Statistics & Probability Letters 16 (1993) 77-83
North-Holland

4 January 1993

Bayesian estimation of hidden Markov chains: A stochastic implementation

Christian P. Robert
LSTA, Université Pierre et Marie Curie, Paris, France

Gilles Celeux
INRIA, Rocquencourt, France

Jean Diebolt
LSTA, Université Pierre et Marie Curie, Paris, France

Received April 1992
Revised June 1992

Abstract: Hidden Markov models lead to intricate computational problems when considered directly. In this paper, we propose an approximation method based on Gibbs sampling which allows an effective derivation of Bayes estimators for these models.

AMS 1990 Subject Classifications: 62M05, 62F15, 62-04.

Keywords: Gibbs sampling; forward-backward recursion formula; ergodicity; stochastic restoration; geometric convergence.

1. Introduction

Hidden Markov models have come recently under closer scrutiny, because they provide a handy extension of i.i.d. mixture models and thus allow for a more accurate modeling of clearly dependent phenomena (see, e.g., Rabiner, 1989, or Titterington, 1990). For instance, they have been used in speech and character recognition, leading to significant improvements in the success rates (see, e.g., Juang and Rabiner, 1991; and Kundu and He, 1991), or in DNA recognition (Churchill, 1989). For a usual mixture model, the observations originate *independently* from the distribution with density

$$\phi(y \mid \theta) = \sum_{k=1}^{K} p_k f(y \mid \xi_k), \tag{1.1}$$

where $f(\cdot \mid \xi)$ belongs to a given parametrized family, the weights $p_k > 0$ add to one and $\theta = (\xi_1, \dots, \xi_K, p_1, \dots, p_{K-1})$. The hidden Markov extension removes the independence assumption by considering that successive observations y_i from (1.1) are correlated through the component k from which they originate. More formally, we can associate to each observation y_i a *missing value* indicator z_i, which represents the component from which y_i is generated, i.e. $z_i = k$ if $y_i \sim f(y \mid \xi_k)$; the assumption made on the z_i's is then that they constitute a Markov chain with transition matrix $\boldsymbol{P} = (p_{k,m})$, with $p_{k,m} = P(z_i = m \mid z_{i-1} = k)$, instead of being generated independently from $(p_1, \dots, p_K)$ as in the inde-

Correspondence to: Prof. Christian P. Robert, LSTA, Boîte 158, Université Paris 6, 4, Place Jussieu, 75252 Paris Cedex 05, France.

pendent mixture model. And, conditionally on the z_i's, the y_i's are independent. As for usual mixture models, one difficulty about inference on hidden Markov chains is that the indicators z_i's are missing (or *hidden*). Obviously, extensions could be considered, where the dependency is modeled through a higher-order Markov chain. For instance, as pointed out in Kundu and He (1991), speech models are usually well-represented by second-order Markov chains. But, notwithstanding the notational complexity of such extensions, they allow for approximation algorithms which are very similar to the one we discuss below.

Parameter estimation in hidden Markov models usually relies on maximum likelihood or Bayesian approaches, moments methods being totally untractable in this setting, but they also face important implementation problems since the dependency structure can only exacerbate the difficulties met in regular mixture estimation (see Titterington et al., 1985); or Diebolt and Robert, 1992). For instance, the EM algorithm (Dempster, Laird and Rubin, 1977) was originally tailored for missing data structures but the dependency between the z_i's adds another problem to the usual drawbacks of EM for mixture estimation (Biscarat, Celeux and Diebolt, 1992), namely the use of a recurrent 'forward–backward' formula which is time-consuming and numerically sensitive, even though adapted algorithms have been designed (see Derin and Elliott, 1987; Devijver, 1985; or Guédon and Cocozza-Thivent, 1990). As pointed out by Qian and Titterington (1991a), a usually tractable alternative to EM is to replace the E-step by a *stochastic restoration* step (see also Robert, 1991). However, while the simulation of the missing data is straightforward for independent structures like (1.1), it is quite difficult to simulate from the distribution of $z = (z_1, \ldots, z_n)$ conditional on $y = (y_1, \ldots, y_m)$, $g(z \mid y, \theta)$, in hidden Markov models. Similarly, Bayesian estimation is already delicate for (1.1) (see Diebolt and Robert, 1992) and the difficulty obviously increases by a factor of magnitude for hidden Markov models.

The approach we advocate in this paper is to provide an efficient Bayesian estimation of the model through *Gibbs sampling*. In an image processing setup, Geman and Geman (1984) introduced Gibbs sampling for Markovian models in a simulated annealing scheme, obtaining pointwise convergence results for this method in finite state structures. Diebolt and Robert (1992) implemented Gibbs sampling for i.i.d. replications from (1.1) with f in an exponential family, leading to considerable improvements in the Bayesian estimation of mixtures; it indeed allows estimation for samples where the implementation of the Bayesian paradigm was previously impossible. In this paper, we generalize their methods to hidden Markov models and derive geometric convergence results to the posterior distribution and related posterior quantities. Cancelling the call to time-consuming forward–backward recursion formulae thus leads to the first effective general Bayesian identification of hidden Markov chains.

Gibbs sampling is an increasingly popular simulation method based on Markov chain theory. Instead of simulating directly from a distribution π, it generates a Markov chain $z^{(m)}$ with stationary distribution π. The sample $z^{(m)}, z^{(m+1)}, \ldots$ thus generated is no longer i.i.d. but an extension of the law of large numbers, *the ergodic theorem*, allows for an approximation of any posterior quantity of interest based upon this sample. (See Tanner (1991), Casella and George (1992), Smith and Roberts (1993), Besag and Green (1993) and Gilks et al. (1993) for details on both implementation and foundations of Gibbs sampling.) In our particular setup, we can show in addition that the convergence to the stationary distribution is uniformly geometric, by a *duality principle* exhibited in Diebolt and Robert (1992), and thus derive more precise conditions for convergence or, equivalently, for validation of the Gibbs approximation.

2. Bayesian estimation

The model introduced in (1.1) is therefore made of r.v.'s $x_i = (y_i, z_i)$ such that $z_i \in \{1, \ldots, K\}$, $z = (z_1, \ldots, z_n)$ is a first-order Markov chain with transition matrix P, i.e.

$$P(z_i = l \mid z_{i-1}, \ldots, z_1) = P(z_i = l \mid z_{i-1}) = p_{z_{i-1},l}, \quad i = 2, \ldots, n,$$

and the y_i's are independent conditionally on the z_i's, with

$$y_i \mid z \sim f(y_i \mid \xi_{z_i}).$$

Only the y_i's are observed and we choose $z_1 = 1$ for identifiability reasons. For simplicity's sake, we also assume that the parametrized densities $f(\cdot \mid \xi)$ belong to an exponential family, i.e. it satisfies $f(y \mid \xi) = h(y)\exp(\xi \cdot t(y) - \psi(\xi))$. In this case, there exists a conjugate prior on ξ, $\pi(\xi)$. The prior distribution we will consider on the rows $\boldsymbol{p}_k$ of the transition matrix $\boldsymbol{P}$ is a product of Dirichlet priors, $\mathscr{D}(\alpha_1^k, \ldots, \alpha_K^k)$, with independence between the rows ($1 \leqslant k \leqslant K$). (A still tractable extension would be to consider a hyperprior on the α^k's.)

Given this setup, the posterior distribution on $\theta = (\xi_1, \ldots, \xi_K, \boldsymbol{P})$ is

$$\pi(\theta \mid \boldsymbol{y}) \propto \left(\sum_{i_2=1}^{K} \sum_{i_3=1}^{K} \cdots \sum_{i_n=1}^{K} p_{1,i_2} f(y_2 \mid \xi_{i_2}) p_{i_2,i_3} f(y_3 \mid \xi_{i_3}) \cdots p_{i_{n-1},i_n} f(y_n \mid \xi_{i_n}) \right) \times f(y_1 \mid \xi_1) \prod_{k=1}^{K} \left\{ \pi(\xi_k) \left(\prod_{l=1}^{K} p_{k,l}^{\alpha_l^k - 1} \right) \right\} \tag{2.1}$$

and is therefore untractable for most values of n, since it involves a sum of K^{n-1} terms. However, the conditional posterior density $\pi(\theta \mid \boldsymbol{y}, \boldsymbol{z})$ is much simpler, since

$$\pi(\theta \mid \boldsymbol{y}, \boldsymbol{z}) \propto \prod_{k=1}^{K} \left\{ \pi(\xi_k) \left(\prod_{l=1}^{K} p_{k,l}^{\alpha_l^k - 1} \right) \right\} \left\{ \prod_{i=2}^{n} p_{z_{i-1},z_i} f(y_i \mid \xi_{z_i}) \right\} f(y_1 \mid \xi_1). \tag{2.2}$$

In particular, we get the following conditional distribution for $\boldsymbol{p}_k = (p_{k,1}, \ldots, p_{k,K})$,

$$\boldsymbol{p}_k \sim \mathscr{D}\left(\alpha_1^k + \sum_{i=2}^{n} \mathbb{I}_{\{z_{i-1}=k\}} \mathbb{I}_{\{z_i=1\}}, \ldots, \alpha_K^k + \sum_{i=2}^{n} \mathbb{I}_{\{z_{i-1}=k\}} \mathbb{I}_{\{z_i=K\}} \right), \tag{2.3}$$

which is straightforward to simulate. Similarly, in the special case when $f(y \mid \xi_k)$ is the density of $\mathscr{N}(\xi_k, 1)$ and the conjugate prior is $\mathscr{N}(\mu_k, 1)$, we have

$$\xi_k \mid \boldsymbol{y}, \boldsymbol{z} \sim \mathscr{N}\left(\frac{\sum_{i=1}^{n} \mathbb{I}_{\{z_i=k\}} y_i + \mu_k}{\sum_{i=1}^{n} \mathbb{I}_{\{z_i=k\}} + 1}, \frac{\sum_{i=1}^{n} \mathbb{I}_{\{z_i=k\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{z_i=k\}} + 1} \right).$$

Contrary to the independent mixture setting, the conditional density

$$g(\boldsymbol{z} \mid \boldsymbol{y}, \theta) \propto \prod_{i=2}^{n} p_{z_{i-1},z_i} f(y_i \mid \xi_{z_i}) \mathbb{I}_{\{1,\ldots,K\}}(z_i) \tag{2.4}$$

is quite involved and simulation from (2.4) requires the time-consuming 'forward–backward' recurrence formulae (see Qian and Titterington, 1991b). This difficulty also prohibits the use of *data augmentation* as in Tanner and Wong (1987), i.e. to simulate according to $g(\boldsymbol{z} \mid \boldsymbol{y}, \theta)$, since

$$P(z_n = j \mid \boldsymbol{y}, \theta, z_{n-1}, \ldots, z_1) = \frac{p_{z_{n-1},j} f(y_n \mid \xi_j)}{\sum_{t=1}^{K} p_{z_{n-1},t} f(y_n \mid \xi_t)},$$

$$P(z_{n-1} = j \mid \boldsymbol{y}, \theta, z_{n-2}, \ldots, z_1) = \frac{p_{z_{n-2},j} f(y_{n-2} \mid \xi_j) \sum_{l=1}^{K} p_{j,l} f(y_{n-1} \mid \xi_l)}{\sum_{t=1}^{K} \sum_{l=1}^{K} p_{z_{n-2},t} p_{t,l} f(y_{n-2} \mid \xi_t) f(y_{n-1} \mid \xi_l)}$$

and complexity grows at each step. However, the conditional distributions are much easier to deal with, since, for $1 < i < n$,

$$g(z_i \mid y, \theta, z_{j \neq i}) = g(z_i \mid y_i, \theta, z_{i-1}, z_{i+1}) = \frac{p_{z_{i-1},z_i} f(y_i \mid \xi_{z_i}) p_{z_i,z_{i+1}}}{\sum_{j=1}^{K} p_{z_{i-1},j} f(y_i \mid \xi_j) p_{j,z_{i+1}}} \tag{2.5a}$$

and

$$g(z_n \mid y, \theta, z_{j<n}) \propto p_{z_{n-1},z_n} f(y_n \mid \xi_{z_n}). \tag{2.5b}$$

Then, as shown in Geyer (1991), Tierney (1991) or Diebolt and Robert (1992), Gibbs sampling, i.e. the iterative simulation of $\theta^{(m)}$ according to $\pi(\theta \mid y, z^{(m-1)})$ from (2.2) and $z^{(m)}$ according to (2.5) (with $\theta = \theta^{(m)}$), produces an homogeneous Markov chain $(\theta^{(m)}, z^{(m)})$. Since the conditional distributions (2.2) and (2.5) are positive, as we are in an exponential setup, this chain is irreducible and aperiodic. Therefore, it has a unique σ-finite invariant measure and, since $\pi(\theta, z \mid y) = g(z \mid y, \theta)\pi(\theta \mid y)$ is stationary, $(\theta^{(m)}, z^{(m)})$ is ergodic. This result implies that the ergodic theorem applies, namely that we can approximate any quantity of interest

$$\mathbb{E}^{\pi}[h(\theta) \mid y] = \int_{\Theta} h(\theta)\pi(\theta \mid y)\, \mathrm{d}\theta$$

by the average

$$\frac{1}{M} \sum_{m=1}^{M} h(\theta^{(m)})$$

for M large enough, whenever $\int |h(\theta)| \pi(\theta \mid y)\, \mathrm{d}\theta < +\infty$.

In this case, more precise convergence results can be obtained. Indeed, the sequence $(z^{(m)})$ is also a Markov chain, with transition kernel density (i.e. conditional density of $z^{(m+1)} = z'$ given $z^{(m)} = z$)

$$K(z, z') = \int_{\Theta} \pi(\theta \mid y, z) g(z_1' \mid \theta, y, z_2, \ldots, z_n) \cdots g(z_n' \mid \theta, y, z_1', \ldots, z_{n-1}')\, \mathrm{d}\theta. \tag{2.6}$$

Since this chain is irreducible and aperiodic, with a finite state space, it is geometrically ergodic and ϕ-mixing. On the contrary, $(\theta^{(m)})$ *is not a Markov chain* even though the couple $(\theta^{(m)}, z^{(m)})$ is a Markov chain (because the generation of $z^{(m+1)}$ between $\theta^{(m)}$ and $\theta^{(m+1)}$ also depends on $z^{(m)}$). Nevertheless, some properties of the chain $(z^{(m)})$ can be transferred to the sequence $\theta^{(m)}$, owing to the *duality principle* exhibited in Diebolt and Robert (1992). (Chan (1991) also took advantage of this property in a setting where both sequences were Markov chains.) Actually, we deduce the uniform geometric convergence to the marginal posterior distribution $\pi(\theta \mid y)$ of θ from the geometric convergence to the marginal posterior distribution $g(z \mid y)$ of z. We denote by $\pi^m(\theta)$ the density of $\theta^{(m)}$ associated with a starting value $\theta^{(0)}$ (it also depends on y) and $\| \cdot \|_1$ the L^1-norm associated with the Lebesgue or the counting measure, depending on the context.

Theorem 1. (i) *There exist constants $C > 0$ and $0 \leqslant \rho < 1$ such that, for each initial value $\theta^{(0)}$,*

$$\| \pi^m(\cdot) - \pi(\cdot \mid y) \|_1 \leqslant C\rho^m. \tag{2.7}$$

(ii) *For any function $h(\theta)$ such that $\mathbb{E}^{\pi}[|h(\theta)| \mid y] < +\infty$ and any initial density $\pi^0(\theta)$, there exists a constant $C' > 0$ such that*

$$\left| \mathbb{E}^{\pi^m}[h(\theta)] - \mathbb{E}^{\pi}[h(\theta) \mid y] \right| \leqslant C'\rho^m. \tag{2.8}$$

(iii) *The process $(\theta^{(m)})$ is geometrically ϕ-mixing.*

Proof. (i) The finite-state homogeneous Markov chain $(z^{(m)})$ is irreducible and aperiodic since $K(z, z')>0$ for all z and z'. Therefore, $(z^{(m)})$ is uniformly geometrically ergodic and geometrically ϕ-mixing. Moreover, its invariant density is

$$g(z\mid y)=\int_{\Theta} g(z\mid y, \theta)\pi(\theta\mid y)\ \mathrm{d}\theta.$$

Thus, there exist constants C and $0 \leqslant \rho<1$ such that

$$\| g^m(\cdot)-g(\cdot\mid y)\|_1 \leqslant C\rho^m$$

for any initial density $g^0(z)$, where g^m is the density of $z^{(m)}$, which also depends on y. Now, since

$$\pi(\theta\mid y)=\int_{\mathscr{Z}} \pi(\theta\mid y, z) g(z\mid y)\ \mathrm{d}z,$$

we have

$$\begin{aligned}
\|\pi^{m+1}-\pi\|_1 &= \int_{\Theta} |\pi^{m+1}(\theta)-\pi(\theta\mid y)|\mathrm{d}\theta \\
&= \int_{\Theta}\left|\int_{\mathscr{Z}}(\pi(\theta\mid y, z) g^m(z)-\pi(\theta\mid y, z) g(z\mid y))\ \mathrm{d}z\right| \mathrm{d}\theta \\
&\leqslant \int_{\Theta}\int_{\mathscr{Z}} \pi(\theta\mid y, z)|g^m(z)-g(z\mid y)|\ \mathrm{d}z\ \mathrm{d}\theta \\
&= \int_{\mathscr{Z}}|g^m(z)-g(z\mid y)|\int_{\Theta}\pi(\theta\mid y, z)\ \mathrm{d}\theta\ \mathrm{d}z \\
&= \| g^m-g\|_1.
\end{aligned}$$

The inequality (2.7) then follows.

(ii) We have, since $g(z\mid y)>0$ for all $z\in\mathscr{Z}$,

$$\begin{aligned}
\left|\mathbb{E}^{\pi^{m+1}}[h(\theta)]-\mathbb{E}^{\pi}[h(\theta)\mid y]\right| &\leqslant \int_{\Theta}|h(\theta)|\int_{\mathscr{Z}}\pi(\theta\mid y, z)|g^m(z)-g(z\mid y)|\ \mathrm{d}z\ \mathrm{d}\theta \\
&= \int_{\Theta}|h(\theta)|\int_{\mathscr{Z}}\pi(\theta\mid y, z) g(z\mid y)\left|\frac{g^m(z)}{g(z\mid y)}-1\right| \mathrm{d}z\ \mathrm{d}\theta \\
&\leqslant \int_{\Theta}|h(\theta)|\int_{\mathscr{Z}}\pi(\theta\mid y, z) g(z\mid y)\ \mathrm{d}z\ \mathrm{d}\theta\ C_0\left\|\frac{g^m}{g}-1\right\|_1 \\
&= C_0\mathbb{E}^{\pi}[|h(\theta)|\mid y]\left\|\frac{g^m}{g}-1\right\|_1 \\
&\leqslant \frac{C_0}{\min_{\mathscr{Z}} g(z\mid y)}\mathbb{E}^{\pi}[|h(\theta)|\mid y]\ \| g^m-g\|_1 \\
&\leqslant C'\rho^m,
\end{aligned}$$

according to (i).

(iii) Geometric ϕ-mixing is established if we show that there exist ϕ and μ such that

$$|\pi^m(\theta)-\pi(\theta\mid y)| \leqslant \phi(m)\mu(\theta)$$

with ϕ geometrically decreasing to 0 and μ the density of a finite positive measure. Since $z^{(m)}$ is a finite-state irreducible aperiodic Markov chain, it follows from Billingsley (1968) that there exists a geometrically decreasing function ϕ_1 such that

$$|g^m(z) - g(z \mid y)| \leqslant \phi_1(m).$$

Now,

$$\begin{aligned}|\pi^{m+1}(\theta) - \pi(\theta \mid y)| &= \left| \int_{\mathscr{Z}} \pi(\theta^{(m)} \mid y, z^{(m)}) \left[g^m(z^{(m)}) - g(z^{(m)} \mid y) \right] \mathrm{d}z^{(m)} \right| \\ &\leqslant \phi_1(m) \int_{\mathscr{Z}} \pi(\theta \mid y, z^{(m)}) \, \mathrm{d}z^{(m)},\end{aligned}$$

and

$$\mu(\theta) = \int_{\mathscr{Z}} \pi(\theta \mid y, z) \, \mathrm{d}z$$

is the density of a finite measure on Θ since $\mathscr{Z}$ is finite. □

The geometric ϕ-mixing property also implies that a central limit theorem holds. Therefore, it is possible to monitor more carefully the convergence of the average

$$\frac{1}{M} \sum_{m=1}^{M} h(\theta^{(m)})$$

to $\mathbb{E}^{\pi}[h(\theta) \mid y]$ (as M goes to infinity) by estimating its variance σ_h^2 (see Geyer, 1991).

Corollary 2. *For a function h such that* $\mathbb{E}^{\pi}[\,|h(\theta)|^2 \mid y] < +\infty$ *and*

$$0 < \sigma_h^2 = \operatorname{Var}^{\pi}(h(\theta)) + 2 \sum_{t=1}^{+\infty} \operatorname{Cov}^{\pi}(h(\theta^{(0)}), h(\theta^{(t)})) < +\infty,$$

there exists a central limit theorem on the average of the $h(\theta^{(m)})$, *i.e.*

$$\frac{1}{\sqrt{M}} \sum_{m=1}^{M} \left(h(\theta^{(m)}) - \mathbb{E}^{\pi}[h(\theta) \mid y] \right) \to \mathscr{N}(0, \sigma_h^2). \quad \square$$

These results show that the duality principle exhibited in Diebolt and Robert (1992) can be of use in settings other than Data Augmentation. When Gibbs sampling is used on the missing values or, more generally, on the artificial parameters, it still allows for a duality principle, i.e. for the transfer of the properties of the chain $z^{(m)}$ to the process $\theta^{(m)}$, even though the latter is not a Markov chain anymore.

3. Conclusion

Stochastic approaches have been previously proposed by Qian and Titterington (1991) for hidden Markov models under the name of stochastic restoration methods. However, as noticed above, the conditional distributions of the missing values, $g(z \mid y, \theta)$, cannot be expressed analytically and, moreover, does not allow for direct simulation since it involves the troublesome forward–backward formulae. In this paper, we propose a restoration type algorithm for Bayesian estimation which avoids the use of these formulae through Gibbs sampling, thus leading to efficient inference in this setting. We are furthermore able to establish detailed convergence results for the Gibbs sampler. Obviously, similar developments could be considered in a non-Bayesian framework, simulating z along the lines of Wei and Tanner (1990) and Biscarat, Celeux and Diebolt (1992).

References

Besag, J. and P.J. Green (1993), Spatial statistics and Bayesian computation (with discussion), *J. Roy. Statist. Soc. Ser. B* **55**.

Billingsley, P. (1968), *Convergence of Probability Measures* (Wiley, New York).

Biscarat, J.C., G. Celeux and J. Diebolt (1992), Stochastic versions of EM, Tech. Rept. LSTA #154, Univ. Paris VI (Paris).

Casella, G. and E.I. George (1992), Explaining the Gibbs sampler, to appear in: *Amer. Statist.*

Chan, K.S. (1991), Asymptotic behavior of the Gibbs sampler, Tech. Rept. No. 294, Dept. of Statist., Univ. of Chicago (Chicago, IL).

Churchill, G. (1989), Stochastic Models for heterogeneous DNA sequences, *Bull. Math. Biol.* **51**, 79–94.

Dempster, A., N. Laird and D. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. Ser. B* **39**, 1–38.

Derin, H. and H. Elliott (1987), Modelling and segmentation of noisy and textured images using Gibbs random fields, *IEEE Trans. Patt. Anal. Mach. Intell.* **PAMI-9**, 39–55.

Devijver, P.A. (1985), Baum's forward–backward algorithm revisited, *Patt. Recogn. Lett.* **3**, 369–373.

Diebolt, J. and C.P. Robert (1992), Estimation of finite mixture distributions through Bayesian sampling, to appear in: *J. Roy. Statist. Soc. Ser. B.*

Geman, S. and D. Geman (1984), Statistical relaxation, Gibbs distribution and the Bayesian restoration of images, *IEEE Trans. Patt. Anal. Mach. Intell.* **PAMI-6**, 721–741.

Geyer, C. (1991), Markov chain Monte-Carlo maximum likelihood, to appear in: *Computer Sciences and Statistics: Proc. 23d Symp. on the Interface.*

Gilks, W.R., D.G. Clayton, D.J. Spiegelhalter, N.G. Best, A.J. McNeil, L.D. Sharples and A.J. Kirby (1993), Modelling complexity: Applications of Gibbs sampling in medicine (with discussion), *J. Roy. Statist. Soc. Ser. B* **55**.

Guédon, Y. and C. Cocozza-Thivent (1990), Explicit state occupancy modeling by hidden semi-Markov models: application of Derin's scheme, *Comput. Speech Lang.* **4**, 167–192.

Juang, B.H. and L.R. Rabiner (1991), Hidden Markov models for speech recognition, *Technometrics* **33**, 251–272.

Kundu, A. and Y. He (1991), On optimal order in modeling sequence of letters in words of common language as a Markov chain, *Patt. Recogn.* **24**, 603–608.

Qian, W. and D.M. Titterington (1991a), Estimation of parameters in hidden Markov models, *Philos. Trans. Roy. Soc. London A* **337**, 407–428.

Qian, W. and D.M. Titterington (1991b), Parameters estimation for hidden Gibbs chains, *Statist. Probab. Lett.* **10**, 49–58.

Rabiner, L.R. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. of the IEEE* **77**, 257–286.

Robert, C.P. (1991), Comments on Meng and Rubin's paper, in: J. Berger, J. Bernardo, R. Dawid and A.F.M. Smith, eds., *Bayesian Statistics, Vol. 4* (Oxford Univ. Press, Oxford).

Smith, A.F.M. and G.O. Roberts (1993), Bayesian computation via the Gibbs sampler and related Markov chains (with discussion), *J. Roy. Statist. Soc. Ser. B* **55**.

Tanner, M. (1991), *Tools for Statistical Inference* (Springer, New York).

Tanner, M. and W. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion), *J. Amer. Statist. Assoc.* **82**, 528–550.

Tierney, L. (1991), Markov chains for exploring posterior distributions, to appear in: *Computer Sciences and Statistics: Proc. 23d Symp. on the Interface.*

Titterington, D.M. (1990), Some recent research in the analysis of mixture distributions, *Statistics* **21**, 619–641.

Titterington, D.M., A.F.M. Smith, and U. Makov (1985), *Statistical Analysis of Finite Mixture Distributions* (Wiley, New York).

Wei, G.C.G. and M. Tanner (1990), A Monte Carlo implementation of the EM algorithm and the poor's man data augmentation algorithms, *J. Amer. Statist. Assoc.* **85**, 699–704.