Computational and Inferential Difficulties With Mixture Posterior Distributions

Gilles CELEUX, Merrilee HURN, and Christian P. ROBERT

This article deals with both exploration and interpretation problems related to posterior distributions for mixture models. The specification of mixture posterior distributions means that the presence of $k!$ modes is known immediately. Standard Markov chain Monte Carlo (MCMC) techniques usually have difficulties with well-separated modes such as occur here; the MCMC sampler stays within a neighborhood of a local mode and fails to visit other equally important modes. We show that exploration of these modes can be imposed using tempered transitions. However, if the prior distribution does not distinguish between the different components, then the posterior mixture distribution is symmetric and standard estimators such as posterior means cannot be used. We propose alternatives for Bayesian inference for permutation invariant posteriors, including a clustering device and alternative appropriate loss functions.

KEY WORDS: Classification; Label switching; Loss functions; Markov chain Monte Carlo; Nonidentifiability; Simulated tempering.

1. INTRODUCTION

Consider the mixture distribution

$$f_{\xi}(x) = \sum_{j=1}^{k} p_j f(x|\zeta_j), \tag{1}$$

where $\xi = (\zeta_1, \ldots, \zeta_k, p_1, \ldots, p_k)$, where the weights satisfy $p_j \geq 0$ with $p_1 + \cdots + p_k = 1$ and the $f(\cdot|\zeta_j)$'s come from some parametric family. An identifiability problem stems from the invariance of (1) under permutation of the indices. This problem is well recognized in the literature (e.g., Titterington, Smith, and Makov 1985) and is usually solved by imposing an identifying ordering constraint on the parameters; for instance, $p_1 \geq \cdots \geq p_k$ (Richardson and Green 1997; Stephens 1997). The identifiability constraint may even be the starting point for a noninformative modelling (e.g., Mengersen and Robert 1996; Roeder and Wasserman 1997). The computational difficulties resulting from working with a posterior distribution based on (1) are also well charted, and various Markov chain Monte Carlo (MCMC) strategies have been proposed (Diebolt and Robert 1994; Liu, Liang, and Wong 1998). Inference on the parameters is then derived from the MCMC sample, with the Bayesian estimates constructed as ergodic averages using diagnostics of convergence analysed by Guihenneuc, Knight, Mengersen, Richardson, and Robert (1999).

Unfortunately, there are difficulties with the model representation at both the *exploration* stage and the *interpretation* stage. First, any influence of the parameter ordering for identifiability is hard to assess and is certainly less benign than previously thought, because it has a bearing on the design and performance of the MCMC sampler and on the resulting inference. Second, exploration of the posterior support by standard MCMC samplers may either be too *local,* in the sense that the MCMC sample is unable to visit the whole range of the posterior modes, or too *unstable,* in the sense that the observations are allocated to all of the components of the model in the course of the simulation, resulting in similar estimates for the (p_j, ζ_j)'s. Such difficulties are particularly obvious when symmetric priors are used on the parameter pairs (p_j, ζ_j), as done by Diebolt and Robert (1994) and Richardson and Green (1997), because the posterior distribution is then also invariant to permutation of the indices. The interpretation of the posterior distribution is thus delicate, because any inference through posterior means is inappropriate, as these means arise through the marginal distributions.

Monitoring of the MCMC sample usually shows that it does not explore the $k!$ equivalent modes of the posterior distribution because the estimates of the marginal posterior distributions are strongly asymmetric in most cases. This is unsatisfactory, because it presents a rare setting where we know that some regions of the parameter space have not been visited by the Markov chain. An equivalent perspective is to consider that these revealed symmetries induce basic control checks that are such that the usual samplers are never stopped against these checks in most cases. Although we may be somewhat presumptuous, we consider that almost the entirety of MCMC samplers implemented for mixture models has failed to converge!

Although from a statistical viewpoint, exploration of the $k!$ modal regions of the posterior distribution is redundant (because the indices j of the components are *not* identifiable), the picture provided by the MCMC sample may be incomplete; that is, the empirical distribution of this sample may be quite different from the true posterior distribution, because the Markov chain almost never leaves the vicinity of one particular local mode. It thus covers only a fraction

Gilles Celeux is Project Manager, INRIA Rhône-Alpes, 38330 Grenoble, France (E-mail: *gilles.celeux@inria.fr*). Merrilee Hurn is Lecturer, Department of Mathematics, University of Bath, BA2 7AY, U.K. (E-mail: *mah@maths.bath.ac.uk*). Christian P. Robert is Professor, Laboratoire de Statistique, CREST-ENSAE, Timbre J340, 92245 Malakoff Cedex, France (E-mail: *robert@ensae.fr*). This work was partially supported by the TMR Network, contract C.E. CT 96-0095. The authors are grateful to the participants in the workshops "McCube" and "MCMC'Ory" for their helpful comments. They also thank Radford Neal for his detailed comments and helpful suggestions on an earlier version of the article, as well as the associate editor and both referees for improving the presentation and helping them focus on major issues.

of the support of the true posterior, even under the identifiability constraint.

For this reason, we feel that truncations of the parameter space may jeopardize the resulting inference (contrary to previous perspectives adopted in Gruet, Philippe, and Robert 1999 and Robert and Mengersen 1998). Even when the truncation does not modify the prior distribution (which is not exactly the case in Mengersen and Robert 1996 or Roeder and Wasserman 1997), the choice of which parameter is used for ordering (i.e., in the Gaussian case, weight vs. mean vs. variance) leads to a particular partition of the complete posterior distribution that may modify the corresponding inference. The point at issue here is that the truncation does not necessarily respect the geometry and shape of the unrestricted posterior distribution; although some orderings may nicely isolate a single mode of this distribution, most will involve parts of several modal regions. A consequence of this is that if a posterior mean is then calculated, the estimate will dwell in a valley between the $k!$ modes rather than close to one of the modes. In addition, removing the ordering constraint from the prior still leaves the exploration problem open. One must decide whether or not the posterior distribution is sufficiently well described by the MCMC sample; that is, whether or not the unexplored part of the support can be recovered by permutations.

Logically, it follows that a novel perspective on the simulation of the posterior distribution and relaxation of the identifiability constraints must also lead to a novel approach at the interpretation stage. Figure 1 illustrates the irrelevance of using posterior means when the identifiability constraint is not enforced, by presenting the case of a three-component exponential mixture where components are sufficiently close to allow a standard Gibbs sampler to move freely between the modes. In this case the three marginal posterior distributions display a high degree of similarity, and this leads to nearly identical posterior means. Similar phenomena has been observed by Gruet et al. (1999) for the constrained case, leading to an estimate where one component overwhelms the others, which become negligible.

The argument that we propose herein is that (a) the unconstrained posterior is the most natural (posterior) distribution to consider, especially in noninformative settings, (b) there exist accelerating strategies that fit naturally in the mixture setting, and (c) there exist alternatives to the posterior means that provide sensible Bayesian answers to the estimation of the parameters of (1). It is worth noting that although informative priors can make identifying constraints redundant by producing distinct distributions on the component parameters ζ_j, these priors most often still allow for some amount of switching between the components, which is not observed for standard algorithms.

In Section 2 we describe the standard available MCMC approaches and show by example that they fail to achieve the required symmetry. We then introduce in Section 3 a tempering acceleration strategy. In Sections 4 and 5 we develop various approaches for deriving Bayes estimates from a componentwise symmetric MCMC sample. We concentrate throughout on normal and exponential mixtures; that is,

$$\sum_{j=1}^{k} p_j \mathcal{N}(\theta_j, \tau_j^2)$$

and

$$\sum_{j=1}^{k} p_j \mathcal{E}\mathrm{xp}(\lambda_j).$$

The approach suggested here extends directly to other types of mixtures (Poisson, Pareto, etc.), and also to other categories of latent variable models (Robert 1998) and nonidentifiability problems. Similarly, we consider only the case of symmetric priors, but modest departures from symme-

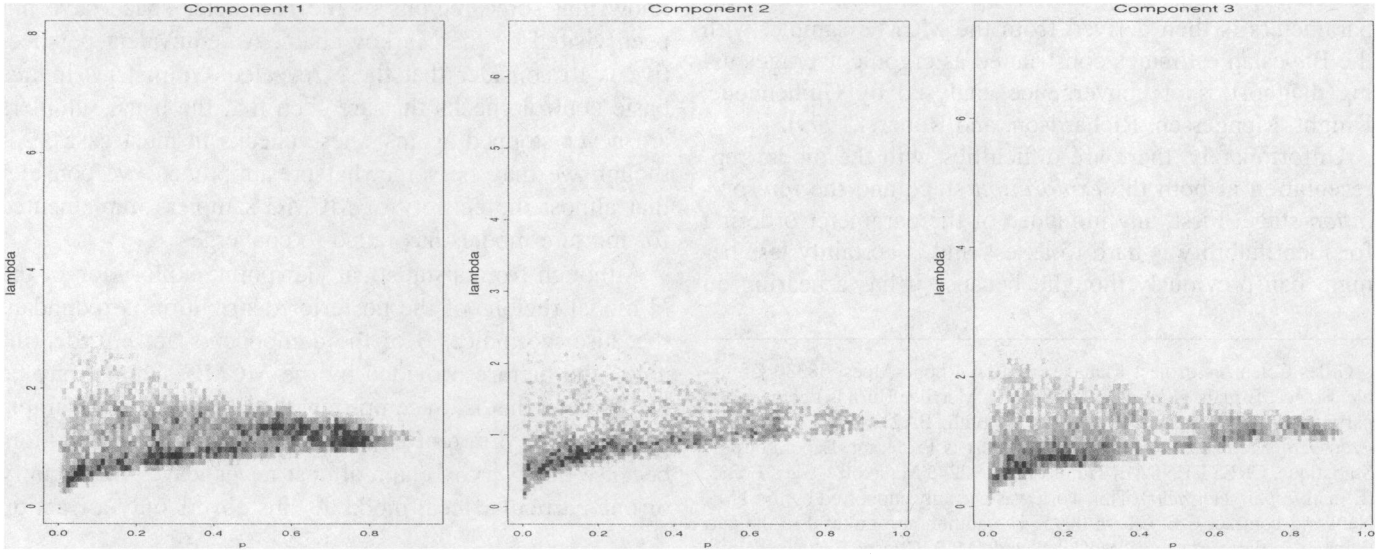

Figure 1. Marginal Posterior Histograms for the Parameters (p_j, λ_j) (j = 1, 2, 3) of a Three-Component Exponential Mixture for the Sample Produced by a Standard Gibbs Sampler for an Unconstrained Symmetric Prior (Uniform Dirichlet on the Vector of p_j's and Exponential $\mathcal{E}xp(1)$ on the λ_j's), Based on a Simulated Sample of Size 100 and 10,000 Iterations. Gray levels scale the height of the histogram; the higher the value of the grid, the darker the hue.

try would lead to similar behavior when the prior does not sufficiently counter the lack of identifiability. Our choice of symmetric proper priors is justified as was done by Richardson and Green (1997); namely, by an empirical determination of location-scale parameters based on the data spread. We use conjugate priors in both cases; that is, $\mathcal{D}(1,\ldots,1)$ distributions on the weights, $\mathcal{E}\text{xp}(1)$ distributions on the scale parameters λ_j and τ_j^2, and $\mathcal{N}(0,10\tau_j^2)$ distributions on the means θ_j.

2. MIXING PROPERTIES OF VARIOUS PROPOSALS

We consider the standard MCMC samplers, in turn showing that each fails to achieve the symmetry inherent to the whole stationary distribution. We note that it is always possible to achieve label switching between the $k!$ modes by the simple addition of a proposal that suggests a random permutation of the labels (which is a Metropolis–Hastings step with acceptance probability 1). Our insistence on searching for an algorithm that can achieve symmetry without such a move is that any concerns over convergence are not necessarily dealt with by such a strategy, which simply alleviates the most obvious symptom.

2.1 The Gibbs Sampler

The Gibbs sampler is the most commonly used approach in mixture estimation following Diebolt and Robert (1990, 1994), and we do not describe it here; (see Chib 1995 or Richardson and Green 1997 for details). The basic feature of the Gibbs sampler in this setup is the augmentation of the parameters of (1) by artificial allocation variables z_i, each one being associated with one of the x_i's to "demix" the observed sample. This allows simulation of the parameters of each component conditionally on the allocations, taking into account only the observations that have been allocated to this component.

The main defect of the Gibbs sampler from our perspective is the ultimate attraction of the local modes; that is, the almost impossible simultaneous reallocation of a group of observations to a different component. In experiments, this behavior was observed in both the unconstrained and the constrained cases; the result in the constrained case being that the parameter is biased in the event that sufficient observations are wrongly allocated. For instance, in a normal unidimensional mixture, if the constraint $\tau_1 < \tau_2$ bears on the variances and if $x_1,\ldots,x_{10}$ are allocated to component "1" and $x_{11},\ldots,x_{25}$ are allocated to component "2", and the opposite holds (i.e., the common variance of $x_1,\ldots,x_{10}$ is larger than the common variance of $x_{11},\ldots,x_{25}$), then the constraint will lead to either an upward bias for τ_2 or a downward bias for τ_1, because an allocation switch will *never* occur.

Figures 2 and 3 illustrate the lack of "mixing" of the chains for the exponential and the normal cases. Although the overall MCMC sample exhibits three zones of importance, the component chains $(p_j^{(t)},\zeta_j^{(t)})$ never switch in the normal case and exhibit only a moderate degree of switching between two components in the exponential case. In many cases, the Gibbs sampler is unable to move the Markov chain to another mode of equal importance because of its inability to step over valleys of low probability. In addition, there is no way that one can judge whether or not the neighborhood of a specific mode has been sufficiently explored, even though the path of the Markov chain can be exploited to provide a rough approximation of the marginal posterior distribution of the component parameters (p_i,ζ_i) $(i=1,\ldots,k)$. This approximation was obtained in the exponential case (Fig. 4) by computing the posterior distribution at each iteration of the Gibbs chain and by averaging, for each square of a 50×50 grid in the (p,λ) space, the values of the posterior distributions within the square. Comparing this figure with the plot of the Gibbs sample in Figure 2, whereas the sample plot covers most of the high regions of the estimated marginal, the ridges between the modal regions are missing. (Note that white regions in Fig. 4 are indicative either of a low posterior probability or of no visit to the corresponding square by the Markov chain.)

2.2 Random Walks

A standard alternative to the Gibbs sampler is the random walk Metropolis–Hastings algorithm. Let $\boldsymbol{\xi}^{(t)}$ denote the whole parameter vector at iteration t; then the next value is proposed as

$$\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^{(t)} + \omega\varepsilon_t, \qquad \varepsilon_t \sim \varphi, \tag{2}$$

where φ is usually chosen as a multivariate normal or Cauchy density and ω is calibrated to achieve a given acceptance rate in the acceptance probability. Following Gelman, Gilks, and Roberts (1996), a good range for the acceptance rate of random walk Metropolis–Hastings algorithms is .1–.3.

Plotting the samples obtained when applying this algorithm to the same simulated data results in figures nearly identical to Figure 2 and Figure 3 for the exponential and normal cases. The similarity shows that they both recover the same features of the posterior distribution but fail to achieve symmetry of the marginals. Although we deliberately chose low acceptance rates for the Metropolis–Hastings algorithm, the other modes remain too far away for the proposal to reach. A reparameterization of the exponential parameters as $(\log(p_i/1-p_i),\log(\lambda_i))$ was used so that the random walk (2) would not be hindered by constraints on the support. But the geometry of the posterior distribution is far from unimodal to allow sufficient acceptance of large moves. In other words, an isotonic random walk, even with a large value of ω, has too slight a chance of encountering one of the remaining $(k!-1)$ modes.

A more adaptive class of samplers are the Langevin diffusion Metropolis–Hastings algorithms proposed by Roberts and Tweedie (1996) and based on a random walk drift proposal,

$$x_{t+1} = x_t + \frac{\sigma^2}{2}\,\nabla \log \pi(x_t) + \sigma\varepsilon_t, \qquad \varepsilon_t \sim \varphi, \tag{3}$$

where π is the nonnormalized density of interest, φ is an arbitrary density, and σ is a simulation scale parameter that

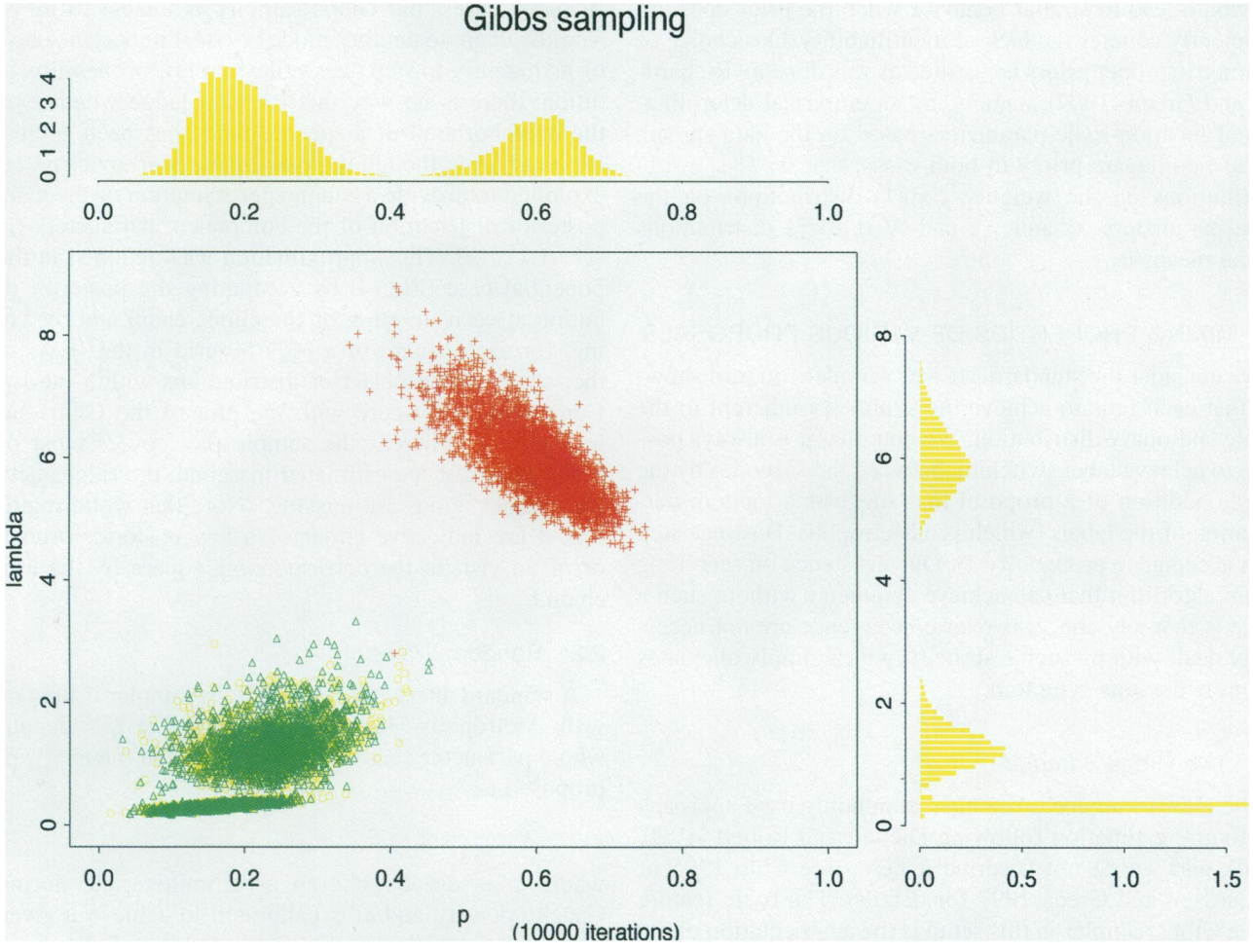

Figure 2. Representation in the (p, λ) Plane of the Gibbs Sample Associated With a Sample of 1,000 Simulated Points From a Three-Component Exponential Mixture. (Circles correspond to component 1; triangles, to component 2; and crosses, to component 3.) The histograms on top and right of the plot are the estimates of the marginal distributions of p_i (top) and λ_i (right).

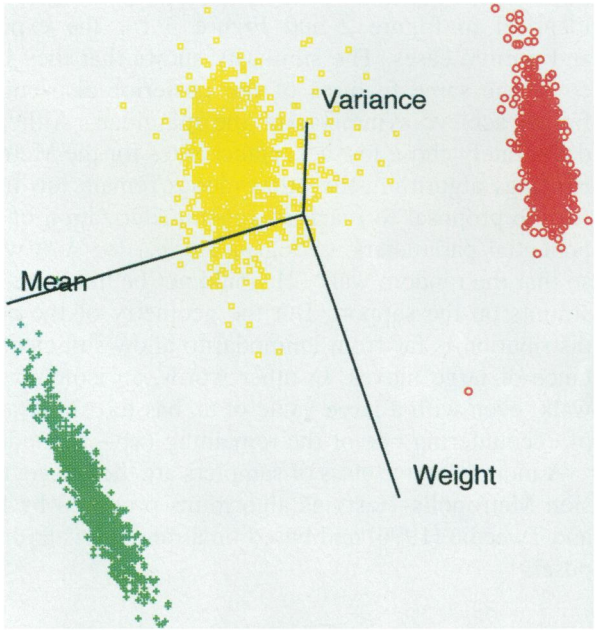

Figure 3. Representation of the Gibbs Sample Associated With a Sample of 500 Simulated Points from a Three-Component Normal Mixture. The symbols are the same as in Figure 2.

can be calibrated against the acceptance rate of the corresponding Metropolis–Hastings algorithm (see Roberts and Rosenthal 1998). This algorithm introduces a drift term, $\sigma^2 \nabla \log \pi(x_t)/2$, which pulls the Markov chain towards the modal zones of the support of π. Although the density φ in (3) is arbitrary, we found better mixing properties for a Cauchy density than for a normal density, a fact in agreement with the literature (see Stramer and Tweedie 1997).

However, as pointed out by Neal (personal communication) and a referee, the presence of many modes on the posterior surface partly cancels the appeal of using this alternative, because the gradient term will overwhelmingly pull the Markov chain back to the closest mode. This inefficient lack of mobility is illustrated by performances similar to those of the random walk sampler for the foregoing simulated datasets, as detailed by Celeux, Hurn, and Robert (1999). The disadvantage of using a Langevin drift as in (3) is even stronger in the tempering steps (see later), and thus we stick to the random walk sampler in the sequel.

3. SIMULATED TEMPERING

Because standard MCMC schemes are not working in our setup, we introduce a tempering scheme, following Neal (1996). We concentrate on one type of tempering strategy,

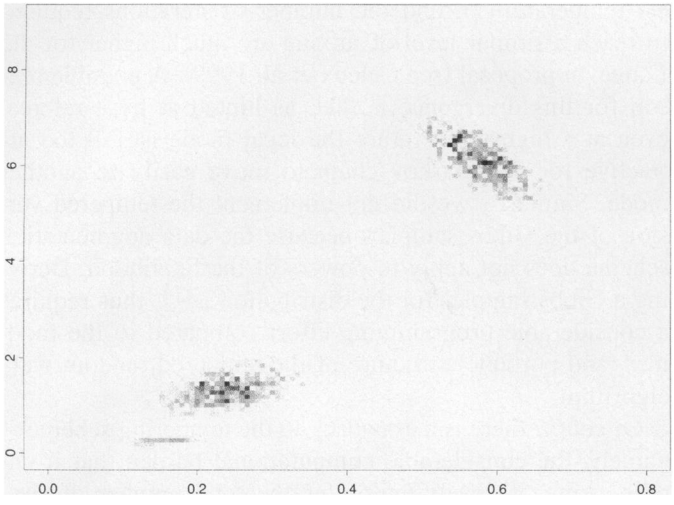

Figure 4. Approximation of the (p, λ) Marginal Posterior Distribution for the Exponential Mixture Sample of Figure 2, Obtained by Averaging the Values of the Joint Posterior at the Points of a Gibbs Sample Occurring in Each Square of a Grid.

using varying powers of the posterior distribution, but note that alternatives such as convolutions with $\mathcal{N}(0, \rho^2)$ distributions could also be used, with ρ playing the role of the temperature. Simulated tempering originates from the same observation as simulated annealing; namely, that the maximizers and minimizers of a function E are identical to those of a monotone transform of the function; for instance, powers of $E, E^{1/\beta}$. Although simulated annealing aims to emphasize minima and maxima of the function, tempering tries the converse, flattening peaks and filling valleys, so that a random walk can move more freely on the surface. But rather than simply simulating from $\pi^{1/\beta}$ with $\beta > 1$, which would give a more complete picture of the surface of π but could not be used directly for estimation from π [except perhaps by Sampling Importance Resampling (SIR) methods], Geyer and Thomson (1995) and Neal (1996) inserted tempering steps within a MCMC sampler to focus on the distribution of interest. This approach is appealing in that it encourages moves between the different modes in full generality. It is to some extent independent of the underlying model and stationary distribution and thus could be used for a wide range of models, including latent variable models (Robert 1998).

3.1 Up-and-Down Power Scheme

Neal (1996) showed that a valid way to insert simulations from $\pi^{1/\beta}$ within a MCMC sampler for π is to use an "up-and-down" scheme. Starting at the configuration $x^{(t)}$, the next configuration $x^{(t+1)}$ is proposed using intermediate MCMC steps that maintain a detailed balance with respect to the flatter $\pi^{1/\beta}$. Because β may have to be very large for the associated distribution to be sufficiently flat for good mixing, the huge difference between π and $\pi^{1/\beta}$ can lead to high rejection rates. Consequently, the number of intermediate steps is increased by using a sequence of closely spaced $\beta_l > \cdots > \beta_1 > 1$, and a corresponding sequence of simulations that operate between consecutive β distributions first down from most peaked to the most flat and then back up again. If we let MCMC(x, π) denote a MCMC kernel with stationary distribution π, then the "up-and-down" scheme is as follows:

$$
\begin{aligned}
&\downarrow \text{Generate } y_1 \sim \text{MCMC}(x^{(t)}, \pi^{1/\beta_1}) \\
&\downarrow \text{Generate } y_2 \sim \text{MCMC}(y_1, \pi^{1/\beta_2}) \\
&\qquad \vdots \\
&\downarrow \text{Generate } y_l \sim \text{MCMC}(y_{l-1}, \pi^{1/\beta_t}) \\
&\uparrow \text{Generate } y_{l+1} \sim \text{MCMC}(y_l, \pi^{1/\beta_{t-1}}) \\
&\uparrow \text{Generate } y_{l+2} \sim \text{MCMC}(y_{l+1}, \pi^{1/\beta_{t-2}}) \\
&\qquad \vdots \\
&\uparrow \text{Generate } y_{2l-1} \sim \text{MCMC}(y_{2l-2}, \pi^{1/\beta_1}).
\end{aligned}
$$

This sequence of y values constitutes the proposal mechanism for the overall MCMC algorithm, a final acceptance step is required to determine whether or not the final one of them will be accepted as the next x value in the main chain:

$$
\begin{aligned}
&\text{Accept } x^{(t+1)} = y_{2l-1} \text{ with probability} \\
&\quad \times \min\left\{1, \frac{\pi^{1/\beta_1}(x^{(t)})}{\pi(x^{(t)})} \cdots \frac{\pi^{1/\beta_l}(y_{l-1})}{\pi^{1/\beta_{l-1}}(y_{l-1})} \right. \\
&\qquad \left. \times \frac{\pi^{1/\beta_{l-1}}(y_l)}{\pi^{1/\beta_l}(y_l)} \cdots \frac{\pi(y_{2l-1})}{\pi^{1/\beta_1}(y_{2l-1})}\right\} \qquad (4)
\end{aligned}
$$

Note that this acceptance probability involves only ratios of each of the distributions π^{1/β_i} and is thus independent of the normalizing constants. When the differences $\beta_i - \beta_{i+1}$ are small, the overall acceptance rate should thus be high, as should the local acceptance rates given the small differences between the π^{1/β_i}'s.

Several modifications to this scheme can be proposed. First, as was done by Neal (1996), additional simulations from the flatest π^{1/β_l} can be generated between stage l and stage $l + 1$, because this facilitates moves between modes without affecting the balance condition or the acceptance probability (3). Second, as the up-and-down scheme is associated with the global exploration of the posterior support, an arbitrary number of simulations from the target distribution π using a nontempered proposal can be inserted between up-and-down proposals to improve the local exploration of the current mode.

3.2 Implementation and Results

Given the freedom in the choice of the "flatter" distributions at level i in the up-and-down scheme, we work with a tempering schedule where at level i, rather than raise the whole posterior to the power $1/\beta_i$, we power up just the likelihood and leave the prior contribution unchanged. This choice was motivated by the uncertainty about the integrability of the pseudodistribution π^{1/β_i} for large values of β_i. The number of levels and the rate of power decrease must be chosen with care to achieve a sufficient rate of label switching at the lowest power. We observed sensitivity of the method to the choice of the lowest power $1/\beta_l$, the number of levels l, and the rate of decrease of the power,

$1/\beta_i - 1/\beta_{i-1}$. First, the lowest power must be chosen sufficiently small to ensure that label switching is frequent at this power; this can be assessed by a preliminary check on the marginal histograms. The flatter the target distribution $\pi^{1/\beta}$, the easier the moves between the labels. Second, we found that in practice, geometric decrease rates perform better than linear decrease rates, whereas increasing the number of levels led to faster mixing at level l.

Figures 5 and 6 give the results of an implementation using random walk proposals with $l = 50$ and $l = 45$ levels in the exponential and normal cases. Both are quite satisfactory in terms of mixing, because the marginal samples for the three components are comparable. The stepwise acceptance rates are all calibrated between .15 and .35, and remain in the same range along the up-and-down iterations. Significantly, the marginals in these figures are quite similar in Figure 5 and Figure 2, which validates a posteriori the output of the original MCMC samplers.

As pointed out in Section 2.2, the Langevin diffusion Metropolis–Hastings algorithm could also be used as a basis for the tempering scheme. However, the similarity with the random walk proposal disappears in this case; the highest temperature β_l and the number of iterations required to reach a similar level of mixing are much higher for the Langevin proposal (see Celeux et al. 1999). A possible reason for this divergence is that, as hinted at by a referee, even at a high temperature the local mode is still too attractive for the Markov chain to move easily to another mode. Similarly, we did not implement the tempered version of the Gibbs sampler, because the data augmentation scheme does not apply to powers of the likelihood. Deriving a Gibbs sampler for the distribution π^{1/β_l} thus requires a considerable programming effort compared to the modular (and portable) structure of the tempered random walk algorithm.

Of course there is a drawback to the tempering scheme—namely, the considerable computational burden that it entails; using 50 levels multiplies the total number of simulations by 100, or, in other words, we use only 1% of the simulations. Moreover, not all of these computationally expensive proposals will be accepted; the low acceptance rate motivates the use of additional steps of the standard Metropolis–Hastings proposal between tempering steps.

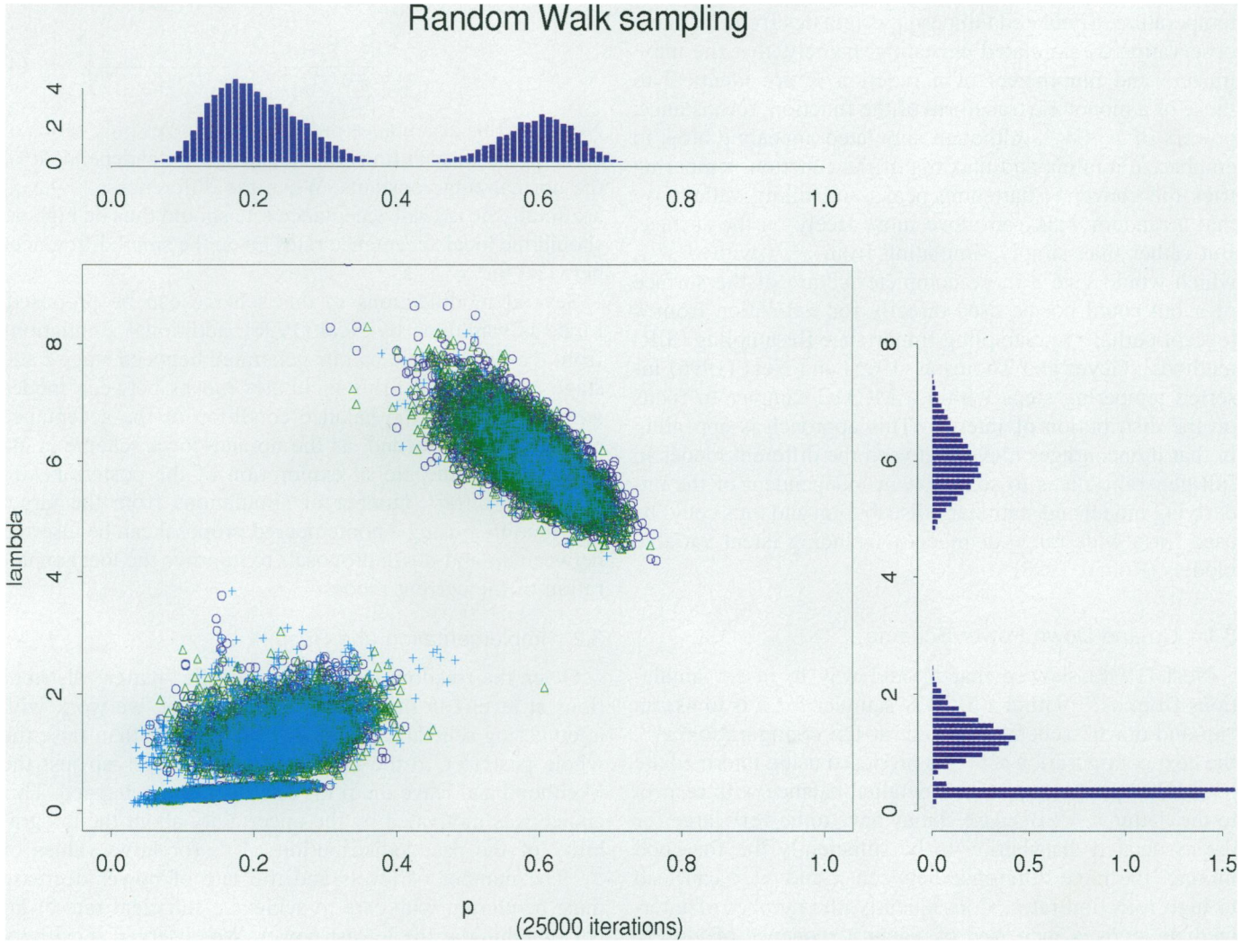

Figure 5. Representation in the (p, λ) Plane of the Markov Chain Monte Carlo Sample Associated With the Distribution π, for the Same Exponential Sample as in Figure 2 and a Tempered Random Walk Metropolis—Hastings Algorithm (l = 50 Levels, Minimum Power .01).

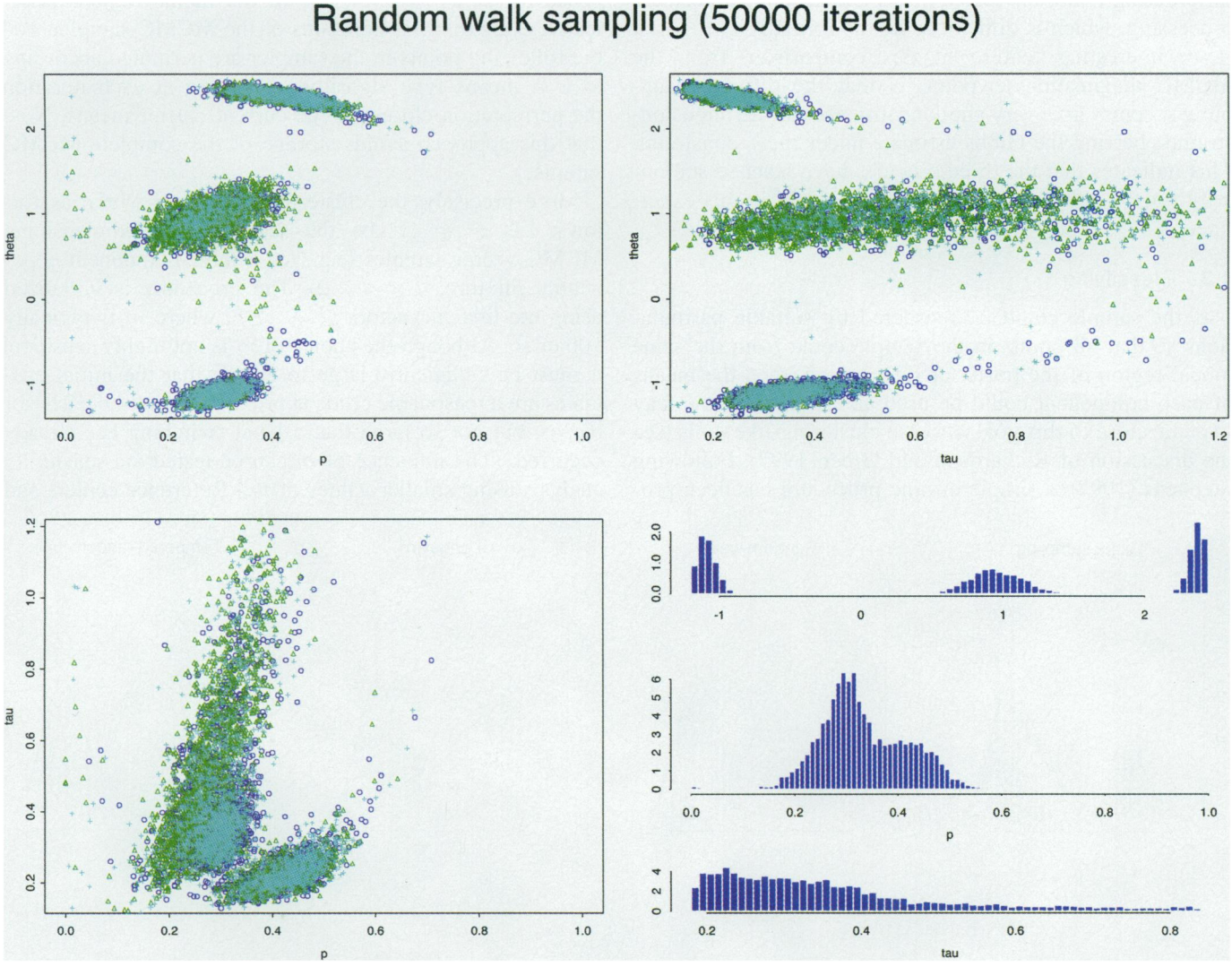

Figure 6. Representation of the Markov Chain Monte Carlo Sample Associated With the Distribution π, and a Tempered Random Walk Metropolis—Hastings Algorithm (l = 45 Levels, Minimum Power .005).

4. INFERENCE USING MARKOV CHAIN MONTE CARLO SAMPLES

4.1 Ordering Constraints

When label switching of the components is a prerequisite of MCMC convergence, automatically a complication is created for inference. The mixing of labels means that it is not possible to form ergodic averages over labels, because if the sampler is behaving properly, then all resulting parameters estimates will be close. By imposing some form of order constraint on the parameters, it is possible to create a k group structure over which averages can be formed. As proven by Stephens (1997, prop. 3.1), the ordering constraint can be imposed ex post; that is, after the simulations have been completed. This is interesting, because it offers a direct and easy comparison of the effects of different possible ordering constraints based on the same MCMC sample.

Applying the different order constraints to the samples obtained via the different samplers of Section 2, the results are disturbing. The estimates for the normal model, shown in Table 1, can be seen to be markedly different for the three orders. Because it is well known that different sets of parameters can lead to the same estimate of the mixture density, as observed by Robert and Mengersen (1998), for instance, a more realistic comparison is based on the estimated "plug-in" densities; that is, the densities where the parameters have been replaced by the estimations of Table 1 (see Fig. 7). Clearly, the choice of the identifying constraint has an influence on the resulting estimate and, ergo, creates a bias. In this case, the best fit is provided by the constraint on the θ_i's, which is natural given that the true θ_i's are well separated (see Table 1). However, this natu-

Table 1. Estimates of the Parameters of a Three-Component Normal Mixture, Obtained for the Same Simulated Sample as in Figure 2 Using a Gibbs Sampler and Reordering According to One of Three Constraints, p: $p_1 < p_2 < p_3$, θ: $\theta_1 < \theta_2 < \theta_3$, or τ: $\tau_1 < \tau_2 < \tau_3$

Order	p_1	p_2	p_3	θ_1	θ_2	θ_3	τ_1	τ_2	τ_3
p	.231	.311	.458	.321	−.55	2.28	.41	.471	.303
θ	.297	.246	.457	−1.1	.83	2.33	.357	.543	.284
τ	.375	.331	.294	1.59	.083	.379	.266	.34	.579
True	.22	.43	.35	1.1	2.4	−.95	.3	.2	.5

ral ordering is not obvious a priori and requires a postdata processing, which is difficult to define formally.

An interesting feature of this comparison, from the MCMC diagnostic viewpoint, is that the different sampling schemes give very similar estimates for the three constraints, barring the Gibbs estimate under the τ constraint. This indicates that the three schemes have reached stationarity for the number of iterations considered in this experiment.

4.2 Classifying

If the sample could be reordered by suitable permutations so that all points in the sample come from the same modal region of the posterior distribution, then the means of each component could be used as estimators. One way to come close to this goal is to use clustering-like tools (see the discussion of Richardson and Green 1997). Following Stephens (1997), a simple on-line procedure has been proposed by Celeux (1998); one of the modal regions is selected using the first iterations of the MCMC sampler, and the following points in the sampler are permuted according to a $k!$ means–type algorithm, selecting at each iteration the permutation closest to the current cluster means. Note that this approach avoids storage of the complete MCMC sample.

More precisely, the clustering procedure works as follows: Let $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots$ be the sequence of d-dimensional MCMC vector samples (e.g., for a three-component exponential mixture, $d = 3 \times 2$). The procedure is initialized using the first m vectors $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^m$, where m is typically 100 or so. Although the choice of m is not highly sensitive, it must be sufficiently large to ensure that the initial estimates are a reasonable crude approximation of the posterior means, but not so large that a label switching has already occurred. (The influence of m can be tested via sensitivity analysis using smaller values of m.) Reference centers and

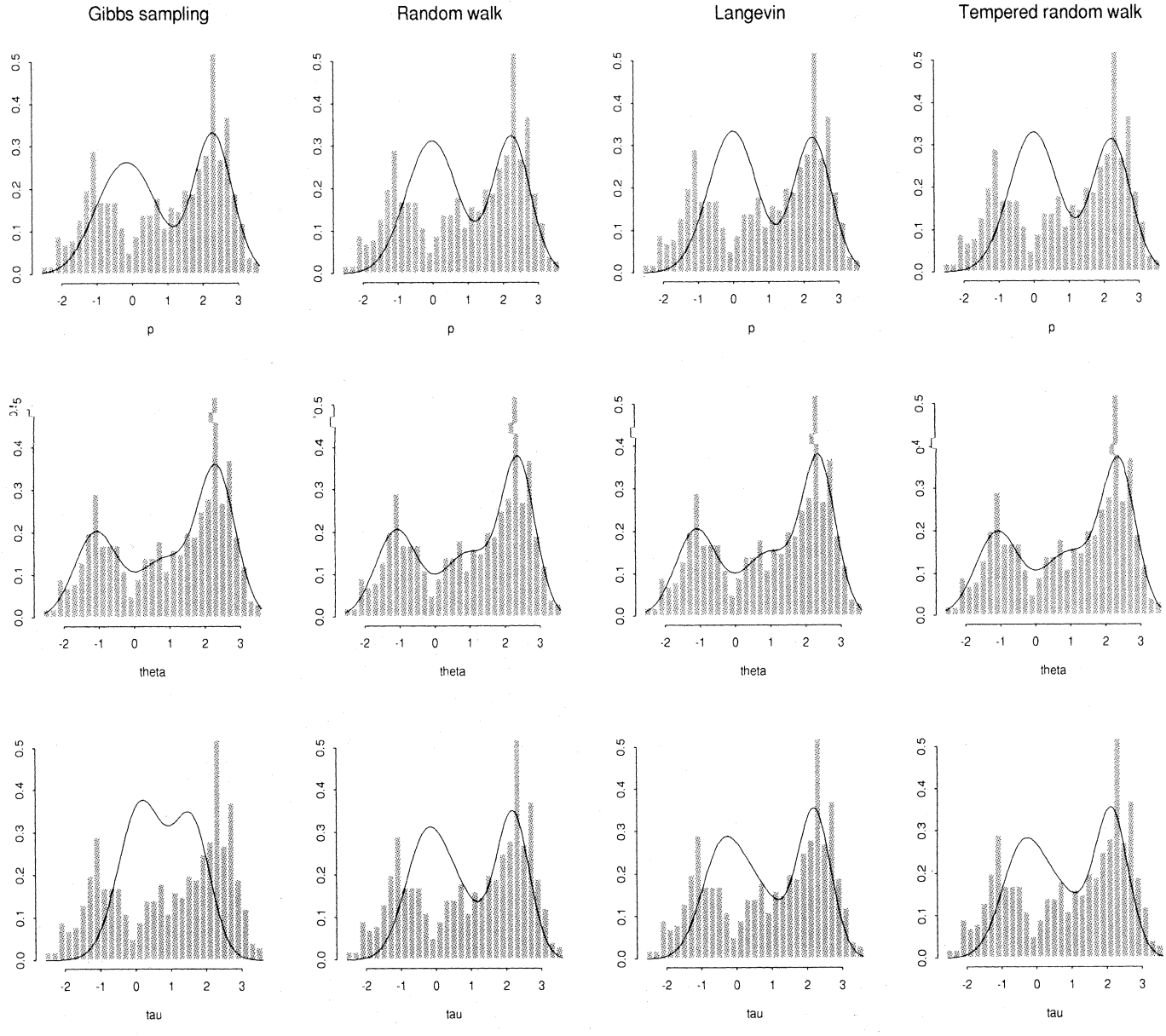

Figure 7. Estimations of the Density of the Sample (Represented by the Histogram), for Four Different Markov Chain Monte Carlo Samplers and the Three Possible Orderings, p: $p_1 < p_2 < p_3$, θ: $\theta_1 < \theta_2 < \theta_3$, and τ: $\tau_1 < \tau_2 < \tau_3$ (50,000 Simulations).

Table 2. Exponential Mixture Estimates (Arranged by Means) Under the Different Estimation Procedures for the Different Simulation Schemes

		True values	Gibbs sampling	Random walk	Langevin	Tempered random walk
Cluster	$\hat{p}_1$	.15	.164	.162	.162	.166
	$\hat{\lambda}_1$	.25	.317	.314	.320	.317
	$\hat{p}_2$	.22	.239	.236	.239	.238
	$\hat{\lambda}_2$	1.1	1.299	1.278	1.288	1.312
	$\hat{p}_3$	.63	.597	.602	.599	.596
	$\hat{\lambda}_3$	5.4	5.993	5.948	5.976	6.000
Local	$\hat{p}_1$	.15	.169	.170	.154	.175
loss	$\hat{\lambda}_1$	.25	.260	.257	.358	.318
	$\hat{p}_2$	.22	.224	.239	.251	.229
	$\hat{\lambda}_2$	1.1	1.294	1.187	1.199	1.340
	$\hat{p}_3$	.63	.607	.592	.596	.596
	$\hat{\lambda}_3$	5.4	6.814	5.128	5.233	5.956
Global	$\hat{p}_1$	.15	.164	.164	.167	.164
loss	$\hat{\lambda}_1$	.25	.306	.307	.308	.307
	$\hat{p}_2$	.22	.227	.226	.228	.225
(L_2)	$\hat{\lambda}_2$	1.1	1.225	1.232	1.258	1.230
	$\hat{p}_3$	.63	.609	.610	.605	.609
	$\hat{\lambda}_3$	5.4	5.863	5.861	5.904	5.865
Global	$\hat{p}_1$	.15	.173	.174	.175	.175
loss	$\hat{\lambda}_1$	.25	.316	.317	.317	.318
	$\hat{p}_2$	.22	.227	.228	.229	.229
(KL)	$\hat{\lambda}_2$	1.1	1.315	1.325	1.334	1.340
	$\hat{p}_3$	.63	.601	.598	.597	.596
	$\hat{\lambda}_3$	5.4	5.922	5.944	5.956	5.959

componentwise variances for $(i = 1, \ldots, d)$ are defined as

$$\bar{\xi}_i = \frac{1}{m} \sum_{j=1}^{m} \xi_i^j,$$

and

$$s_i = \frac{1}{m} \sum_{j=1}^{m} (\xi_i^j - \bar{\xi}_i)^2.$$

We set $s_i^{[0]} = s_i, i = 1, \ldots, d$. If we denote $\bar{\xi}_1^{[0]} = \bar{\xi}$, then the $(k! - 1)$ other centers $\bar{\xi}_2^{[0]}, \ldots, \bar{\xi}_{k!}^{[0]}$ can be deduced from $\bar{\xi}_1^{[0]}$ by permuting the labeling of the mixture components. After this initialization stage, the rth iteration of the clustering procedure runs as follows:

1. Allocate ξ^{m+r} to the cluster j^* that minimizes

$$\|\xi^{m+r} - \bar{\xi}_j^{[r-1]}\|^2 = \sum_{i=1}^{d} \frac{(\xi_i^{m+r} - \bar{\xi}_{ij}^{[r-1]})^2}{s_i^{[r-1]}}, \qquad (5)$$

where $\bar{\xi}_{ij}^{[r-1]}$ is the ith coordinate of $\bar{\xi}_j^{[r-1]}$.

If $j^* \neq 1$, permute the coordinates of ξ^{m+r} to get $j^* = 1$.

2. Update the $k!$ centers and the d normalizing coefficients: compute

$$\bar{\xi}_1^{[r]} = \frac{m + r - 1}{m + r} \bar{\xi}_1^{[r-1]} + \frac{1}{m + r} \xi^{m+r}.$$

Derive the $k! - 1$ other centers by permutation, and take

$$s_i^{[r]} = \frac{m + r - 1}{m + r} s_i^{[r-1]} + \frac{m + r - 1}{m + r} (\bar{\xi}_{i1}^{[r-1]} - \bar{\xi}_{i1}^{[r]})^2 + \frac{1}{m + r} (\xi_i^{m+r} - \bar{\xi}_{i1}^{[r]})^2.$$

The mode of reference corresponds to $j = 1$ at each iteration. Note that the normalization of the distances (5) makes the procedure independent of location-scale transformations of the parameters, even though the resulting estimates depend on the parameterization of ξ. Tables 2 and 3 display the resulting estimates for the exponential and normal mixtures examples, based on the MCMC samples obtained in Sections 2 and 3. The variation between the estimates is wider on parameter λ_3 in Table 2, but this is to be expected, given the poor identifiability of exponential mixtures in the tails. As shown in Figure 8, the variation between the estimates does not jeopardize the estimated density.

5. ALTERNATIVE BAYES ESTIMATORS

Another possibility is to use a loss function $L(\xi, \hat{\xi})$ for

Table 3. Normal Mixture Estimates (Arranged by Means) Under the Different Estimation Procedures for the Different Simulation Schemes

		True values	Gibbs sampling	Random walk	Langevin	Tempered random walk
Cluster	$\hat{p}_1$	.35	.299	.303	.300	.299
	$\hat{\theta}_1$	−.95	−1.031	−1.057	−1.080	−1.039
	$\hat{\tau}_1^2$	.5	.371	.354	.341	.360
	$\hat{p}_2$	.22	.247	.269	.277	.278
	$\hat{\theta}_2$	1.1	.778	.906	.924	.893
	$\hat{\tau}_2^2$	.3	.530	.507	.526	.591
	$\hat{p}_3$	.43	.454	.428	.423	.423
	$\hat{\theta}_3$	2.4	2.319	2.368	2.386	2.365
	$\hat{\tau}_3^2$	.2	.284	.230	.220	.231
Local	$\hat{p}_1$	.35	.319	.323	.293	.284
loss	$\hat{\theta}_1$	−.95	−1.259	−1.238	−1.184	−1.169
	$\hat{\tau}_1^2$	.5	.361	.355	.395	.463
	$\hat{p}_2$	.22	.206	.233	.298	.316
	$\hat{\theta}_2$	1.1	.982	1.032	.794	.879
	$\hat{\tau}_2^2$	.3	.807	.823	.340	.277
	$\hat{p}_3$	.43	.475	.444	.408	.400
	$\hat{\theta}_3$	2.4	2.228	2.253	2.235	2.186
	$\hat{\tau}_3^2$	.2	.209	.165	.216	.231
Global	$\hat{p}_1$	.35	.303	.304	.303	.301
loss	$\hat{\theta}_1$	−.95	−1.010	−1.102	−1.096	−1.103
	$\hat{\tau}_1^2$	.5	.357	.340	.349	.348
(L_2)	$\hat{p}_2$	.22	.232	.263	.270	.277
	$\hat{\theta}_2$	1.1	.826	.932	.945	.954
	$\hat{\tau}_2^2$	.3	.461	.480	.489	.544
	$\hat{p}_3$	.43	.465	.433	.427	.423
	$\hat{\theta}_3$	2.4	2.330	2.383	2.390	2.388
	$\hat{\tau}_3^2$	.2	.282	.225	.218	.223
Global	$\hat{p}_1$	.35	.305	.306	.304	.287
loss	$\hat{\theta}_1$	−.95	−1.094	−1.091	−1.010	−1.131
	$\hat{\tau}_1^2$	.5	.359	.351	.341	.334
(KL)	$\hat{p}_2$	.22	.224	.253	.265	.334
	$\hat{\theta}_2$	1.1	.814	.914	.936	1.063
	$\hat{\tau}_2^2$	.3	.434	.447	.474	.791
	$\hat{p}_3$	.43	.471	.441	.431	.379
	$\hat{\theta}_3$	2.4	2.323	2.375	2.387	2.410
	$\hat{\tau}_3^2$	.2	.286	.230	.221	.209

which the labeling is immaterial, and then find the corresponding Bayes estimator $\hat{\boldsymbol{\xi}}^*$,

$$\hat{\boldsymbol{\xi}}^* = \arg\min_{\hat{\xi}} \mathbb{E}_{\xi|x} L(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}). \tag{6}$$

Depending on the loss function, the complexity of the problem may mean that calculation of the estimator is not analytically feasible. Consequently, the choice of loss function has usually been restricted to those for which the form of the estimator is known; for example, the squared loss function $L(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}) = \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}\|^2$, which leads to the posterior marginal means as estimates. It has recently been demonstrated, however, that apparently intractable estimators in Bayesian image analysis can be approximated using MCMC (Frigessi and Rue 1997; Rue 1995). We consider a similar approach here, suggesting two types of loss function: the first when inference for the parameters is the issue and the second when the predictive mixture distribution is of more interest.

5.1 A Loss Function for the Parameters

We return to Figures 2 and 3 to motivate our approach; here the mixture models are represented by points in the (p, λ) or (p, θ, τ) spaces. Removing any labeling of the points, as would be natural from a point process perspective, suggests that the components could be viewed as points in a unlabeled point process (in these cases fixed dimension point processes). So in formulating an appropriate loss function what we are looking for is a way of measuring distance between two point configurations. Clearly, many of the obvious solutions would rely on labelling of the points; for example, the sum of squared distances between pairs of points, one from each configuration, with the same label. Because we are explicitly removing all labeling information, it is more complicated to find a suitable measure of discrepancy; we suggest the following, loosely based on the Baddeley Δ metric (Baddeley 1992) used by Frigessi and Rue (1997) in an image analysis context.

The Δ metric measures the distance between two binary images; every pixel in the image grid contributes the squared difference between the shortest distances from it to the closest foreground pixel in the images to be compared. We are considering a continuous rather than a discrete space, and so will need to modify the definition while trying to maintain the same spirit. The idea is to have a collection of reference points, and for each of these to calculate the distance to the closest parameter point in the two mixture configurations. We denote the collection of reference points $t_1, \ldots, t_n$, where these lie in the same space as the components of $\boldsymbol{\xi} = (\boldsymbol{\xi}_j)$; that is, in the exponential case, the t_i's are in (p, λ) space and in the normal case, they are in (p, θ, τ^2) space. For each t_i, define $d(t_i, \boldsymbol{\xi})$ to be the distance between t_i and the closest of the $\boldsymbol{\xi}_j$'s $(j = 1, \ldots, k)$, where the distance d is Euclidean but applies to the transformed variables $\ln(p/(1-p)), \ln\lambda, \theta, \ln\sqrt{\tau^2}$ (see Sec. 2.2 for the reason behind this transformation). Then let

$$L(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}) = \sum_{i=1}^{n} (d(t_i, \boldsymbol{\xi}) - d(t_i, \hat{\boldsymbol{\xi}}))^2. \tag{7}$$

That is, for each of the fixed points t_i, there is a contribution to the loss if the distance from t_i to the nearest $\boldsymbol{\xi}_j$ is not the same as the distance from t_i to the nearest $\hat{\boldsymbol{\xi}}_j$.

Clearly, the choice of the t_i's plays an important role, because we want $L(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}) = 0$ only if $\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}$, and for the loss function to respond appropriately to changes in the two point configurations. To avoid the possibility of zero loss between two configurations which actually differ, it must be possible to determine $\boldsymbol{\xi}$ from the $\{t_i\}$ and the corresponding $\{d(t_i, \boldsymbol{\xi})\}$. To do this in k-dimensional space requires distances from $k+1$ distinct locations; thus in the exponential case, each component of $\boldsymbol{\xi}$ needs to be the closest component to at least three t_i's, whereas for the normal mixture model, which has a parameter space one dimension higher, at least four t_i's are required for each component of that $\boldsymbol{\xi}$. For the second desired property, the t_i's are best positioned in high posterior density regions of the $\boldsymbol{\xi}_j$'s space.

We have chosen to divide our simulation effort so that the first half of the simulations are allocated to selecting a suitable set of t_i locations by randomly selecting one of the components in each realization as t_i (to avoid correlation). The second half of the simulations are used to estimate the corresponding $\mathbb{E}_{\boldsymbol{\xi}|x}[d(t_i, \boldsymbol{\xi})]$ using the two-step procedure proposed by Rue (1995). Note that we can write

$$\mathbb{E}_{\boldsymbol{\xi}|x}\left[\sum_{i=1}^{n} (d(t_i, \boldsymbol{\xi}) - d(t_i, \hat{\boldsymbol{\xi}}))^2\right] = \sum_{i=1}^{n} (\mathbb{E}_{\boldsymbol{\xi}|x}[d(t_i, \boldsymbol{\xi})^2] - 2d(t_i, \hat{\boldsymbol{\xi}})\mathbb{E}_{\xi|x}[d(t_i, \boldsymbol{\xi})] + d(t_i, \hat{\boldsymbol{\xi}})^2),$$

because the expectation is with respect to the posterior $\pi(\boldsymbol{\xi}|x)$ and $\hat{\boldsymbol{\xi}}$ is not involved. It is possible to estimate $\gamma_i = \mathbb{E}_{\xi|x}[d(t_i, \boldsymbol{\xi})]$ by simulating from the posterior (i.e., using the MCMC sample) and using the usual ergodic averaging. (There is no issue of labeling here, because we are always talking about the closest point for a fixed t_i.) This leaves the minimization task, which, ignoring terms not involving $\hat{\boldsymbol{\xi}}$, is

$$\begin{aligned} \hat{\boldsymbol{\xi}}^* &= \arg\min_{\hat{\xi}} h(\hat{\boldsymbol{\eta}}) \\ &= \arg\min_{\hat{\xi}} \sum_{i=1}^{n} (-2\hat{\gamma}_i d(t_i, \hat{\boldsymbol{\xi}}) + d(t_i, \hat{\boldsymbol{\xi}})^2). \end{aligned} \tag{8}$$

Simulated annealing can be used for this minimization. At this stage, any proposed $\hat{\boldsymbol{\xi}}$ without the requisite number of t_i points in its neighborhood can be flagged, and, if necessary, further t_i's can be added and associated γ_i's can be estimated from the existing simulations. Tables 2 and 3 give the resulting estimates for the exponential and normal mixtures examples.

5.2 Loss Functions for the Predictive Distribution

When the object of inference is the predictive distribution, more global loss functions can be devised to measure distributional discrepancies. (This was the approach adopted

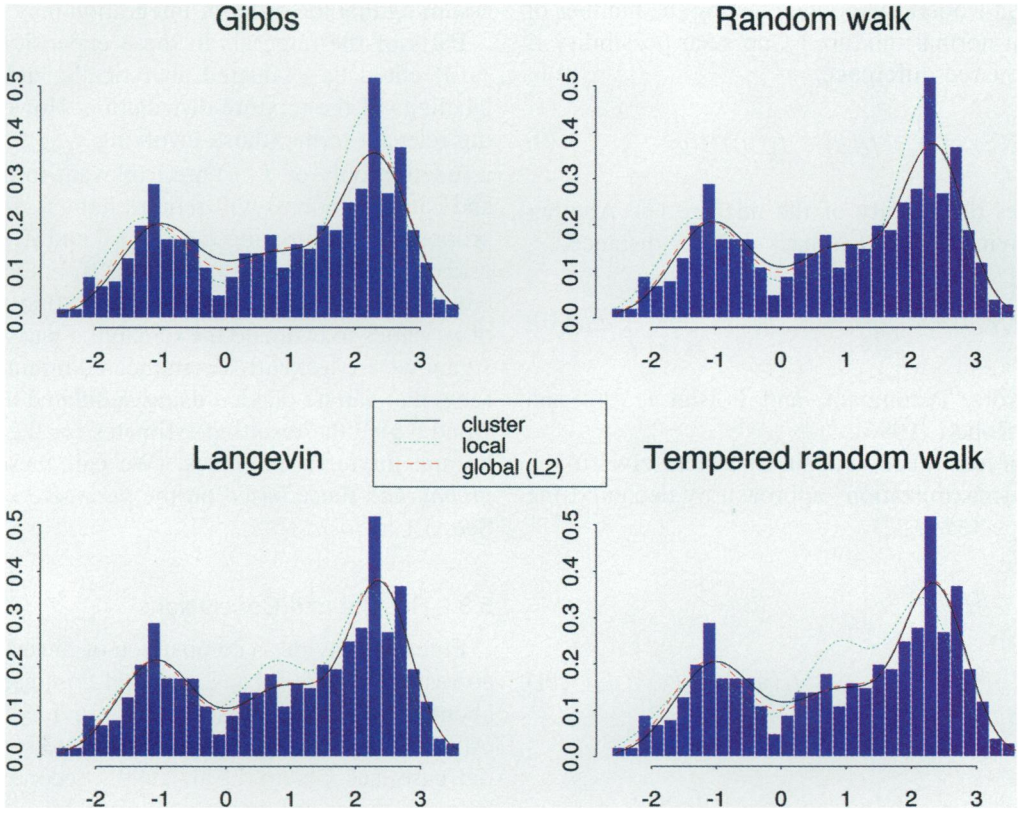

Figure 8. Estimations by the Clustering and the Loss Function Procedures of the Density of the Same Normal Sample as in Figure 2 (Represented by the Histogram), for the Different Samplers. Solid lines represent the clustering procedure; short dashes, for the local loss function; and long dashes, for the global loss function.

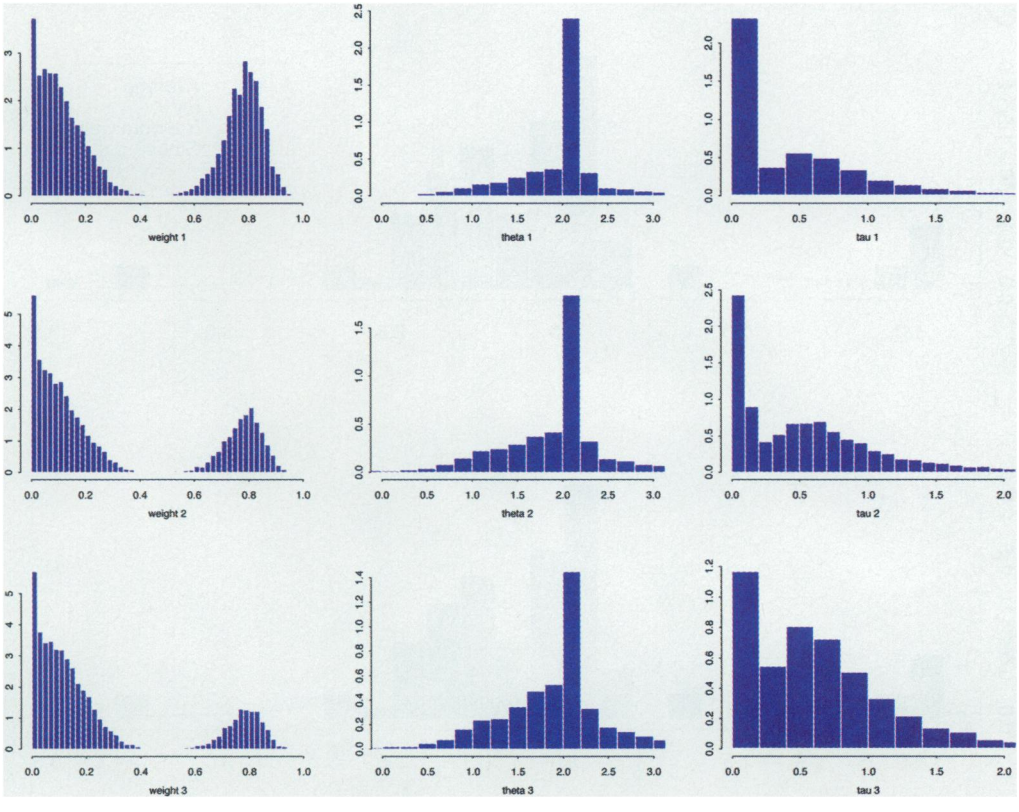

Figure 9. Histograms for the MCMC Samples of the Parameters of a Three-Component Normal Mixture, for the Galaxy Dataset and a Random Walk Sampler (100,000 Iterations).

in Mengersen and Robert 1996 when testing the number of components in a normal mixture.) One such possibility is the integrated squared difference,

$$L(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}) = \int_{\mathcal{R}} (f_{\boldsymbol{\xi}}(y) - f_{\hat{\boldsymbol{\xi}}}(y))^2 \, dy, \tag{9}$$

where $f_{\boldsymbol{\xi}}$ denotes the density of the mixture (1). Another possibility is a symmetrized Kullback–Leibler distance,

$$L(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}) = \int_{\mathcal{R}} \left(f_{\boldsymbol{\xi}}(y) \log \frac{f_{\boldsymbol{\xi}}(y)}{f_{\hat{\boldsymbol{\xi}}}(y)} + f_{\hat{\boldsymbol{\xi}}}(y) \log \frac{f_{\hat{\boldsymbol{\xi}}}(y)}{f_{\boldsymbol{\xi}}(y)} \right) dy, \tag{10}$$

as used by Carota, Parmigiani, and Polson (1996) and Mengersen and Robert (1996).

These forms of loss function again lend themselves to the "estimation-then-maximization" approach by decomposing the posterior expected loss,

$$\mathbb{E}_{\boldsymbol{\xi}|x} \left[\int_{\mathcal{R}} (f_{\boldsymbol{\xi}}(y) - f_{\hat{\boldsymbol{\xi}}}(y))^2 \, dy \right]$$

$$= \int_{\mathcal{R}} (\mathbb{E}_{\boldsymbol{\xi}|x}[f_{\boldsymbol{\xi}}(y)^2] - 2 f_{\hat{\boldsymbol{\xi}}}(y) \mathbb{E}_{\boldsymbol{\xi}|x}[f_{\boldsymbol{\xi}}(y)] + f_{\hat{\boldsymbol{\xi}}}(y)^2) \, dy \tag{11}$$

and

$$\mathbb{E}_{\boldsymbol{\xi}|x} \left[\int_{\mathcal{R}} \left(f_{\boldsymbol{\xi}}(y) \log \frac{f_{\boldsymbol{\xi}}(y)}{f_{\hat{\boldsymbol{\xi}}}(y)} + f_{\hat{\boldsymbol{\xi}}}(y) \log \frac{f_{\hat{\boldsymbol{\xi}}}(y)}{f_{\boldsymbol{\xi}}(y)} \right) dy \right]$$

$$= \int_{\mathcal{R}} (\mathbb{E}_{\boldsymbol{\xi}|x}[f_{\boldsymbol{\xi}}(y) \log f_{\boldsymbol{\xi}}(y)] - \log f_{\hat{\boldsymbol{\xi}}}(y) \mathbb{E}_{\boldsymbol{\xi}|x}[f_{\boldsymbol{\xi}}(y)]$$

$$+ f_{\hat{\boldsymbol{\xi}}}(y) \log f_{\hat{\boldsymbol{\xi}}}(y) - f_{\hat{\boldsymbol{\xi}}}(y) \mathbb{E}_{\boldsymbol{\xi}|x}[\log f_{\boldsymbol{\xi}}(y)]) \, dy, \tag{12}$$

assuming that the order of integration may be interchanged.

Parts of the integrals in these expansions not involving $f_{\boldsymbol{\xi}}(y)$ could be evaluated analytically, independent of the labelings of the mixture distribution. However, the remaining relevant terms (those involving $f_{\hat{\boldsymbol{\xi}}}(y)$, as terms that are a function only of $f_{\boldsymbol{\xi}}(y)$ are irrelevant to the minimization and can be ignored) will require numerical techniques. We propose first estimating $\mathbb{E}_{\boldsymbol{\xi}|x}[f_{\boldsymbol{\xi}}(y)]$ and $\mathbb{E}_{\boldsymbol{\xi}|x}[\log f_{\boldsymbol{\xi}}(y)]$ on a grid of y_i values, using simulations from the posterior distribution. We then use numerical integration on the same grid of y_i values to estimate the remaining parts of the integrals. Again we are left with a complicated minimization problem for $\hat{\boldsymbol{\xi}}$ that can be tackled using simulated annealing. Tables 2 and 3 give the resulting estimates for the exponential and normal mixtures examples. (We call these loss functions *global,* and those based on the parameters as developed in Sec. 5.1 as *local.*)

5.3 Results and Conclusions

Figure 8 provides a comparison of the estimated densities corresponding to Table 2. Several comments can be made about these graphs. First, the fit provided by the different estimates is reasonable (and even more so in the exponential case; see Celeux et al. 1999). Second, the differences between the samplers are negligible for a given estimation method, even though the corresponding parameters (p_j, ζ_j) seem to vary considerably (this is particularly true for the clustering procedure). Third, we can detect a strong similarity between the clustering and the global loss function approaches, even though intuition would suggest a similarity

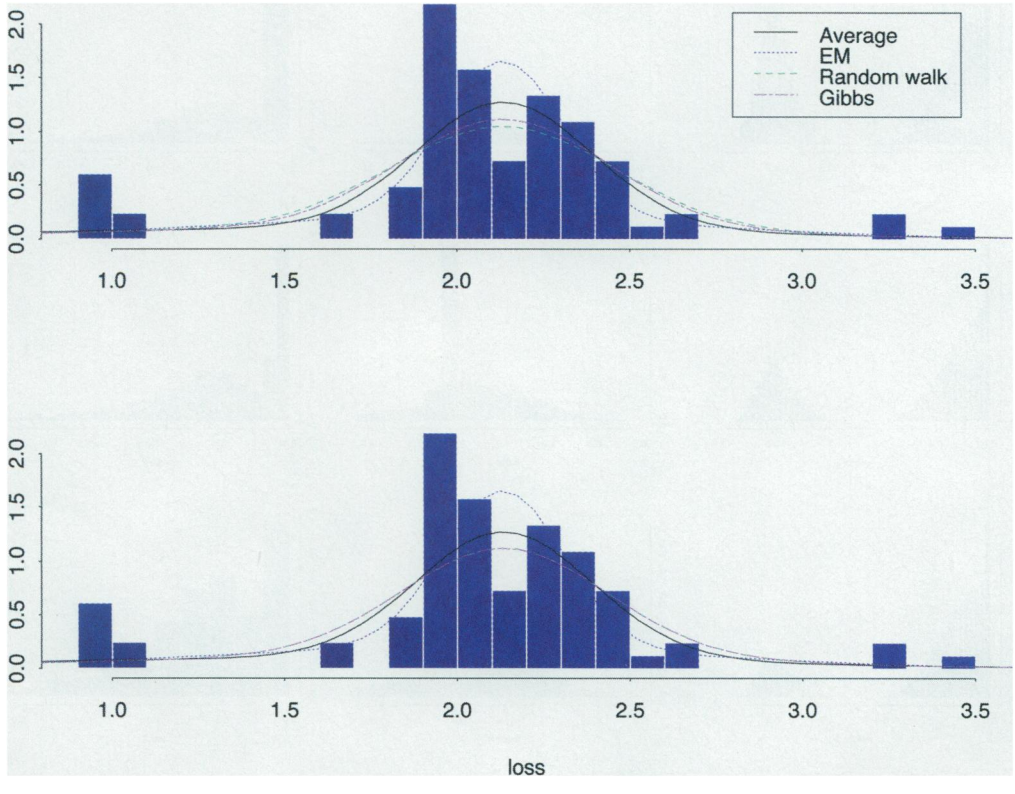

Figure 10. Estimations by the Clustering (a) and the Loss Function (b) Procedures of the Density of the Galaxy Dataset (Represented by the Histogram), for a Two-Component Normal Mixture Modeling, and Comparison With the EM and Crude Averaging Estimates.

with the local loss function. This correspondence is interesting, because it indicates that the clustering technique is a good approximation to a method with decision-theoretic foundations and also that it is not overly dependent on the choice of the distance (5). In particular, Tables 2 and 3 show that both global distances (9) and (10) give very similar estimates (which is why the estimated densities corresponding to the Kullback–Leibler distance have not been reproduced in Figure 9). The agreement with the global distances provides a sounder justification than previously given for the clustering procedure, which is less time-consuming than the loss approach, because the later often requires calls to simulated annealing steps.

6. DISCUSSION

We conclude with an application of our approach to the benchmark galaxy dataset, already used by, for example, Richardson and Green (1997), among many others. In this case, it appears that both the Gibbs sampler and the random walk Metropolis–Hastings algorithms provide well-mixed MCMC samplers, as illustrated in Figure 9 for the three-component case and the random walk Metropolis–Hastings algorithm. The three histograms are quite similar for the weights p_i and the mean θ_i, and are on a common scale for the variances σ_i^2. This setup is thus quite interesting, because it shows that tempering is not always necessary and, more important, that standard datasets such as the galaxy dataset may require advanced processing, such as those developed in this article, because of the label switching phenomenon. Figure 10 shows, for two-component modeling, how both clustering and loss approaches give similar results for both Gibbs sampler and random walk Metropolis–Hastings algorithms (for the loss functions, both graphs cannot be distinguished); the figure also includes a comparison with the EM estimate and the crude average obtained by averaging the estimated densities over iterations,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k} p_i^{(t)}{\sigma_i^{(t)}}^{-1}\exp\{-(x-\theta_i^{(t)})^2/2\sigma_i^{(t)}\}.$$

Note the similarity with the loss estimate. The case $k = 3$ led to very similar estimates, meaning that the normal modeling with moderately vague priors could not capture the tail behavior of the dataset. [Robert and Mengersen (1998) had to use huge values of their dispersion hyperparameter to get a multimodal estimate.]

The overall impression of this article may be that we complicate rather than clarify certain aspects of mixture modeling. We argue that the picture is indeed more complicated than previously thought, but we are equally convinced that this level of complexity is necessary if we want to present a thorough analysis of the mixture problem. We have demonstrated that identifiability constraints can have a potentially detrimental effect on the estimates, with the corresponding implication that the full posterior distribution must be simulated. We have also shown that the standard MCMC samplers are not guaranteed to overcome the attraction effect of local modes, and established the appeal of tempered algorithms. On the inferential side of the problem, we have suggested two approaches to the statistical analysis of both symmetric and nonsymmetric MCMC samples, one based on a clustering approximation and the other based on new choices of loss functions. Although the overall message is indeed one of increased complexity, we do not see this as a deterrent, because mixtures of distributions (and other latent variable models) are a complex area and that calls for advanced procedures.

[Received March 1999. Revised November 1999.]

REFERENCES

Baddeley, A. (1992), "Errors in Binary Images and a L^p Version of the Hausdorff Metric," *Nieuw Archief Voor Wiskunde*, 10, 157–183.

Carota, C., Parmigiani, G., and Polson, N. (1996), "Diagnostic Measures for Model Criticism," *Journal of the American Statistical Association*, 91, 753–762.

Celeux, G. (1998), "Bayesian Inference for Mixtures: The Label Switching Problem," in *COMPSTAT 98*, eds. R. Payne and P. Green, Berlin: Physica-Verlag, pp. 227–232.

Celeux, G., Hurn, M., and Robert, C. P. (1999), "Computational and Inferential Difficulties With Mixture Posterior Distributions," Rapport de recherche INRIA 3627, INRIA, Montbonnot, France.

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Diebolt, J., and Robert, C. P. (1990), "Estimation des Paramètres d'un Mélange par Échantillonnage Bayésien," *Notes aux Comptes-Rendus de l'Académie des Sciences I*, 311, 653–658.

——— (1994), "Estimation of Finite Mixture Distributions by Bayesian Sampling," *Journal of the Royal Statistical Society*, Ser. B, 56, 363–375.

Frigessi, A., and Rue, H. (1997), "Bayesian Image Classification Using Baddeley's Delta Loss," *Journal of Computational and Graphical Statistics*, 6, 55–73.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1996), "Efficient Metropolis Jumping Rules," in *Bayesian Statistics 5*, eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 599–608.

Geyer, C. J., and Thompson, E. A. (1995), "Annealing Monte Carlo Maximum Likelihood With Application to Pedigree Analysis," *Journal of the American Statistical Association*, 90, 909–920.

Gruet, M. A., Philippe, A., and Robert, C. P. (1999), "MCMC Control Spreadsheets for Exponential Mixture Estimation," *Journal of Computational and Graphical Statistics*, 8, 298–317.

Guihenneuc, C., Knight, S., Mengersen, K. L., Richardson, S., and Robert, C. P. (1999), "MCMC Diagnostics in Action," technical report, CREST, Paris.

Liu, J. S., Liang, F., and Wong, W. H. (1998), "The Use of Multiple-Try Method and Local Optimization in Metropolis Sampling," technical report, Stanford University, Dept. of Statistics.

Mengersen, K., and Robert, C. P. (1996), "Testing for Mixtures: A Bayesian Entropic Approach," in *Bayesian Statistics 5*, eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley, and A. F. M. Smith, London: Oxford University Press, pp. 255–276.

Neal, R. (1996), "Sampling From Multimodal Distributions Using Tempered Transitions," *Statistics and Computing*, 4, 353–366.

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 731–792.

Robert, C. P. (1996), "Inference in Mixture Models," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 441–464.

——— (1997), Discussion of "On Bayesian Analysis of Mixtures With an Unknown Number of Components," by S. Richardson and P. J. Green, *Journal of the Royal Statistical Society*, Ser. B, 59, 758–764.

——— (1998), "Specifics of Latent Variable Models," in *COMPSTAT 98*, eds. R. Payne and P. G. Green, Berlin: Physica-Verlag, pp. 101–112.

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

Robert, C. P., and Mengersen, K. L. (1998), "Reparameterization Issues in

Mixture Estimation and Their Bearings on the Gibbs Sampler," *Computations and Statistical Data Analysis*, 29, 325–343.

Roberts, G. O., and Rosenthal, J. S. (1998), "Optimal Scaling of Discrete Approximations to Langevin Diffusions," *Journal of the Royal Statistical Society*, Ser. B, 60, 255–268.

Roberts, G. O., and Tweedie, R. L. (1996), "Exponential Convergence for Langevin Diffusions and Their Discrete Approximations," *Bernoulli*, 2, 341–363.

Roeder, K., and Wasserman, L. (1997), "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of the American Statistical Association*, 92, 894–902.

Rue, H. (1995), "New Loss Functions in Bayesian Imaging," *Journal of the American Statistical Association*, 90, 900–908.

Stephens, M. (1997), "Bayesian Methods for Mixtures of Normal Distributions," unpublished Ph.D. thesis, Oxford University.

Stramer, O., and Tweedie, R. L. (1997), "Geometric and Subgeometric Convergence of Diffusions With Given Stationary Distributions, and Their Discretizations," technical report, University of Iowa.

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