Fast Bayesian reconstruction of chaotic dynamical systems via extended Kalman filtering

Renate Meyer[1,*] and Nelson Christensen[2,†]
[1]*Department of Statistics, The University of Auckland, Auckland, New Zealand*
[2]*Physics and Astronomy, Carleton College, Northfield, Minnesota, 55057*
(Received 19 July 2001; revised manuscript received 17 September 2001; published 14 December 2001)

We present an improved Markov chain Monte Carlo (MCMC) algorithm for posterior computation in chaotic dynamical systems. Recent Bayesian approaches to estimate the parameters of chaotic maps have used the Gibbs sampler which exhibits slow convergence due to high posterior correlations. Using the extended Kalman filter to compute the likelihood function by integrating out all unknown system states, we obtain a very efficient MCMC technique. We compare the new algorithm to the Gibbs sampler using the logistic, the tent, and the Moran-Ricker maps as applications, measuring the performance in terms of CPU and integrated autocorrelation time.

DOI: 10.1103/PhysRevE.65.016206 PACS number(s): 05.45.Tp, 02.70.Rr, 06.20.Dk, 02.60.Pn

I. INTRODUCTION

In the physical sciences, experimental data often show an irregular, complicated, and ostensibly random time dependence. This led to the use of chaotic dynamical processes in order to explain and model the observed irregularities [1–4]. In this paper we address the problem of reconstructing the nonlinear dynamic equations assumed to be underlying an observed noisy time series. These observations can stem from laboratory experiments in the physical sciences or ''real world'' systems.

Previous work on nonlinear noise reduction from a dynamical systems perspective uses probabilistic models to account for uncertainties in the measurements [5–8]. It is generally assumed that the observations, y_i, are conditionally independent random variables given unknown system states x_i, $i=1,\ldots,N$. The time evolution of the systems states is determined by a parametric nonlinear function $x_i = f(x_{i-1},a)$ that depends on the previous state and an unknown p-dimensional parameter vector a. Least-squares methods [9,10] to estimate the unknown model parameters that minimize the sum of squared one-step prediction errors, systematically under- or overestimate the parameters because they do not take into account that the values of the ''independent'' variable are subject to measurement errors. Total least-squares methods [11,12], introduced by Kostelich [13] to reduce this so-called errors-in-variables bias of LS, suffer from so-called time-series bias since they ignore the serial correlation between successive observations. As shown in [14], both errors-in-variables bias and time-series bias can be eliminated by allowing for stochastic errors in the dynamics, thus casting the problem into the framework of nonlinear state-space modeling. As shown for instance in [14] and [15], the Bayesian approach [16] to parameter estimation can quantify both process and observation errors through the posterior distribution of the model parameters and difficulties with Bayesian posterior computation can be overcome using computer-intensive Markov chain Monte Carlo (MCMC) methods [17].

The Gibbs sampler is used in [14] to generate a sample from the joint posterior distribution of unknown parameters *and* unknown system states. However, due to the temporal dependencies between consecutive states, there are high posterior correlations that cause the Markov chain to traverse the state space in only very tiny steps and thus to mix inefficiently. Therefore, convergence of the Markov chain to the equilibrium distribution is slow, a large number of iterations are required to achieve a satisfactory precision of parameter estimates, and the estimation procedure becomes very time consuming. A far more efficient MCMC method can be developed by first integrating out the unknown states. This reduces the problem of sampling vectors in a high $(N+p)$-dimensional space to that of sampling in a low (p)-dimensional space. If the state transitions were *linear*, this integration could be performed using the Kalman filter [18,19]. Due to the nonlinear chaotic dynamics, however, this is not feasible here. Thus we suggest an approach that combines the extended Kalman filter [20] with the Laplace approximation [21]. The extended Kalman filter (EKF) has been developed for nonlinear non-Gaussian state-space models whereas the Laplace approximation has a long tradition in Bayesian computation as an asymptotic approximation to the posterior distribution [22]. The proposed technique is not restricted to Gaussian errors but can also be applied to make models robust by allowing for outlying observations through heavy-tailed error distributions. This yields an extremely effective and fast MCMC technique that provides a unified, practical likelihood-based framework for the analysis of nonlinear dynamical systems.

The outline of the paper is as follows. In Sec. II we describe the theory underlying the calculation of the likelihood function via extended Kalman filtering and Laplace approximation. MCMC techniques to sample from the posterior distribution are detailed in Sec. III. In Sec. IV we illustrate the new technique using the logistic, the tent, and the Moran-Ricker maps. Its performance is compared to that of the Gibbs sampler. We measure performance in terms of CPU time, integrated autocorrelation time, and a variety of other diagnostic measures. We conclude in Sec. V with a discussion on the efficiency of this approach.

*Email address: meyer@stat.auckland.ac.nz
†Email address: nchriste@carleton.edu

II. EXTENDED KALMAN FILTERING FOR NONLINEAR STATE-SPACE MODELS

Following the notation of [14], we model the noisy observations y_i, $i=1,\ldots,N$, of a time series as being conditionally independent Gaussian random variables given unobserved sufficient true states x_i, i.e.,

$$y_i|x_i = x_i + v_i, \quad v_i \overset{\text{iid}}{\sim} N(0,\epsilon^2), \quad i=1,\ldots,N, \tag{1}$$

with known error variance ϵ^2 and where iid denotes independent and identically distributed. The time evolution of the system states is itself assumed to be Markovian,

$$x_i|x_{i-1},a = f(x_{i-1},a) + u_i, \quad u_i \overset{\text{iid}}{\sim} N(0,\tau^2),$$
$$i=1,\ldots,N, \tag{2}$$

where $f(x_{i-1},a)$ is a nonlinear function of x_{i-1}, a is a p-dimensional parameter, and x_0 a starting value. For ease of notation, we assume that the observations as well as the states are one dimensional, but it is straightforward to extend this to the d-dimensional case.

Here, the focus is on estimating the unknown parameters $\boldsymbol{\theta}=(a,\tau^2,x_0)$ given the observations y_i, with the parameter (vector) a that defines the nonlinear function being the main parameter (vector) of interest. A fully Bayesian approach specifies the joint distribution of all observables $[\mathbf{y}=(y_1,\ldots,y_N)]$ and parameters $[\boldsymbol{\theta}=(a,\tau^2,x_0)]$. The joint probability density function (PDF) $p(\mathbf{y},\boldsymbol{\theta})$ can be factorized into the product of the PDF of parameters, $p(\boldsymbol{\theta})$, referred to as the *prior* PDF, and the conditional PDF of the observations given the parameters, $p(\mathbf{y}|\boldsymbol{\theta})$, referred to as the sampling distribution or *likelihood*, i.e., $p(\boldsymbol{\theta},\mathbf{y})=p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. The prior PDF contains all pre-experimental information about the parameters stemming from substantive knowledge and expert opinion. After observing the data, prior knowledge about the parameters, as quantified through the *prior* PDF of $\boldsymbol{\theta}$, is updated to the *posterior* PDF, $p(\boldsymbol{\theta}|\mathbf{y})$, via the Bayes theorem:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{3}$$

where $p(\mathbf{y})=\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal PDF of $\mathbf{y}$. Due to the conditioning on unobserved states in a state-space model, the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed form but requires N-dimensional integration over the state vector $\mathbf{x}=(x_1,\ldots,x_N)$ as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\boldsymbol{\theta},\mathbf{x})p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}. \tag{4}$$

Taking the temporal structure of the observations into account, we can factorize the likelihood by successive conditioning into

$$p(\mathbf{y}|\boldsymbol{\theta}) = p(y_1|\boldsymbol{\theta})\prod_{i=2}^{N} p(y_i|\mathbf{y}_{i-1},\boldsymbol{\theta}), \tag{5}$$

where $\mathbf{y}_{i-1}=(y_1,\ldots,y_{i-1})$ collects all the observable information obtained up until time $i-1$. Thus, the N-dimensional integration in Eq. (4) can be reduced to N successive 1-dimensional integrations, starting with

$$p(y_1|\boldsymbol{\theta}) = \int p(y_1|x_1,\boldsymbol{\theta})p(x_1|\boldsymbol{\theta})dx_1 \tag{6}$$

and, subsequently, for $i=2,\ldots,N$:

$$p(y_i|\mathbf{y}_{i-1}) = \int p(y_i|x_i,\boldsymbol{\theta})p(x_i|\mathbf{y}_{i-1},\boldsymbol{\theta})dx_i. \tag{7}$$

This also implies that the data can be processed in a single sweep, updating knowledge about states as we receive more information. For instance, in the light of just the first observation y_1, we update the prior $p(x_1|\boldsymbol{\theta})$ of the unknown state x_1 to the *filtering* PDF via Bayes theorem

$$p(x_1|y_1,\boldsymbol{\theta}) = \frac{p(y_1|x_1,\boldsymbol{\theta})p(x_1|\boldsymbol{\theta})}{p(y_1|\boldsymbol{\theta})} \tag{8}$$

[where the denominator is just the first factor in the likelihood decomposition in Eq. (4), given in Eq. (6)]. As both likelihood, $p(y_1|x_1,\boldsymbol{\theta})$, and prior, $p(x_1|\boldsymbol{\theta})$, are Gaussian $N(x_1,\epsilon^2)$ and $N(f(x_0,a),\tau^2)$ PDF's, respectively, the posterior filtering PDF $p(x_1|y_1,\boldsymbol{\theta})$ is again Gaussian with mean $\hat{x}_1$ and variance $\hat{\sigma}_1^2$ given by

$$\hat{x}_1 = \frac{\tau^2}{\epsilon^2+\tau^2}y_1 + \frac{\epsilon^2}{\epsilon^2+\tau^2}f(x_0,a), \tag{9}$$

$$\hat{\sigma}_1^2 = \left(\frac{1}{\epsilon^2}+\frac{1}{\tau^2}\right)^{-1}, \tag{10}$$

respectively. Furthermore, the denominator in Eq. (8) is

$$p(y_1|\boldsymbol{\theta}) = \int p(y_1|x_1,\boldsymbol{\theta})p(x_1|\boldsymbol{\theta})dx_1 \tag{11}$$

$$= \sqrt{2\pi}\mathrm{e}^{-\psi_1(y_1,\hat{x}_1)}|D^2\psi_1(y_1,\hat{x}_1)|^{-1/2}, \tag{12}$$

where

$$\psi_1(y_1,x_1) = -\log[p(y_1|x_1,\boldsymbol{\theta})p(x_1|\boldsymbol{\theta})]$$
$$= \frac{1}{2}\log(2\pi\epsilon^2) + \frac{1}{2\epsilon^2}(y_1-x_1)^2$$
$$+ \frac{1}{2}\log(2\pi\tau^2) + \frac{1}{2\tau^2}[x_1-f(x_0,a)]^2,$$

and $D^2\psi_1(y_1,x_1)$ denotes the second-order derivative of the function $\psi_1(y_1,x_1)$ with respect to x_1. Note, that $\hat{x}_1 = \mathrm{argmin}_{x_1}\psi_1(y_1,x_1)$ and $\hat{\sigma}_1^2=|D^2\psi_1(y_1,\hat{x}_1)|^{-1}$. If either

likelihood $[p(y_1|x_1,\boldsymbol{\theta})]$ or prior $[p(x_1|\boldsymbol{\theta})]$ were not Gaussian, the identity (12) would become an approximation to the integral (11), the so-called *Laplace approximation*, an asymptotic approximation of the posterior distribution that dates back to the work of Laplace in the eighteenth century [21,22]. This is easily seen by a second-order Taylor series expansion of $\psi_1(y_1,x_1)$ at $\hat{x}_1=\mathrm{argmin}_{x_1}\,\psi_1(y_1,x_1)$.

We now learn about a state at time i, successively for $i=2,\ldots,N$, given comtemporaneously available information. This is done repeatedly in a two-stage procedure by on-line extended Kalman filtering. In the first stage of the extended Kalman filter, after observing $\mathbf{y}_{i-1}$ but before observing y_i, the *predictive* PDF of $x_i|\mathbf{y}_{i-1},\boldsymbol{\theta}$ is approximated by a normal PDF $\tilde{p}(x_i|\mathbf{y}_{i-1},\boldsymbol{\theta})$ with mean and variance given by

$$\beta_i=f(\hat{x}_{i-1},a) \tag{13}$$

and

$$\gamma_i^2=[f'(\hat{x}_{i-1},a)]^2\hat{\sigma}_{i-1}^2+\tau^2, \tag{14}$$

respectively, using a first-order Taylor series expansion of $f(x_{i-1},a)$ at the mean $\hat{x}_{i-1}$ of $x_{i-1}|\mathbf{y}_{i-1},\boldsymbol{\theta}$. Here, $f'(x,a)$ denotes the first derivative of $f(x,a)$ with respect to x. In the second stage, after observing y_i, the *filtering* PDF $p(x_i|\mathbf{y}_i,\boldsymbol{\theta})$ is updated via Bayes theorem to

$$\begin{aligned} p(x_i|\mathbf{y}_i,\boldsymbol{\theta}) &\propto p(y_i|x_i,\boldsymbol{\theta})p(x_i|\mathbf{y}_{i-1},\boldsymbol{\theta}) \\ &\approx p(y_i|x_i,\boldsymbol{\theta})\tilde{p}(x_i|\mathbf{y}_{i-1},\boldsymbol{\theta}) \end{aligned} \tag{15}$$

and approximated by a normal distribution with mean and variance given by

$$\hat{x}_i=\frac{\gamma_i^2}{\epsilon^2+\gamma_i^2}y_i+\frac{\epsilon^2}{\epsilon^2+\gamma_i^2}\beta_i, \tag{16}$$

$$\hat{\sigma}_i^2=\left(\frac{1}{\epsilon^2}+\frac{1}{\gamma_i^2}\right)^{-1}. \tag{17}$$

Using the Laplace approximation then yields an approximation to the ith likelihood contribution in Eq. (7)

$$\begin{aligned} p(y_i|\mathbf{y}_{i-1},\boldsymbol{\theta}) &= \int p(y_i|x_i,\boldsymbol{\theta})p(x_i|\mathbf{y}_{i-1},\boldsymbol{\theta})dx_i \\ &\approx \int p(y_i|x_i,\boldsymbol{\theta})\tilde{p}(x_i|\mathbf{y}_{i-1},\boldsymbol{\theta})dx_i \\ &= \sqrt{2\pi}\mathrm{e}^{-\psi_i(y_i,\hat{x}_i)}|D^2\psi_i(y_i,\hat{x}_i)|^{-1/2}, \end{aligned} \tag{18}$$

where

$$\begin{aligned} \psi_i(y_i,x_i) &= -\log[p(y_i|x_i,\boldsymbol{\theta})\tilde{p}(x_i|\mathbf{y}_{i-1},\boldsymbol{\theta})] \\ &= \frac{1}{2}\log(2\pi\epsilon^2)+\frac{1}{2\epsilon^2}(y_i-x_i)^2 \\ &\quad+\frac{1}{2}\log(2\pi\gamma_i^2)+\frac{1}{2\gamma_i^2}(x_i-\beta_i)^2. \end{aligned}$$

Note that

$$\hat{x}_i=\mathrm{argmin}_{x_i}\,\psi_i(y_i,x_i), \tag{19}$$

$$\hat{\sigma}_i^2=|D^2\psi_i(y_i,\hat{x}_i)|^{-1}. \tag{20}$$

Completion of this sequential two-stage procedure yields a closed-form approximate expression for the likelihood of Eq. (5) that no longer depends on the latent system states $\mathbf{x}$. More precisely, this likelihood is given by

$$\begin{aligned} \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) = \exp\Bigg\{ &-\frac{N}{2}\log(2\pi\epsilon^2)-\frac{1}{2\epsilon^2}\sum_{i=1}^{N}(y_i-\hat{x}_i)^2 \\ &-\frac{1}{2}\sum_{i=1}^{N}\log(2\pi\gamma_i^2)-\sum_{i=1}^{N}\frac{1}{2\gamma_i^2}(\hat{x}_i-\beta_i)^2 \\ &-\sum_{i=1}^{N}\log\left(\frac{1}{\hat{\sigma}_i^2}\right)\Bigg\} \end{aligned} \tag{21}$$

with $\beta_1=f(x_0,a)$ and $\gamma_1^2=\tau^2$. From Eq. (3) we then obtain the posterior PDF up to normalization constant:

$$\tilde{p}(\boldsymbol{\theta}|\mathbf{y})\propto p(\boldsymbol{\theta})\tilde{p}(\mathbf{y}|\boldsymbol{\theta}). \tag{22}$$

III. METROPOLIS-HASTINGS ALGORITHM

Various techniques are feasible to obtain a sample from the posterior (22), e.g., importance resampling and MCMC algorithms. We suggest the Metropolis-Hastings (MH) algorithm, developed by Metropolis *et al.* [23] and generalized by Hastings [24]. It is a MCMC method which means that it generates a Markov chain whose equilibrium distribution is just the target posterior distribution. The MH algorithm shares the concept of a generating PDF with the well-known simulation technique of *rejection sampling* [22]. However, the *candidate generating PDF* $q(\boldsymbol{\theta}|\boldsymbol{\theta}_c)$, $\int q(\boldsymbol{\theta}|\boldsymbol{\theta}_c)d\boldsymbol{\theta}=1$, can now depend on the current state $\boldsymbol{\theta}_c$ of the sampling process. A new candidate $\boldsymbol{\theta}^*$ is accepted with a certain *acceptance probability* $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}_c)$ also depending on the current state $\boldsymbol{\theta}_c$, and chosen such that the transition probability $p(\boldsymbol{\theta}_c,\boldsymbol{\theta}^*)=q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_c)\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}_c)$ satisfies detailed balance. This is met by setting

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}_c)=\min\left\{\frac{\tilde{p}(\boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}_c|\boldsymbol{\theta}^*)}{\tilde{p}(\boldsymbol{\theta}_c|\mathbf{y})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_c)},1\right\}.$$

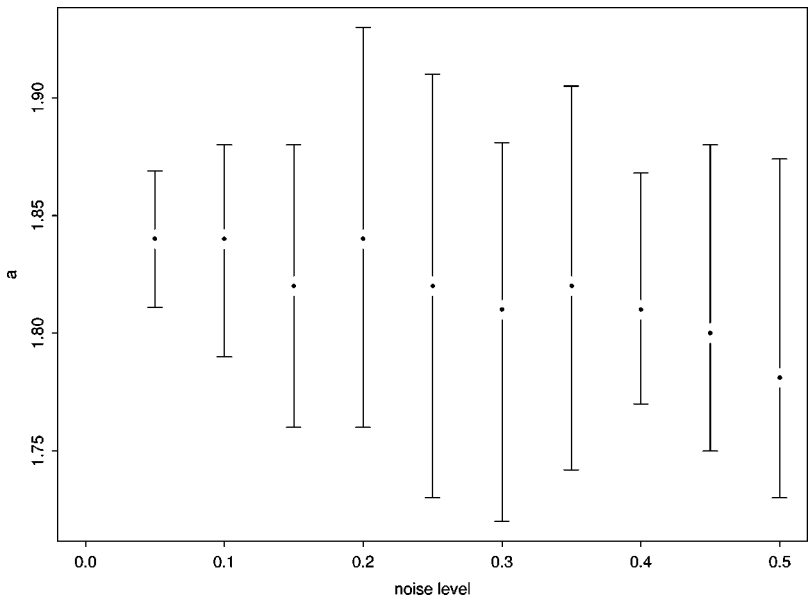

FIG. 1. Posterior means and 95% posterior probability intervals for increasing noise levels obtained using the EKF, based on 100 observations from the logistic map with true parameters $a=1.85$ and $x_0=0.3$.

The steps of the MH algorithm are therefore

(1) Step 0: Start with an arbitrary value $\boldsymbol{\theta}_0$.

(2) Step $k+1$: Generate $\boldsymbol{\theta}^*$ from $q(\cdot|\boldsymbol{\theta}_k)$ and u from $U(0,1)$. If $u \leqslant \alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}_k)$ set $\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}^*$ (acceptance). If $u > \alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}_k)$ set $\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}_k$ (rejection).

Note that the MH algorithm does not require the normalization constant of the target PDF. The outcomes from the MH algorithm can be regarded as a sample from the invariant PDF only after a certain ''burn-in'' period. A menu-driven collection of SPLUS functions, CODA [25], is available for analyzing the samples obtained from MCMC. Besides trace plots and convergence diagnostics based on [26], CODA calculates statistical summaries of the posterior distributions and kernel density estimates. CODA can be downloaded from site in Ref. [35].

The efficiency of the MH algorithm depends crucially on the choice of the proposal PDF. Similar to rejection sampling, the efficiency can be improved by choosing a proposal that is ''close'' to the posterior PDF. Once more, we make use of the Laplace approximation to $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$ to determine a good proposal PDF. This means that we use a multivariate normal PDF with mean $\boldsymbol{\mu}$ equal to the posterior mode, and covariance matrix $\boldsymbol{\Sigma}$ equal to the inverse of the Hessian matrix of the log posterior, i.e., defining

$$\phi(\boldsymbol{\theta}) = -\log[p(\boldsymbol{\theta})\tilde{p}(\mathbf{y}|\boldsymbol{\theta})],$$

the mean and covariance matrix are

$$\boldsymbol{\mu} = \mathrm{argmin}_{\theta}\, \phi(\boldsymbol{\theta}),$$

$$\boldsymbol{\Sigma} = |D^2\phi(\boldsymbol{\mu})|^{-1}.$$

The covariance matrix is dynamically scaled until a reasonable acceptance rate in the MH algorithm is observed.

Thus, to determine the multivariate normal proposal PDF, we need to find the posterior mode, or alternatively minimize $\phi(\boldsymbol{\theta})$. To this end, we employ the Newton-Raphson algorithm [27], and make use of automatic differentiation [28] to calculate the first- and second-order partial derivatives of $\phi(\boldsymbol{\theta})$. This can be done to the same degree of accuracy as the function evaluation itself. We use automatic differentiation implemented in a C^{++} class library which combines an array language with the reverse mode of automatic differentiation supplemented with precompiled adjoint code for the derivatives of common array and matrix operations [29].

IV. EXAMPLES

A. Logistic map

In order to compare results to those in [5] and [14], we simulated $N=100$ observations from Eq. (1) and underlying system evolution given by the logistic map $x_i = 1 - a x_{i-1}^2$ with true parameters $a=1.85$, $x_0=0.3$, and noise levels $l = \sigma_{noise}/\sigma_{signal}$ ranging from 0.05 to 0.5. Assuming *a priori* independence of the parameters a, x_0, and τ^2, we specified a prior uniform distribution for a on [0,4], a uniform distribution for x_0 on [0,1], and a diffuse inverse-gamma distribution for τ^2 with mean 0.005 and standard deviation 0.05. Combining this with the likelihood calculated by the EKF in Eq. (21), we performed 6000 MCMC iterations using the MH algorithm as described in Sec. III. We discarded the first 1000 observations as a burn-in period so that estimates are based on a final sample size of 5000. These will be referred to as Bayesian EKF estimates in the sequel.

Figure 1 displays the posterior means of the parameter a together with 95% credibility intervals for varying degrees of noise levels. A comparison with Fig. 2 of [14] shows an equivalent precision of the Bayesian EKF estimates compared to the one in [14] using the Gibbs sampler implementation in BUGS [30]. BUGS (Bayesian inference using Gibbs sampling) is a software package for Bayesian posterior simulation using the Gibbs sampler. It is freely available and can be downloaded from the site of Ref. [35]. Note that parameter estimates in [14] were based on 100 000 iterations of the Gibbs sampler after a burn-in period of 10 000. This large sample size was necessary because of the slow convergence of the single-update Gibbs sampler. However, the Gibbs sampler seems to handle larger signal-to-noise ratios better.

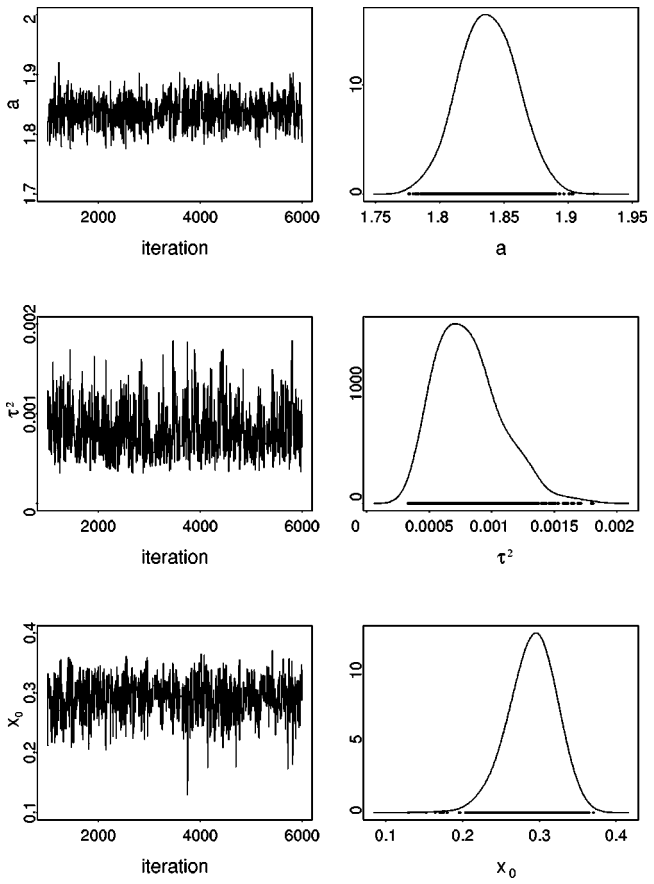

FIG. 2. Trace (left) and kernel density (right) plots of the marginal posterior distributions generated by EKF for the parameters a, τ^2, and x_0 based on 100 observations from the logistic map with true parameters $a=1.85$, $x_0=0.3$, and noise level 0.1.

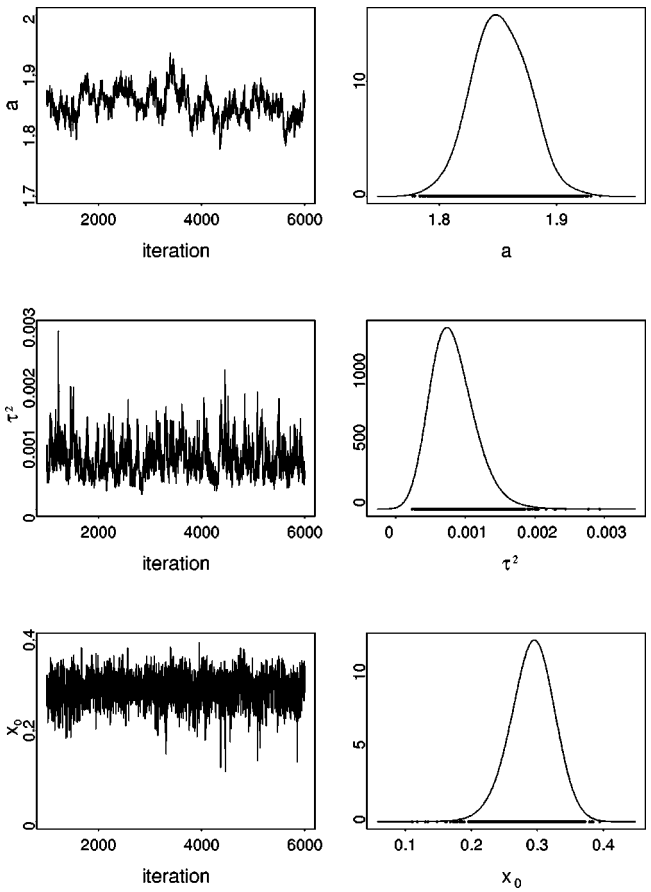

FIG. 3. Trace (left) and kernel density (right) plots of the marginal posterior distributions generated by BUGS for the parameters a, τ^2, and x_0 based on 100 observations from the logistic map with true parameters $a=1.85$, $x_0=0.3$, and noise level 0.1.

For a more detailed comparison of the efficiency of the EKF to that of the Gibbs sampler, we selected the simulated time series for noise level 0.1 and performed 6000 iterations of the Gibbs sampler as described in [14] using WINBUGS, the BUGS version for the WINDOWS operating system. Again, the first 1000 iterations were discarded. Figure 2 and Fig. 3 display trace plots and kernel density estimates of the three parameters a, τ^2, and x_0 based on 5000 iterations of the extended Kalman filter and the Gibbs sampler, respectively. We base the comparison on a variety of convergence diagnostics detailed in the sequel.

The Markov chain generated by the EKF passed the Heidelberger and Welsh [31] stationarity and halfwidth test, but the Gibbs sampling chain failed, indicating that the number of iterations needs to be increased by an order of magnitude to achieve convergence to the stationary distribution. However, 5000 iterations are sufficient for the EKF.

We used the Raftery and Lewis [32] convergence diagnostic to provide a sample size estimate needed to achieve a certain accuracy of estimated quantiles of parameters. For instance, to obtain an estimate of the 2.5th quantile of the parameter a to an accuracy of ± 0.01 with a probability of 0.9, one would need a minimum of $n=4764$ iterations of the EKF but a minimum of $n=28\,932$ iterations of the Gibbs sampler. If one could generate *independent* samples, only $n_{min}=660$ values would be required. Thus, the so-called *dependence factor* $I=n/n_{min}$ that measures the increase in the number of iterations needed to reach convergence due to dependence between the samples in the Markov chain equals 7.2 for the EKF but 43.8 for the Gibbs sampler.

Figure 4 and Fig. 5 display the correlograms, i.e., the graphs of the autocorrelation functions within each chain for each of the three parameters. The autocorrelation function of a time series x_t, $t=1,\ldots,N$, is a function of the time distances or *lags* $\tau=0,1,\ldots,N$, defined by $c(\tau)=\Sigma_t(x_t-\overline{x})(x_{t+\tau}-\overline{x})/\Sigma_t(x_t-\overline{x})^2$. High autocorrelations indicate slow mixing which will be reflected by plots of sample traces which ''snake'' slowly up and down, as opposed to showing more rapid fluctuations over the sample space. Such a feature can be clearly discerned from Fig. 3 for the Gibbs sampling chain. Also, while the lag 50 autocorrelation for the Gibbs sampling chain for parameter a is still 0.566, it is merely 0.0163 for the EKF chain.

The integrated autocorrelation time (IACT or τ_c) [33], also referred to as ''autocovariance time,'' ''autocorrelation time,'' and ''inefficiency factor,'' is the number of correlated samples with the same variance-reducing power as one independent sample. This is seen as follows: the estimate of the posterior mean of a parameter x is the average of n *corre-*

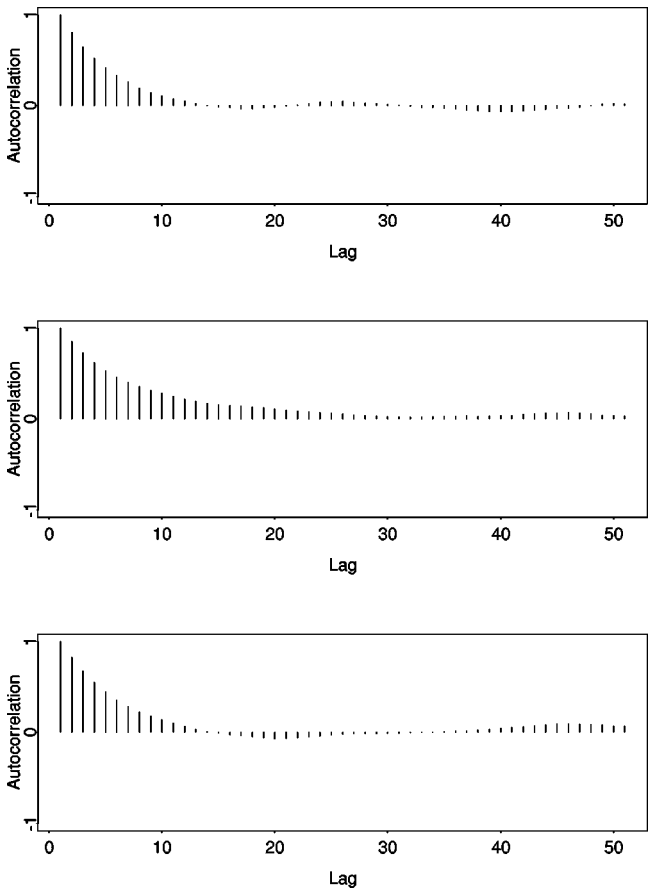

FIG. 4. Correlograms generated by EKF for the parameters a, τ^2, and x_0 based on 100 observations from the logistic map with true parameters $a=1.85$, $x_0=0.3$, and noise level 0.1.

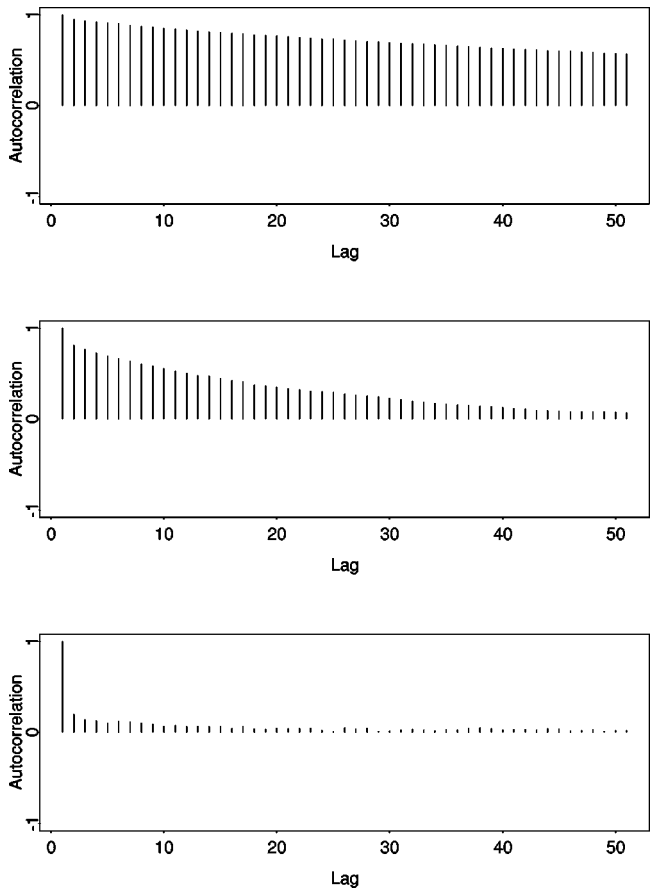

FIG. 5. Correlograms generated by BUGS for the parameters a, τ^2, and x_0 based on 100 observations from the logistic map with true parameters $a=1.85$, $x_0=0.3$, and noise level 0.1.

lated samples from a Markov chain, its variance is a factor of IACT larger than the variance of the sample mean based on the same number of independent samples, i.e.,

$$\mathrm{var}(\overline{x}_{MC}) = \tau_c \frac{\mathrm{var}(x)}{n}.$$

A reasonable estimate of the TIAC can be obtained by dividing the estimated squared Monte Carlo standard error (MCSE) of $\overline{x}$ by the the estimated standard deviation and multiplying by the sample size (here, $n=5\,000$). We calculated the Monte Carlo standard error by Geweke's [34] method, often referred to as ''numerical standard error'' or ''time-series standard error'' which is based on estimating the spectral density.

Table I compares the posterior means, time series standard errors, posterior standard deviations of the parameters, integrated autocorrelation times, and CPU time of the EKF with the Gibbs sampling chain. All computations were performed on a Pentium III, 700-MHz PC.

The computational efficiency of an algorithm is determined principally by its autocorrelation time. If one wishes to compare two alternative MCMC algorithms, the better is the one with smaller IACT. For parameter a, the IACT is a factor of almost 5 higher for the Gibbs sampler than for EKF. Of course, in practice there may arise a tradeoff between ''physical'' autocorrelation time (i.e., IACT measured in *iterations*) and computational complexity *per iteration*. But even here, the CPU time for EKF is a third lower than CPU time for the Gibbs sampler.

More striking is the difference in efficiency if estimation is based on 1000 instead of 100 observations. The Gibbs sampler now has to sample from the full conditional posterior distributions of 1003 instead of 103 parameters. This causes an increase in CPU time by a factor of ≈ 10 as seen in Table II. Table II compares the results obtained from using the extended Kalman filter with those using the Gibbs sampler as implemented in WINBUGS on the basis of 1000 observations from the logistic map for a noise level of 0.1. Whereas the IACT for parameter a for the EKF increases only marginally from 6.5 to 7.3 and even decreases from 8.9 to 7.5 for τ^2, it almost doubles and triples for the Gibbs sampling chain for a and τ^2, respectively. The CPU time for EKF also increases but only to 30 s instead of 78 s for the Gibbs sampler.

Overall, all convergence diagnostics demonstrate a much improved efficiency of the EKF over the Gibbs sampler.

B. Other Maps

We also simulated observations with other underlying system evolutions. Here we report on results obtained for the

TABLE I. Comparison of Bayesian estimates obtained using the extended Kalman filter with those using the Gibbs sampler (WINBUGS) based on 100 observations from the logistic map for noise level 0.1.

	BUGS				EKF			
	Mean	MCSE	IACT	SD	Mean	MCSE	IACT	SD
a	1.84	1.55×10^{-3}	29.2	2.03×10^{-2}	1.84	7.74×10^{-4}	6.5	2.15×10^{-2}
τ^2	8.21×10^{-4}	1.43×10^{-5}	13.4	2.77×10^{-4}	8.13×10^{-4}	1.09×10^{-5}	8.9	2.58×10^{-4}
x_0	2.93×10^{-1}	5.85×10^{-4}	1.8	3.06×10^{-2}	2.90×10^{-1}	1.08×10^{-3}	6.8	2.92×10^{-2}
Time (s)		6				4		

tent and Moran-Ricker map. Since our results are consistent with those from the logistic map we will only briefly summarize the results.

The so-called ''tent map'' has much in common with the logistic map [8]. We simulated $N=100$ observations from Eq. (1) and underlying system evolution given by

$$x_i=\begin{cases} ax_{i-1}, & \text{if } 0\leqslant x_{i-1}<0.5, \\ a(1-x_{i-1}), & \text{if } 0.5\leqslant x_{i-1}\leqslant 1 \end{cases}$$

with true parameters $x_0=0.25$, $a=2$, and noise level $l=\sigma_{noise}/\sigma_{signal}=0.05$. Using the same prior as in the previous example, we performed 6000 MCMC iterations using the EKF and BUGS. Both techniques estimated the parameters with reasonably high accuracy. Our results demonstrate that the Gibbs sampler requires a longer burn-in period. It has not reached equilibrium until about 3000 iterations. Again, autocorrelations are much lower for the EKF than the Gibbs sampler. The lag 50 autocorrelation for the main parameter of interest a, for instance, is 0.0528 for the EKF as opposed to 0.641 for the Gibbs sampler. This is also reflected in the IACT of 6.9 for EKF versus 12.4 for the Gibbs sampler.

A similar gain in efficiency could be observed when the EKF was applied to the Moran-Ricker map. Again, we observed that the Gibbs sampler seems to cope better with higher (>0.5) signal-to-noise ratios. We simulated $N=100$ observations from Eq. (1) and underlying system evolution given by the Moran-Ricker map $x_i=x_{i-1}\exp[a(1-x_{i-1})]$ with true parameters $x_0=0.5$, $a=3.7$, and noise level $l=\sigma_{noise}/\sigma_{signal}=0.1$.

Again, both EKF and Gibbs sampler are capable of accurately estimating the unknown parameters. Our results reveal that the Gibbs sampler requires a much larger burn-in period and has not reached equilibrium as quickly as the EFK chain. The correlograms reveal a higher efficiency of the EKF. The lag 50 autocorrelation for the main parameter of interest a, for instance, is 0.0072 for the EKF as opposed to 0.867 for the Gibbs sampler. In this example, the IACT for the parameter a is 8.1 for the EKF compared to 17.9 for the Gibbs sampler.

V. DISCUSSION

The single-update Gibbs sampler for posterior computation in nonlinear chaotic state-space models [14] is very simple to implement, reliable, and easily generalizable. However, it can have poor convergence properties that prompted us to develop a more efficient sampler that exploits the time-series structure of the model.

The Gibbs sampler samples from the *joint* posterior distribution of the states *and* the model parameters. Slow convergence of the Gibbs sampler is due to high posterior correlations between the unknown system states that cause the Markov chain to make only tiny moves from one iteration to the next. Thus, it will take a long time to traverse the whole state space. By integrating out the latent states and sampling from the *marginal* posterior distribution of the parameters of interest (a,τ^2,x_0) the dimension of the problem can be reduced enormously. Instead of sampling from the joint $N+p$-dimensional posterior PDF, one only needs to sample from the marginal p dimensional posterior PDF. Thus, the gain in efficiency is even greater the larger the sample size N, as demonstrated for the logistic map in Sec. IV.

For the integration, we make use of the time-series structure. Although nonlinearity in the state equation prohibits the use of the Kalman filter for sequential integration, a version of the extended Kalman filter in combination with a Laplace approximation performs well. The simulations conducted in this paper show that our proposed method can achieve significant efficiency gains over the Gibbs sampler for estimating the parameters of nonlinear chaotic dynamics.

The formulas in the paper are not restricted to the Gauss-

TABLE II. Comparison of Bayesian estimates obtained using the extended Kalman filter with those using the Gibbs sampler (WINBUGS) based on 1000 observations from the logistic map for noise level 0.1.

	BUGS				EKF			
	Mean	MCSE	IACT	SD	Mean	MCSE	IACT	SD
a	1.84	4.28×10^{-5}	41.2	4.71×10^{-3}	1.83	1.99×10^{-5}	7.3	5.22×10^{-3}
τ^2	2.31×10^{-4}	3.06×10^{-6}	33.5	3.74×10^{-5}	2.51×10^{-4}	1.69×10^{-6}	7.5	4.35×10^{-5}
x_0	2.97×10^{-1}	3.27×10^{-4}	2.2	1.55×10^{-2}	2.97×10^{-1}	5.72×10^{-5}	7.1	1.51×10^{-2}
Time (s)		78				30		

ian error distribution but can be applied to other distributions. One important extension would be the *t* distribution or a mixture of a normal and a *t* distribution (with low degrees of freedom) to allow for outliers in the observations. In this case, the defining Eq. (16) and Eq. (17) would have to be substituted by Eq. (19) and Eq. (20), respectively. The required minimum could be obtained using the Newton-Raphson algorithm. Using a heavy-tailed error distribution has the potential of making the parameter estimates robust with respect to crude measurement errors. The benefits of such models still have to be explored and this will be a topic for further research.

ACKNOWLEDGMENTS

This work was supported by the Royal Society of New Zealand Marsden Fund, the University of Auckland Research Committee, National Science Foundation, Grant No. PHY-0071327, and Carleton College.

[1] P. Collet and J.-P. Eckmann, *Iterated Maps on the Interval as Dynamical Systems* (Birkhäuser, Berlin, 1980).
[2] J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).
[3] D. Ruelle, *Chaotic Evolution and Strange Attractors* (Cambridge University Press, New York, 1989).
[4] R. L. Devaney, *Introduction to Chaotic Dynamical Systems* (Benjamin-Cummings, Menlo Park, CA, 1989).
[5] P. E. McSharry and L. A. Smith, Phys. Rev. Lett. **83**, 4285 (1999).
[6] E. J. Kostelich and J. A. Yorke, Phys. Rev. A **38**, 1649 (1988).
[7] S. M. Hammel, Phys. Lett. A **148**, 421 (1990).
[8] M. Berliner, J. Am. Stat. Assoc. **86**, 938 (1991).
[9] P. Grassberger, T. Schreiber, and C. Schaffrath, Int. J. Bifurcation Chaos Appl. Sci. Eng. **1**, 521 (1991).
[10] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. Sh. Tsimring, Rev. Mod. Phys. **65**, 1331 (1993).
[11] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem* (SIAM, Philadelphia, 1991).
[12] E. J. Kostelich and T. Schreiber, Phys. Rev. E **48**, 1752 (1993).
[13] E. J. Kostelich, Physica D **58**, 138 (1992).
[14] R. Meyer and N. L. Christensen, Phys. Rev. E **62**, 3535 (2000).
[15] M. E. Davies, Chaos **8**, 775 (1998).
[16] B. P. Carlin and T. A. Louis, *Bayes and Emperical Bayes Methods for Data Analysis* (Chapman and Hall, London, 1996).
[17] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman and Hall, London, 1996).
[18] R. E. Kalman, J. Basic Eng. **82**, 34 (1960).
[19] R. J. Meinhold and N. D. Singpurwalla, Am. Stat. **37**, 123 (1983).
[20] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter* (Cambridge University, Cambridge, 1990).
[21] P. S. Laplace, Stat. Sci. **1**, 364 (1986).
[22] D. Gamerman, *Markov Chain Monte Carlo, Stochastic Simulation for Bayesian Inference* (Chapman & Hall, London, 1997).
[23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[24] W. K. Hastings, Biometrika **57**, 97 (1970).
[25] N. G. Best, M. K. Cowles, and S. K. Vines, *CODA Manual Version 0.30* (MRC Biostatistics Unit, Cambridge, 1995).
[26] M. K. Cowles and B. P. Carlin, J. Am. Stat. Assoc. **91**, 883 (1996).
[27] L. Fahrmeir and G. Tutz, *Multivariate Statistical Modelling Based on Generalized Linear Models* (Springer-Verlag, New York, 1994).
[28] A. Griewank and G. F. Corliss, *Automatic Differentiation of Algorithms: Theory, Implementation, and Application* (SIAM, Philadelphia, 1991).
[29] D. Fournier, *AD Model Builder, Version 5.0.1*. (Otter Research Ltd, Canada, 2000).
[30] D. J. Spiegelhalter, A. Thomas, N. Best, and W. R. Gilks, *BUGS 0.5, Bayesian Inference using Gibbs Sampling, Manual* (Version ii) (MRC Biostatistics Unit, Cambridge, 1996).
[31] P. Heidelberger and P. Welch, Oper. Res. **31**, 1109 (1983).
[32] A. L. Raftery and S. Lewis, Stat. Sci. **7**, 493 (1992).
[33] A. D. Sokal, in Lectures at the Cargese Summer School on ''Functional Integration: Basics and Applications,'' 1996.
[34] J. Geweke, in *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, edited by J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F M. Smith (Oxford Univerity Press, 1992).
[35] http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml