

# Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century

Steven L. SCOTT

---

Markov chain Monte Carlo (MCMC) sampling strategies can be used to simulate hidden Markov model (HMM) parameters from their posterior distribution given observed data. Some MCMC methods used in practice (for computing likelihood, conditional probabilities of hidden states, and the most likely sequence of states) can be improved by incorporating established recursive algorithms. The most important of these is a set of forward-backward recursions calculating conditional distributions of the hidden states given observed data and model parameters. I show how to use the recursive algorithms in an MCMC context and demonstrate mathematical and empirical results showing a Gibbs sampler using the forward-backward recursions mixes more rapidly than another sampler often used for HMMs. I introduce an augmented variables technique for obtaining unique state labels in HMMs and finite mixture models. I show how recursive computing allows the statistically efficient use of MCMC output when estimating the hidden states. I directly calculate the posterior distribution of the hidden chain's state-space size by MCMC, circumventing asymptotic arguments underlying the Bayesian information criterion, which is shown to be inappropriate for a frequently analyzed dataset in the HMM literature. The use of log-likelihood for assessing MCMC convergence is illustrated, and posterior predictive checks are used to investigate application specific questions of model adequacy.

KEY WORDS: Forward-backward recursion; Gibbs sampler; Kalman filter; Local computation; Markov chain Monte Carlo.

---

A hidden Markov model (HMM) is a mixture model whose mixing distribution is a finite state Markov chain. HMMs have been successfully applied to problems in a variety of fields, including signal processing (Juang and Rabiner 1991; Andrieu and Doucet 2000), biology (Fredkin and Rice 1992; Leroux and Puterman 1992), genetics (Churchill 1989; Liu, Neuwald, and Lawrence 1999), ecology (Guttorp 1995), image analysis (Romberg, Choi, and Baraniuk 2000), economics (Hamilton 1989, 1990; Albert and Chib 1993), and network security (Scott 1999, 2001). The book by MacDonald and Zucchini (1997) illustrates several applications of HMMs and provides an extended bibliography.

For much of their history, HMMs have been implemented using recursive algorithms developed for parameter estimation (Baum, Petrie, Soules, and Weiss 1970) and for restoring the hidden Markov chain (Viterbi 1967). Today's Markov chain Monte Carlo (MCMC) techniques allow researchers to implement HMMs without the traditional recursive algorithms, which are viewed as "black boxes" by many statisticians. Contrary to this perception, the recursions have intuitive probabilistic interpretations and can improve many MCMC methods. Through the use of recursive computing, HMM users can take advantage of rapidly mixing MCMC algorithms, more efficiently use MCMC output to estimate the hidden chain, and use model selection techniques and convergence diagnostics that are otherwise unavailable. The algorithms also appeal to an aesthetic that says that relying on simulation to produce directly computable answers is time-consuming, inelegant, needlessly inaccurate, and generally not a good idea. This article reviews the most common recursive algorithms for HMMs and explains their role in a modern MCMC environment.

The article is structured around important tasks facing the Bayesian modeler: obtaining the posterior distribution of model parameters, estimating the hidden Markov chain, deter-

mining the size of the hidden chain's state space, and using diagnostics to assess model validity and MCMC convergence. Section 1 provides background on HMMs, including two closely linked recursive procedures for evaluating the HMM likelihood (the *likelihood recursion*) and the posterior distribution of each hidden state given observed data and model parameters (the *forward-backward recursion*). Section 2 discusses methods for sampling HMM parameters from their posterior distribution given observed data, with particular emphasis on two Gibbs samplers: the forward-backward Gibbs sampler (FB), which capitalizes on recursive techniques from Section 1, and the direct Gibbs sampler (DG), which samples each state in the hidden Markov chain given the most recent draws of its neighbors. FB requires more computer time per iteration than DG, but mixes more rapidly and produces useful quantities as byproducts. Section 2 also discusses a label-switching issue that HMMs share with finite mixture models. Section 3 describes Bayes and empirical Bayes methods for estimating the hidden Markov chain. Section 4 considers methods for determining  $S$ , the size of the hidden chain's state space. Section 5 uses the likelihood recursion to diagnose MCMC convergence and illustrates the use of posterior predictive checks to determine a model's capacity to produce specific features seen in the data. Section 6 provides a concluding summary.

## 1. BACKGROUND

Hidden Markov models assume that the distribution of an observed data point  $d_t, t = 1, \dots, n$ , depends on an unobserved (hidden) state  $h_t \in \mathcal{S} = \{0, \dots, S-1\}$ . The elements of  $\mathbf{h} = (h_1, \dots, h_n)$  follow a Markov chain with stationary transition matrix  $\mathbf{Q} = (q(r, s))$  and initial distribution  $\pi_0$ , often taken to be the stationary distribution of  $\mathbf{Q}$ . Formally,

$$p(h_t | h_1^{t-1}, \mathbf{Q}) = q(h_{t-1}, h_t), \quad (1)$$

---

Steven L. Scott is Assistant Professor of Statistics, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: [sls@usc.edu](mailto:sls@usc.edu)). The author thanks Art Dempster, John Rolph, and Padhraic Smyth for helpful conversations, and the editors and referees for suggestions improving the article.

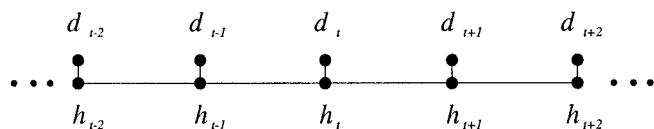


Figure 1. Graphical Depiction of HMM Dependencies. The conditional distribution of the value at any node, given values at all other nodes, depends only on nodes to which it is connected by an edge.

where  $h_{1:t-1} = (h_1, \dots, h_{t-1})$ . HMMs do not require that  $\mathbf{h}$  exist in a physical sense, but the model is more compelling if the hidden states have a physical interpretation. Later sections illustrate how scientific insight about  $\mathbf{h}$  can simplify otherwise difficult problems, such as choosing  $S$ , preventing label switching, and selecting model diagnostics.

The full conditional distribution of  $d_t$  is

$$p(d_t | d_{-t}, \mathbf{h}, \theta, \mathbf{Q}, \pi_0) = P_{h_t}(d_t | \theta), \quad (2)$$

where  $d_{-t} = \{d_{t'} : t' \neq t\}$  and  $\theta$  is a parameter vector for the probability distributions  $P_0, \dots, P_{S-1}$ . Sometimes  $\mathbf{Q}$  is a function of  $\theta$ , but it is usually treated as a separate parameter. In words, (2) says that  $d_t$  is conditionally independent of all other missing and observed data given  $h_t$  and model parameters, a concept illustrated in Figure 1.

HMMs are closely related to other well-known classes of models. Finite mixture models (Everitt and Hand 1981; Titterton, Smith, and Makov 1985) are HMMs in which all rows of  $\mathbf{Q}$  are equal. State-space models (West and Harrison 1997) assume that  $\mathbf{h}$  follows a Gaussian process. The Markov modulated Poisson process (Scott 1999, 2001) extends HMMs to continuous time. For clarity of exposition, we restrict attention to the model defined by (1) and (2) and note generalizations where appropriate. The following examples illustrate the notation and indicate some of the ways that (1) and (2) can be usefully expanded.

**Example 1 (Fetal Lamb Movements).** A popular dataset in the HMM literature is the record of fetal lamb movements over 240 consecutive 5-second intervals given by Leroux and Puterman (1992) and replicated here in Figure 2(a). The hidden state  $h_t$  describes the fetal activity level during interval  $t$ . The activity levels are usually assumed to be  $S = [\text{passive}, \text{active}]$  or  $S = [\text{passive}, \text{somewhat active}, \text{very active}]$ . The fetal activity category evolves according to a stationary Markov chain with transition matrix  $\mathbf{Q}$ . The number of movements in interval  $t$ ,  $d_t$ , is Poisson with mean  $\theta_{h_t}$ .

**Example 2 (The Business Cycle).** Hamilton (1989, 1990) modeled the U.S. gross national product (GNP) using an autoregressive integrated moving average (ARIMA) process evolving around means determined by a hidden Markov chain, that is,  $y_t - \mu_{h_t} \sim \text{ARIMA}(p, d, q)$ . The hidden state  $h_t$  indicates whether the economy is expanding ( $h_t = 1$ ) or contracting ( $h_t = 0$ ) at time  $t$ . Here  $\theta$  contains the ARIMA parameters and the mean GNP growth rates during economic expansions and contractions. The memory in the underlying business cycle is captured by  $\mathbf{Q}$ . For this example, more links must be added to the top part of Figure 1 to reflect dependence in the observed data even given  $\mathbf{h}$ .

**Example 3 (Network Intrusion).** Scott (1999, 2001) used HMMs to detect criminal intrusion into a telephone account. The observed data,  $d_t$ , is the placement time of the  $t$ th telephone call on the account, which can be augmented by observed call characteristics as in Figure 2(b). The customer responsible for the account generates traffic according to a (possibly nonhomogeneous) Poisson process. Criminals break into and leave the account according to a two-state continuous-time Markov process. When a criminal is present, he generates additional traffic according to a second Poisson process, independent of the customer. When all processes are homogeneous, data from the account may be represented as a discrete-time HMM. The hidden state  $h_t$  indicates the fraud status of the  $t$ th call and the criminal's presence or absence during the

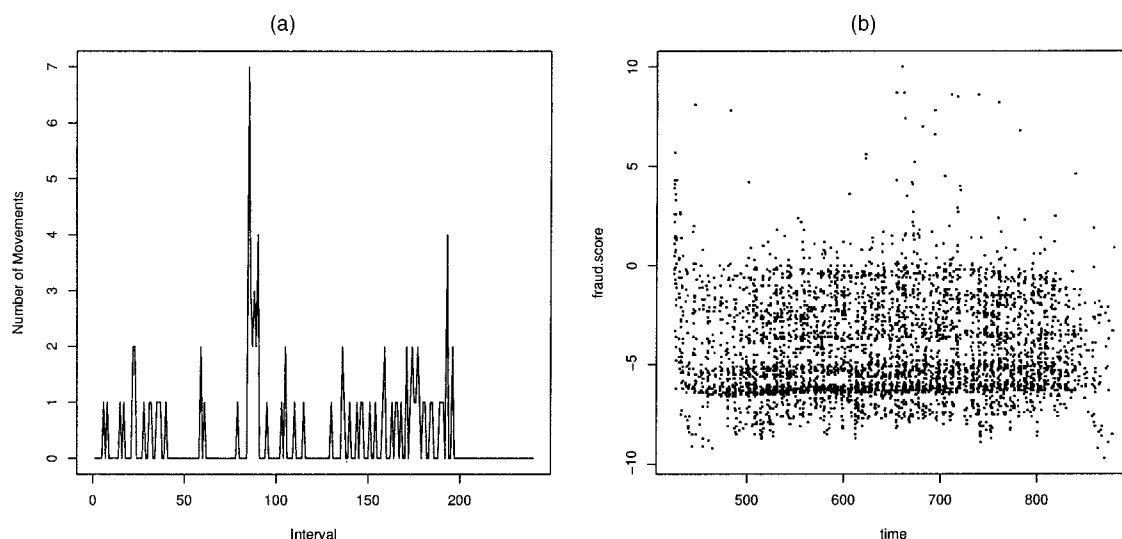


Figure 2. (a) Number of Fetal Lamb Movements in 240 Consecutive 5-Second Intervals and (b) Output of a Fraud Detection System for a Telephone Account. The vertical axis for panel (b) is a score indicating how closely a telephone call matches the customer's historical calling pattern. Higher scores indicate more unusual calls. Time is in days since January 1, 1994.

$t$ th interval between calls. The parameters  $\theta$  and  $\mathbf{Q}$  are functions of the calling rates for the customer and the criminal, of parameters describing the distribution of call characteristics for the customer and the criminal, and of the rates at which criminals break into and leave the system.

Returning to the general case, extend  $\theta$  to include  $\mathbf{Q}$ , and for arbitrary vectors  $\mathbf{x} = (x_1, \dots, x_n)$ , write  $x_j^k$  for  $(x_j, \dots, x_k)$ . The likelihood function for an HMM with observed data  $d_1^n$  is

$$p(d_1^n | \theta) = \sum_{\mathbf{h} \in \mathcal{S}^n} \pi_0(h_1) P_{h_1}(d_1 | \theta) \prod_{t=2}^n q(h_{t-1}, h_t) P_{h_t}(d_t | \theta). \quad (3)$$

The sum in (3) is over  $S^n$  elements, so it quickly becomes infeasible to compute even for small values of  $S$  as  $n$  grows to moderate size. A method for evaluating (3) is needed so the many statistical procedures that depend on likelihood (including likelihood ratio testing, Bayesian model selection, and certain Metropolis–Hastings samplers) can be used. The next two sections describe recursive procedures developed to sidestep computational difficulties posed by (3). Section 1.1 explains a technique for quickly evaluating the HMM likelihood, and Section 1.2 discusses a related set of forward-backward recursions originally developed to implement an EM algorithm maximizing (3) over  $\theta$ . Slightly modified forward-backward recursions lead to a Gibbs sampler that is central to the rest of the article.

### 1.1 The Likelihood Recursion

The likelihood recursion is a procedure for calculating (3) in  $O(S^2n)$  steps, instead of the  $O(S^n)$  steps suggested by direct evaluation. Define the *forward variable*  $\ell_t(h_t) \equiv p(d_1^t, h_t | \theta)$ . In words,  $\ell_t(s)$  is the joint likelihood contribution of  $d_1^t$  and the event  $\{h_t = s\}$  averaging over  $h_1^{t-1}$ . The likelihood contribution from  $d_1^t$  is  $\ell_t^* = \sum_{s=0}^{S-1} \ell_t(s)$ , yielding (3) when  $t = n$ . The likelihood recursion computes  $\ell_t(s)$  from  $\{\ell_{t-1}(r) : r \in \mathcal{S}\}$  as follows:

$$\begin{aligned} \ell_t(s) &= \sum_{r=0}^{S-1} p(d_t, d_1^{t-1}, h_t = s, h_{t-1} = r | \theta) \\ &= P_s(d_t | \theta) \sum_{r=0}^{S-1} q(r, s) \ell_{t-1}(r). \end{aligned} \quad (4)$$

For each  $s$ , computing  $\ell_t(s)$  requires the sum of  $S$  quantities. Thus a single step in the recursion is  $O(S^2)$ , and evaluating (3) is  $O(S^2n)$ , as indicated earlier.

Chib (1996) observed that (4) is unstable because it calculates likelihood rather than log-likelihood, but stable modifications of (4) exist. Define  $\pi_t(s | \theta) = \ell_t(s) / \ell_t^* = p(h_t = s | d_1^t, \theta)$  and let  $M_t = \max_s \log \{P_s(d_t | \theta) \sum_r q(r, s) \pi_{t-1}(r | \theta)\}$ . It is easy to show that  $\log \ell_t^*$  obeys the recursive relationship

$$\begin{aligned} \log \ell_t^* &= \log \ell_{t-1}^* + M_t \\ &+ \log \left( \sum_{s=0}^{S-1} \exp \left[ \log P_s(d_t | \theta) \right. \right. \\ &\quad \left. \left. + \log \left( \sum_{r=0}^{S-1} \pi_{t-1}(r | \theta) q(r, s) \right) - M_t \right] \right). \end{aligned} \quad (5)$$

Equation (5) scales each component of  $\log \pi_t(s | \theta)$  before exponentiation to prevent computer overflow, then increments  $\log \ell_t^*$  by the scaling factor and the log of the normalizing constant for  $\pi_t(s | \theta)$ .

### 1.2 The Forward-Backward Recursions

The forward-backward recursions were developed by Baum et al. (1970) to implement an EM algorithm (Dempster, Laird, and Rubin 1977) maximizing (3) over  $\theta$ . The recursions produce the *backward variable*  $\pi'_t(s | \theta) \equiv \Pr(h_t = s | d_1^n, \theta)$ , for all  $s$  and  $t$ , which is needed for the E step of the EM algorithm. Readers familiar with state-space models will note the correspondence between the forward-backward recursions and the prediction and smoothing steps of the Kalman filter (Kalman 1960). The forward recursion accumulates information about the distribution of  $\mathbf{h}$  as it moves down the hidden Markov chain. The backward recursion updates the distribution of  $\mathbf{h}$  calculated in the forward step once information has been collected from all observed data.

The forward-backward recursions are traditionally derived using (4) as the forward recursion. An alternative representation generates the matrices  $\mathbf{P}_2, \dots, \mathbf{P}_n$ , where  $\mathbf{P}_t = (p_{trs})$  and  $p_{trs} = p(h_{t-1} = r, h_t = s | d_1^t, \theta)$ . In words,  $\mathbf{P}_t$  is the joint distribution of  $(h_{t-1}, h_t)$  given model parameters and observed data up to time  $t$ . One computes  $\mathbf{P}_t$  from  $\mathbf{P}_{t-1}$  as

$$\begin{aligned} p_{trs} &\propto p(h_{t-1} = r, h_t = s, d_t | d_1^{t-1}, \theta) \\ &= \pi_{t-1}(r | \theta) q(r, s) P_s(d_t | \theta), \end{aligned} \quad (6)$$

where proportionality is reconciled by  $\sum_r \sum_s p_{trs} = 1$ . Notice that  $\pi_t(s | \theta) = \sum_r p_{trs}$  can be computed once  $\mathbf{P}_t$  is known, setting up the next step in the recursion. Calculating  $\pi_t$  as a margin of  $\mathbf{P}_t$  is equivalent to calculation by (5), so the log-likelihood can be stably computed as a byproduct of the forward recursion. There is no need to perform a separate likelihood calculation.

The backward recursion replaces  $\mathbf{P}_2, \dots, \mathbf{P}_n$  with  $\mathbf{P}'_2, \dots, \mathbf{P}'_n$ , where  $\mathbf{P}'_t = (p'_{trs})$  is the conditional distribution of  $(h_{t-1}, h_t)$  given model parameters and  $d_1^n$ . Note that  $\mathbf{P}'_t$  conditions on all of the observed data, whereas  $\mathbf{P}_t$  conditions only on data observed up to time  $t$ . Clearly,  $\mathbf{P}_n = \mathbf{P}'_n$ , and one obtains  $\mathbf{P}'_t$  from  $\mathbf{P}_t$  and  $\mathbf{P}'_{t+1}$ , using Bayes's rule,

$$\begin{aligned} p'_{trs} &= p(h_{t-1} = r | h_t = s, d_1^n, \theta) p(h_t = s | d_1^n, \theta) \\ &= p(h_{t-1} = r | h_t = s, d_1^t, \theta) \pi'_t(s | \theta) \\ &= p_{trs} \frac{\pi'_t(s | \theta)}{\pi_t(s | \theta)}, \end{aligned} \quad (7)$$

where  $\pi'_t(s | \theta)$  is computed as the appropriate margin of  $\mathbf{P}'_{t+1}$ .

Focusing on the forward-backward matrices  $\mathbf{P}_t$  and  $\mathbf{P}'_t$  rather than on the forward-backward variables  $\ell_t$  and  $\pi'_t$  helps strip away some of the mystery of the algorithm. HMMs can be understood using the more general theory of graphical models (Lauritzen and Spiegelhalter 1988; Whittaker 1990; Cowell 1999), where “graph” is understood in the sense of Figure 1. Marginal distributions for variables in a graphical model can be efficiently calculated using a local computation algorithm based on a “junction tree” describing relationships between

cliques on the graph. A clique is a set of variables whose members are all neighbors on the graph, so the cliques in Figure 1 are the transitions  $A_t = (h_{t-1}, h_t)$  and the completed data  $B_t = (h_t, d_t)$ . The local computation algorithm operates by calculating the marginal distributions of cliques. Marginal distributions of individual variables in the cliques may then be derived using low-dimensional sums or integrals. The junction tree for an HMM is a chain  $B_1 \rightarrow A_2 \rightarrow B_2 \rightarrow \dots$ , and  $\mathbf{P}_t$  and  $\mathbf{P}'_t$  represent the  $A_t$  cliques being marginalized. A general forward recursion for graphical models calculates the marginal distribution of each clique by recursively averaging over the marginal distribution of its parents in the junction tree. The backward recursion updates the distribution of each clique using information accumulated in the forward step to ensure that the marginal distributions cohere with the joint distribution of all variables in the model. Cowell et al. (1999) provided details of the general algorithm.

Forward-backward recursions for extensions of HMMs are easily derived using the general local computation method for graphical models. The computational gain from the recursions depends on the size of the cliques. As a simple example, suppose that  $h_t$  obeys a second-order Markov chain  $p(h_t | h_{t-1}^{t-1}) = p(h_t | h_{t-1}, h_{t-2})$ . An  $O(S^3n)$  set of forward-backward recursions is immediately available by expanding  $\mathbf{P}_t$  to an array of order 3. See Romberg et al. 2000 and Scott 2001 for examples in which modified forward-backward recursions are applied to more complicated graphs than Figure 1. The recursions may also be extended to hidden semi-Markov models, where  $h_t$  persists in state  $s$  for a duration  $w$  determined by a distribution  $f_s(w)$ . Hidden semi-Markov models have been accommodated by running forward-backward recursions conditional on the vector of durations  $\mathbf{w} = (w_1, w_2, \dots)$  and then marginalizing over  $\mathbf{w}$  (Schmidler, Liu, and Brutlag 2000, 2001; Ge and Smyth 2000). The procedure for hidden semi-Markov models is much more computationally intensive than that for HMMs.

## 2. POSTERIOR SAMPLING OF $\theta$ BY MARKOV CHAIN MONTE CARLO

This section discusses MCMC methods for sampling  $\theta$  from its posterior distribution given  $d_1^n$ . HMM's missing-data structure naturally admits posterior samplers that alternate between simulating  $\mathbf{h}$  given  $\theta$  and  $d_1^n$  and simulating  $\theta$  given complete data. An obvious way to sample  $\mathbf{h}$  is by DG (Albert and Chib 1993; Robert, Celeux, and Diebolt 1993; Robert and Titterton 1998). DG draws each  $h_t$  from its full conditional distribution,

$$p(h_t = s | h_{-t}, d_1^n, \theta) \propto q(h_{t-1}, s)q(s, h_{t+1})P_s(d_t | \theta), \quad (8)$$

for  $t = 1, \dots, n$ , with appropriate adjustments for end effects. When paired with Gibbs or Metropolis-Hastings steps for sampling  $\theta$  from  $p(\theta | d_1^n, \mathbf{h})$ , (8) produces a sequence  $\{(\theta, \mathbf{h})^{(j)} : j = 1, 2, \dots\}$  from a Markov chain whose limiting distribution is  $p(\theta, \mathbf{h} | d_1^n)$ .

FB (Chib 1996; Scott 1999; Robert, Rydén, and Titterton 1999) is our preferred alternative to DG. FB modifies (8) by using a stochastic version of the forward-backward recursions to sample  $\mathbf{h}$  directly from  $p(\mathbf{h} | \theta, d_1^n)$ . The modification leads to a more rapidly mixing algorithm because fewer

components are introduced into the Gibbs Markov chain. The forward recursion for FB produces  $\mathbf{P}_2, \dots, \mathbf{P}_n$  exactly as in Section 1.2. The *stochastic backward recursion* begins by drawing  $h_n$  from  $\pi_n(\cdot | \theta)$ , then recursively draws  $h_t$  from the distribution proportional to column  $h_{t+1}$  of  $\mathbf{P}_{t+1}$ . To understand the stochastic backward recursion factor  $p(\mathbf{h} | \theta, d_1^n)$  as

$$p(\mathbf{h} | d_1^n, \theta) = p(h_n | d_1^n, \theta) \prod_{t=1}^{n-1} p(h_{n-t} | h_{n-t+1}^n, d_1^n, \theta), \quad (9)$$

and notice (Fig. 1 may help) that

$$\begin{aligned} p(h_{n-t} = r | h_{n-t+1}^n, d_1^n, \theta) &= p(h_{n-t} = r | h_{n-t+1}, d_1^{n-t+1}, \theta) \\ &\propto p_{n-t+1, r, h_{t+1}}. \end{aligned} \quad (10)$$

Although the stochastic backward recursion replaces (7), Section 3 explains how combining the two backward recursions leads to improved estimates of  $\mathbf{h}$ . It is easy to implement the stochastic and nonstochastic backward recursions simultaneously once the forward recursion has been run.

Recursive computing also opens the door for other MCMC procedures that would otherwise be impossible. As an example, consider a Metropolis-Hastings sampler (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953; Hastings 1970) proposing candidate values  $\theta_*^{(j+1)}$  from a multivariate normal distribution centered close to the current draw  $\theta^{(j)}$ . One of  $\theta_*^{(j+1)}$  or  $\theta^{(j)}$  is promoted to  $\theta^{(j+1)}$  depending on a Hastings probability determined by the relative likelihood of  $\theta^{(j)}$  and  $\theta_*^{(j+1)}$  under  $p(\theta | d_1^n)$  and the candidate distribution. This sampler avoids drawing  $\mathbf{h}$  altogether by using the likelihood recursion to integrate over  $\mathbf{h}$  when calculating Hastings probabilities. Improved candidate distributions may be obtained by modifying the likelihood recursion to produce derivatives of log-likelihood suitable for use in a Langevin diffusion (Gilks and Roberts 1996). The modification involves recursively applying the derivative operator to (4) or (5).

Metropolis-Hastings procedures that average over  $\mathbf{h}$  are useful (Celeux, Hurn, and Robert 2000), but there are reasons to prefer sampling  $\mathbf{h}$  instead of integrating it out. First, Metropolis-Hastings algorithms tend to perform poorly when the dimension of  $\theta$  is large. Most candidate draws in high-dimensional problems are rejected unless the variance of the candidate distribution is small, in which case the algorithm moves slowly through the parameter space (although see Cappé, Robert, and Rydén 2001). The usual remedy of partitioning  $\theta$  and applying smaller Metropolis-Hastings algorithms to the elements of the partition is unattractive for HMMs, because partitioning  $\theta$  requires a separate run of the likelihood recursion for each Hastings probability that must be computed. A second reason to prefer sampling  $\mathbf{h}$  is that highly correlated elements of  $\theta$  under  $p(\theta | d_1^n)$  are often nearly independent under  $p(\theta | d_1^n, \mathbf{h})$ , so including  $\mathbf{h}$  in the sampling algorithm provides an expanded parameter space that may accelerate mixing with respect to  $\theta$  if  $\mathbf{h}$  can be efficiently drawn (van Dyk and Meng 2001). This is the idea behind augmented variable samplers from spatial statistics (Besag and Green 1993; Higdon 1998; Damian, Wakefield, and Walker 1999). Finally, some function  $g(\mathbf{h})$  is often the object of scientific interest, so sampled  $\mathbf{h}$ 's may be used for inference and

as diagnostics for model adequacy and MCMC convergence (Robert et al. 1999).

Later subsections focus on FB and DG to the exclusion of other sampling schemes. In particular, I omit discussion of approximate sampling methods not involving MCMC (e.g., Liu and Chen 1998). Section 2.1 explains why FB mixes more rapidly than DG when drawing from  $p(\mathbf{h} | d_1^n, \theta)$ . Section 2.2 discusses a label-switching issue that must be dealt with when drawing from  $p(\theta | d_1^n, \mathbf{h})$ . Label switching is a problem regardless of which sampling scheme is used. Section 2.3 uses the model from Example 1 to illustrate a method for preventing label switching and to compare FB and DG.

## 2.1 Autocovariance of the Forward-Backward Gibbs and Direct Gibbs Samplers

FB reduces the dependence of  $h_i$  on other hidden states drawn in previous Gibbs iterations by sampling  $\mathbf{h}$  directly from  $p(\mathbf{h} | d_1^n, \theta)$ . Faster mixing for  $\mathbf{h}$  translates into faster mixing for  $\theta$  by analogy with a duality principle introduced by Diebolt and Robert (1994) for finite mixture models. This section shows that the autocovariance of complete data sufficient statistics drawn by DG is the FB autocovariance plus a penalty term. The penalty term tends to be positive and increases with the posterior covariance of the hidden states. The theory presented here is a special case of Liu, Wong, and Kong (1994, 1995), in whose terminology FB is a “grouped” sampler and DG is an “ungrouped” sampler.

Assume that  $p(\theta | \mathbf{h}, d_1^n)$  depends on a vector of complete-data sufficient statistics  $T = \sum_{j=1}^m T_j$ . Dependence on  $d_1^n$  henceforth is suppressed in this section. For interpretation purposes, imagine that  $m = n$  and  $T_j$  is an  $S^2$  vector of indicator variables describing  $(h_{j-1}, h_j)$ , although the notation allows greater generality. Both FB and DG produce sequences  $\{(\theta^{(\tau)}, \mathbf{T}^{(\tau)}) : \tau = 1, 2, \dots\}$ , where  $\mathbf{T}^{(\tau)} = (T_1^{(\tau)}, \dots, T_m^{(\tau)})$ , from Markov chains whose stationary distribution is  $p(\theta, \mathbf{T})$ . Under FB,  $\mathbf{T}^{(\tau)}$  is conditionally independent of  $\mathbf{T}^{(\tau-1)}$  given  $\theta^{(\tau)}$ . DG satisfies the weaker assumption that  $T_j^{(\tau)}$  is conditionally independent of  $T_k^{(\tau-1)}$  given  $\theta^{(\tau)}$ ,  $\{T_{j'}^{(\tau-1)} : j' > j\}$ , and  $\{T_{j'}^{(\tau)} : j' < j\}$ , for  $k \leq j$  (Fig. 3).

Once FB and DG have achieved stationarity, it is easily established that  $(\mathbf{T}^{(\tau-1)}, \theta^{(\tau)})$  and  $(\mathbf{T}^{(\tau)}, \theta^{(\tau)})$  have the same

marginal distribution. It follows that

$$\begin{aligned} \text{cov}(T^{(\tau-1)}, T^{(\tau)}) &= E\{\text{cov}(T^{(\tau-1)}, T^{(\tau)} | \theta^{(\tau)})\} \\ &\quad + \text{cov}\{E(T^{(\tau-1)} | \theta^{(\tau)}), E(T^{(\tau)} | \theta^{(\tau)})\} \\ &= E\{\text{cov}(T^{(\tau-1)}, T^{(\tau)} | \theta^{(\tau)})\} \\ &\quad + \text{var}\{E(T^{(\tau)} | \theta^{(\tau)})\}, \end{aligned} \quad (11)$$

with  $E$ ,  $\text{cov}$ , and  $\text{var}$  defined under either sampler. The first term in (11) is 0 under FB. The second is the same for either sampler because FB and DG have the same stationary distribution. Consequently,

$$\begin{aligned} \text{cov}_{\text{DG}}(T^{(\tau-1)}, T^{(\tau)}) &= \text{cov}_{\text{FB}}(T^{(\tau-1)}, T^{(\tau)}) \\ &\quad + E_{\text{DG}}\{\text{cov}_{\text{DG}}(T^{(\tau-1)}, T^{(\tau)} | \theta^{(\tau)})\}. \end{aligned} \quad (12)$$

In words, (12) says that the autocovariance of the complete-data sufficient statistics under direct Gibbs is the FB autocovariance plus a penalty term. To understand the penalty term, notice that  $\text{cov}_{\text{DG}}(T^{(\tau-1)}, T^{(\tau)} | \theta^{(\tau)})$  decomposes as

$$\begin{aligned} \text{cov}_{\text{DG}}(T^{(\tau-1)}, T^{(\tau)} | \theta^{(\tau)}) &= \sum_{j=1}^m \sum_{k=j}^m \text{cov}_{\text{DG}}(T_j^{(\tau-1)}, T_k^{(\tau)} | \theta^{(\tau)}) \\ &\quad + \sum_{j=2}^m \sum_{k=1}^{j-1} \text{cov}_{\text{DG}}(T_j^{(\tau-1)}, T_k^{(\tau)} | \theta^{(\tau)}) \\ &= UT + LT. \end{aligned}$$

Here  $UT$  and  $LT$  represent sums over the upper and lower triangles of  $\text{cov}_{\text{DG}}(\mathbf{T}^{(\tau-1)}, \mathbf{T}^{(\tau)} | \theta^{(\tau)})$ . For DG to outperform FB,  $UT + LT$  must be negative (in a matrix sense). However, it can be shown that  $UT$  is nonnegative definite, so the autocovariance under DG tends to be greater than under FB. Furthermore,  $\text{cov}_{\text{DG}}(T_j^{(\tau-1)}, T_k^{(\tau)} | \theta^{(\tau)}) = \text{cov}_p(T_j, T_k | \theta^{(\tau)})$  for  $k < j$ , so DG performs worst when the elements of  $\mathbf{h}$  are highly related in their posterior distribution. The last observation mathematically expresses a well-known feature of DG from spatial statistics (Besag and Green 1993; Higdon 1998), namely, it is hard for DG to move from one configuration of  $\mathbf{h}$  to another when there is strong dependence between  $h_i$  and its neighbors.

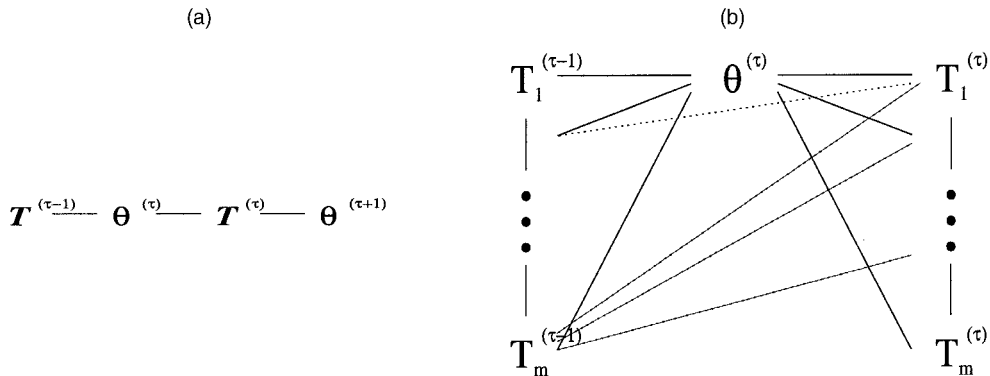


Figure 3. Conditional Independence Graphs for the FB (a) and DG (b) Samplers. The conditional distribution of a variable given all other quantities depends only on variables to which it is connected by an edge. Each  $T_j^{(\tau)}$  in DG depends on  $\theta, T_{j+1}^{(\tau-1)}, \dots, T_m^{(\tau-1)}$ , and  $T_1^{(\tau)}, \dots, T_{j-1}^{(\tau)}$ .

## 2.2 Label Switching, State Collapsing, and Prior Modeling

Sampling  $\theta$  from its complete data posterior should be trivial once  $\mathbf{h}$  is drawn, but the draw is complicated by an identifiability issue known as *label switching*. This term describes the fact that the HMM likelihood is invariant under arbitrary permutations of the state labels. Most literature on label switching works in the context of finite mixture models, but HMMs face the same issues. Consider the case where  $S = 2$ . Swap values of  $\theta_0$  and  $\theta_1$ , relabel all points currently in category 0 as category 1, and vice versa. The complete-data likelihood achieves exactly the same value under the new labels as under the old. If  $\theta_0$  and  $\theta_1$  are exchangeable in their prior distribution, then their marginal posterior densities are identical. Label switching is an obstacle for HMMs regardless of which sampling strategy is used. Its effect can be seen in Figure 4, which shows posterior density estimates for the mean parameters of a two-state Gaussian HMM [i.e.,  $P_s = \mathcal{N}(\cdot | \mu_s, \sigma^2)$ ] applied to an economic time series. This model is a simplification of the model discussed in Example 2.

Because the data contain no information about the order of state labels, labels may only be identified in the posterior distribution by assumptions in the prior. For example, it is common to assume that  $\theta$  obeys a set of constraints that would be violated by permuting the state labels. Typical constraints involve ordering the means or variances of the mixture components. Parameter constraints can be enforced by explicitly choosing a prior with zero mass on regions where the ordering is violated. Such priors lead to Metropolis–Hastings algorithms that reject draws of  $\theta$  violating the constraint (Richardson and Green 1997). Constraints can also be enforced by reparameterizing the model (Robert and Titterton 1998; Robert 1998).

Care is needed when choosing parameter constraints, because priors that truncate the support of  $\theta$  can be informative, and enforcing constraints by reparameterizing  $\theta$  can produce a different model. Celeux et al. (2000) provided examples

where ordering different sets of model parameters produces noticeably different posterior means of  $\theta$ . Also note from Figure 4(b) that a weakly informative prior on  $\theta$  does little to prevent label switching. Scientific insight about  $\mathbf{h}$  may therefore provide useful prior information by suggesting an order for  $\theta$ . “Scientific insight” and “prior information” are often discussed in abstract terms, but here I mean something concrete. For example, the model for U.S. GNP in Figure 4 is intended to associate  $h_t = 0$  with recessions and  $h_t = 1$  with economic expansions. Thus  $\mu_1 < \mu_0$  is nonsensical and should be removed from consideration. Similar logic can often be applied in other circumstances as long as  $\mathbf{h}$  has a physical interpretation.

Selecting appropriate prior distributions can also help curb state collapsing, a special case of label switching where no observations are allocated to a particular state during a Gibbs iteration. Parameters for a collapsed state are drawn from their prior distribution, so weak priors produce draws of  $\theta$  far from values justified by data. Hierarchical priors allowing elements of  $\theta$  to borrow strength from one another can help the sampler recover from a collapsed state. Priors constraining mean parameters to lie within the range of the data are also helpful. Frequent state collapsing is evidence that the model is overparameterized. Thus variable-dimension Monte Carlo methods discussed in Section 4.2 can help deal with collapsed states by allowing a state to “die” and later be reborn in a sensible location.

Scientific insight about  $\mathbf{h}$  might not be available in all cases. For example, an HMM might be fit to “nonparametrically” obtain the predictive distribution of  $d_{n+1}$  given  $d_1^n$  or to cluster the observations of a time series when no intuition exists about the source of the clustering. Stephens (1997, 2000b) proposed a method for identifying mixture components when prior information about  $\mathbf{h}$  is unavailable. Stephens’ method runs MCMC with no identifiability constraints to sample from

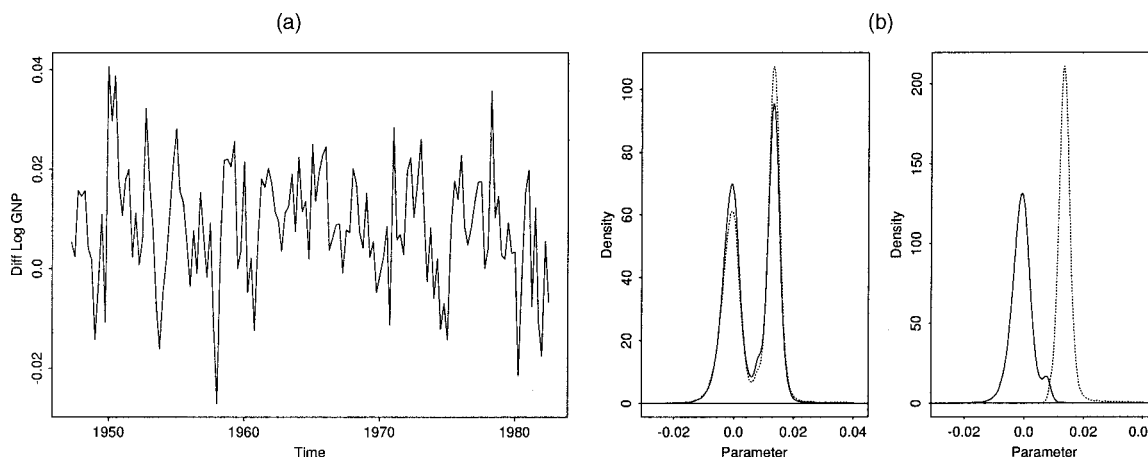


Figure 4. (a) Quarterly Differences in Log U.S. GNP and (b) Posterior Distributions of Mean Parameters in a Two-State Gaussian HMM Applied to the GNP Series. Both panels in (b) assume prior distribution  $p(\mu_s | \tau^2) = \mathcal{N}(\cdot | M_s, \tau^2)$  and  $p(1/\tau^2) = \Gamma(DF/2, SS/2)$ , where  $M_s$  was estimated from an EM algorithm and DF and SS were chosen so the prior mean for  $\tau^2$  was comparable to the sample variance. Both panels also truncate  $p(\mu_s | \tau^2)$  at the maximum and minimum data values. The prior distribution for the right panel further constrains  $\mu_1 > \mu_0$ . The unconstrained parameters in the left panel exhibit label switching.

a posterior distribution containing  $S!$  symmetric modes. At the end of the MCMC run, the method searches for a permutation of the state labels leading to marginal distributions of  $\theta$  that minimize a specified loss function. Loss functions have been developed to promote unimodal marginal distributions of  $\theta$  and efficient clustering of  $d_1^n$ .

### 2.3 Example: Constrained Markov–Poisson Hidden Markov Model

This section uses a constrained Markov–Poisson HMM,  $P_s = \text{Poisson}(\theta_s)$  with  $\theta_0 < \dots < \theta_{S-1}$ , to illustrate a general method for ordering the variance parameters of HMMs and finite mixture models. The method works when  $P_0, \dots, P_{S-1}$  are distributional forms that are invariant under addition, such as (multivariate) Gaussian, Poisson, gamma with common scale parameter, and binomial with common  $p$ . The key is to view  $d_t = \sum_s d_{st}$  as the sum of contributions from various “regimes.” Regimes  $0, \dots, h_t$  are active, but  $d_{st} = 0$  for  $s > h_t$ . The result is an ordered variance model, because adding two nondegenerate random variables increases their variance. Let  $\mathbf{d} = (\mathbf{d}_1, \dots, \mathbf{d}_n)$ , where  $\mathbf{d}_t = (d_{0t}, \dots, d_{S-1,t})$  contains the regime contributions.

For the Poisson case, define  $\lambda_s = \theta_s - \theta_{s-1}$ , where  $\lambda_0 = \theta_0$ . If  $\boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_{S-1})$ , then

$$p(\mathbf{d}, \mathbf{h} | \boldsymbol{\lambda}, \mathbf{Q}) \propto \pi_0(h_1) \left( \prod_{r=0}^{S-1} \prod_{s=0}^{S-1} q(r, s)^{n_{rs}} \right) \times \left( \prod_{s=0}^{S-1} \lambda_s^{\sum_t d_{st}} \exp(-n_s \lambda_s) \right), \quad (13)$$

where  $n_s = \sum_{t=1}^n I(h_t \geq s)$  is the number of times that regime  $s$  was active and  $n_{rs}$  is the number of transitions from  $r$  to  $s$ . Assume prior density  $p(\theta, \mathbf{Q}) = \prod_{r=0}^{S-1} \Gamma(\lambda_r | a_r, b_r) \mathcal{D}(\mathbf{Q}_r | \mathbf{v}_r)$ , where  $\Gamma$  and  $\mathcal{D}$  are the gamma and Dirichlet densities and  $\mathbf{Q}_r$  is the  $r$ th row of  $\mathbf{Q}$ . The rows of  $\mathbf{Q}$  and elements of  $\boldsymbol{\lambda}$  are independent in their posterior distribution given  $\mathbf{d}$  and  $\mathbf{h}$ , with

$$p(\lambda_s | \mathbf{d}, \mathbf{h}) = \Gamma\left(a_s + \sum_t d_{st}, b_s + n_s\right),$$

$$p(\mathbf{Q}_r | \mathbf{d}, \mathbf{h}) = \mathcal{D}(\mathbf{v}_r + \mathbf{n}_r),$$

where  $\mathbf{n}_r = (n_{rs})$  is the vector of transitions out of state  $r$ .

To apply FB, note that  $p(\mathbf{d}, \mathbf{h} | \boldsymbol{\lambda}, \mathbf{Q}, d_1^n) = p(\mathbf{h} | \boldsymbol{\lambda}, \mathbf{Q}, d_1^n) p(\mathbf{d} | \mathbf{h}, \boldsymbol{\lambda}, \mathbf{Q}, d_1^n)$ . FB simulates  $\mathbf{h}$  from  $p(\mathbf{h} | \boldsymbol{\lambda}, \mathbf{Q}, d_1^n)$  using the stochastic forward-backward recursions. Given  $\mathbf{h}$  and  $\boldsymbol{\lambda}$ ,  $(\mathbf{d}_1, \dots, \mathbf{d}_n)$  are conditionally independent. If  $h_t = r$ , then no state higher than  $r$  can contribute to  $d_t$ , so  $d_{st} = 0$  for all  $s > r$ . The full conditional distribution of  $(d_{0t}, \dots, d_{rt})$  is multinomial with total  $d_t$  and probability vector proportional to  $(\lambda_0, \dots, \lambda_r)$ .

Reckless use of DG leads to trouble, because the full conditional of  $h_t$  depends on  $\mathbf{d}_t$ ,

$$p(h_t = s | \boldsymbol{\lambda}, \mathbf{Q}, h_{-t}, \mathbf{d}) \propto q(h_{t-1}, s) q(s, h_{t+1}) \prod_{r=0}^s \lambda_r^{d_{rt}} \exp(-\lambda_r) \prod_{r=s+1}^{S-1} I(d_{rt} = 0). \quad (14)$$

If  $d_{st} > 0$  for any  $s > r$ , then (14) places probability 0 on the events  $\{h_t = r\}$ , making it difficult for the sampler to move from larger to smaller values of  $h_t$ . A compromise defines the missing data for time  $t$  as  $(h_t, \mathbf{d}_t)$  and samples  $(h_t, \mathbf{d}_t)$  by nested conditioning. The full conditional factors as  $p(h_t, \mathbf{d}_t | \boldsymbol{\lambda}, \mathbf{Q}, h_{-t}, \mathbf{d}_{-t}, d_1^n) = p(h_t | \boldsymbol{\lambda}, \mathbf{Q}, h_{-t}, \mathbf{d}_{-t}, d_1^n) p(\mathbf{d}_t | \boldsymbol{\lambda}, \mathbf{Q}, h_t, \mathbf{d}_{-t}, d_1^n)$ . First draw  $h_t$  from

$$p(h_t = s | \boldsymbol{\lambda}, \mathbf{Q}, h_{-t}, \mathbf{d}_{-t}, d_1^n) \propto q(h_{t-1}, s) q(s, h_{t+1}) P_s(d_t | \theta).$$

Then draw  $\mathbf{d}_t$  as in FB.

Figure 5 displays marginal posterior density estimates calculated from FB and DG for the two-state Markov–Poisson HMM applied to the fetal lamb data. There is little difference between the densities, because both samplers have the same stationary distribution. Their difference is highlighted in Figure 6, which shows empirical autocorrelation functions for Gibbs draws of  $\theta_0$  under FB, the DG sampler drawing  $h_t$  from (14), and the DG sampler drawing  $(h_t, \mathbf{d}_t)$  by nested conditioning. Figure 6 illustrates the cost of introducing unnecessary Gibbs components into the sampling algorithm. Figure 6(a) is from a rapidly mixing sampler with only two components. Mixing suffers as Figure 6(b) increases the number of Gibbs components to  $n+1$ . Figure 6(c) includes  $n$  additional Gibbs components that are highly correlated with other missing data drawn by the sampler, causing a terrible waste of the computer’s resources.

Computationally, DG is an  $O(Sn)$  algorithm, whereas FB is  $O(S^2n)$ . Thus DG’s speed can be an advantage as  $S$  increases. Table 1 shows the CPU time needed for FB and DG to generate 1,000 draws from the posterior distribution of  $\theta$  given data simulated from Markov–Poisson HMMs with different sample sizes and state spaces. Directly comparing CPU time for FB and DG is not completely fair, because FB calculates the likelihood associated with each drawn  $\theta$ , a quantity needed in Sections 4 and 5. Therefore, simulation times are shown for the DG algorithm with and without the likelihood recursion. Table 1 illustrates the increasing toll exacted by the forward-backward recursions as  $S$  increases. When  $S = 2$ ,

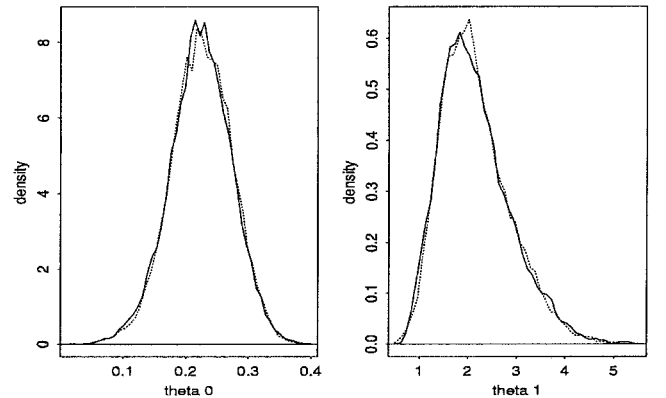


Figure 5. Marginal Distributions for  $\theta_0$  and  $\theta_1$  Computed by the FB (---) and DG (—) Samplers for the Two-State Markov–Poisson HMM Applied to the Fetal Lamb Data. Density estimates are based on Gibbs runs of 10,000 iterations, with the first 100 dropped. There is little difference between the FB and DG density estimates, because FB and DG sample from the same target density.

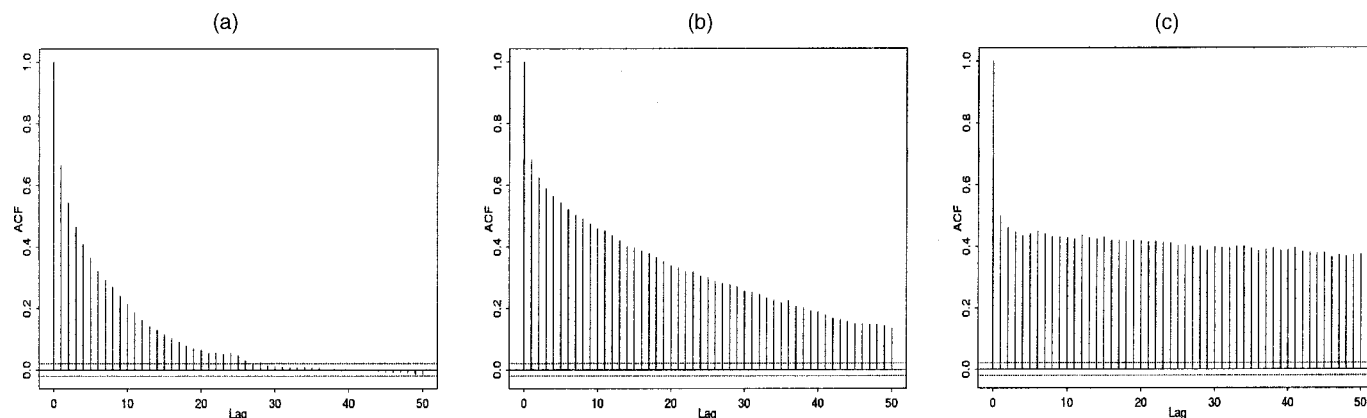


Figure 6. Autocorrelation Functions for Gibbs Draws of  $\theta_0$  Under (a) the FB Recursions, (b) DG Sampling Where  $\mathbf{d}_t$  is Drawn Conditional on  $h_t$ , and (c) DG. All are based on 10,000 draws from the constrained two-state Markov-Poisson HMM applied to the fetal lamb data. DG mixes much slower than the other two algorithms because the additional missing data.

FB takes about 30% longer than the direct Gibbs algorithm not computing the likelihood. When  $S = 8$ , FB takes more than twice as long. Pairing DG with the likelihood recursion eliminates DG's speed advantage when  $S = 2$ , but when  $S = 8$ , the algorithm is comfortably between FB and raw DG. Unless the number of states is exceptionally large, I prefer FB because of its superior mixing properties, because it calculates likelihood "for free" as part of the forward recursion, and because it allows more efficient estimation of  $\mathbf{h}$ , as discussed in Section 3.

### 3. ESTIMATING THE HIDDEN STATES

Estimating  $\mathbf{h}$  is often the central question in applied problems, such as the fraud detection problem from Example 3. All Bayes estimates of  $\mathbf{h}$  derive from its posterior distribution  $p(\mathbf{h} | d_1^n)$ , a high-dimensional distribution that must be summarized to be understood. For most applications, summarizing  $p(\mathbf{h} | d_1^n)$  with its marginal distributions  $\pi'_t(s) = \Pr(h_t = s | d_1^n)$  is sufficient. Sometimes the overall configuration of  $\mathbf{h}$  is more interesting than the value of an individual state. For example, in gene sequencing (Liu et al. 1999),  $h_t$  might represent a DNA base element where  $\mathbf{h}$  represents a gene; or in speech recognition (Juang and Rabiner, 1991),  $h_t$  may be a phoneme in the word  $\mathbf{h}$ . Maximum a posteriori (MAP) estimation selects  $\hat{\mathbf{h}}$  to maximize  $p(\mathbf{h} | d_1^n)$ , ensuring a coherent reconstruction of the hidden chain that may differ from a corresponding reconstruction from its marginals. Recursive computing improves

both marginal and MAP summaries of  $p(\mathbf{h} | d_1^n)$ , as discussed in Sections 3.1 and 3.2.

#### 3.1 Marginal Distribution of $h_t$

Let  $\theta^{(1)}, \dots, \theta^{(m)}$  be Gibbs draws of  $\theta$  and let  $\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(m)}$  be Gibbs draws of  $\mathbf{h}$  with  $\mathbf{h}^{(j)} = (h_1^{(j)}, \dots, h_n^{(j)})$ . An obvious estimate of  $\pi'_t(s)$  is

$$\tilde{\pi}'_t(s) = 1/m \sum_{j=1}^m I(h_t^{(j)} = s). \quad (15)$$

The forward-backward recursions improve (15) through Rao-Blackwellization (Gelfand and Smith 1990; Casella and Robert 1996). The Rao-Blackwellized estimate

$$\hat{\pi}'_t(s) = 1/m \sum_{j=1}^m \pi'_t(s | \theta^{(j)}) \quad (16)$$

sheds a layer of Monte Carlo variability by averaging probabilities rather than events simulated with those probabilities. Calculating (16) requires that the nonstochastic backward recursion produce  $\pi'_t(s | \theta^{(j)})$ , but simultaneously running both the stochastic and nonstochastic backward recursions requires little effort once the forward recursion has been implemented.

Figure 7 compares  $\hat{\pi}'_t$  to  $\tilde{\pi}'_t$  and to the empirical Bayes probability of economic recession based on the two-state Gaussian model for the U.S. GNP data from Figure 4. Figure 7(c) shows the difference in posterior standard devi-

Table 1. CPU Time (in Seconds) Required to Generate 1,000 Gibbs Draws for the FB and DG Samplers Applied to Data Simulated from Markov-Poisson Mixtures of Length  $n$

	S = 2			S = 4			S = 8		
	DG	DGL	FB	DG	DGL	FB	DG	DGL	FB
$n = 500$	4.7	6.5	6.2	7.9	12.2	13.8	24.4	32.1	50.1
$n = 1,000$	8.6	12.2	11.5	15.7	22.6	26.6	48.8	62.5	99.8
$n = 1,500$	13.0	17.4	17.9	22.9	33.2	40.3	63.2	84.9	144.4
$n = 2,000$	16.7	23.0	21.9	31.3	44.2	53.6	82.5	111.9	188.2

NOTE: DGL is the DG algorithm running the likelihood recursion. The true parameter values in the simulation are  $\theta_0 = 1$ ,  $\theta_1 = 2$ , and  $\theta_s = \theta_{s-1} + \theta_{s-2}$  for  $s = 2, \dots, S-1$ ,  $q(s, s) = .6$  and  $q(s, s+1) = q(s, s-1) = .2$  for  $s = 1, \dots, S-2$ ,  $q(0, 0) = q(S-1, S-1) = .8$ , and  $q(0, 1) = q(S, S-1) = .2$ .



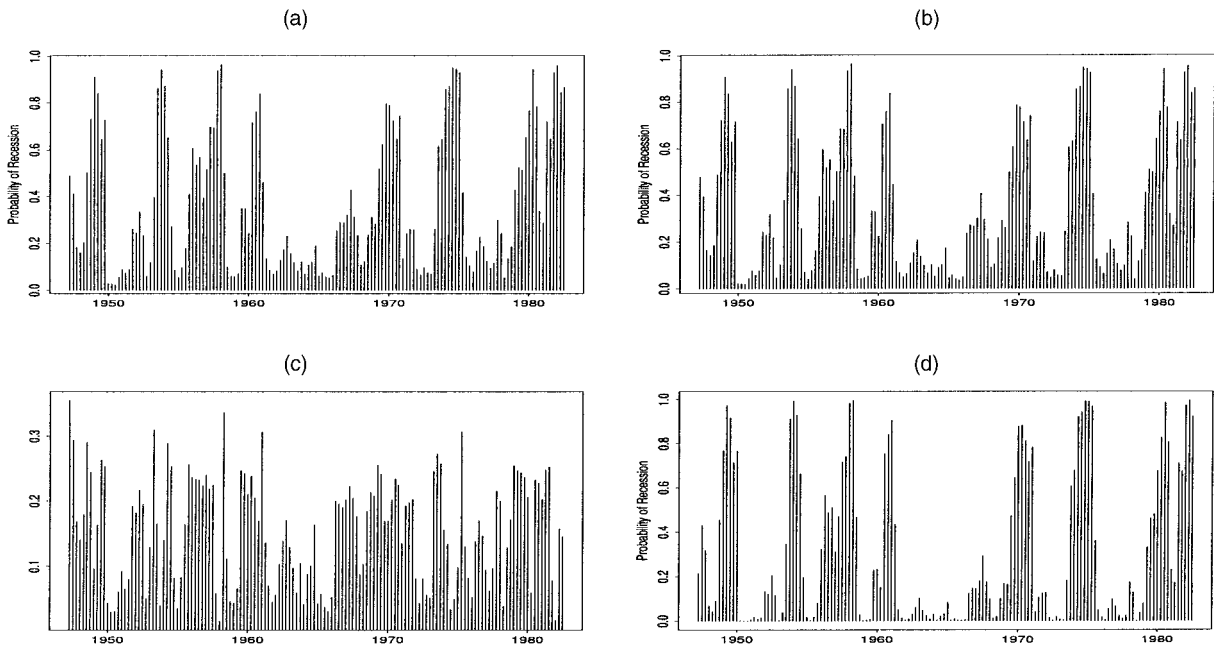


Figure 7. Marginal Probabilities of Recession Based on 100,000 Gibbs Iterations for the Two-State Gaussian HMM Applied to the U.S. GNP Data. (a) Rao-Blackwellized estimates,  $\hat{\pi}'_t(0)$ ; (b) brute force estimates,  $\tilde{\pi}'_t(0)$ ; (c) difference:  $SD(\hat{\pi}'_t(0 | \theta)) - SD(I(h_t = 0))$ ; (d) empirical Bayes estimates,  $\pi'_t(0 | \hat{\theta})$ . These estimates agree closely with official business cycle data from <http://www.nber.org/cycles.html>.

ations for Gibbs draws of  $\pi'_t(0 | \theta^{(j)})$  and  $I(h_t^{(j)} = 0)$ . If the MCMC output had been independently sampled, then the standard deviations represented by Figure 7(c) would be divided by  $\sqrt{m}$  to produce standard errors for the estimates in Figures 7(a) and (b). The extra variability in  $\tilde{\pi}'_t$  can clearly be overcome by running the sampler longer, but  $\hat{\pi}'_t$  produces more information content per iteration and is a virtually cost-less addition to the FB sampler.

The two Bayes estimates in Figure 7 are more similar to one another than to the empirical Bayes estimate  $\pi'_t(0 | \hat{\theta})$ , which can be calculated only by the forward-backward recursions. Because the empirical Bayes probabilities ignore uncertainty about  $\theta$ , they tend to be closer to 0 and 1. Empirical Bayes probabilities are inferior to full Bayes, but they are useful when

the cost of MCMC calculations is prohibitive. For example, the fraud detection model described in Example 3 is designed to monitor real time data from several million accounts like the account in Figure 2(b). MCMC may be used sparingly, perhaps once per month, to update estimates of  $\theta$  for each account, but MCMC is far too slow to monitor the entire network of accounts in real time. Figure 8(a) presents the empirical Bayes probability of fraud for each telephone call on the account, which may be quickly computed by a single run of the forward-backward recursions based on a past estimate of  $\theta$ . Without the forward-backward recursions, empirical Bayes calculations for an HMM would be intractable. Instead, one might ignore the effect of time and screen for fraud using a finite mixture model. Figure 8(b) compares the

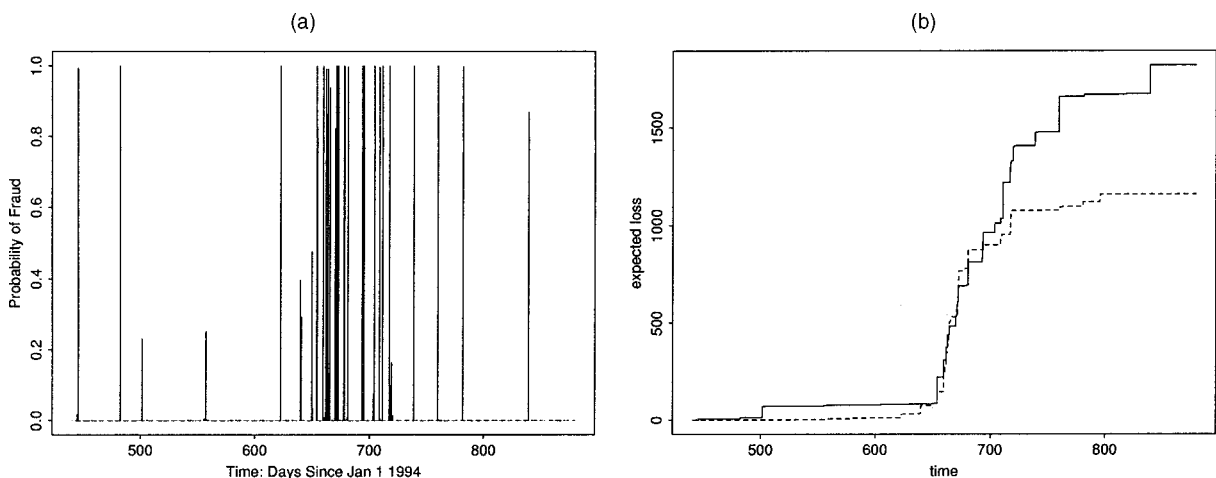


Figure 8. (a) Probability of Fraud for Telephone Calls on the Account Presented in Figure 2(b) and (b) Expected Loss due to Fraud Based on the HMM (.....) and a Mixture Model not Considering Time (—).

cumulative expected financial loss due to fraud computed from an HMM and from a finite mixture model. Figure 8(b) shows that the HMM is a clearly superior model, because it allows information about fraud to be shared across time. The HMM understands that occasional calls with large fraud scores represent weaker evidence of fraud than a cluster of such calls, so it generates fewer “false alarms” than the finite mixture model.

### 3.2 Maximum a Posteriori Estimation of $\mathbf{h}$

The *Viterbi algorithm* (Viterbi 1967; Fredkin and Rice 1992) is a method of finding the empirical Bayes *most likely trajectory*,  $\mathbf{h} = (h_1, \dots, h_n)$ , maximizing  $p(\mathbf{h} | d_1^n, \theta)$ . The algorithm uses a forward-backward strategy similar to that in Section 1.2, but with maximizations replacing averages. Begin by observing that the  $\mathbf{h}$  optimizing  $p(\mathbf{h} | d_1^n, \theta)$  also optimizes  $p(\mathbf{h}, d_1^n | \theta)$ , so the problem reduces to maximizing the complete-data likelihood. Define  $L_1(s) = \pi_0(s)P_s(d_1 | \theta)$  and  $L_t(h_t) = \max_{h_1, \dots, h_{t-1}} p(h_1^t, d_1^t | \theta)$ . In words,  $L_t(s)$  is the largest contribution to the complete-data likelihood that can be obtained if  $h_t = s$ . The quantities  $L_t(\cdot)$  are computed recursively through

$$L_t(s) = \max_r [L_{t-1}(r)q(r, s)]P_s(d_t | \theta). \quad (17)$$

To find  $\hat{\mathbf{h}}$ , run (17) for  $t = 1, \dots, n$ , storing each  $L_t(\cdot)$  as the algorithm progresses. Choose  $h_t$  to maximize  $L_n(s)$ , and for  $t = n-1, \dots, 1$ , choose

$$\hat{h}_t = \arg \max_{r \in S} L_t(r)q(r, \hat{h}_{t+1}) \quad (18)$$

to maximize the complete-data likelihood conditional on  $\hat{h}_{t+1}$ . Note that  $\hat{h}_1, \dots, \hat{h}_t$  are determined when  $\arg \max_r L_t(r)q(r, s)$  is the same for all  $s$ , a frequent event when  $S$  is small. When this occurs, we say that the estimation has *converged*. At each convergence, select  $\hat{h}_t$  to maximize  $L_t(r)q(r, s)$  over  $r$ , and use (18) to compute all previously undetermined elements of  $(\hat{h}_1, \dots, \hat{h}_{t-1})$ . Earlier values of  $L_t(\cdot)$  are no longer necessary and may be discarded. Updating  $\mathbf{h}$  at each convergence limits the memory demands of the Viterbi algorithm and allows the algorithm to run in nearly real time.

The Viterbi algorithm suffers from the same stability issues as the forward-backward recursions, with the same remedies. Renormalizing  $L_t$  at each iteration and transforming the algorithm to the log scale may be accomplished by analogy with (5). The algorithm may be expanded to accommodate more general HMMs by modifying the general local computation method for marginalization discussed in Section 1.2 (see Cowell et al. 1999 for details).

There is no Viterbi-style algorithm for maximizing  $p(\mathbf{h} | d_1^n)$ , because averaging over  $\theta$  destroys the model's Markov structure. Simply evaluating the objective function requires Monte Carlo,

$$p(\mathbf{h} | d_1^n) = \int p(\mathbf{h} | d_1^n, \theta)p(\theta | d_1^n)d\theta \approx 1/m \sum_{j=1}^m p(\mathbf{h} | d_1^n, \theta^{(j)}). \quad (19)$$

Equation (19) is difficult to optimize, because the optimal  $\mathbf{h}$  need not be the most likely trajectory for any particular  $\theta$ . A primitive solution is to select the most commonly occurring  $\mathbf{h}^{(j)}$  in the Gibbs run. Random MCMC fluctuations in  $\mathbf{h}^{(j)}$  will require extremely long MCMC runs, along with ample memory and hard disk storage, for this approach to be effective. An approximate but preferable solution can be obtained by viewing  $\hat{\mathbf{h}}$  as a function of  $\theta$  and selecting  $\hat{\mathbf{h}}(\theta^{(j)})$  from the MCMC run that is “most often most likely.” The set of distinct  $\hat{\mathbf{h}}(\theta^{(j)})$ 's produced by the sampler will be much smaller than the set of distinct  $\mathbf{h}^{(j)}$ 's. Figure 9(a) shows the most likely trajectories for the fetal lamb data multiplied by their frequency in the MCMC output. Marginal probabilities are included for reference in Figure 9(b). Note that some states with high marginal probability of fetal activity, like intervals 22 and 23, may not be represented when  $\mathbf{h}$  is estimated by MAP.

### 4. SELECTING THE STATE-SPACE SIZE

Up to now,  $S$  has been assumed known from the context of the application, but sometimes determining  $S$  is a question of scientific interest. Absent firm a priori theories, choosing

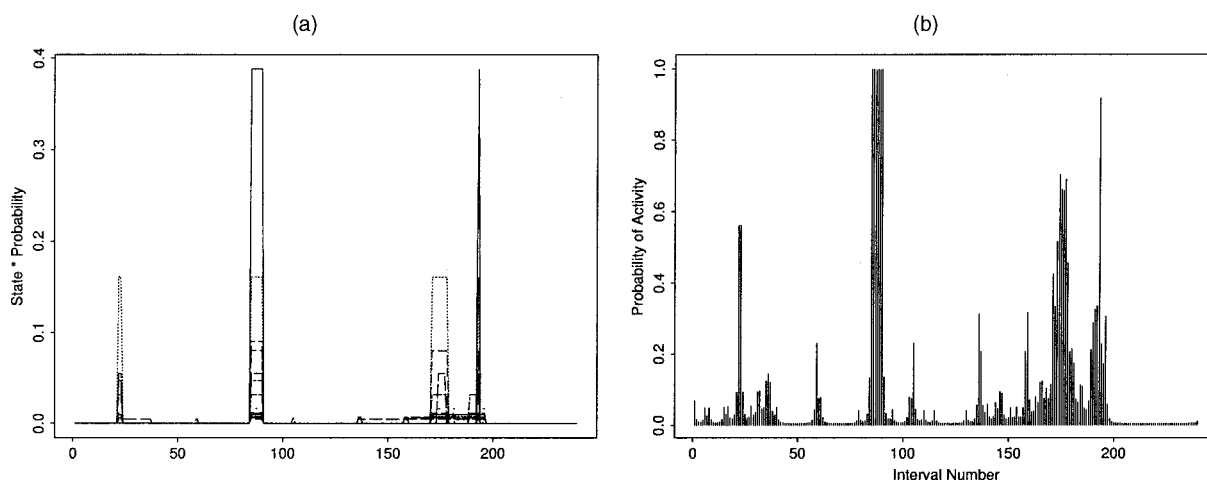


Figure 9. Estimating  $\mathbf{h}$  for the Fetal Lamb Data. (a) Most likely trajectories times their MCMC frequency. (b) Marginal probabilities  $\hat{\pi}'_t(1)$ .

$S$  is a model selection problem that views  $S$  as a random variable with a posterior distribution given  $d_1^n$ . Recursive computing is important for model selection, because standard Bayesian model selection and model-averaging techniques require the ability to compute likelihoods for the models being considered. Section 4.1 explains how to calculate  $p(S | d_1^n)$  by MCMC, and Section 4.2 discusses recently developed variable-dimension Monte Carlo methods that may be used instead of standard model selection theory.

#### 4.1 Calculating $p(S | d_1^n)$ by Markov Chain Monte Carlo

For practical reasons, assume that  $S \in \{1, \dots, S_{\max}\}$ . Bounding  $S$  is a mild restriction, because choosing  $S_{\max}$  to be the number of distinct values in  $d_1^n$  allows each data value to have its own mixture component. Smaller values of  $S_{\max}$  can be chosen, but no larger state space is identifiable.

Redefine  $\theta = (\theta_1, \dots, \theta_{S_{\max}})$ , where  $\theta_s$  is the parameter of an HMM with state space  $\mathcal{S} = \{0, \dots, S-1\}$ . Formally,  $p(d_1^n | \theta, S) = p(d_1^n | \theta_s, S)$ . The dimension of  $\theta$  is  $\sum_{s=1}^{S_{\max}} k_s$ , where  $k_s$  is the number of free parameters in  $\theta_s$ . Assume that

$$p(\theta, S) = p(S) \prod_{r=1}^{S_{\max}} p(\theta_r). \quad (20)$$

It follows from (20) that  $\theta_1, \dots, \theta_{S_{\max}}$  are conditionally independent in their posterior distribution given  $d_1^n$  and  $S$ , and may be independently sampled by  $S_{\max}$  parallel Gibbs samplers. Let  $\theta^{(1)}, \dots, \theta^{(m)}$  represent  $m$  draws of  $\theta$  from the samplers, where  $\theta^{(j)} = (\theta_1^{(j)}, \dots, \theta_{S_{\max}}^{(j)})$ . A Monte Carlo calculation of  $p(S | d_1^n)$  is

$$\begin{aligned} p(S | d_1^n) &= \int p(S | d_1^n, \theta) p(\theta | d_1^n) d\theta \\ &\approx 1/m \sum_{j=1}^m p(S | d_1^n, \theta^{(j)}), \end{aligned} \quad (21)$$

where  $p(S | d_1^n, \theta^{(j)}) \propto p(d_1^n | \theta_s^{(j)}, S) p(S)$ . The normalizing constant for  $p(S | d_1^n, \theta^{(j)})$  is easily computed, because  $S$  takes values on a finite set. Up to proportionality, (21) averages the  $S_{\max}$  likelihoods corresponding to each  $\theta^{(j)}$  over the life of the Gibbs sampler.

Calculating  $p(S | d_1^n)$  by MCMC is an improvement over the Schwarz criterion (Schwarz 1978; Kass and Raftery 1995), an asymptotic approximation to  $\log p(S | d_1^n)$ . Minus twice the Schwarz criterion is the Bayesian information criterion (BIC), which has been used to select  $S$  for hidden Markov models (e.g., Leroux and Puterman 1992). The Schwarz criterion is  $C(S) = \log \hat{\ell} - k_S \log(n)/2$ , where  $\hat{\ell}$  is the likelihood  $p(d_1^n | \theta_s, S)$  maximized over  $\theta_s$ ,  $n$  is the sample size, and  $k_S$  is the number of free parameters in  $\theta_s$ . The Schwarz criterion (or, equivalently, BIC) assumes that  $p(S)$  is uniform and approximates  $p(\theta_s | d_1^n, S)$  with a multivariate normal to be integrated over using Laplace's approximation (Tierney and Kadane 1986).

Figure 10 compares MCMC estimates of  $p(S | d_1^n)$  with BIC for Markov–Poisson HMMs applied to the fetal lamb data. The two methods lead to different conclusions. BIC suggests a two-state model, but the three-state model is most likely under  $p(S | d_1^n)$ . BIC penalizes the larger models more than it should based on Figure 10(a), suggesting that the sample size of 240 is not large enough to justify BIC's asymptotics. Figure 11 shows the marginal posterior densities of  $q(r, s)$  under the two- and three-state Markov–Poisson models applied to the fetal lamb data. Indeed, BIC's normality assumptions are violated by asymmetry, heavy tails, and proximity to boundaries. The posterior density of  $q(1, 1)$  when  $S = 3$ , shown in the central plot of Figure 11(b), is a specific example. It is highly probable that  $q(1, 1)$  lies in a narrow interval centered around .93, although there is a small but appreciable probability that it falls somewhere between 0 and .8. Roughly speaking, BIC approximates the posterior density of  $q(1, 1)$  by a normal distribution centered at .93 with variance chosen to cover the “narrow interval.” BIC integrates over the approximating normal density, ignoring the probability mass between 0 and .8.

#### 4.2 Variable-Dimension Monte Carlo

Variable-dimension Monte Carlo (VDMC) methods are sampling schemes that allow  $S$ , and thus the dimension of  $\theta$ , to vary within the sampling algorithm so that the stationary distribution of  $S$  is  $p(S | d_1^n)$ . The growing list of VDMC algorithms that have been applied to finite mixture models includes jump diffusions (Grenander and Miller 1994;

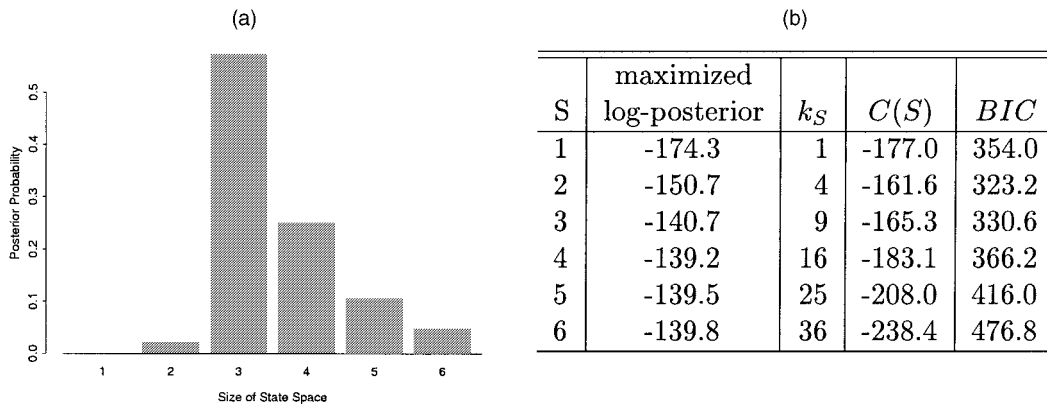


Figure 10. Comparing  $P(S | D)_1^n$  and BIC for Markov–Poisson Mixtures Applied to the Fetal Lamb Data. (a) Posterior probability of  $S$  based on an MCMC run of 30,000 iterations; (b) maximized log posterior, number of free parameters, Schwarz criterion, and BIC for each  $S$ . The two methods give different conclusions.

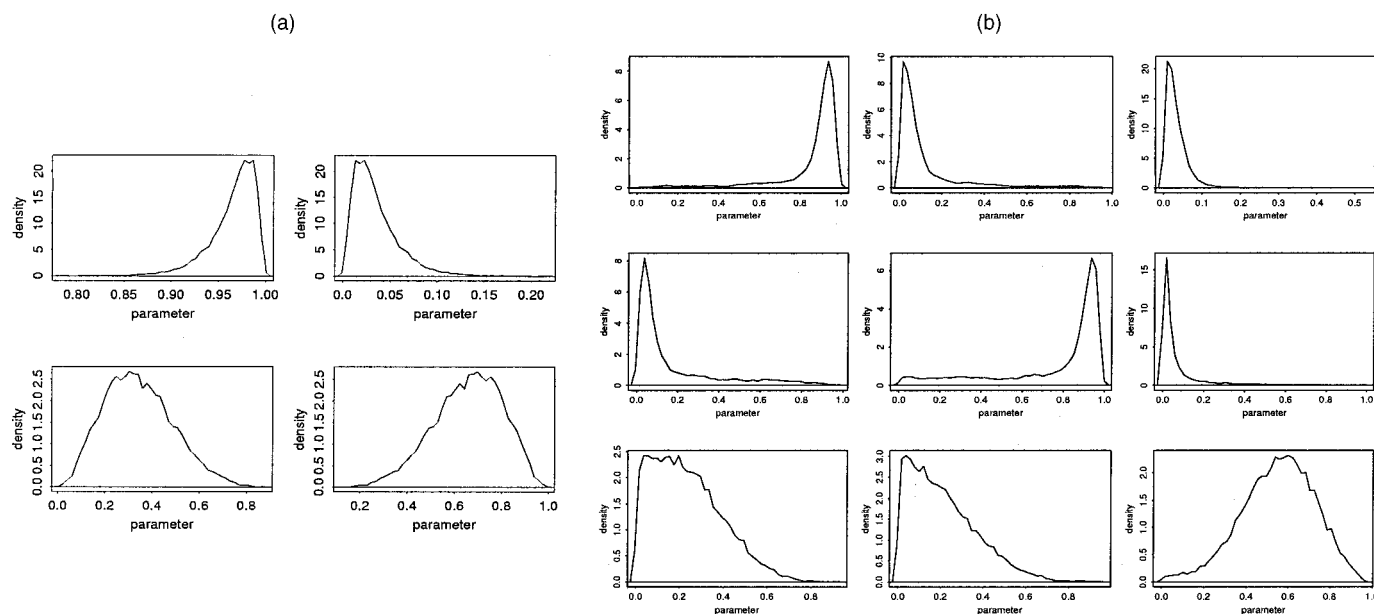


Figure 11. Marginal Posterior Densities for Elements of  $\mathbf{Q}$  Under the Two-State (a) and Three-State (b) Markov-Poisson HMMs Applied to the Fetal Lamb Data. The figures are arranged in the same order as elements of  $\mathbf{Q}$ , so that the first row, second column shows the posterior density of  $q(0,1)$ . Results are based on MCMC runs of 10,000 iterations, with the first 1,000 deleted. The densities are far from normal, especially in (b).

Phillips and Smith 1996), reversible-jump MCMC (Green 1995; Richardson and Green 1997), and birth-death sampling (Stephens 2000a). Reversible-jump MCMC (Robert et al. 2000) and birth-death samplers (Cappé et al. 2001) have been applied to HMMs. Jump diffusions could be applied similarly. Reversible-jump MCMC is distinct from the other two methods, because it accommodates sampling  $\mathbf{h}$  and  $\theta$  together (although a referee has noted that the birth-death sampler could be expanded to sample  $\mathbf{h}$  as well). The forward-backward recursions can improve reversible-jump MCMC by speeding mixing for  $\theta_S$  and  $\mathbf{h}$  during periods where  $S$  remains constant. Currently existing jump diffusion and birth-death algorithms integrate over  $\mathbf{h}$  much like the Metropolis-Hastings algorithm discussed in Section 2. Recursive computing is a requirement for these methods, because they must be able to evaluate the likelihood of proposed jumps, births, and deaths.

A difference between (21) and VDMC is the definition of the parameter space. If  $\theta_s \in \Theta_s$  for  $s \in \{1, \dots, S_{\max}\}$ , then (21) simulates  $\theta$  from  $\Theta = \Theta_1 \times \dots \times \Theta_{S_{\max}}$ . VDMC samples from  $\Theta^* = \Theta_1 \cup \dots \cup \Theta_{S_{\max}}$ . The distinction between  $\Theta$  and  $\Theta^*$  is important, because the scientific interpretation of  $\theta$  often changes with its dimension, in much the same way as regression coefficients are interpreted differently when new variables are added to a regression model.

As  $S_{\max}$  grows, the need to run  $S_{\max}$  Gibbs samplers in (21) becomes a burden. VDMC methods require careful tuning and clever choices of potential moves between parameter spaces. In return, they escape the computational burden of (21) for large  $S_{\max}$  by rarely sampling  $(S, \theta)$  from regions of small posterior probability. The advantages of (21) are that it is easily implemented when  $S_{\max}$  is small and it computes  $p(S | d_1^n, \theta^{(j)})$  directly at each iteration rather than by simulation, much in the spirit of Section 3.1.

## 5. DIAGNOSTICS

This section briefly discusses diagnostics for MCMC convergence and model adequacy. Cowles and Carlin (1996) and Mengersen, Robert, and Guihenneuc-Jouyaux (1999) have reviewed the large body of work on MCMC convergence diagnostics. Robert et al. (1999) dealt specifically with HMMs. Most MCMC convergence diagnostics seek to determine whether the distribution of some function  $g(\theta)$  has stabilized under the long-run frequency properties of the MCMC chain. The function  $g$  is typically left unspecified. The user may either apply the diagnostic to each component of  $\theta$  or select  $g$  to be some global summary of the model parameters. In many ways, the most natural global summary of a multidimensional parameter  $\theta$  is its value under its posterior distribution  $p(\theta | d_1^n)$ . Thus, a final benefit of the likelihood and forward-backward recursions is that they provide a global summary of the model that may be monitored to deduce MCMC convergence. Figure 12 shows log-posterior values from the FB sampler applied to the two- and three-state Markov-Poisson HMMs for the fetal lamb data. Both samplers rapidly achieve regions of high posterior density. The second mode apparent in Figure 12(b) is responsible for the nonnormal posterior distribution of  $\mathbf{Q}$  noted in Figure 11. Frequent switching between modes in Figure 12(b) suggests that the mixing weights for the modes are well estimated.

The posterior distribution of  $\mathbf{h}$  can be an excellent diagnostic for model adequacy when the hidden states are interpretable. If the hidden states are not behaving as expected, then the model probably needs more structure. Figure 13(a) shows empirical Bayes probabilities  $\pi'_i(h_i = 1 | \theta)$  for the fraud detection model from Example 3 applied to an account known to not contain fraud. The hidden states are supposed to represent a criminal's presence or absence. Instead,

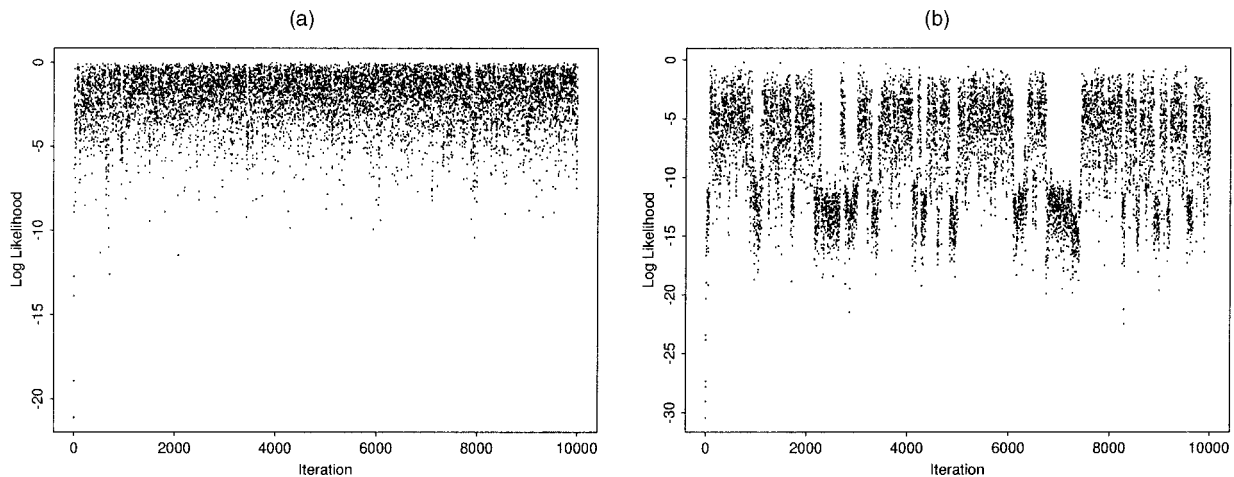


Figure 12. Log-Posteriors From Gibbs Runs of Two-State (a) and Three-State (b) Markov-Poisson HMMs on Fetal Lamb Data. Only the first 10,000 values are shown, but log-posterior values are scaled by the maximum achieved in the full run of 100,000 iterations. The chains for both models converge very quickly.

Figure 13(a) suggests that they describe whether the company's office is closed or open for business. An expanded model that considers the company's hourly and daily calling behavior produces the more believable estimates shown in Figure 13(b).

Appropriate diagnostics for model adequacy will depend on the application at hand, but two generally useful diagnostics are the autocorrelation function (ACF), advocated by MacDonald and Zucchini (1997), and the predictive distribution of the data, advocated by Robert et al. (1999). The ACF can be used to determine whether the model requires additional structure. For example, the model fit to the U.S. GNP data in Sections 2 and 3 ignores possible dependence in the residuals ( $d_t - \mu_{h_t}$ ), which are functions of  $\theta$ . Figure 14(a) shows the posterior distribution of the ACF for the residuals. All marginal distributions of the ACF easily cover 0, suggesting that dependence in the residuals is weak enough to be safely ignored.

The predictive distribution of the data,  $p(d | d_1^n)$ , is often used as a diagnostic for finite mixture models. HMMs have a more complicated predictive distribution because of the added time dimension. As with finite mixture models, the marginal distribution of  $d_t$  may be obtained by weighting  $P_0, \dots, P_{S-1}$  according to the stationary distribution of  $\mathbf{Q}$ . More complicated summaries of  $p(d | d_1^n)$  may be achieved by posterior predictive checks (Rubin 1984; Meng 1994; Gelman, Carlin, Stern, and Rubin 1995). That is, for each  $\theta^{(j)}$  drawn by the MCMC algorithm, simulate a dataset  $d^{(j)}$  from  $p(d_1^n | \theta^{(j)})$  and compute the summary for  $d^{(j)}$ . For example, the fetal lamb dataset ends with a run of 0's much longer than other such runs in the dataset. Figure 14(b) calculates the posterior predictive distribution of the longest run of 0's for the two- and three-state Poisson models. The three-state model is much more capable of producing long runs of 0's than the two-state model.

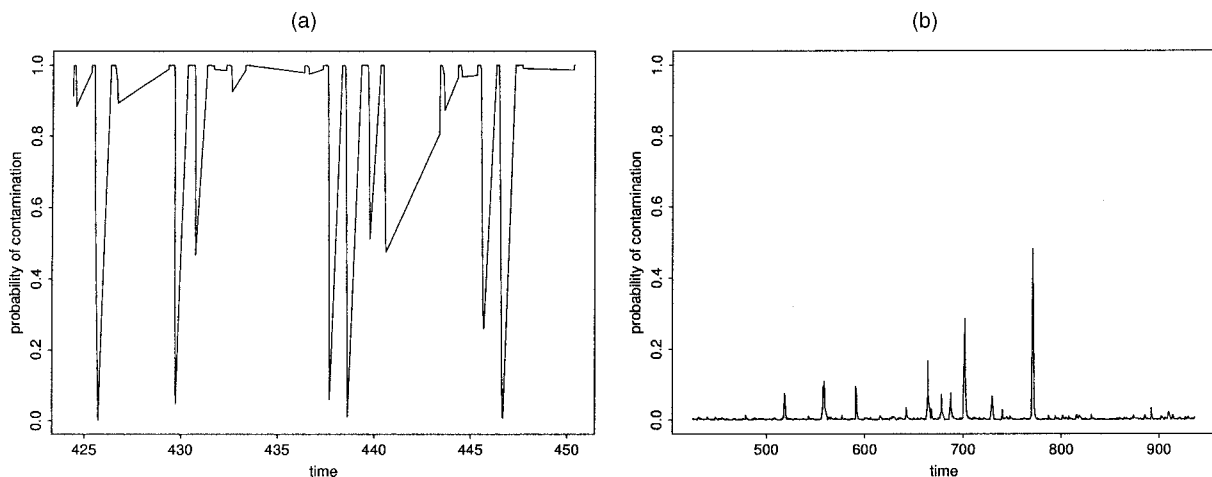


Figure 13. Empirical Bayes Probability of Criminal Activity as a Function of Time for an Account Known to not Contain Fraud. (a) Based on the HMM discussed in Example 3; (b) a more elaborate model considering the accounts hourly and daily calling patterns. Panel (b) more closely matches the intended behavior of the model. Time is measured in days since January 1, 1994. Panel (a) is shown on a restricted time scale to prevent overplotting.

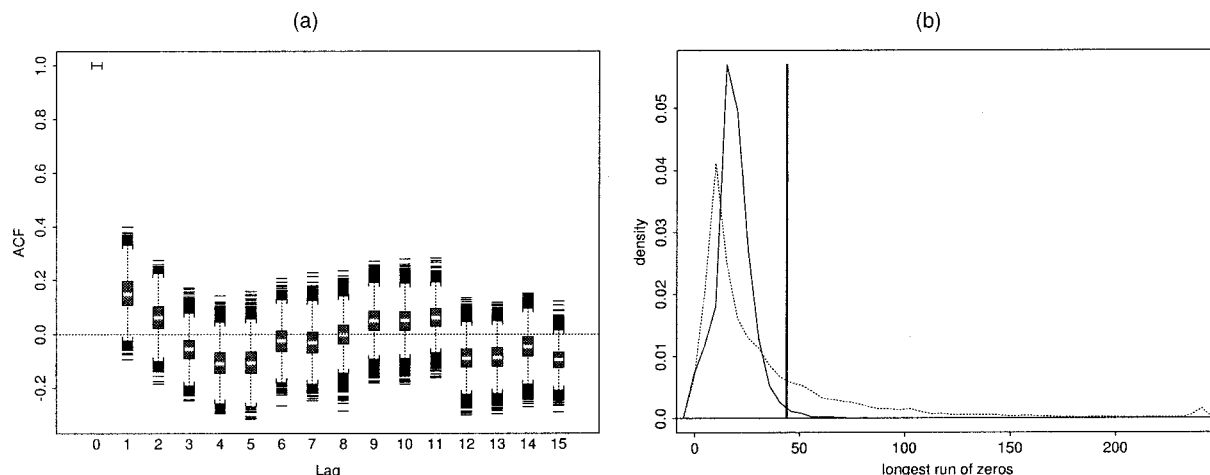


Figure 14. (a) Posterior Distribution of the Autocorrelation Function for the Residuals From the Two-Parameter Gaussian Model Applied to the U.S. GNP Data From Figure 4 and (b) Posterior Predictive Distribution for the Longest Run of 0's for the Two-State (—) and Three-State (....) Poisson Models Applied to the Fetal Lamb Data. The vertical line indicates the longest run of 0's in the data.

## 6. CONCLUSION

MCMC allows implementation of HMMs without using recursive computing, but the likelihood, forward-backward, and Viterbi recursions bring a richness to the models that would not otherwise exist. The forward-backward recursions lead to a Gibbs sampler that mixes faster than its natural competitor, and the likelihood recursion opens the door to more general samplers that would be impossible without a tractable method for computing HMM likelihoods. The nonstochastic backward recursion allows Bayes estimates of  $\mathbf{h}$  to be Rao-Blackwellized and empirical Bayes estimates of  $\mathbf{h}$  to be used when MCMC is not an option. The Viterbi algorithm allows approximate MAP reconstructions of  $\mathbf{h}$  to be constructed much more efficiently than can be done by naive MCMC sampling. The likelihood recursion is indispensable for classical Bayesian model selection, either by MCMC or by approximate methods, and it allows the use of VDMC techniques that depend on likelihood. Likelihood also provides an automatically computed summary of model parameters that can be monitored to deduce MCMC convergence. Finally, Section 2.3 illustrates that the ability to draw  $\mathbf{h}$  directly from  $p(\mathbf{h} | d_1^n, \theta)$  can simplify the task of drawing other missing data which is no part of the hidden Markov chain.

A second theme of the article, which has nothing to do with recursive computing, is that there are advantages to understanding the physical meaning of  $\mathbf{h}$ . Selecting appropriate diagnostics and preventing label switching and are both simplified through scientific understanding of  $\mathbf{h}$ . Informative prior assumptions are needed to defeat the intrinsic identifiability problems that cause label switching. The assumptions may be applied directly through the prior distribution, or indirectly by reparameterizing the model or choosing loss functions for estimation after an unidentified MCMC run. Intuition about how  $\mathbf{h}$  should behave allows the use of the posterior distribution of  $\mathbf{h}$  as a model diagnostic.

I conclude with a note about the historical significance of the article by Baum et al. (1970). That article contained two innovations developed to implement HMMs: an inequality and a recursion. Both foreshadowed much more general

techniques. Dempster et al. (1977) showed that the inequality could be applied to a much broader set of problems when they introduced the general EM algorithm. The recursion is a special case of a general local computation algorithm for graphical models, an area of current research in both statistics and computer science. The computing landscape has changed considerably since the work of Baum et al. (1970), but their ideas remain relevant in today's MCMC world.

[Received August 1999. Revised July 2001.]

## REFERENCES

- Albert, J. H., and Chib, S. (1993), "Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts," *Journal of Business and Economic Statistics*, 11, 1–15.
- Andrieu, C., and Doucet, A. (1999), "Joint Bayesian Model Selection and Estimation of Noisy Sinusoids via Reversible Jump MCMC," *IEEE Transactions on Signal Processing*, 47, 2667–2676.
- Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, 41, 164–171.
- Besag, J., and Green, P. (1993), "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society, Ser. B*, 55, 25–37.
- Cappé, O., Robert, C. P., and Rydén, T. (2001), "Reversible Jump MCMC Converging to Birth-and-Death MCMC and More General Continuous Time Samplers," technical report, CREST, [www.ceremade.dauphine.fr/~xian/ctrjrmcmc.ps.gz](http://www.ceremade.dauphine.fr/~xian/ctrjrmcmc.ps.gz).
- Casella, G., and Robert, C. P. (1996), "Rao-Blackwellisation of Sampling Schemes," *Biometrika*, 83, 81–94.
- Celeux, G., Hurn, M., and Robert, C. P. (2000), "Computational and Inferential Difficulties With Mixture prior distributions," *Journal of the American Statistical Association*, 95, 957–970.
- Chib, S. (1996), "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models," *Journal of Econometrics*, 75, 79–97.
- Churchill, G. A. (1989), "Stochastic Models for Heterogeneous DNA Sequences," *Bulletin of Mathematical Biology*, 51, 79–94.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems*, New York: Springer.
- Cowles, M. K., and Carlin, B. P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 883–904.
- Damian, P., Wakefield, J., and Walker, S. (1999), "Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical models by Using Auxiliary Variables," *Journal of the Royal Statistical Society, Ser. B*, 61, 331–344.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–22.

- Diebolt, J., and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society, Ser. B*, 56, 363–375.
- Everitt, B. S., and Hand, D. J. (1981), *Finite Mixture Distributions*, New York: Chapman and Hall.
- Fredkin, D. R., and Rice, J. A. (1992), "Bayesian Restoration of Single-Channel Patch Clamp Recordings," *Biometrics*, 48, 427–448.
- Ge, X., and Smyth, P. (2000), "Deformable Markov Model Templates for Time-Series Pattern Matching," Technical Report 00-10, University of California Irvine, Dept. of Information and Computer Science. <http://www.ics.uci.edu/~datalab/papers/trseghmm.pdf>.
- Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, New York: Chapman and Hall.
- Gilks, W. R., and Roberts, G. O. (1996), "Strategies for Improving MCMC," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman and Hall, pp. 89–114.
- Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.
- Grenander, U., and Miller, M. I. (1994), "Representations of Knowledge in Complex Systems," *Journal of the Royal Statistical Society, Ser. B*, 56, 549–581.
- Guttorp, P. (1995), *Stochastic Modeling of Scientific Data*, New York: Chapman and Hall.
- Hamilton, J. D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.
- (1990), "Analysis of Time Series Subject to Changes in Regime," *Journal of Econometrics*, 45, 39–70.
- Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.
- Higdon, D. M. (1998), "Auxiliary Variable Methods for Markov Chain Monte Carlo With Applications," *Journal of the American Statistical Association*, 93, 585–595.
- Jaung, B. H., and Rabiner, L. R. (1991), "Hidden Markov Models for Speech Recognition," *Technometrics*, 33, 251–272.
- Kalman, R. E. (1960), "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, 82, 35–45.
- Kass, R. E., and Raftery, A. E. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.
- Lauritzen, S. L., and Spiegelhalter, D. J. (1988), "Local Computations With Probabilities on Graphical Structures and Their Application to Expert Systems," *Journal of the Royal Statistical Society, Ser. B*, 50, 157–194.
- Leroux, B. G., and Puterman, M. L. (1992), "Maximum-Penalized-Likelihood Estimation for Independent and Markov-Dependent Mixture Models," *Biometrics*, 48, 545–558.
- 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., Neuwald, A. F., and Lawrence, C. E. (1999), "Markovian Structures in Biological Sequence Alignments," *Journal of the American Statistical Association*, 94, 1–15.
- 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.
- (1995), "Covariance Structure and Convergence Rate of the Gibbs Sampler With Various Scans," *Journal of the Royal Statistical Society, Ser. B*, 57, 157–169.
- MacDonald, I. L., and Zucchini, W. (1997), *Hidden Markov and Other Models for Discrete-Valued Time Series*, New York: Chapman and Hall.
- Meng, X.-L. (1994), "Posterior Predictive  $p$ -Values," *The Annals of Statistics*, 22, 1142–1160.
- Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999), "MCMC Convergence Diagnostics: A Review," in *Bayesian Statistics 6*, eds. J. Berger, J. Bernardo, A. Dawid, D. Lindley, and A. Smith, New York: Oxford University Press, pp. 415–440.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1092.
- Phillips, D. B., and Smith, A. F. (1996), "Bayesian Model Comparison via Jump Diffusions," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman and Hall, pp. 215–239.
- Richardson, S., and Green, P. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components," *Journal of the Royal Statistical Society, Ser. B*, 59, 731–792.
- Robert, C. P. (1998), "MCMC Specifics for Latent Variable Models," in *COMPSTAT—Proceedings in Computational Statistics*, 13th Symposium, Heidelberg: Physica-Verlag, pp. 101–112.
- Robert, C. P., Celeux, G., and Diebolt, J. (1993), "Bayesian Estimation of Hidden Markov Chains: A Stochastic Implementation," *Statistics & Probability Letters*, 16, 77–83.
- Robert, C. P., Rydén, T., and Titterton, D. M. (1999), "Convergence Controls for MCMC Algorithms, With Applications to Hidden Markov Chains," *Journal of Statistical Computation and Simulation*, 64, 327–355.
- (2000), "Bayesian Inference in Hidden Markov Models Through the Reversible Jump Markov Chain Monte Carlo Method," *Journal of the Royal Statistical Society, Ser. B*, 62, 57–76.
- Robert, C. P., and Titterton, D. M. (1998), "Reparametrization Strategies for Hidden Markov Models and Bayesian Approaches to Maximum Likelihood Estimation," *Statistics and Computing*, 8, 145–158.
- Romberg, J. K., Choi, H., and Baraniuk, R. G. (2001), "Bayesian Tree-Structured Image Modeling Using Wavelet-Domain Hidden Markov Models," *IEEE Transactions on Image Processing*, 10, 1056–1068.
- Rubin, D. B. (1984), "Bayesian Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.
- Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2000), "Bayesian Segmentation of Protein Secondary Structure," *Journal of Computational Biology*, 7, 233–248.
- (2001), "Bayesian Protein Structure Prediction," in *Case Studies in Bayesian Statistics*, Vol. 5, eds. C. Gatsonis, R. E. Kass, B. Carlin, A. Carriquiry, A. Gelman, B. Virdinelli, and M. West, New York: Springer.
- Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.
- Scott, S. L. (1999), "Bayesian Analysis of a Two-State Markov Modulated Poisson Process," *Journal of Computational and Graphical Statistics*, 8, 662–670.
- (2002), "Detecting Network Intrusion Using the Markov Modulated Nonhomogeneous Poisson Process," *Journal of the American Statistical Association*, to come.
- Stephens, M. (1997), Discussion of "On Bayesian Analysis of Mixtures With an Unknown Number of Components," by S. Richardson and P. Green, *Journal of the Royal Statistical Society, Ser. B*, 59, 768–769.
- (2000a), Bayesian Analysis of Mixtures With an Unknown Number of Components—An Alternative to Reversible Jump Methods," *The Annals of Statistics*, 28, 40–74.
- (2000b), "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society, Ser. B*, 62, 795–810.
- Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.
- Titterton, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.
- van Dyk, D. A., and Meng, X.-L. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–50.
- Viterbi, A. J. (1967), "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, 13, 260–269.
- West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, New York: Springer.
- Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, New York: Wiley.