J. R. Statist. Soc. B (1994)
56, *No.* 2, *pp.* 363–375

Estimation of Finite Mixture Distributions through Bayesian Sampling

By JEAN DIEBOLT and CHRISTIAN P. ROBERT†

Université Pierre et Marie Curie, Paris, France

[Received November 1990. Final revision January 1993]

SUMMARY

A formal Bayesian analysis of a mixture model usually leads to intractable calculations, since the posterior distribution takes into account all the partitions of the sample. We present approximation methods which evaluate the posterior distribution and Bayes estimators by Gibbs sampling, relying on the missing data structure of the mixture model. The *data augmentation* method is shown to converge geometrically, since a duality principle transfers properties from the discrete missing data chain to the parameters. The fully conditional Gibbs alternative is shown to be ergodic and geometric convergence is established in the normal case. We also consider non-informative approximations associated with improper priors, assuming that the sample corresponds exactly to a k-component mixture.

Keywords: BAYESIAN COMPUTATION; CONJUGATE PRIORS; DATA AUGMENTATION; EM ALGORITHM; GIBBS SAMPLING; MARKOV CHAINS; MONTE CARLO METHOD; NON-INFORMATIVE MODELLING

1. INTRODUCTION

1.1. *Mixture Estimation*

The estimation of the parameters of finite mixture distributions has recently come under close scrutiny (for example, Everitt and Hand (1981), Titterington *et al.* (1985) and West (1992)). Mixture models provide an interesting alternative to nonparametric modelling, while being less restrictive than the usual distributional assumptions. Computational methods have also appeared, including the *EM algorithm* proposed by Dempster *et al.* (1977), which allows for maximum likelihood estimation in models larger than before.

Given a proper prior, a Bayesian approach to the mixture estimation problem always provides estimators which can be written explicitly for conjugate priors. None-the-less, a Bayesian analysis of mixtures has barely been developed, mainly because of major computational obstacles: although closed form expressions are available, the computing time is so prohibitive that they cannot be used for reasonable sample sizes. Other set-ups also lead to difficulties in implementing the Bayesian paradigm and approximations already exist. However, the case of mixtures, where the posterior distribution considers *all* possible partitions of the sample, is very peculiar and calls for adequate approximation methods. In the special case where only the proportions of the mixture are unknown, sequential approximations have been provided by Smith and Makov (1978) and Bernardo and Girón (1988).

†*Address for correspondence*: Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, Paris VI, 4 place Jussieu, 75252 Paris Cedex 05, France.

0035–9246/94/56363

In Sections 3 and 4, we propose two approximations based on *Bayesian sampling* and show that these simulation techniques are justified in a mixture setting, i.e. that they converge to the posterior distribution. We also exhibit geometric convergence for the *data augmentation* algorithm, based on a *duality principle* which requires minimal assumptions about the prior distribution. Section 5 illustrates through two simulations the potential interest of our approximations, since they allow for Bayesian estimation in set-ups where it was previously impossible.

An additional difficulty in the treatment of mixtures is that *improper priors* cannot be used. Here we propose a non-informative approximation based on Jeffreys priors and on the assumption that enough observations have been generated from each component of the mixture for the posterior distribution to be well defined. This approximation can also be implemented by using Bayesian sampling and convergence results still hold.

1.2. *General Solutions*

The mixture model that we consider is given by the observation of n independent random variables $x_1, \ldots, x_n$, from a k-component mixture,

$$f(x_i) = \sum_{j=1}^{k} p_j f_j(x_i), \qquad i=1, \ldots, n, \tag{1.1}$$

where the densities f_j $(1 \leqslant j \leqslant k)$ are known *or* are known up to a parameter and the proportions $0 < p_j < 1$ satisfy $\Sigma_{j=1}^{k} p_j = 1$. We denote the unknown parameters of the mixture by $\theta = (\theta_1, \ldots, \theta_r)$, including $p = (p_1, \ldots, p_k)$. As pointed out in Dempster *et al.* (1977), a mixture model can always be expressed in terms of *missing* (or *incomplete*) *data*. If, for $1 \leqslant i \leqslant n$, z_i is a k-dimensional vector indicating to which component x_i belongs, such that $z_{ij} \in \{0, 1\}$ and $\Sigma_{j=1}^{k} z_{ij} = 1$, the density of the *completed data* (x_i, z_i) is

$$\prod_{j=1}^{k} p_j^{z_{ij}} f_j^{z_{ij}}(x_i). \tag{1.2}$$

From this perspective, the model is *hierarchical* with, on top, the true parameters of the mixture, θ, then the missing data whose distribution depends on θ, $z \sim f(z|\theta)$, and, at the bottom, the observed data x, with distribution depending on z and θ, $x \sim f(x|\theta, z)$.

In this case, the EM method works iteratively, replacing each missing data z_{ij} by its expectation

$$z_{ij}^{(m)} = p_j^{(m)} f_j^{(m)}(x_i) \Big/ \sum_{t=1}^{k} p_t^{(m)} f_t^{(m)}(x_i), \tag{1.3}$$

where m indexes the current iteration step and $p_j^{(m)}$ and $f_j^{(m)}$ are the current evaluations of the parameters. New maximum likelihood estimates are then derived from the *pseudocompleted data* $(x_i, z_i^{(m)})$. To avoid some defects of the EM method, Celeux and Diebolt (1985, 1992) introduced a stochastic version called *SEM*. Instead of estimating the missing data, they simulate the z_i' from their conditional distribution, i.e. multinomial distributions with weights (1.3).

Given a prior distribution π on the parameter θ, a logical Bayesian extension of this technique is to simulate $\theta \sim \pi(\theta|x,z)$, completing the corresponding SEM simulation of z. If we iterate the simulations, this procedure converges to simulate from $\pi(\theta|x)$. This method is now well known as *Gibbs sampling* or Bayesian sampling. Tanner and Wong (1987) introduced it under the name of data augmentation for missing value problems but it has been generalized since into what are called *Markov chain Monte Carlo methods* (Geyer, 1991). Comprehensive descriptions of these techniques are provided in Besag and Green (1993), Smith and Roberts (1993) and Tierney (1993). A recent illustration in the mixture set-up is given in Escobar and West (1991).

1.3. *Bayesian Sampling*

Data augmentation corresponds to the simulation method described above, namely to generate iteratively the parameters $\theta^{(m)}$ and the missing data $z^{(m)}$ according to $\pi(\theta|x,z^{(m)})$ and $f(z|x, \theta^{(m+1)})$. The method then produces a sequence $(\theta^{(m)})$ of parameter evaluations.

In general, Gibbs sampling allows for any possible partition of $\eta=(\theta,z)$ into $(\eta_1, \ldots, \eta_t)$ and then iteratively generates η according to the conditional distributions

$$\eta_1^{(m+1)} \sim \pi(\eta_1|x, \eta_2^{(m)}, \ldots, \eta_t^{(m)}), \ldots, \eta_t^{(m+1)} \sim \pi(\eta_t|x, \eta_1^{(m+1)}, \ldots, \eta_{t-1}^{(m+1)}).$$

Data augmentation thus appears as a special case of Gibbs sampling induced by the natural hierarchical structure of the model, so that each group (z and θ) corresponds to a hierarchical level. To allow for arbitrary grouping obviously leads to a wider range of application, since $\pi(\theta|x,z)$ is not always available.

Bayesian sampling produces a Markov chain $(\theta^{(m)})$, usually ergodic with stationary distribution $\pi(\theta|x)$. Therefore, after m_0 initial steps ('warming up'), the random variables $\theta^{(m_0+1)}, \ldots, \theta^{(m_0+K)}$ can be considered to be approximately distributed according to $p(\theta|x)$. The resulting sample can then lead to an approximation of any well-defined posterior quantity by

$$\mathbf{E}^{\pi}[h(\theta)|x] \approx \frac{1}{K}\sum_{k=1}^{K} h(\theta^{(m_0+k)}), \tag{1.4}$$

by the ergodic theorem. Gelfand and Smith (1990) propose an improvement on this approximation, namely to use instead the average of the conditional expectations $\mathbf{E}^{\pi}[h(\theta)|x, z^{(m_0+k)}]$, based on the *Rao-Blackwell theorem.*

Since the $\theta^{(m_0+k)}$ are correlated, several modifications of the algorithm have been proposed. See Robert (1992), Besag and Green (1993), Smith and Roberts (1993) and Tierney (1993) for detailed discussion on implementation; they will not be considered here because of lack of space.

In this paper, considering specifically mixture distributions, we derive general geometric convergence results for data augmentation from a duality principle which shows that the properties of the finite state Markov chain $(z^{(m)})$ can be transferred to the other chain $(\theta^{(m)})$ without imposing any requirement on the prior π. This principle is a conceptual foundation of the simulation methodology and is of interest in other set-ups as long as the study of the chain $(z^{(m)})$ is straightforward. We also derive ergodicity results for Gibbs sampling, as well as geometric convergence in the normal case. Note that geometric convergence guarantees fast convergence to the posterior distribution and furthermore ensures a central limit theorem (CLT).

2. INFORMATION PARADOX OF BAYESIAN MIXTURE ESTIMATION

2.1. *Conjugate Priors*

In this section, we show that a Bayesian approach can be implemented in a formal way, with closed form expressions for the Bayes estimators, but leads to an 'information paradox', i.e. the fact that, when the number of observations becomes large, even moderately, the Bayes estimators cannot be computed in practice.

Consider the mixture model (1.1) and assume, for simplicity, that the components f_j are all of the same exponential family, namely $f_j(x) = f(x|\gamma_j)$. If $\mathscr{F}_0$ is a class of conjugate priors for $f(x|\gamma)$, the mixtures of products of elements of $\mathscr{F}_0$ form a class of conjugate priors for the estimation of the γ_j. For a given prior distribution, since the likelihood is

$$\prod_{i=1}^{n} \sum_{j=1}^{k} p_j f(x_i|\gamma_j), \tag{2.1}$$

the posterior distribution is a mixture with k^n terms, corresponding to the partitions of $\{x_1, \ldots, x_n\}$ into at most k groups. The conjugate priors on p are Dirichlet distributions, $\mathscr{D}(\alpha_1, \ldots, \alpha_k)$.

2.1.1. *Example 1*

Bernardo and Girón (1988) consider the case where only the weights are unknown. The posterior distribution on p is then

$$\sum_{i_1+\ldots+i_k=n} q_{i_1\ldots i_k}(x)\, \mathscr{D}(\alpha_1+i_1, \ldots, \alpha_k+i_k)$$

with weights

$$q_{i_1\ldots i_k}(x) \propto \binom{\alpha_0+n}{\alpha_1+i_1 \;\ldots\; \alpha_k+i_k}^{-1} \sum_{(k_l)} \prod_{l=1}^{i_1} f_1(x_{k_l}) \prod_{l=i_1+1}^{i_1+i_2} f_2(x_{k_l}) \ldots \prod_{l=n-i_k+1}^{n} f_k(x_{k_l}),$$

where the sum is taken over the partitions (k_l) of $\{1, \ldots, n\}$ into k groups of sizes $i_1, i_2, \ldots, i_k$ and $\alpha_0 = \alpha_1 + \ldots + \alpha_k$. For instance, if $k=2$ and $p \sim \mathrm{Be}(\frac{1}{2}, \frac{1}{2})$, the estimate

$$\hat{p}(x) = \sum_{i=0}^{n} \omega_i(x) \frac{i+\frac{1}{2}}{n+1}$$

is the weighted sum of the possible values $(i+\frac{1}{2})/(n+1)$ according to the posterior probabilities $\omega_i(x)$ that i of the n observations belong to the first population.

2.1.2. *Example 2*

Consider a normal mixture

$$f(x|p, \mu, \sigma) = \sum_{j=1}^{k} p_j \phi(x|\mu_j, \sigma_j),$$

where $\phi(x|\mu_j,\sigma_j)$ is the density of $\mathcal{N}(\mu_j,\sigma_j^2)$. The conjugate prior for (μ_j,σ_j^2) is $\mathcal{N}(\xi_j,\ \sigma_j^2/n_j)$ for $\pi(\mu_j|\sigma_j^2)$ and an *inverted gamma* distribution for σ_j^2, $\mathrm{IG}(\nu_j,s_j^2)$, i.e.

$$\pi(\sigma_j) \propto \sigma_j^{-\nu_j-1}\exp(-s_j^2/2\sigma_j^2), \tag{2.2}$$

where the hyperparameters are known. The posterior distribution of (μ,σ), $\pi(\mu,\ \sigma|x)$, is then a mixture over all the partitions of $\{1, 2, \ldots, n\}$ into at most k groups of

$$\prod_{j=1}^{k}\left\{\mathcal{N}\left(\frac{n_j\xi_j+i_j\bar{x}_j}{n_j+i_j},\ \frac{\sigma_j^2}{n_j+i_j}\right)\mathrm{IG}\left(\nu_j+i_j,\ s_j^2+\hat{s}_j^2+\frac{n_ji_j}{n_j+i_j}(\xi_j-\bar{x}_j)^2\right)\right\},$$

where i_j is the cardinal of the jth member of the partition and, if $i_j \neq 0$, $\bar{x}_j$ and $\hat{s}_j^2$ are respectively the average and the sum of the squared errors for the observations attributed to the jth member. The weights are of the form

$$\prod_{j=1}^{k}\Gamma(\alpha_j+i_j)(n_j+i_j)^{-1/2}\Gamma\left(\frac{\nu_j+i_j}{2}\right)\left\{s_j^2+\hat{s}_j^2+\frac{n_ji_j}{n_j+i_j}(\xi_j-\bar{x}_j)^2\right\}^{-(\nu_j+i_j)/2}, \tag{2.3}$$

modulo a normalization; $\pi(p|x)$ is again a mixture over the partitions of the distributions $\mathcal{D}(\alpha_1+i_1, \ldots, \alpha_k+i_k)$. The posterior distribution thus computes the probability that a given partition is the correct partition and, conditionally on this partition, it independently actualizes the distributions of the parameters of each component.

2.2. *Improper Priors*

Mixture models are not *identifiable*: if p_j is the weight of the jth component, a sample of size n contains no observation from this component with probability $(1-p_j)^n$. Therefore, *mixture models do not allow for improper priors.*

We circumvent this major difficulty in the extension to non-informative settings by using regular Jeffreys priors and by rejecting from the posterior mixture every partition where too few observations become allocated to some components. Such an *ad hoc* approach is necessary for using improper priors, unless we obtain some observations from each component in addition. An alternative approximation would be to use restricted parameter spaces with (truncated) Jeffreys priors, even though the choice of the bounds may sometimes be delicate. The following example illustrates the implementation of our modification.

2.2.1. *Example 3*

Consider the normal mixture of example 2. A possible non-informative prior is

$$\pi(\mu,\sigma,p) = \prod_{j=1}^{k}\frac{1}{\sigma_j^2}\mathbf{I}_{\{p_j\leqslant p_{j+1}\}}\mathbf{I}_{\{\Sigma_{l=1}^{k}p_l=1\}}, \tag{2.4}$$

if $\mathbf{I}$ denotes the indicator function and $p_{k+1}=1$. We have to exclude from the marginal distribution of x partitions which attribute fewer than two observations to any component. The *pseudoposterior distribution* $\tilde{\pi}\{(\mu,\sigma,p)|x\}$ is then a weighted sum over all the *remaining* partitions of products of $\mathcal{N}(\bar{x}_j,\ \sigma_j^2/i_j)$ ×

IG$(i_j, \hat{s}_j^2)$ and of $\mathscr{D}(i_1+1, \ldots, i_k+1)$. The weights can be derived from expression (2.3) via appropriate modifications of the hyperparameters.

The computational problems related to the use of these pseudoposteriors are obviously the same as in the informative case, since the number of terms in the posterior mixture is roughly the same.

3. DATA AUGMENTATION

3.1. *The Algorithm*

In this section, we describe a first approximation of the posterior distribution. Starting with an initial value $\theta^{(0)}$, the algorithm is run the following way: at step m,

$$\begin{aligned} &\text{(a) generate } z^{(m)} \sim f(z|x, \theta^{(m)}), \\ &\text{(b) generate } \theta^{(m+1)} \sim \pi(\theta|x, z^{(m)}). \end{aligned} \tag{3.1}$$

The implementation is straightforward when only the weights have to be estimated: the former step corresponds to the generation of the missing values, according to a multinomial distribution with probabilities (1.3), and the latter step to the generation of the weights under the distribution

$$p^{(m+1)} \sim \mathscr{D}\left(\alpha_1 + \sum_{i=1}^{n} z_{i1}^{(m)}, \ldots, \alpha_k + \sum_{i=1}^{n} z_{ik}^{(m)}\right).$$

When θ involves other parameters than the weights p_j, the second step is usually broken into several unidimensional simulations.

3.1.1. *Example 2 (continued)*

For a k-component normal mixture with prior distributions

$$p \sim \mathscr{D}(\alpha_1, \ldots, \alpha_k), \qquad \mu_j|\sigma_j^2 \sim \mathscr{N}(\xi_j, \sigma_j^2/n_j), \qquad \sigma_j^2 \sim \mathrm{IG}(\nu_j, s_j^2),$$

the mth simulation step for the posterior distribution $\pi\{(\mu, \sigma, p)|x, z^{(m)}\}$ is

(a) generate $p \sim \mathscr{D}(\alpha_1 + i_1, \ldots, \alpha_k + i_k)$,

(b) for $j = 1, \ldots, k$,

(i) generate $\sigma_j^2 \sim \mathrm{IG}\left(\nu_j + i_j,\ s_j^2 + \hat{s}_j^2 + n_j i_j \left\{\sum_{i=1}^{n} z_{ij}^{(m)}(x_i - \xi_j)\right\}^2 \Big/ (n_j + i_j)\right)$,

(ii) generate $\mu_j \sim \mathscr{N}\left(\dfrac{n_j\xi_j + \sum_{i=1}^{n} z_{ij}^{(m)} x_i}{n_j + i_j}, \dfrac{\sigma_j^{(m+1)2}}{n_j + i_j}\right)$.

Therefore, data augmentation produces a sequence $(\theta^{(m)})$ which is a Markov chain with stationary distribution $\pi(\theta|x)$. Thus, if we are mainly interested in parameter estimation, the ergodic theorem gives an approximation of $\mathrm{E}^{\pi}[h(\theta)|x]$ by averaging the $h(\theta^{(m)})$; *Rao–Blackwellization* substitutes the conditional expectations $\mathrm{E}^{\pi}[h(\theta)|x, z^{(m)}]$ to $h(\theta^{(m)})$.

3.2. *Convergence*

We now consider the convergence properties of algorithm (3.1). Since it can be described as the iterative action of two *dual* Markov kernels with densities

$$H(z'|z) = \int f(z'|x,\theta)\,\pi(\theta|x,z)\,\mathrm{d}\theta, \qquad K_{\mathrm{T}}(\theta'|\theta) = \int \pi(\theta'|x,z) f(z|x,\theta)\,\mathrm{d}z,$$

the posterior distributions at step m, π^m and f^m, are derived from these kernels:

$$\begin{aligned} \pi^m(\theta'|x) &= \int K_{\mathrm{T}}(\theta'|\theta)\,\pi^{m-1}(\theta|x)\,\mathrm{d}\theta, \\ f^m(z'|x) &= \int H(z'|z) f^{m-1}(z|x)\,\mathrm{d}z. \end{aligned} \tag{3.2}$$

It is straightforward to see that $f(z|x)$ and $\pi(\theta|x)$ are *stationary points* for the transformations (3.2). They are actually the unique invariant distributions.

Theorem 1. The sequences $(z^{(m)})$ and $(\theta^{(m)})$ are ergodic Markov chains, with respective invariant distributions $f(z|x)$ and $\pi(\theta|x)$. Moreover, the convergence is uniformly geometric, i.e there exist $0 < \rho < 1$ and $C > 0$ such that

$$\int_\Theta |\pi^m(\theta|x) - \pi(\theta|x)|\,\mathrm{d}\theta \leqslant C\rho^m.$$

Similarly, there is geometric convergence for every function h such that $\mathbf{E}^\pi[\|h(\theta)\|\;|x] < \infty$, i.e. there exists $C_h > 0$ such that

$$\|\mathbf{E}^{\pi^m}[h(\theta)|x] - \mathbf{E}^\pi[h(\theta)|x]\| \leqslant C_h\rho^m.$$

A detailed proof of this result is given in Diebolt and Robert (1990). Its main feature is that *it does not require any condition on the prior* π other than $p_i^{(m)} > 0$ for every i, and this is always the case for Dirichlet priors. This theorem relies on a duality principle, relating the distributions of the two chains $z^{(m)}$ and $\theta^{(m)}$ by

$$\pi^m(\theta|x) = \int \pi(\theta|x,\, z) f^m(z|x)\,\mathrm{d}z.$$

Owing to this duality, most properties of $(z^{(m)})$ can be transferred to $(\theta^{(m)})$. For instance, geometric convergence of the second chain is derived from

$$\begin{aligned} \int_\Theta |\pi^m(\theta|x) - \pi(\theta|x)|\,\mathrm{d}\theta &\leqslant \int_\Theta\int_{\mathcal{Z}} |f^m(z|x) - f(z|x)|\,\pi(\theta|x,\, z)\,\mathrm{d}z\,\mathrm{d}\theta \\ &= \int_{\mathcal{Z}} |f^m(z|x) - f(z|x)|\,\mathrm{d}z \end{aligned}$$

and geometric convergence of $(z^{(m)})$. In fact, it follows from Billingsley (1968) that, since $(z^{(m)})$ has a finite support and is aperiodic and irreducible, it is geometrically ergodic and ϕ mixing. The duality principle can also be exhibited for other Markov chain simulation methods.

3.3. *Extensions*

If, instead of approximating $\mathbf{E}^{\pi^m}[\theta|x]$ as in Tanner and Wong (1987), we are averaging some elements of the generated chain as in approximation (1.4) (or by Rao–Blackwellization), the ergodic theorem overcomes the correlations between the $\theta^{(m)}$.

Corollary 1. For $(\theta^{(m)})$ in its stationary regime and h such that $\mathbf{E}^{\pi}[\|h(\theta)\|\,|x] < \infty$, $M^{-1}\Sigma_{1\leqslant m\leqslant M}h(\theta_i^{(m)})$ converges almost surely to $\mathbf{E}^{\pi}[h(\theta_i)|x]$. Moreover,

$$\tau_i = \mathrm{var}^{\pi}(\theta_i|x) + 2\sum_{l=1}^{+\infty}(\mathbf{E}^{\pi}[\theta_i^{(0)}\theta_i^{(l)}|x] - \mathbf{E}^{\pi}[\theta_i^{(0)}|x]^2) < \infty$$

and, if $\tau_i > 0$, a CLT holds:

$$\frac{1}{\sqrt{M}}\sum_{m=1}^{M}(\theta_i^{(m)} - \mathbf{E}^{\pi}[\theta_i|x]) \overset{\mathscr{L}}{\to} \mathscr{N}(0, \tau_i).$$

The CLT follows from the ϕ-mixing property of the Markov chain $(z^{(m)})$ (see Diebolt and Robert (1990)). It also appears in Hastings (1970) in a discrete setting.

3.4. *Non-informative Settings*

The above results hold when the posterior distributions, $f(z|x,\theta)$ and $\pi(\theta|x,z)$, are proper distributions. However, as noted in Section 2.2, these distributions are not defined for improper priors. The suggested approximation to non-informative posteriors is given by

$$\tilde{\pi}_A(\theta,z|x) = \pi(\theta)f(x,z|\theta)\,\mathbf{I}_A(z)\Big/\int\int_A \pi(\theta)f(x,z|\theta)\,\mathrm{d}z\,\mathrm{d}\theta,$$

where A is defined so that $\int\pi(\theta)f(x,z|\theta)\,\mathrm{d}\theta$ is finite on A. We can then simulate from the truncated version of $f(z|x,\theta^{(m)})$ and from $\pi(\theta|x,\ z^{(m)})$ and data augmentation is still valid here, in that $(\theta^{(m)})$ is ergodic with unique invariant distribution

$$\tilde{\pi}_A(\theta|x) = \int_A \tilde{\pi}_A(\theta,z|x)\,\mathrm{d}z$$

and theorem 1 and corollary 1 still hold.

Obviously, this approach can be questioned since $\tilde{\pi}_A$ does not correspond to any prior. However, data augmentation can be used here with no modification. Moreover, it appeared in the simulations that we ran that, when the components are not too intricate, truncation never works (in that no extreme imputation is ever sampled). This indicates that $\tilde{\pi}_A$ is a good approximation to the 'true' posterior distribution, $\pi_A(\theta|x)$, derived from the modified sampling distribution

$$f_A(x,\ z|\theta) = f(x,\ z|\theta)\,\mathbf{I}_A(z)/P_\theta(z\in A), \tag{3.3}$$

which cannot be easily simulated by Bayesian sampling. For instance, for a two-component normal mixture with unknown weights and means, $P_\theta(z\in A)$ is $1-\{p^n+(1-p)^n\}$, essentially equal to 1 for most p. Furthermore, a posterior distribution similar to $\tilde{\pi}_A$ occurs in discrimination cases, where some additional observations are attached to each component of the mixture. (See Lavine and West (1991) for extensions in this discrimination setting.)

3.4.1. *Example 3 (continued)*

For a two-component normal mixture, the non-informative priors are $\pi(p) = \mathbf{I}_{[0,1/2]}(p)$ and $\pi(\mu_j, \sigma_j^2) = 1/\sigma_j^2$. Then, for a generation of $z^{(m)}$ such that each component contains at least *two* observations, we have the following conditional posterior distributions:

$$p|x, z^{(m)} \sim \mathrm{Be}(i_1, i_2)\, \mathbf{I}_{[0,1/2]}(p), \qquad \sigma_j^2|x, z^{(m)} \sim \mathrm{IG}(i_j - 1, \hat{s}_j^2),$$
$$\mu_j|x,\ z^{(m)},\ \sigma_j^{(m+1)} \sim \mathscr{N}(\bar{x}_j, \sigma_j^{(m+1)2}/i_j).$$

4. GIBBS SAMPLING GENERALIZATION

4.1. *The Algorithm*

Gibbs sampling generalizes data augmentation by increasing the number of conditional simulations, i.e. by breaking down θ into subvectors $\theta_1, \ldots, \theta_s$. At each step, the generation of θ_j is conditional on all the other parameters whereas, for data augmentation, there is a dichotomy between z and θ. Starting with an initial value $\theta^{(0)}$, the algorithm simulates the following way: at step m,

$$\begin{array}{ll} \text{(a)} & \text{generate } z^{(m)} \sim f(z|x, \theta^{(m)}), \\ & \quad \text{(1)} \;\; \text{generate } \theta_1^{(m+1)} \sim \pi(\theta_1|x, z^{(m)}, \theta_2^{(m)}, \ldots, \theta_s^{(m)}), \\ & \quad \;\;\vdots \\ & \quad (s) \;\; \text{generate } \theta_s^{(m+1)} \sim \pi(\theta_s|x, z^{(m)}, \theta_1^{(m+1)}, \ldots, \theta_{s-1}^{(m+1)}). \end{array} \tag{4.1}$$

Although the method is equivalent to data augmentation in terms of computation time and even identical for $s = 1$, the convergence study of Gibbs sampling is more difficult and produces less complete results. Moreover, Gelfand and Smith (1990) have pointed out that this algorithm does not perform as well as data augmentation.

4.1.1. *Example 4*

Consider again a k-component normal mixture with conjugate priors. The simulation steps are the same as in data augmentation, except for the scale factors σ_j^2, which are generated according to

$$\mathrm{IG}\left(\nu_j + \sum_{i=1}^{n} z_{ij}^{(m)} + 1, s_j^2 + \sum_{i=1}^{n} z_{ij}^{(m)} (x_i - \mu_j^{(m+1)})^2 + n_j (\xi_j - \mu_j^{(m+1)})^2\right).$$

A similar change takes place in the non-informative case.

4.2. *Convergence*

The convergence results that we obtain for this algorithm do not go as far as for data augmentation. This difficulty with Gibbs sampling was already apparent in Geman and Geman (1984) and we cannot use their results since they consider a finite state model. In our case, the above simulation steps can be described through Markov kernel densities. In particular, the transition kernel from $\theta^{(m)}$ to $\theta^{(m+1)}$ is

$$K_G(\theta'|\theta) = \int f(z'|x,\theta)\,\pi(\theta_1'|x, z', \theta_2, \ldots, \theta_s)\,\pi(\theta_2'|x, z', \theta_1', \theta_3, \ldots, \theta_s) \ldots$$
$$\pi(\theta_s'|x, z', \theta_1', \ldots, \theta_{s-1}')\,\mathrm{d}z'.$$

However, $(z^{(m)})$ *is not a Markov chain* in general. Thus, the duality principle on which Section 3 relies does not apply here.

The following result gives general convergence properties of the Gibbs sampling algorithm. More precise characterizations can only be obtained for specific mixtures (see Section 4.3).

Proposition 1. If the conditional distributions in algorithm (4.1) are positive almost entirely on the parameter space, $(\theta^{(m)})$ is a homogeneous ergodic Markov chain, with invariant density $\pi(\theta|x)$. If $\int\|h(\theta)\|\pi(\theta|x)\,\mathrm{d}\theta < \infty$, then, for almost every $\theta^{(0)}$,

$$\mathbf{E}^{\pi^m}[h(\theta)\,|x,\theta^{(0)}] \xrightarrow[m\to\infty]{} \mathbf{E}^{\pi}[h(\theta)\,|x], \tag{4.2}$$

where π^m is the density of $\theta^{(m)}$ if $\theta^{(0)}$ has been taken as the initial value, and also

$$\frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)}) \xrightarrow[M\to\infty]{} \mathbf{E}^{\pi}[h(\theta)\,|x] \qquad \text{almost surely.} \tag{4.3}$$

This result follows from the irreducibility of $(\theta^{(m)})$, since $\pi(\theta|x)$ is the unique invariant distribution (see Diebolt and Robert (1990)). The ergodic theorem justifies our approximations, since a convergence result similar to result (4.3) holds for Rao–Blackwellization.

4.3. *Normal Mixtures*

Additional results can be obtained if we consider the special case introduced in example 4 under the supplementary assumption that $\sigma_j \in [\sigma_{\min}, \sigma_{\max}]$. Truncated conjugate priors are still conjugate and thus allow for explicit expressions. In practice, sampling can be done by generating $\sigma_j^{(m)}$ from the posterior distribution until $\sigma_j^{(m)} \in [\sigma_{\min}, \sigma_{\max}]$. Note that $\sigma_{\min}$ can be chosen to be very small and $\sigma_{\max}$ very large.

Proposition 2. If the conditional marginal densities defining $K_G(\theta'|\theta)$ are those considered in example 4, $(\theta^{(m)})$ is geometrically ergodic and ϕ mixing.

Therefore, under the assumption $\sigma_j \in [\sigma_{\min}, \sigma_{\max}]$, Gibbs sampling is also geometrically convergent. It is possible to obtain similar results for other functions of θ and a CLT as well, and they also hold for Rao-Blackwellized estimators (Diebolt and Robert, 1990). Schervish and Carlin (1990) obtained an L_2 geometric convergence result whereas Roberts and Polson (1990) provide L_1 geometric convergence for the Gibbs sampler under conditions on K_G which are difficult to verify in usual settings. (See also Tierney (1993).)

5. SIMULATION RESULTS

In this section, we briefly discuss the practical problems related to the sampling methods described above. As an illustration, we consider an ‘informative’ and a 'non-informative' normal example. In each case, we compare data augmentation and Gibbs sampling with the 'ideal' EM solution, which gives the maximum likelihood consistent solution (obtained by using the true θ as starting value). This is not a Monte Carlo study; we only consider the results of the Bayes approach for a *single* simulated sample. However, replications of these experiments have shown the reliability of the sampling methods. In general, stationarity occurs rather quickly and $m_0 = 100$ is a reasonable upper bound for the warming-up process. The number of iterations is $M = 200$.

5.1. *Example 5*

We simulated the following informative case,

$$x \sim 0.5\{\phi(x|0, 1) + \phi(x|2, 1)\},$$

with the prior distributions $\mathcal{N}(0, \sigma_1^2/2)$ and $\mathcal{N}(2, \sigma_2^2/2)$ on the means and non-informative priors on the other parameters (Table 1). Furthermore, we initiated the parameters at $\mu_1^{(0)} = 0$, $\sigma_1^{(0)} = 2$, $p^{(0)} = 0.5$, $\mu_2^{(0)} = 4$ and $\sigma_2^{(0)} = 2$. The estimates of the parameters for the 50 *complete* observations were $\hat{\mu}_1 = 0.2$, $\hat{\sigma}_1 = 1$, $\hat{p} = 0.48$, $\hat{\mu}_2 = 1.8$ and $\hat{\sigma}_2 = 0.7$.

Here, data augmentation is clearly preferable to Gibbs sampling. This shows that the additional variability of this method is a drawback for small sample sizes. When the sample size increases, the two estimators are essentially the same.

5.2. *Example 3 (Continued)*

As pointed out earlier, the non-informative case has to be considered with particular care. Using the algorithms introduced previously, we observed that the estimators were quite unstable and even, in intricate cases, that they were converging to degenerate solutions of the likelihood equations. Therefore, to produce more reliable estimators, we had to modify the original algorithms.

The first step was to reduce the variability of the iterations by replacing the simulations of $\sigma_j^{(m)}$ by a direct attribution through the posterior conditional mean, i.e.

$$\sigma_j^{(m)} = \frac{\Gamma(i_j/2)}{\Gamma\{(i_j+1)/2\}} \left\{ \sum_{i=1}^{n} z_{ij}^{(m-1)} (x_i - \bar{x}_j)^2 \right\}^{1/2},$$

TABLE 1
Approximated Bayes estimators for a sample of size 50, compared with the maximum likelihood estimator†

Method	μ_1	σ_1	p	μ_2	σ_2
Data augmentation	0.198	0.88	0.57	2.02	0.55
	(0.23)	(0.19)	(0.1)	(0)	(0.12)
Gibbs sampling	0.46	1.07	0.73	1.98	0.32
	(0.35)	(0.23)	(0.2)	(0.11)	(0.3)
Maximum likelihood estimation	0.32	0.92	0.61	2.11	0.49

†Standard deviations are given in parentheses.

for data augmentation, with $\mu_j^{(m)}$ replacing $\bar{x}_j$ for Gibbs sampling.

Another modification was suggested by these simulations. It appeared that the restriction on the weight, $p \leqslant \frac{1}{2}$, although theoretically necessary for identifiability, could lead to poor estimates of p. As shown in the previous sections, the Markov chain generated by the algorithm is ergodic and the stationary distribution is independent of the initial values of the parameters. However, if these initial values are in the 'wrong order' (e.g. $\mu_1 < \mu_2$ whereas $\mu_1^{(0)} > \mu_2^{(0)}$), the stabilizing steps are slow to converge and, for some intricate cases, cannot invert this order and lead to degenerate maximum likelihood estimators. Therefore, we withdrew the order condition on p and obtained a noticeable improvement in the results. This shows that the choice of the initial values of the parameters reflects somewhat a prior specification that is incompatible with the order restriction.

Table 2 gives an example for a sample of size 100. The original parameters were $\mu_1=0$, $\sigma_1=1$, $p=0.2$, $\mu_2=2$ and $\sigma_2=0.5$. The estimated parameters of the completed sample were $\hat{\mu}_1=0.1$, $\hat{\sigma}_1=0.9$, $\hat{p}=0.2$, $\hat{\mu}_2=2$ and $\hat{\sigma}_2=0.4$. The initial values were $\mu_1^{(0)}=-5$, $\sigma_1^{(0)}=5$, $p^{(0)}=0.5$, $\mu_2^{(0)}=5$ and $\sigma_2^{(0)}=5$ for both methods.

For this sample, the different methods lead to equivalent results. This is not very surprising with regard to the stabilizing modifications introduced in the two Bayesian algorithms. Owing to these modifications, the standard errors are not theoretically reliable for building confidence intervals.

6. CONCLUSION

We have shown that Bayesian sampling provides a valid and practical solution to the problem of Bayesian estimation for *finite mixture distributions*. It allows statisticians to overcome the computational inapplicability of formal Bayes estimators, while maintaining the strength of the Bayesian approach. Our theoretical developments reinforce previous work on the validity and the convergence of Bayesian sampling. We show that the duality principle leads to stronger and more general results about the convergence of the simulated Markov chains and of the related moments. In many cases, Bayesian sampling provides an efficient and easily implementable alternative when exact Bayes estimators cannot be obtained. It will therefore be used extensively in the near future.

Some aspects of Bayesian sampling obviously have to be studied further. First, although satisfactory convergence results for the Markov chain $(\theta^{(m)})$ have been

TABLE 2

Approximated Bayes estimators for a non-informative prior and a sample of size 100, compared with the maximum likelihood estimator†

Method	μ_1	σ_1	p	μ_2	σ_2
Data augmentation	−0.19 (0.14)	0.51 (0.05)	0.17 (0.04)	1.96 (0.05)	0.45 (0.01)
Gibbs sampling	−0.19 (0.13)	0.53 (0.05)	0.17 (0.04)	1.91 (0.05)	0.45 (0.01)
Maximum likelihood estimation	−0.2	0.49	0.17	1.97	0.45

†Standard deviations are given in parentheses.

obtained here, a stationarity stopping rule has still be to designed. Secondly, the importance of our non-informative approach has still to be assessed, in comparison with alternative methods.

ACKNOWLEDGEMENTS

We are grateful to G. Celeux and C. Goutis for their valuable comments. The suggestions of the referees led to this much shortened and improved version of the paper.

REFERENCES

Bernardo, J. M. and Girón, J. (1988) A Bayesian analysis of simple mixture problems. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 67-78. Oxford: Oxford University Press.

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc.* B, **55**, 25-37.

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

Celeux, G. and Diebolt, J. (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Q.*, **2**, 73-82.

——(1992) A stochastic approximation type EM algorithm. *Stochastics*, to be published.

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

Diebolt, J. and Robert, C. P. (1990) Bayesian estimation of finite mixture distributions: part II, Sampling implementation. *Technical Report 111.* Laboratoire de Statistique Théorique et Appliquée, Université Paris VI, Paris.

Escobar, M. D. and West, M. (1991) Bayesian prediction and density estimation. *J. Am. Statist. Ass.*, to be published.

Everitt, B. and Hand, D. (1981) *Finite Mixture Distributions.* London: Chapman and Hall.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398-409.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721-740.

Geyer, C. J. (1991) Markov chain Monte Carlo maximum likelihood. In *Computer Science and Statistics: Proc. 23rd Symp. Interface* (ed. E. M. Keramidas). Fairfax Station: Interface Foundation.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains, and their applications. *Biometrika*, **57**, 97-109.

Lavine, M. and West, M. (1991) Bayesian calculations for normal mixtures. *Can. J. Statist.*, to be published.

Robert, C. P. (1992) *L'Analyse Statistique Bayésienne.* Paris: Economica.

Roberts, G. and Polson, N. (1990) A note on the geometric convergence of the Gibbs sampler. *Technical Report.* Department of Mathematics, University of Nottingham, Nottingham.

Schervish, M. J. and Carlin B. P. (1990) On the convergence of successive substitution sampling. *J. Comput. Graph. Statist.*, **1**, 111-127.

Smith, A. F. M. and Makov, U. E. (1978) A quasi-Bayes sequential procedure for mixtures. *J. R. Statist. Soc.* B, **40**, 106-112.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3-23.

Tanner, M. D. and Wong, W. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Ass.*, **82**, 528-550.

Tierney, L. (1993) Markov chains for exploring posterior distributions. *Ann. Statist.*, to be published.

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

West, M. (1992) Modelling with mixtures. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.