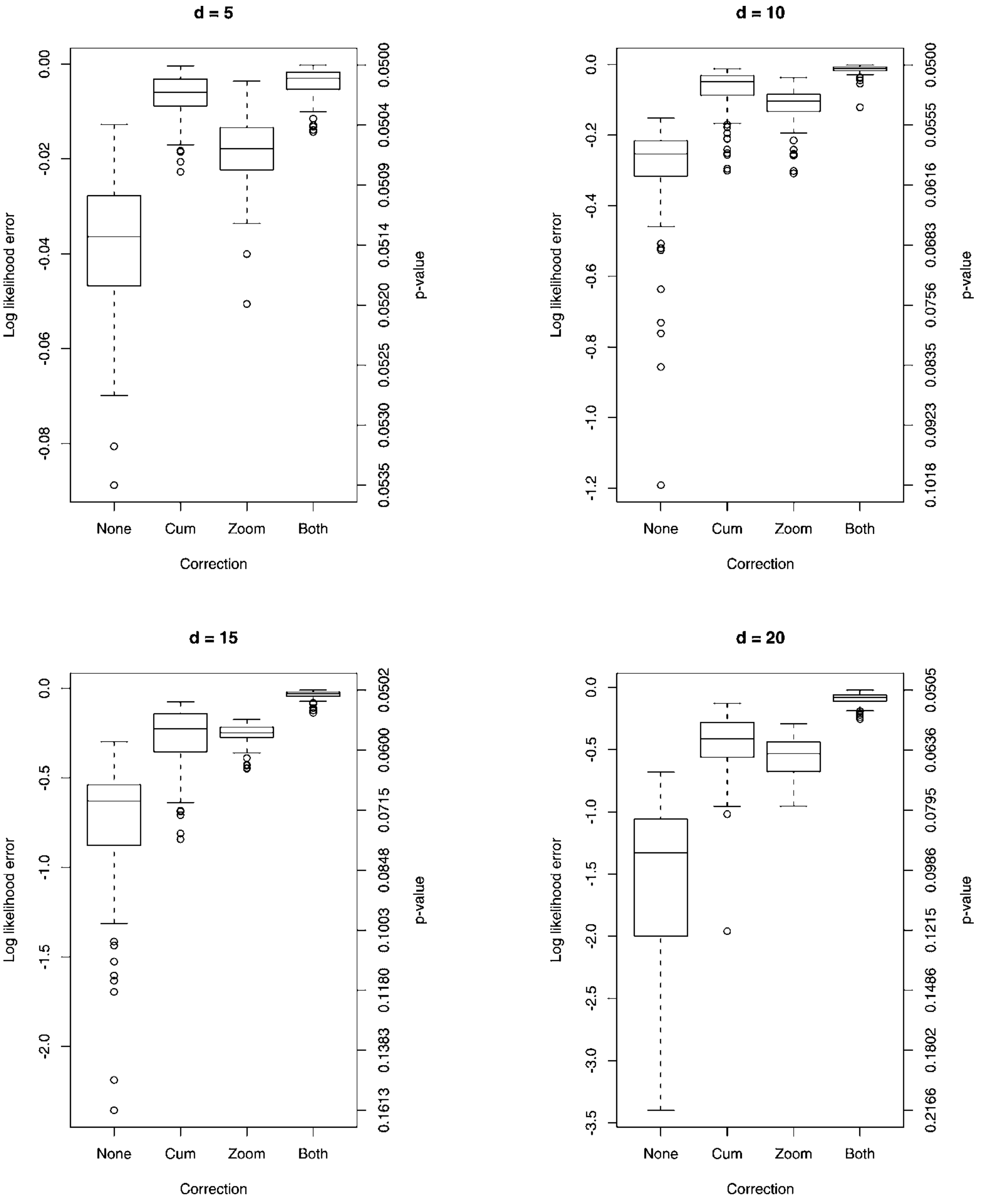

Figure 5. Distributions of Log Maximum Likelihood Error for Example 2. For dimensionalities $d = 5, 10, 15, 20$, box-and-whisker plots from 100 simulated datasets are plotted with no correction ("None"), cumulant correction ("Cum"), one iteration of zooming ("Zoom"), and zooming with cumulant correction ("Both"). Right-side axes show p values that would be estimated by a χ_d^2 test if the true p value is .05 (see text).

for each proposal. The primary difficulty would still be the parameter–noise or parameter–state correlations, and it is unclear whether this alternative formulation would have led to a more general approach to that difficulty.

Implementation issues aside, application of MC state-space likelihood ratio tests to population dynamics experiments offers the potential for significant insight into complex ecological dynamics. Many such experiments are conducted every year and are typically analyzed with ANOVA models that do not incorporate relevant biological processes, with a handful of exceptions that are in various ways specialized (e.g., Dennis, Desharnais, Cushing, and Costantino 1995; Dennis et al. 2001; Ives et al. 1999; Gibson, Gilligan, and Kleczkowski 1999; Bjørnstad, Sait, Stenseth, Thompson, and Begon 2001). State-space likelihood methods offer increased statistical power, closer connections between hypothesized processes and statistical analyses, and the potential to analyze novel kinds of experiments.

[Received January 2002. Revised September 2003.]

REFERENCES

Berzuini, C., Best, N. G., Gilks, W. R., and Larizza, C. (1997), "Dynamic Conditional Independence Models and Markov Chain Monte Carlo Methods," *Journal of the American Statistical Association*, 92, 1403–1412.

Berzuini, C., and Gilks, W. (2001), "RESAMPLE-MOVE Filtering With Cross-Model Jumps," in *Sequential Monte Carlo Methods in Practice*, eds. A. Doucet, N. de Freitas, and N. Gordon, New York: Springer, pp. 117–138.

Bjørnstad, O. N., Fromentin, J.-M., Stenseth, N. C., and Gjosaeter, J. (1999), "Cycles and Trends in Cod Populations," *Proceedings of the National Academy of Sciences USA*, 96, 5066–5071.

Bjørnstad, O. N., Sait, S. M., Stenseth, N. C., Thompson, D. J., and Begon, M. (2001), "The Impact of Specialized Enemies on the Dimensionality of Host Dynamics," *Nature*, 409, 1001–1006.

Booth, J. G., and Hobert, J. P. (1999), "Maximizing Generalized Linear Mixed Model Likelihoods With an Automated Monte Carlo EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 61, 265–285.

Carlin, B. P., Polson, N. G., and Stoffer., D. S. (1992), "A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling," *Journal of the American Statistical Association*, 87, 493–500.

Carpenter, S. R., and Kitchell, J. F. (1993), *The Trophic Cascade in Lakes*, Cambridge, U.K.: Cambridge University Press.

Caswell, H. (2001), *Matrix Population Models: Construction, Analysis, and Interpretation* (2nd ed.), Sunderland, MA: Sinauer Associates.

Chan, K., and Ledolter, J. (1995), "Monte Carlo EM Estimation for Time Series Models Involving Counts," *Journal of the American Statistical Association*, 90, 242–252.

Chen, M.-H. (1994), "Importance-Weighted Marginal Bayesian Posterior Density Estimation," *Journal of the American Statistical Association*, 89, 818–824.

Clayton, D. (1996), "Generalized Linear Mixed Models," in *Markov Chain Monte Carlo in Practice*, eds. W. Gilks, S. Richardson, and D. Spiegelhalter, Boca Raton, FL: Chapman & Hall/CRC, pp. 275–301.

Costantino, R. F., Desharnais, R. A., Cushing, J. M., and Dennis, B. (1997), "Chaotic Dynamics in an Insect Population," *Science*, 275, 389–391.

Dennis, B., Desharnais, R. A., Cushing, J. M., and Costantino, R. (1995), "Nonlinear Demographic Dynamics: Mathematical Models, Statistical Methods, and Biological Experiments," *Ecological Monographs*, 65, 261–281.

Dennis, B., Desharnais, R. A., Cushing, J. M., Henson, S. M., and Costantino, R. F. (2001), "Estimating Chaos and Complex Dynamics in an Insect Population," *Ecological Monographs*, 71, 277–303.

de Valpine, P. (2003), "Better Inferences From Population Dynamics Experiments Using Monte Carlo State-Space Likelihood Methods," *Ecology*, 84, 3064–3077.

Doucet, A., de Freitas, N., and Gordon, N. (2001a), "An Introduction to Sequential Monte Carlo Methods," in *Sequential Monte Carlo Methods in Practice*, eds. A. Doucet, N. de Freitas, and N. Gordon, New York: Springer-Verlag, pp. 3–14.

——— (eds.) (2001b), *Sequential Monte Carlo Methods in Practice*, New York: Springer-Verlag.

Downing, A. L., and Leibold, M. A. (2002), "Ecosystem Consequences of Species Richness and Composition in Pond Food Webs," *Nature*, 416, 837–839.

Durbin, J., and Koopman, S. J. (1997), "Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State-Space Models," *Biometrika*, 84, 669–684.

——— (2000), "Time Series Analysis of Non-Gaussian Observations Based on State-Space Models From Both Classical and Bayesian Perspectives," *Journal of the Royal Statistical Society*, Ser. B, 62, 3–56.

Durham, G. B., and Gallant, R. (2002), "Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes," *Journal of Business & Economic Statistics*, 20, 297–316.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Ellner, S. P., McCauley, E., Kendall, B. E., Briggs, C. J., Hosseini, P. R., Wood, S. N., Janssen, A., Sabelis, M. W., Turchin, P., Nisbet, R. M., and Murdoch, W. W. (2001), "Habitat Structure and Population Persistence in an Experimental Community," *Nature*, 412, 538–543.

Gause, G. (1934), *The Struggle for Existence*, Baltimore, MD: Williams & Wilkins.

Geyer, C. J. (1994), "Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo," Technical Report 568, University of Minnesota, School of Statistics.

——— (1996), "Estimation and Optimization of Functions," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman & Hall, pp. 241–258.

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data," *Journal of the Royal Statistical Society*, Ser. B, 54, 657–699.

Gibson, G., Gilligan, C., and Kleczkowski, A. (1999), "Predicting Variability in Biological Control of a Plant–Pathogen System Using Stochastic Models," *Proceedings of the Royal Society of London*, Ser. B, 266, 1743–1753.

Gilks, W. R., and Berzuini, C. (2001), "Following a Moving Target—Monte Carlo Inference for Dynamic Bayesian Models," *Journal of the Royal Statistical Society*, Ser. B, 63, 127–146.

Givens, G. H., and Raftery, A. E. (1996), "Local Adaptive Importance Sampling for Multivariate Densities With Strong Nonlinear Relationships," *Journal of the American Statistical Association*, 433, 132–141.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993), "Novel Approach to Nonlinear Non-Gaussian Bayesian State Estimation," *IEE Proceedings F*, 140, 107–113.

Grund, B., and Hall, P. (1995), "On the Minimisation of L^p Error in Mode Estimation," *The Annals of Statistics*, 23, 2264–2284.

Gurney, W., Blythe, S., and Nisbet, R. (1980), "Nicholson's Blowflies Revisited," *Nature*, 287, 17–21.

Gurney, W., and Nisbet, R. (1998), *Ecological Dynamics*, New York: Oxford University Press.

Gurney, W., Nisbet, R., and Lawton, J. (1983), "The Systematic Formulation of Tractable Single-Species Population Models Incorporating Age Structure," *Journal of Animal Ecology*, 52, 479–495.

Hairston, N. G., Sr. (1989), *Ecological Experiments: Purpose, Design, and Execution*, Cambridge, U.K.: Cambridge University Press.

Holyoak, M. (2000), "Effects of Nutrient Enrichment on Predator–Prey Metapopulation Dynamics," *Journal of Animal Ecology*, 69, 985–997.

Huffaker, C. B. (1958), "Experimental Studies on Predation: Dispersion Factors and Predator–Prey Oscillations," *Hilgardia*, 27, 343–383.

Hürzeler, M., and Künsch, H. R. (1998), "Monte Carlo Approximations for General State-Space Models," *Journal of Computational and Graphical Statistics*, 7, 175–193.

——— (2001), "Approximating and Maximising the Likelihood for a General State-Space Model," in *Sequential Monte Carlo Methods in Practice*, eds. A. Doucet, N. de Freitas, and N. Gordon, New York: Springer-Verlag, pp. 159–175.

Ives, A. R., Carpenter, S. R., and Dennis, B. (1999), "Community Interaction Webs and Zooplankton Responses to Planktivory Manipulations," *Ecology*, 80, 1405–1421.

Jones, M., and Signorini, D. (1997), "A Comparison of Higher-Order Bias Kernel Density Estimators," *Journal of the American Statistical Association*, 92, 1063–1073.

Karban, R., English-Loeb, G., and Hougen-Eitzman, D. (1997), "Mite Vaccinations for Sustainable Management of Spider Mites in Vineyards," *Ecological Applications*, 7, 183–193.

Kaunzinger, C. M. K., and Morin, P. J. (1998), "Productivity Controls Food-Chain Properties in Microbial Communities," *Nature*, 395, 495–497.

Kitagawa, G. (1996), "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State-Space Models," *Journal of Computational and Graphical Statistics*, 5, 1–25.

——— (1998), "A Self-Organizing State-Space Model," *Journal of the American Statistical Association*, 93, 1203–1215.

Kitagawa, G., and Sato, S. (2001), "Monte Carlo Smoothing and Self-Organizing State-Space Model," in *Sequential Monte Carlo Methods in Practice*, eds. A. Doucet, N. de Freitas, and N. Gordon, New York: Springer, pp. 177–195.

Klug, J. L., Fischer, J., Ives, A., and Dennis, B. (2000), "Compensatory Dynamics in Planktonic Community Responses to pH Perturbations," *Ecology*, 81, 387–398.

Kunsch, H. R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241.

Levine, R. A., and Casella, G. (2001), "Implementations of the Monte Carlo EM Algorithm," *Journal of Computational and Graphical Statistics*, 10, 422–439.

Liu, J., and West, M. (2001), "Combined Parameter and State Estimation in Simulation-Based Filtering," in *Sequential Monte Carlo Methods in Practice*, eds. A. Doucet, N. de Freitas, and N. Gordon, New York: Springer-Verlag, pp. 197–223.

Liu, J. S., and Chen, R. (1998), "Sequential Monte Carlo Methods for Dynamic Systems," *Journal of the American Statistical Association*, 93, 1032–1044.

Liu, J. S., and Sabatti, C. (2000), "Generalised Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation," *Biometrika*, 87, 353–369.

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

Luckinbill, L. S. (1973), "Coexistence in Laboratory Populations of Paramecium Aurelia and Its Predator Didinium Nasutum," *Ecology*, 54, 1320–1327.

McCauley, E., Nisbet, R. M., Murdoch, W. W., de Roos, A. M., and Gurney, W. S. C. (1999), "Large-Amplitude Cycles of Daphnia and Its Algal Prey in Enriched Environments," *Nature*, 402, 653–656.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170.

Metz, J., and Diekmann, O. (eds.) (1986), *The Dynamics of Physiologically Structured Populations*, Berlin: Springer-Verlag.

Meyer, R., and Millar, R. B. (1999), "Bayesian Stock Assessment Using a State-Space Implementation of the Delay Difference Model," *Canadian Journal of Fisheries and Aquatic Sciences*, 56, 37–52.

Mignani, S., and Rosa, R. (1995), "The Moving Block Bootstrap to Assess the Accuracy of Statistical Estimates of Ising Model Simulations," *Computer Physics Communications*, 92, 203–213.

Millar, R. B., and Meyer, R. (2000), "Non-Linear State Space Modelling of Fisheries Biomass Dynamics by Using Metropolis–Hastings Within-Gibbs Sampling," *Applied Statistics*, 49, 327–342.

Murdoch, W., Nisbet, R. M., McCauley, E., de Roos, A. M., and Gurney, W. S. C. (1998), "Plankton Abundance and Dynamics Across Nutrient Levels: Tests of Hypotheses," *Ecology*, 79, 1339–1356.

Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H., and Woodfin, R. M. (1994), "Declining Biodiversity Can Alter the Performance of Ecosystems," *Nature*, 368, 734–737.

Nicholson, A., and Bailey, V. (1935), "The Balance of Animal Populations, Part I," *Proceedings of the Zoological Society, London*, 3, 551–598.

Pace, M. L., Cole, J. J., Carpenter, S. R., and Kitchell, J. F. (1999), "Trophic Cascades Revealed in Diverse Ecosystems," *Trends in Ecology and Evolution*, 14, 483–488.

Pitt, M. K., and Shephard, N. (1999), "Filtering via Simulation: Auxiliary Particle Filters," *Journal of the American Statistical Association*, 94, 590–599.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in C: The Art of Scientific Computing* (2nd ed.), Cambridge, U.K.: Cambridge University Press.

Resetarits, W. J., and Bernardo, J. (eds.) (1998), *Experimental Ecology: Issues and Perspectives*, New York: Oxford University Press.

Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Roberts, G. O., and Sahu, S. (1997), "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler," *Journal of the Royal Statistical Society*, Ser. B, 59, 291–317.

Romano, J. (1988), "On Weak Convergence and Optimality of Kernel Density Estimates of the Mode," *The Annals of Statistics*, 16, 629–647.

Rosenheim, J. A. (2001), "Source-Sink Dynamics for a Generalist Insect Predator in Habitats With Strong Higher-Order Predation," *Ecological Monographs*, 71, 93–116.

Rosenheim, J. A., Kaya, H. K., Ehler, L. E., Marois, J. J., and Jaffee, B. A. (1995), "Intraguild Predation Among Biological Control Agents: Theory and Evidence," *Biological Control*, 5, 303–335.

Scott, D. (1992), *Multivariate Density Estimation*, New York: Wiley.

Severini, T. A. (2000), *Likelihood Methods in Statistics*, New York: Oxford University Press.

Shephard, N., and Pitt, M. K. (1997), "Likelihood Analysis of Non-Gaussian Measurement Time Series," *Biometrika*, 84, 653–667.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman & Hall.

Snyder, W. E., and Ives, A. R. (2001), "Generalist Predators Disrupt Biological Control by a Specialist Parasitoid," *Ecology*, 82, 705–716.

Stavropoulos, P., and Titterington, D. M. (2001), "Improved Particle Filters and Smoothing," in *Sequential Monte Carlo Methods in Practice*, eds. A. Doucet, N. de Freitas, and N. Gordon, New York: Springer-Verlag, pp. 295–317.

Strong, D. R., Whipple, A. V., Child, A. L., and Dennis, B. (1999), "Model Selection for a Subterranean Trophic Cascade: Root-Feeding Caterpillars and Entomopathogenic Nematodes," *Ecology*, 80, 2750–2761.

Stuart, A., and Ord, J. K. (1994), *Kendall's Advanced Theory of Statistics: Distribution Theory* (6th ed.), Vol. 1, London: Edward Arnold.

Tanizaki, H., and Mariano, R. S. (1998), "Nonlinear and Non-Gaussian State Space Modeling With Monte Carlo Simulations," *Journal of Econometrics*, 83, 263–290.

Tuljapurkar, S., and Caswell, H. (eds.) (1997), *Structured-Population Models in Marine, Terrestrial, and Freshwater Systems*, New York: Chapman & Hall.

Underwood, A. (1997), *Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance*, Cambridge, U.K.: Cambridge University Press.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

Wei, G. C. G., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85, 699–704.

West, M. (1993), "Approximating Posterior Distributions by Mixture," *Journal of the Royal Statistical Society*, Ser. B, 55, 409–422.

Wootton, J. T. (1994), "The Nature and Consequences of Indirect Effects in Ecological Communities," *Annual Review of Ecology and Systematics*, 25, 443–466.