Statistics and Computing (1998) **8**, 145–158

Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation

CHRISTIAN P. ROBERT[1] and D. M. TITTERINGTON[2]

[1] *Laboratoire de Statistique, CREST, INSEE, Timbre J340, 92245 Malakoff cedex, France*
[2] *Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland, UK*

Received August 1996 and accepted October 1997

This paper synthesizes a global approach to both Bayesian and likelihood treatments of the estimation of the parameters of a hidden Markov model in the cases of normal and Poisson distributions. The first step of this global method is to construct a non-informative prior based on a reparameterization of the model; this prior is to be considered as a penalizing and bounding factor from a likelihood point of view. The second step takes advantage of the special structure of the posterior distribution to build up a simple Gibbs algorithm. The maximum likelihood estimator is then obtained by an iterative procedure replicating the original sample until the corresponding Bayes posterior expectation stabilizes on a local maximum of the original likelihood function.

Keywords: Bounded likelihood, identifiability, Gibbs sampler, non-informative prior, normal distribution, Poisson distribution, prior feedback, proper posterior distribution, simulated annealing

1. Introduction

Although prevalent in a large variety of domains (speech recognition, image processing, epidemiological studies, genetics, etc.), hidden Markov chain models face severe difficulties in terms of estimation and inference, in particular when using the maximum likelihood approach. Archer and Titterington (1997) present a survey of the existing methods and draw the conclusion that, at least in certain simple problems, the EM algorithm of Dempster *et al.* (1977) is yet the most reliable approach.

In parallel with maximum likelihood estimation, Bayesian estimation has been considerably facilitated in recent years by the introduction of new computational techniques like the MCMC algorithms, and Bayes estimators for hidden Markov models can currently be computed at little cost in terms of programming and computing time, as shown by Robert *et al.* (1993), Shephard (1994) or Chib (1996). This paper proposes an alternative to Robert *et al.* (1993) in which the prior distribution is arguably less informative. The foundation of our method is a reparameterization of the hidden Markov chain model that naturally leads to a single parameter prior distribution, while providing increased stability of the Gibbs sampler and good convergence properties. This approach was first proposed by Mengersen and Robert (1996) in order to test for the presence of a mixture of two normal distributions, and Robert and Mengersen (1995) showed how it could be extended to an arbitrary number of components for estimation purposes. Further extension to the hidden Markov chain model is the topic of this paper.

Bayesian analysis of the model can moreover lead to a derivation of the maximum likelihood estimators by the so-called 'prior feedback' method developed by Robert (1993) for exponential families and regular mixture distributions. This approach is quite similar to *simulated annealing*, in the sense that the likelihood is considered with increasingly high powers, while the prior distribution stays the same, but it also takes advantage of the MCMC implementation by using replications of the original sample.

We thus achieve in this paper a unified implementation of the Bayesian and likelihood paradigms, which allows us to use the same algorithm for the derivation of estimators in both cases. We start the paper with a description of the

reparameterization technique and of the associated non-informative prior distribution in Section 2. We then construct a more efficient Gibbs sampling implementation than in Robert and Mengersen (1995) in Section 3 and show how it can be used to derive the maximum likelihood estimators in Section 4.1. Sections 4.2 and 4.3 provide simulated and real-life illustrations. Throughout, we consider both normal and Poisson settings.

2. Reparameterization of hidden Markov models

2.1. *The normal case*

Mengersen and Robert (1996) reparameterize the usual two-component normal mixture distribution

$$p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2), \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean, μ, and variance, σ^2, in terms of a *global location-scale* parameter, (θ_1, τ_1),

$$p\mathcal{N}(\theta_1, \tau_1^2) + (1-p)\mathcal{N}(\theta_1 + \tau_1\theta_2, \tau_1^2\tau_2^2), \tag{2}$$

where, for identifiability reasons, the global location-scale parameters correspond to those of the first normal component of the mixture.

Robert and Mengersen (1995) extend Equation 2 to cover a larger number of components while preserving the 'local' aspect of this model, i.e. by expressing each component in terms of the location-scale parameter of the previous component. Starting from the two-component mixture, Equation 2, we can replace the second component, $\mathcal{N}(\theta_1 + \tau_1\theta_2, \tau_1^2\tau_2^2)$, by a mixture

$$\begin{aligned} &p\mathcal{N}(\theta_1, \tau_1^2) + (1-p)\left[q_1\mathcal{N}(\theta_1 + \tau_1\theta_2, \tau_1^2\tau_2^2)\right. \\ &\quad \left. + (1-q_1)\mathcal{N}(\theta_1 + \tau_1\theta_2 + \tau_1\tau_2\theta_3, \tau_1^2\tau_2^2\tau_3^2)\right], \end{aligned}$$

where $(\theta_1 + \tau_1\theta_2, \tau_1\tau_2)$ thus becomes the location-scale parameter for the third component. The extension of this parameterization to the k component mixture is then

$$\begin{aligned} &q_1\mathcal{N}(\theta_1, \tau_1^2) + \sum_{i=2}^{k-1}(1-q_1)\cdots(1-q_{i-1}) \\ &\quad \times q_i\mathcal{N}(\theta_1 + \cdots + \tau_1\cdots\tau_{i-1}\theta_i, \tau_1^2\cdots\tau_i^2) \\ &\quad + (1-q_1)(1-q_2)\cdots(1-q_{k-2}) \\ &\quad \times \mathcal{N}(\theta_1 + \cdots + \tau_1\cdots\tau_{k-1}\theta_k, \tau_1^2\cdots\tau_k^2). \end{aligned} \tag{3}$$

Since Equation 3 is invariant under permutation of the indices, Robert and Mengersen (1995) impose in addition the identifiability constraint that the variances, σ_i are decreasing with i, i.e.

$$\tau_2 \leq 1, \ldots, \tau_k \leq 1, \tag{4}$$

which allows for less informative prior modelling on $\theta = (\theta_1, \ldots, \theta_k)$ and $\tau = (\tau_1, \ldots, \tau_k)$, and induces more stable behaviour on the part of the associated Gibbs sampler.

This reparameterization, called 'splitting' by Robert and Mengersen (1995), can be extended successfully to the case of hidden Markov models with normal observables, namely when the members of the sample $(x_1, \ldots, x_n)$ are conditionally independent, given a set of latent variables, $z = (z_1, \ldots, z_n)$, with

$$\begin{aligned} &x_i|z, x_j\ j \neq i\ \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2), \\ &P(z_i = u|z_1, \ldots, z_{i-1}) = p_{z_{i-1}u}, \qquad (u = 1, \ldots, k); \end{aligned} \tag{5}$$

i.e. the latent variables form a Markov chain. We assume in addition that the first variable, z_1, is distributed from the stationary distribution associated with the transition matrix $\mathbb{P} = (p_{uv})$. Each observation, x_i, is then marginally distributed from a mixture distribution, Equation 3, whose probability weights are based on the stationary distribution for the matrix $\mathbb{P}$, the difference from an ordinary mixture model being that the observations x_i, exhibit a dependence represented by the unobserved states, z_i. As a result of this marginal mixture representation, we can incorporate the above parameterization

$$\begin{aligned} &\mu_j = \theta_1 + \tau_1\theta_2 + \cdots + \tau_1\cdots\tau_{j-1}\theta_j, \\ &\sigma_j = \tau_1\cdots\tau_j \qquad (j = 1, \ldots, k) \end{aligned}$$

on the means, $\mu = (\mu_1, \ldots, \mu_k)$ and standard deviations, $\sigma = (\sigma_1, \ldots, \sigma_k)$, of the different components, while keeping the standard representation of the Markov chain transition matrix, $\mathbb{P}$, because a decomposition of the weights, p_{uv}, as $(1-q_{u1}), \ldots, [1-q_{u(v-1)}]q_{uv}$ forces the same order of importance on all the states u and is moreover redundant if we impose the identifiability constraint, Equation 4 [see Robert and Mengersen (1995), for a discussion on the restrictive effects of this parameterization].

2.2. *The Poisson case*

Consider now a hidden Markov model based on Poisson distributions. Given the latent variable, z_i, the observation, x_i, now follows a Poisson $\mathcal{P}(\lambda_{z_i})$ distribution and, as in Equation 5, z is a Markov chain with transition matrix $\mathbb{P}$. Because the λ_us are usually considered as scale parameters, particularly from a Bayesian point of view, the more natural reparameterization is then to write $\lambda = (\lambda_1, \ldots, \lambda_k)$ as

$$\lambda_u = \tau_1\cdots\tau_u; \qquad (u = 1, \ldots, k); \tag{6}$$

with $\tau_2 \leq 1, \ldots, \tau_k \leq 1$ to provide identifiability.

3. Bayesian estimation of hidden Markov models

3.1. *The normal case*

The hidden Markov model, Equation 5, allows for straightforward Bayesian processing when conjugate-like

priors are used, as shown in Robert *et al.* (1993). The new parameterization is, however, interesting from a Bayesian point of view because it allows for less informative a priori modelling, although fully non-informative priors do not exist in such settings. As shown in Robert and Mengersen (1995), the link between the different components created by the reparameterization dispenses with using a proper prior on (θ_1, τ_1) and allows for the following prior distribution:

$$\pi(\theta_1, \tau_1) = 1/\tau, \qquad \tau_u \sim \mathscr{U}_{(0,1)}, \qquad \theta_u \sim \mathscr{N}(0, \zeta^2) \quad (u = 2, \ldots, k), \tag{7}$$

independently, where the hyperparameter, ζ, is to be specified, although it usually has little bearing on the estimate. Other proper priors can be found on the τ_us and θ_us, for instance $\mathscr{B}\exp(\alpha, 1)$ priors on the τ_us. However, the practical implications of this choice are usually limited, given that it operates on the posterior distribution through factors of the form $\alpha + n_u$, where n_u is of the order of $p_u n$ (see below). We show in Appendix 1 that this improper prior distribution is acceptable, in the sense that the corresponding posterior distribution is proper. The prior on the Markov transition matrix, $\mathbb{P}$, is made of an independent Dirichlet prior, $\mathscr{D}_k(1, \ldots, 1)$ on each of its rows.

Contrary to the approach adopted in Mengersen and Robert (1996), we consider a full Gibbs algorithm for the simulation of the posterior distribution of $(\theta, \tau, \mathbb{P})$. Indeed, when we consider the corresponding version of the prior, Equation 7, for the standard parameterization, in terms of (μ_i, σ_i), the Jacobian of the transform is

$$\mathscr{J} = 1/\sigma_1^2\sigma_2^2 \cdots \sigma_{k-1}^2.$$

The posterior distribution on $(z, \mu, \sigma, \mathbb{P})$ is then

$$\prod_{i=2}^{n} p_{z_{i-1}z_i} \prod_{u=1}^{k} \sigma_u^{-n_u} \exp\Big\{-[n_u(\bar{x}_u - \mu_u)^2 + s_u]/2\sigma_u^2\Big\}\sigma_k^2\sigma_1^{-3} \times \prod_{u=2}^{k} \sigma_u^{-2} \exp[-(\mu_u - \mu_{u-1})^2/2\zeta^2\sigma_{u-1}^2],$$

where n_u, $\bar{x}_u$ and s_u are the size, the average and the sum of squares, respectively, for the observations allocated to component u, i.e. observations such that $z_i = u$. The quantity, n_{uv} is the number of is such that $z_i = u$ and $z_{i+1} = v$; i.e. in terms of indicator functions,

$$n_u = \sum_{i=1}^{n} \mathbb{I}_u(z_i), \quad n_{uv} = \sum_{i=1}^{n-1} \mathbb{I}_u(z_i)\mathbb{I}_v(z_{i+1});$$
$$n_u\bar{x}_u = \sum_{i=1}^{n} \mathbb{I}_u(z_i)x_i, \quad s_u = \sum_{i=1}^{n} \mathbb{I}_u(z_i)(x_i - \bar{x}_u)^2.$$

The conditional distributions for the parameters $\mu, \sigma, \mathbb{P}$ are therefore $(u = 1, \ldots, k)$

$$\pi(p_{u1}, \ldots, p_{uk}) \propto p_{u1}^{n_{u1}} \cdots p_{uk}^{n_{uk}};$$
$$\pi(\sigma_u) \propto \sigma_u^{-n_u - \delta_{u1}} \exp\Big\{-\Big[n_u(\bar{x}_u - \mu_u)^2 + s_u + \delta_{u2}(\mu_{u+1} - \mu_u)^2\zeta^{-2}\Big]\Big/2\sigma_u^2\Big\} \times \mathbb{I}_{(\delta_{u3}\sigma_{u+1}, \sigma_{u-1}/\delta_{u3})}(\sigma_u);$$
$$\pi(\mu_u) \propto \exp\Big[-n_u(\bar{x}_u - \mu_u)\Big/2\sigma_u^2\Big] \times \exp\Big[-\delta_{u3}(\mu_u - \mu_{u-1})^2/2\zeta^2\sigma_{u-1}^2\Big] \times \exp\Big[-\delta_{u2}(\mu_u - \mu_{u+1})^2/2\zeta^2\sigma_u^2\Big];$$

where $\mathbb{I}_A$ denotes the indicator function and the $\delta_{u\ell}$ $(\ell = 1, 2, 3)$ are the indicator variables

$$\delta_{u1} = 3\mathbb{I}_{u=1}, \qquad \delta_{u2} = \mathbb{I}_{u\neq k}, \qquad \delta_{u3} = \mathbb{I}_{u\neq 1};$$

i.e. δ_{u1} is 0 if $u > 1$, $\delta_{k2} = 0$ and $\delta_{13} = 0$. These conditional distributions can therefore be easily simulated because $(1 \leq u \leq k)$

$$(p_{u1}, \ldots, p_{uk}) \sim \mathscr{D}(1 + n_{u1}, \ldots, 1 + n_{uk})$$

and $(1 < u < k)$:

$$\sigma_u^{-2} \sim \mathscr{G}a\left[\frac{n_u + \delta_{u1} - 1}{2}, \frac{n_u(\bar{x}_u - \mu_u)^2 + s_u^2 + \delta_{u2}(\mu_{u+1} - \mu_u)^2\zeta^{-2}}{2}\right]\mathbb{I}_{(\delta_{u2}\sigma_{u+1}, \sigma_{u-1})/\delta_{u3}}(\sigma_u),$$
$$\mu_u \sim \mathscr{N}\left[\frac{n_u\bar{x}_u + \zeta^{-2}(\delta_{u2}\mu_{u+1} + \delta_{u3}\mu_{u-1}\sigma_u^2/\sigma_{u-1}^2)}{n_u + \zeta^{-2}(\delta_{u2} + \delta_{u3}\sigma_u^2/\sigma_{u-1}^2)}, \frac{\sigma_u^2}{n_u + \zeta^{-2}(\delta_{u2} + \delta_{u3}\sigma_u^2/\sigma_{u-1}^2)}\right].$$

The simulation of the z_is follows from Robert *et al.* (1993), the conditional distributions being then $(1 \leq u \leq k)$

$$P(z_1 = u|z_2, \ldots, \mathbb{P}) \propto p_{uz_2} f(x_1|\mu_u, \sigma_u);$$
$$P(z_i = u|, \ldots, z_{i-1}, z_{i+1}, \ldots, \mathbb{P}) \propto p_{z_{i-1}u}p_{uz_{i+1}} f(x_i|\mu_u, \sigma_u), \quad (1 < i < k);$$
$$P(z_k = u|, \ldots, z_{k-1}, \ldots, \mathbb{P}) \propto p_{z_{k-1}u} f(x_k|\mu_u, \sigma_u); \tag{9}$$

where $f(\cdot|\mu, \sigma)$ denotes the density of the normal $\mathscr{N}(\mu, \sigma^2)$ distribution.

The Gibbs sampler is then run by simulating successively from the distributions in Equation 8 and 9, and replacing the conditioning parameters by their current values (see Gelfand and Smith, 1990). This implementation is thus clearly simpler than the one proposed in Mengersen and Robert (1996), which requires Metropolis steps to simulate some parameters. Because both algorithms aim at simulating from the same posterior distribution, a direct comparison is possible, to the advantage of the present implementation in terms of programming complexity and convergence speed.

As in the case of mixtures of distributions, the identifiability constraint, Equation 4, may retard convergence. Indeed, the algorithm sometimes produces *homogeneous inverted subsamples*, in the sense that the subsample is correctly identified as distributed from the same component but the labelling of this component is incorrect. For instance, if $k = 2$, the algorithm may divide the sample into $(x_{i_1}, \ldots, x_{i_{n_1}})$ and $(x_{i_{n_1+1}}, \ldots, x_{i_n})$ and associate them with $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, although they were generated from $\mathcal{N}(\mu_2, \sigma_2^2)$ and $\mathcal{N}(\mu_1, \sigma_1^2)$ respectively. The parameter (μ_1, σ_1) is then estimated from the wrong subsample, $(x_{i_1}, \ldots, x_{i_{n_1}})$ and this inversion leads to a value of τ_2 that is close to one and generally to a positively biased estimator of σ_1. Apart from visual checking procedures (close variance estimators and/or components with very large variances), this bias can be eliminated by a warm-up algorithm, with a Metropolis step every γ iterations of the Gibbs sampler, which proposes a random permutation of the subsamples and the corresponding simulation of the parameter. Mengersen and Robert (1996) adopt the alternative of running a preliminary warm-up step without imposing the constraint in Equation 4 on the σ_is.

Figure 1 illustrates the performance of the Gibbs sampler, based on 341 observations from a normal hidden Markov model with transition matrix,

$$\mathbb{P} = \begin{pmatrix} 0.62 & 0.25 & 0.13 \\ 0.09 & 0.18 & 0.73 \\ 0.21 & 0.62 & 0.13 \end{pmatrix},$$

and corresponding components $\mathcal{N}[0.25, (2.28)^2]$, $\mathcal{N}[3.37, (0.61)^2]$ and $\mathcal{N}[2.45, (0.56)^2]$. Along with the histogram of the sample, Fig. 1 displays the estimated density of the form, Equation 3, when mean and variance parameters are replaced with the posterior expectations and the weights of the mixture are derived as the stationary distribution for the estimated matrix,

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.51 & 0.34 & 0.15 \\ 0.20 & 0.10 & 0.70 \\ 0.19 & 0.73 & 0.08 \end{pmatrix}.$$

The fit of the estimated density is therefore quite satisfactory. (A larger number of iterations in the Gibbs sampler produces identical estimates, as do modifications in the starting values or the hyperparameter, ζ.) Figure 1 contains as an insert a comparison of the true state (known from the simulation) and the most likely state for each observation. Table 1 provides an evaluation of the predictive properties of the posterior distribution by comparing the observed rates of correct allocations, based on the modes of the posterior distributions of the z_is, with the same rates based on the true values of the parameter.

3.2. *The Poisson case*

A similar treatment applies for this hidden Markov model because the prior distribution can be chosen to be almost as non-informative as one might desire. For instance, the natural non-informative prior

$$\pi(\tau_1) = 1/(\tau_1)^{1/2}, \qquad \tau_u \sim \mathcal{U}_{(0,1)} \qquad (u = 2, \ldots, k)$$

is acceptable, in the sense that the posterior is proper (see Appendix 1), if used along with Dirichlet priors, $\mathcal{D}(1, \ldots, 1)$, on the rows of $\mathbb{P}$. The Jacobian of the transform is

Table 1. *Rates of correct allocations of the observations to the states based on the estimated parameters compared with the optimal rates based on the true parameters*

Case	Sample size	Observed rate	Optimal rate
Normal	341	247	252
Poisson	129	83	94
Poisson	555	344	337

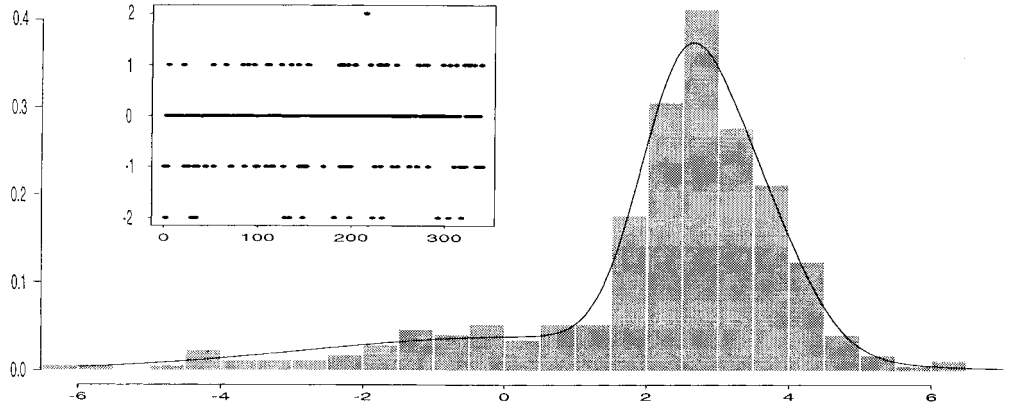

Fig. 1. *Estimation of a simulated normal model, including the histogram for the 341 data points and graph of the estimated marginal density, based on 5000 iterations of the Gibbs sampler. The final estimates are* $\hat{\mu} = (0.207, 3.34, 2.44)$ *and* $\hat{\sigma}^2 = (8.38, 0.571, 0.357)$. *The insert represents the difference of the true states and of the most probable states*

$$\mathscr{J} = 1/\lambda_1 \cdots \lambda_{k-1};$$

and the posterior distribution on $(z, \lambda, \mathbb{P})$ is then

$$\prod_{i=2}^{n} p_{z_{i-1}z_i} \prod_{u=1}^{k} \lambda_u^{s_u - 1} \exp(-n_u\lambda_u) \frac{\lambda_k}{\lambda_1^{1/2}} \mathbb{I}_{\lambda_1 \geq \cdots \geq \lambda_k};$$

where n_u denotes the number of observations associated with $z_i = u$ and s_u is the sum of these observations (with the convention that $s_u = 0$ when $n_u = 0$). The conditional distributions for the Gibbs sampler follow immediately:

$$\begin{aligned}
(p_{u1}, \ldots, p_{uk}) &\sim \mathscr{D}(1 + n_{u1}, \ldots, 1 + n_{uk}), \qquad (1 \leq u \leq k);\\
\lambda_1 &\sim \mathscr{G}a(s_1 - 1/2, n_1)\mathbb{I}_{(\lambda_2, \infty)};\\
\lambda_u &\sim \mathscr{G}a(s_u, n_u)\mathbb{I}_{(\lambda_{u+1}, \lambda_{u-1})}, \qquad (1 < u < k);\\
\lambda_k &\sim \mathscr{G}a(s_k + 1, n_k)\mathbb{I}_{(0, \lambda_{k-1})};
\end{aligned} \tag{10}$$

where n_{uv} is the number of is with $z_i = u$ and $z_{i+1} = v$. The simulation of the z_is is similarly straightforward:

$$\begin{aligned}
P(z_1 = u | z_2, \ldots, \mathbb{P}) &\propto p_{uz_2}\lambda_u^{x_1} \exp(-\lambda_u);\\
P(z_i = u |, \ldots, z_{i-1}, z_{i+1}, \ldots, \mathbb{P}) &\propto p_{z_{i-1}u}p_{uz_{i+1}}\lambda_u^{x_i} \exp(-\lambda_u),\\
&\qquad (1 < i < n);\\
P(z_n = u |, \ldots, z_{n-1}, \ldots, \mathbb{P}) &\propto p_{z_{n-1}u}\lambda_u^{x_n} \exp(-\lambda_u).
\end{aligned} \tag{11}$$

There is thus no difficulty with the Gibbs sampler in this setup. For instance, we considered a simulated data set of 129 observations from the three Poisson distributions $\mathscr{P}(4.5)$, $\mathscr{P}(1.2)$ and $\mathscr{P}(0.43)$, associated with the transition matrix

$$\mathbb{P} = \begin{pmatrix} 0.33 & 0.64 & 0.03 \\ 0.47 & 0.14 & 0.37 \\ 0.18 & 0.30 & 0.52 \end{pmatrix}.$$

The corresponding posterior expectations, $\hat{\lambda}_1 = 3.89$, $\hat{\lambda}_2 = 1.12$ and $\hat{\lambda}_3 = 0.39$ are quite close to the true values. The estimated matrix,

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.47 & 0.29 & 0.24 \\ 0.37 & 0.28 & 0.35 \\ 0.30 & 0.31 & 0.39 \end{pmatrix},$$

does not capture the true Markovian structure of the model very well because it suggests a model that is almost an independent Poisson mixture. The Gibbs sampler is not to blame because larger numbers of iterations do not show significant changes in the estimates. The discrete nature of the Poisson distribution together with the small sample size are actually responsible for the apparently limited performance of the Bayes estimates in this case. However, Table 1, which displays the rates of correct allocations, shows that the performance of the posterior distribution is quite reasonable.

A second example is based on 555 observations from four possible components, $\mathscr{P}(8.0)$, $\mathscr{P}(5.0)$, $\mathscr{P}(2.5)$ and $\mathscr{P}(0.5)$, with transition matrix,

$$\mathbb{P} = \begin{pmatrix} 0.33 & 0.64 & 0.02 & 0.01 \\ 0.37 & 0.43 & 0.14 & 0.06 \\ 0.08 & 0.10 & 0.46 & 0.36 \\ 0.03 & 0.05 & 0.64 & 0.28 \end{pmatrix}.$$

The estimated parameters are $\hat{\lambda}_1 = 7.16$, $\hat{\lambda}_2 = 3.77$, $\hat{\lambda}_3 = 2.19$ and $\hat{\lambda}_4 = 0.44$, which are within range of the true values. The transition matrix is not particularly well estimated, though, as

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.62 & 0.31 & 0.05 & 0.02 \\ 0.35 & 0.30 & 0.21 & 0.14 \\ 0.12 & 0.20 & 0.23 & 0.45 \\ 0.07 & 0.17 & 0.53 & 0.23 \end{pmatrix}.$$

(Note the confusion between the two first states and between the two last states.) The allocation of the observations to the most probable state is nonetheless comparable with the allocation based on the true value of the parameters (see Table 1).

4. The prior feedback approach to maximum likelihood estimation

4.1. *Presentation*

Several approaches have exploited Gibbs sampling methods to come up with new approximation techniques for maximum likelihood estimation, thus circumventing the EM algorithm approach (see for instance Geyer and Thomson, 1992). The method we consider in this paper is based on a remark, recurrent in the literature (Pincus, 1968; Rubinstein, 1981; Aitkin, 1991), that the effect of the prior distribution fades away when the likelihood function is taken to a high enough power. When the computation of the Bayesian posterior expectation is straightforward, it makes sense to consider this alternative approach to likelihood maximization. Robert (1993) develops a version of this Bayesian approach to maximum likelihood estimation under the name of *prior feedback* and obtains some theoretical results about the validity of the method for mixture estimation (see also Robert and Soubiran, 1993), while stressing the connections with the more general technique of simulated annealing.

In the present context of hidden Markov models, the prior feedback can be implemented by running the Gibbs sampler with an increasing number of replications of the original sample $(x_1, \ldots, x_n)$, acting as if these were iid observations from Equations 5 or 6, until the Bayesian posterior expectations stabilize. More precisely, the sample $(x_1, \ldots, x_n)$ is replicated into copies $(x_1^{(1)}, \ldots, x_n^{(1)}), \ldots, (x_1^{(m)}, \ldots, x_n^{(m)})$. Each replication $(x_1^{(t)}, \ldots, x_n^{(t)})$ is then associated with a separate set of allocations, $(z_1^{(t)}, \ldots, z_n^{(t)})$. The modified Gibbs sampler then consists of two steps. First, for each replication of the sample, generate the

corresponding $z_i^{(t)}$s according to Equation 9 and derive the sufficient statistics $(u, v = 1, \ldots, k)$

$$n_u = \sum_{i=1}^{n}\sum_{t=1}^{m} \mathbb{I}_u\left(z_i^{(t)}\right), \quad n_{uv} = \sum_{i=1}^{n-1}\sum_{t=1}^{m} \mathbb{I}_u\left(z_i^{(t)}\right)\mathbb{I}_v\left(z_{i+1}^{(t)}\right),$$

$$n_u\bar{x}_u = \sum_{i=1}^{n}\sum_{t=1}^{m} \mathbb{I}_u\left(z_i^{(t)}\right)x_i, \quad s_u = \sum_{i=1}^{n}\sum_{t=1}^{m} \mathbb{I}_u\left(z_i^{(t)}\right)(x_i - \bar{x}_u)^2.$$

Secondly, generate the parameters $(\mu, \sigma, \mathbb{P})$ as in Equation 8.

The theoretical justification of the prior feedback method is that the posterior distribution, π_m, is associated with m replications of the same sample, i.e.

$$\pi_m(\theta) \propto \pi(\theta)\mathrm{L}(\theta)^m,$$

where θ denotes the generic parameter and $\mathrm{L}(\theta)$ the original likelihood function, is bound to converge to a Dirac mass located at the maximum of $\mathrm{L}(\theta)$ as m goes to infinity, under some regularity conditions (see Rubinstein, 1981). The index size, m, can then be interpreted as the temperature factor in simulated annealing (see Duflo, 1996). However, the regularity conditions do not immediately hold *in the normal case* because the likelihood function, $\mathrm{L}(\mu, \sigma, \mathbb{P})$, is unbounded and, therefore, there is no proper maximum likelihood estimator.

This difficulty can be circumvented by considering the cause of the unboundedness. The likelihood function, $\mathrm{L}(\mu, \sigma, \mathbb{P})$, goes to infinity when a σ_u goes to zero and the corresponding μ_u converges to one of the x_i (see Titterington *et al*., 1985). If the σ_us can be prevented from going to zero then the likelihood stays bounded and attention can focus on interior and local maxima. This goal can be achieved by a modification of the prior distribution, Equation 7. Indeed, as it stands,

$$\pi(\mu, \sigma) \propto \frac{\prod_{u=2}^{k-1} \exp[-(\mu_u - \mu_{u-1})^2/2\zeta^2\sigma_{u-1}^2]}{\sigma_1^3\sigma_2^2\cdots\sigma_{k-1}^2}\ \mathbb{I}_{\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_k} \tag{12}$$

puts too much weight on the neighbourhood of $\sigma_u = 0$ to ensure boundedness of π_m, although it leads to a proper posterior distribution. We thus need to change the denominator of Equation 12 in order to avoid the unbounded likelihood at zero. An acceptable modification to Equation 12 is

$$\pi(\mu, \sigma) \propto \prod_{u=2}^{k-1} \exp\left[-(\mu_u - \mu_{u-1})^2/2\zeta^2\sigma_{u-1}^2\right] \times \sigma_1^{-2k+1}\sigma_k\ \mathbb{I}_{\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_k}, \tag{13}$$

because it provides both a proper posterior distribution and a bounded likelihood function (see Appendices 2 and 3). The modification from Equation 11 to 12 is minor in terms of consequences for the Gibbs sampler, because it only changes the power of the γ distributions in Equation 8. Obviously, the resulting prior cannot be justified directly from a 'non-informative' point of view. Instead, it must be considered as a computational device whose role is to provide access to the maximum likelihood estimator, rather than from a Bayesian perspective.

The changes justifying the prior feedback method are very minor in the *Poisson case* because, if there is a single observation different from zero, the posterior distribution, i.e. the penalized likelihood, is bounded (see Appendix 2) and, therefore, the original prior distribution of Section 3.2 can be used. When all the observations are zero, the prior on $(\lambda_1, \ldots, \lambda_k)$ must be modified to

$$\pi^*(\lambda_1, \ldots, \lambda_k) = \frac{1}{\lambda_1\cdots\lambda_{k-1}},$$

which still provides a proper posterior distribution (see Appendices 2 and 3). (This unusual case is of hardly any practical interest because there is little to be inferred from an uninterrupted sequence of zeros!)

The prior feedback method is then applied with these modified non-informative priors, because the modification ensures that the penalized likelihood remains bounded and thus that a maximum does exist. In principle, the posterior expectation associated with π_m should converge to the global maximum of L (within some truncated version of the support). However, the multimodal structure of the likelihood function and the difficulty the Gibbs sampler has in leaving the neighbourhood of a strongly attractive mode lead to the conclusion that the prior feedback estimate will converge to a local maximum of L. This lack of guaranteed convergence can be seen in practice if we start the procedure with several replications of the sample, because the method thus usually produces suboptimal estimates. Although unfortunate, this feature is common to recursive algorithms in multimodal settings, including EM. In Example 4.3 below we obtain, for instance, estimates of the parameters that are different from the EM estimates, while the two corresponding values of the likelihood are very close.

4.2. *Simulated examples*

The normal example used in Section 3.1 can be processed by the prior feedback method. Figure 2 shows that convergence may be slow for some posterior expectations of the means and variances of the components: after 30 iterations of the prior feedback method, each involving 5000 iterations of the Gibbs sampler, some prior feedback estimates are not yet stable, although close to the true values of the parameters. At the end of the 30 iterations, the estimated transition matrix is

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.65 & 0.27 & 0.09 \\ 0.20 & 0.07 & 0.73 \\ 0.16 & 0.82 & 0.02 \end{pmatrix},$$

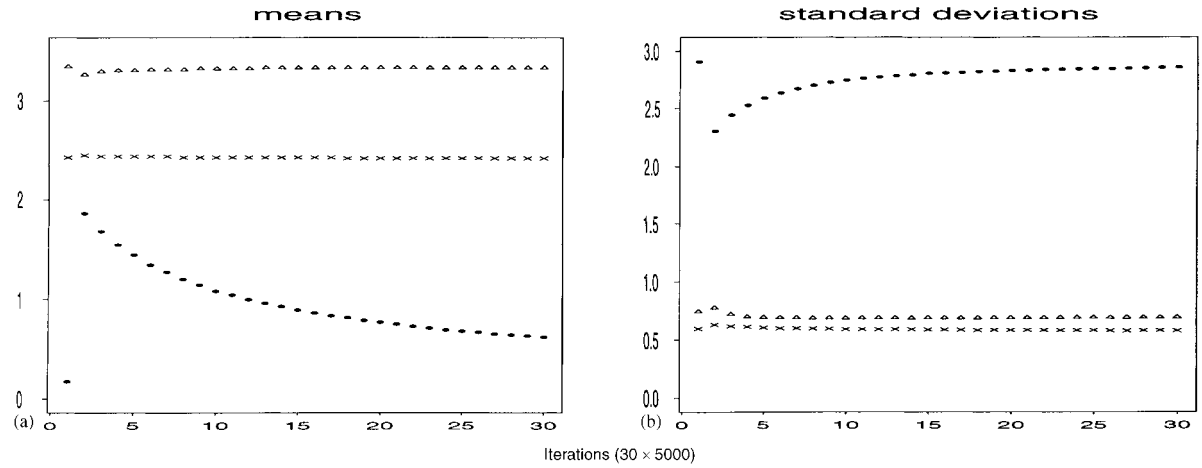

Fig. 2. *Convergence of the prior feedback estimates of the parameters of the normal hidden Markov model corresponding to the data of Fig. 1. The limiting values after 30 replications of the original sample are 0.61, 2.42 and 3.34 for the means (a), compared with the true values 0.25, 2.45 and 3.37, and 2.86, 0.58 and 0.70 for the standard deviations (b), compared with 2.28, 0.56 and 0.61*

which does not differ considerably from the corresponding Bayes estimate (see Section 3.1), while being closer to the true matrix.

In the first Poisson example (129 observations), the convergence of the three estimates of the λ_us is described by Fig. 3a, which shows fast stabilization of the method. The estimate of the transition matrix takes more iterations to stabilize, however, and the prior feedback estimate after 30 iterations is

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.44 & 0.55 & 0.01 \\ 0.41 & 0.01 & 0.59 \\ 0.18 & 0.76 & 0.07 \end{pmatrix}.$$

The second simulated example with 555 observations leads to a similar conclusion, as shown by Fig. 3b. The stable estimates are in this case closer to the true values than the Bayes posterior expectation of Section 3.2, because $\hat{\lambda}_1 = 7.86$, $\hat{\lambda}_2 = 4.87$, $\hat{\lambda}_3 = 2.45$ and $\hat{\lambda}_4 = 0.361$. The estimate of the transition matrix,

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.19 & 0.81 & 0.00 & 0.00 \\ 0.64 & 0.02 & 0.33 & 0.01 \\ 0.01 & 0.26 & 0.27 & 0.45 \\ 0.06 & 0.00 & 0.74 & 0.20 \end{pmatrix},$$

is rather less satisfactory, given the zero entries that do not exist in the original matrix $\mathbb{P}$ (although the corresponding true values are quite small) and also given $\hat{p}_{22} = 0.02$, which estimates $p_{22} = 0.43$.

4.3. *Real-life data sets*

Example 4.1: A frequently analysed data set is the sunspots series, which is available in S-Plus. Although the original series is far from being a stationary Markov chain, because of long-term dependencies, a hidden Markov model can be fitted to the transform $y_t = L^Y[L^M(x_t)]$, where $L^M(x_t) = x_t - x_{t-1}$ and $L^Y(x_t) = x_t - x_{t-13}$, because it

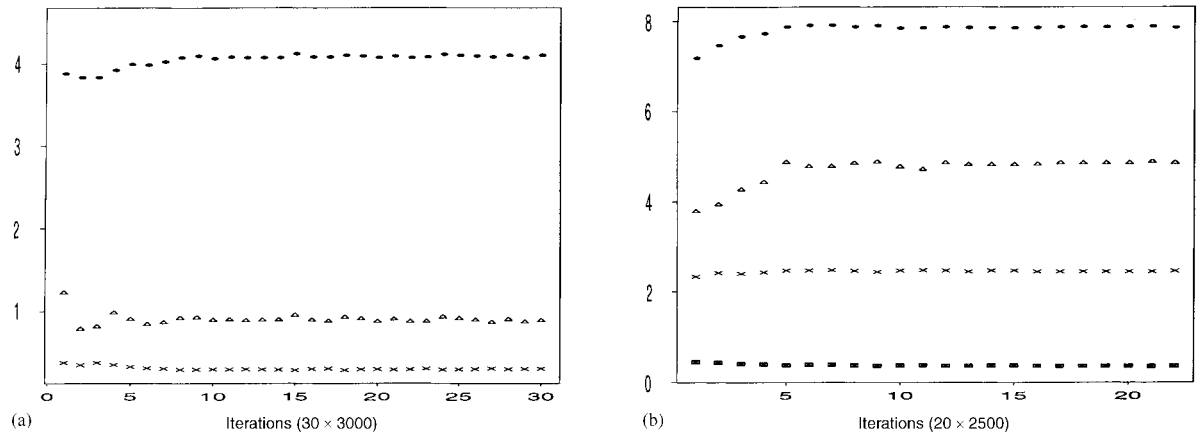

Fig. 3. *Convergence of the prior feedback estimates of the scale parameters of the two Poisson hidden Markov models considered in Section 3.1*

eliminates the seasonal periodicity. The new series of 2720 values, y_t, is represented in Figs. 4 and 5 gives the histogram of the series along with the prior feedback estimate of the density, based on two components in the hidden Markov model. The convergence graphs for the ten iterations of the prior feedback algorithm are not reproduced because the estimates are quite stable.

Example 4.2: Another standard series available in S-Plus consists of 299 consecutive observations of the waiting times between eruptions of the Old Faithful geyser in the Yellowstone National Park (see also Azzalini and Bowman, 1990). These authors concentrated more on the duration times of the eruptions, but noted the predictably high correlation between an eruption duration and the next waiting time; thus, qualitatively similar models should apply to both series. We chose the waiting times because they were complete. Although Azzalini and Bowman (1990) provide some physical motivation for fitting a two-state Markov chain to the durations data set, they showed that the associated partial autocorrelation functions did not display the correct behaviour. Instead, they fitted a second-order Markov chain model, with somewhat greater success. The states corresponded to the following combinations of duration times for two consecutive eruptions, each discretized as 'short' or 'long': (short, long), (long, short) and (long, long).

We thus compare the fits of the waiting-times series assuming two-state and three-state normal hidden Markov chains, the result being that the fit is noticeably superior with a three-state Markov chain, and with the additional bonus that the prior feedback method converges faster for the three-state case. The final result for the two-state model is given in Fig. 6, and the sequence of prior feedback estimates for the three-state model is presented in Fig. 7, against the histogram of the data set. Note the overall stability of the estimated density. The estimated parameters are $\hat{\mu}_1 = 55.4$, $\hat{\mu}_2 = 84.9$, $\hat{\mu}_3 = 75.4$, $\hat{\sigma}_1^2 = 35.8$, $\hat{\sigma}_2^2 = 29.9$ and $\hat{\sigma}_3^2 = 14.4$, while the estimated transition matrix is then

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.001 & 0.995 & 0.004 \\ 0.667 & 0.062 & 0.271 \\ 0.306 & 0.123 & 0.571 \end{pmatrix}.$$

The stability of the solution and the attraction of the central mode ($\hat{\mu}_3 = 75.4$ and $\hat{\sigma}_3^2 = 14.4$) support the observation of Azzalini and Bowman (1990) that 'a low level

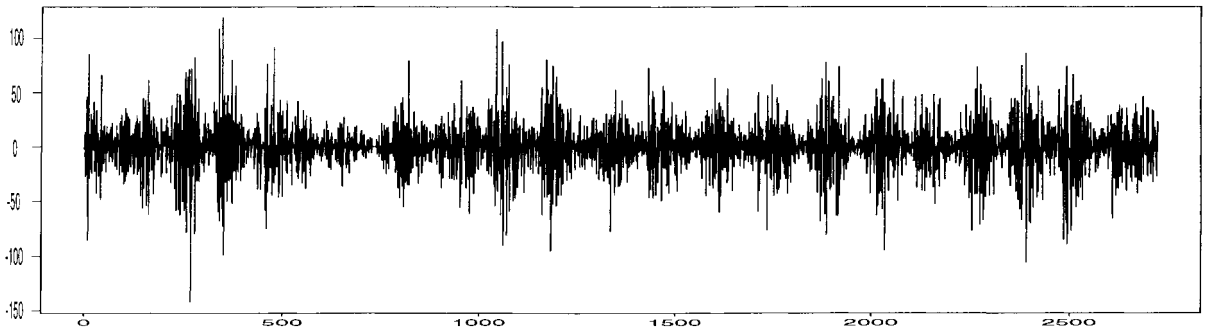

Fig. 4. *Transformed differences of the sunspot data set from 1750 to 1973*

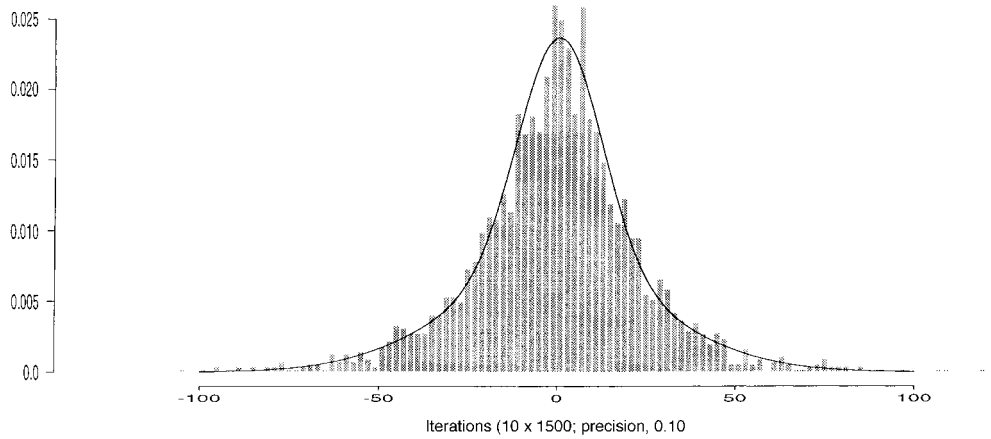

Fig. 5. *Graph of the estimated marginal density and histogram for the transformed sunspot data set*

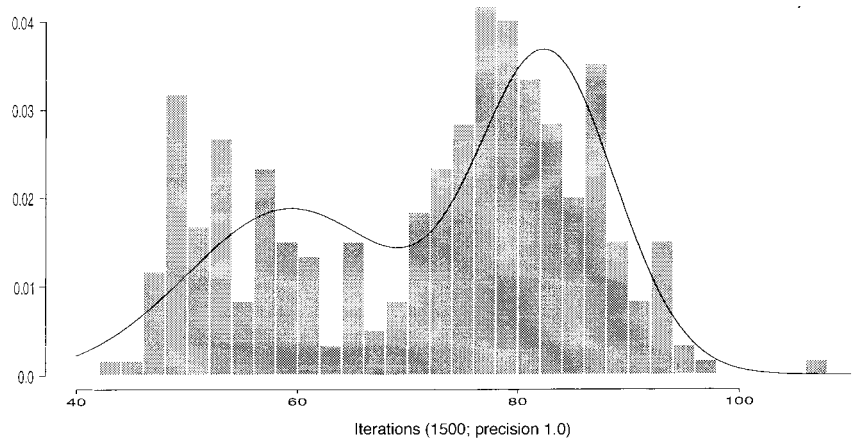

Fig. 6. *Prior feedback estimate of the density for the Old Faithful geyser data and a two-state Markov chain representation*

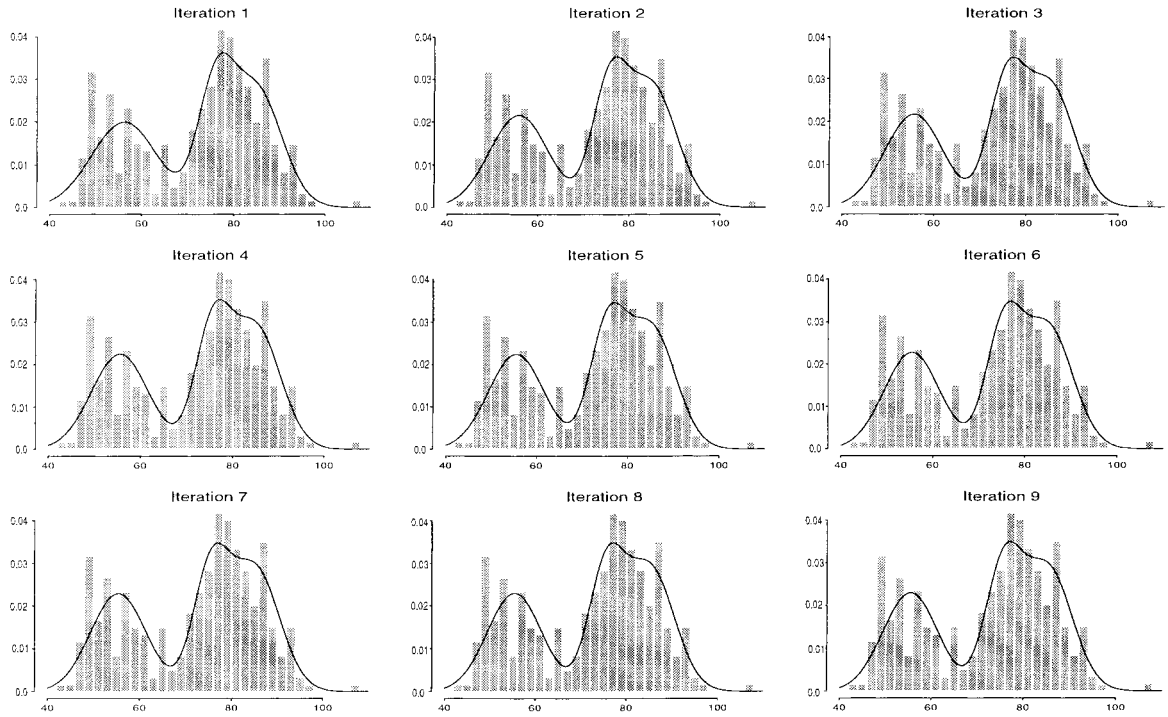

Fig. 7. *Convergence of the prior feedback estimate of the density for the Old Faithful geyser data and a three-state Markov chain representation (2500 Gibbs iterations, precision 1.00)*

is always followed by a high level, and a high level is often, but not always, followed by a low level'. The first row of the transition matrix reinforces the first assertion and the rest of the model accommodates two 'high-level' states, from the lower of which one is likely to move into one or other of the other two states, in line with their second assertion. Note that an extension of the reversible jump method implemented in Richardson and Green (1997) for the analysis of normal mixtures could reinforce this evaluation of the numbers of components in a more quantitative way.

Example 4.3: Chib (1996) considers the foetal movement data, analysed in Leroux and Puterman (1992), which consists of 240 consecutive numbers of foetal lamb movements. The underlying model is a Poisson hidden Markov model with two or three components. Chib (1996) uses a proper conjugate prior approach to derive Bayes estimators for the scale parameter and transition probabilities, but our improper non-informative prior leads to the same estimates, both for two and three components. Figure 8 shows the convergence of the prior feedback estimates for two components and Fig. 9 does the same for three com-

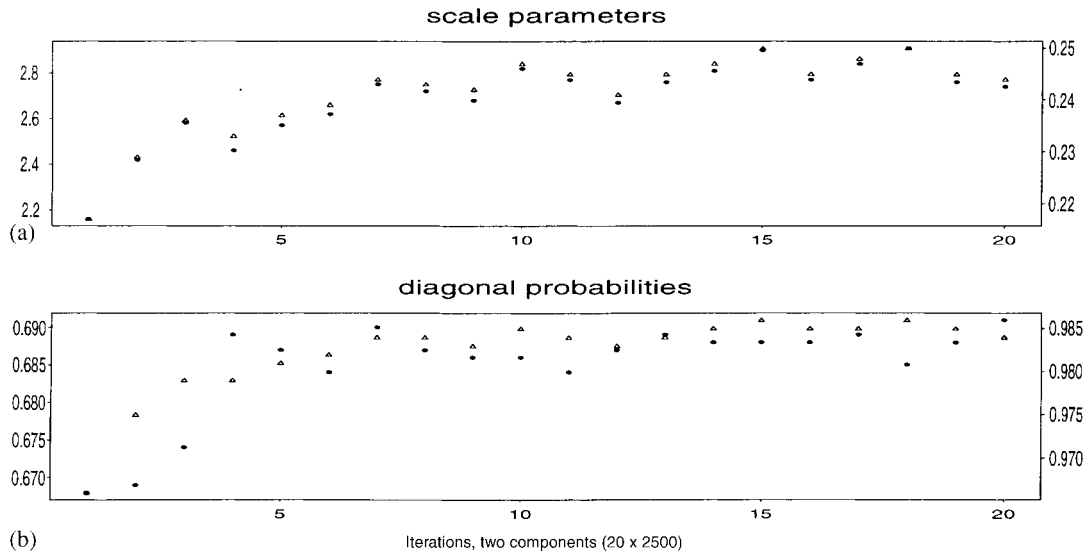

Fig. 8. *Convergence of the prior feedback estimates for (a) scale parameters and (b) diagonal probabilities of a two component Poisson hidden Markov based on the data of Leroux and Puterman (1992): (•) first and (△) second components, respectively*

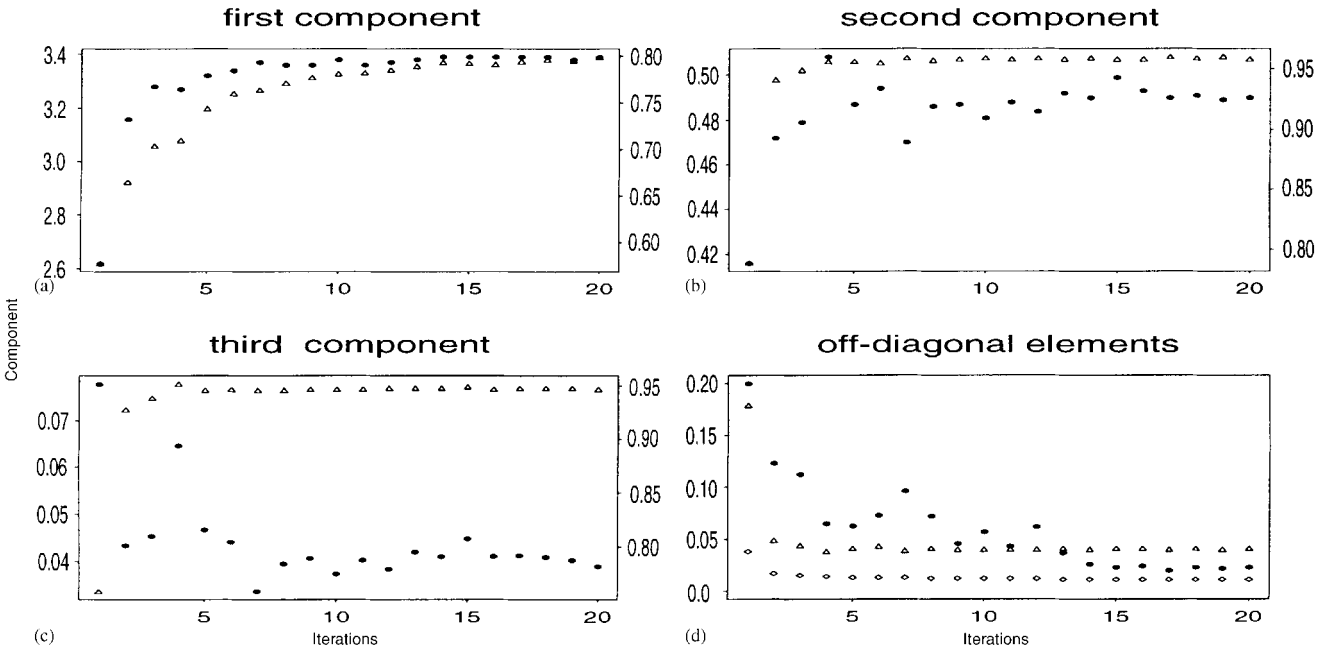

Fig. 9. *Convergence of the prior feedback estimates (•) of the parameters of a three component parameters of a three component Poisson hidden Markov for the same data set. For (a–c): (△) estimates of diagonal probabilities. For (d):(•) p_{12}, (△) p_{23}, and (◇) p_{31}*

ponents. While the resulting estimates are very close to Chib's (1996) SEM estimates they are not identical and we can only conjecture that the difference can be explained by either convergence to another mode of the likelihood function or incomplete convergence in the case of Chib's (1996) algorithm. Table 2 provides Chib's (1996) estimate as well as our estimate, which remains identical for a series of random starting values and allocations. Note that the exact likelihoods, which can be computed by the backward algorithm (see Baum *et al.*, 1970), are also very close.

Table 2. *Comparison of Chib's (1996) SEM and prior feedback estimates, with the likelihood values (multiplied by 10^{78}).*

Parameters	SEM	Prior feedback[a]
λ_1	2.93	2.84
λ_2	0.26	0.25
p_{11}	0.72	0.68
p_{22}	0.99	0.985
L	7.539	7.686

[a] The prior feedback estimates are based on Gibbs samples of size 5000 and on 20 iterations of the prior feedback method

Acknowledgements

This research was partially supported by a visiting grant from the European Science Foundation, through the Highly Structured Stochastic Systems Programme, October–November 1995. The referees' comments on the original version of the paper were much appreciated.

References

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society*, *Series B*, **53**, 111–42.

Archer, G. E. B. and Titterington, D. M. (1997) *Parameter estimation for hidden Markov Chains*. Technical Report, University of Glasgow, Department of Statistics (submitted).

Azzalini, A. and Bowman, A. W. (1990) A look at some data on the Old Faithful geyser. *Applied Statistics*, **39**, 357–65.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–71.

Chib, S. (1996) Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75**, 79–97

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *Series B*, **39**, 1–38.

Duflo, M (1996) *Algorithmes Stochastiques*. Coll. Mathématiques et Applications, Vol. **23**, Berlin: Springer-Verlag.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of Royal Statistical Society*, *Series B*, **54**, 657–99.

Leroux, B. G. and Puterman, M. L. (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, **48**, 545–58.

Mengersen, K. and Robert, C. P. (1996) Testing for mixtures: a Bayesian entropic approach. In J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith (eds), *Bayesian Statistics*, Vol. 5, 255–76, Oxford University Press.

Pincus, M. (1968) A closed form solution of certain programming problems. *Operational Research*, **18**, 1225–8.

Richardson, S. T. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–92.

Robert, C. P. (1993) Prior Feedback: Bayesian tools for maximum likelihood estimation. *Journal of Computational Statistics*, **8**, 279–94.

Robert, C. P., Celeux, G. and Diebolt, J. (1993) Bayesian estimation of Hidden Markov Models: a stochastic implementation. *Statistics and Probability Letters*, **16**(1), 77–83.

Robert, C. P. and Mengersen, K. L. (1995) *Reparameterization Issues in Mixture Modelling and their Bearings on the Gibbs Sampler*. Technical Report No. 9538, CREST, Insee, Paris.

Robert, C. P. and Soubiran, C. (1993) Estimation of a mixture model through Bayesian sampling and prior feedback. *Test*, **2**, 125–46.

Rubinstein, R. Y. (1981) *Simulation and the Monte Carlo Method*. New York: Wiley.

Shephard, N. (1994) Partial non-Gaussian state space. *Biometrika*, **81**, 115–31.

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

Appendix 1: Proof that the posterior is indeed proper

Normal case: The posterior distribution is a sum over all possible partitions of the sample into k subsamples. It is therefore sufficient to prove that the integral of an arbitrary term in this sum is finite. Consider, therefore,

$$\prod_{u=1}^{k} \sigma_u^{-n_u} \exp\left[-\frac{n_u(\bar{x}_u-\mu_u)^2+s_u}{2\sigma_u^2}\right]\pi(\mu,\sigma), \tag{A1}$$

where the terms in $\mathbb{P}$ have been integrated out. The n_us satisfy $n_1+\cdots+n_k=n$ and the subsample mean and sum of squared differences $\bar{x}_u$ and s_u are put to zero when $n_u=0$. (A convention here is that C will denote a generic constant.) The prior distribution, $\pi(\mu,\sigma)$, is derived from

$$\pi(\theta,\tau)=\frac{1}{\tau_1}\exp\left(-\sum_{u=2}^{k}\theta_u^2/2\zeta^2\right),$$

by computing the Jacobian, $\mathcal{J}$, of the transform from (θ,τ) to (μ,σ):

$$\begin{aligned}\mathcal{J} &= \tau_1(\tau_1\tau_2),\ldots,(\tau_1,\ldots,\tau_{k-1})\tau_1(\tau_1\tau_2),\ldots,(\tau_1,\ldots,\tau_{k-1})\\ &= \sigma_1^2,\ldots,\sigma_{k-1}^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\pi(\mu,\sigma) &= \sigma_1^{-3}\sigma_2^{-2}\cdots\sigma_{k-1}^{-2}\\ &\quad\times\exp\left[-\frac{1}{2}\sum_{u=2}^{k}\frac{(\mu_u-\mu_{u-1})^2}{\zeta^2\sigma_{k-1}^2}\right]\mathbb{I}_{\sigma_k\le\sigma_{k-1}\le\cdots\le\sigma_1}.\end{aligned}$$

The integration of Equation A1 is achieved by successive integration of the (μ_u,σ_u)s. First, for $u=k$,

$$\begin{aligned}&\int_{\sigma_k\le\sigma_{k-1}} \sigma_k^{-n_k}\exp\left\{-\frac{1}{2\sigma_k^2}\left[n_k(\bar{x}_k-\mu_k)^2+s_k\right]\right\}\\ &\qquad\times\exp\left[-\frac{(\mu_k-\mu_{k-1})^2}{2\sigma_{k-1}^2\zeta^2}\right]\mathrm{d}\mu_k\mathrm{d}\sigma_k\\ &= C\int_{\sigma_k\le\sigma_{k-1}} \sigma_k^{-n_k}\frac{\sigma_k\sigma_{k-1}\exp(-s_k/2\sigma_k^2)}{\left[\sigma_k^2/(n_k+\zeta^2\sigma_{k-1}^2)\right]^{1/2}}\\ &\qquad\times\exp\left\{-\frac{(\bar{x}_k-\mu_{k-1})^2}{2\left[\sigma_k^2/(n_k+\zeta^2\sigma_{k-1}^2)\right]}\right\}\mathrm{d}\sigma_k\end{aligned}$$

$$\leq C\exp\left[-\frac{(\overline{x}_k-\mu_{k-1})^2}{2\zeta^2\sigma_{k-1}^2}\right]\int_{\sigma_k\leq\sigma_{k-1}}\sigma_k^{-n_{k+1}}\exp(-s_k/2\sigma_k)\mathrm{d}\sigma_k$$
$$\leq C\exp\left[-\frac{(\overline{x}_k-\mu_{k-1})^2}{2\zeta^2\sigma_{k-1}^2}\right]\sigma_{k-1}\exp(-s_k/4\sigma_{k-1}^2),$$

if $n_k>1$. A direct computation shows that the majorization extends to $n_k=1$ (with $s_k=0$) and, for $n_k=0$, the upper bound is $C\sigma_{k-1}^2$. If we integrate out (μ_{k-1},σ_{k-1}), the upper bound on the corresponding term depends on n_k and n_{k-1}. For instance, if $(n_{k-1},n_k)=(1,0)$, the partial integral satisfies

$$\int_{\sigma_{k-1}\leq\sigma_{k-2}}\sigma_{k-1}\exp\left[-\frac{(\overline{x}_{k-1}-\mu_{k-1})^2}{2\sigma_{k-1}^2}\right]\times\exp\left[-\frac{(\mu_{k-1}-\mu_{k-2})^2}{2\zeta^2\sigma_{k-2}^2}\right]\mathrm{d}\mu_{k-1}\mathrm{d}\sigma_{k-1}$$
$$\leq C\sigma_{k-2}^2\exp\left[-\frac{(\overline{x}_{k-1}-\mu_{k-2})^2}{2\sigma_{k-2}^2}\right].$$

Similarly, if $(n_{k-1},n_k)=(0,0)$, an upper bound is $C\sigma_{k-2}^2$, while, if $n_{k-1}=0$ and $n_k>0$, it is

$$C\sigma_{k-1}\exp\left[-\frac{(\overline{x}_k-\mu_{k-2})^2}{2\zeta^2\sigma_{k-2}^2}\right]\exp\left(-s_k/8\sigma_{k-2}^2\right).$$

Straightforward assessment of all possible cases reveals that the upper bound on the integral involves an expression in (μ_{k-2},σ_{k-2}) of the form

$$\sigma_{k-2}^{\gamma}\exp\left[-\delta\frac{(\overline{x}_k-\mu_{k-2})^2}{2\sigma_{k-2}^2}\right]\exp\left(-\alpha/2\sigma_{k-2}^2\right),$$

where γ is either zero, one or two, δ is zero or ζ^{-2} and $\alpha\geq 0$. The crux of the proof relies on the fact that this expression can be transferred from component u to component $u-1$ with a decrease in γ if $n_u>1$, or a slight increase otherwise. Therefore, for $1<u<k$, the integral in (μ_u,σ_u) is bounded by $(v\geq u+1)$

$$\int_{\sigma_u\leq\sigma_{u-1}}\sigma_u^{-n_u-2+\gamma}\exp\left\{-\frac{1}{2\sigma_u^2}\left[n_u(\overline{x}_u-\mu_u)^2+s_u+\delta(\overline{x}_v-\mu_u)^2\right]-\frac{(\mu_u-\mu_{u-1})^2}{2\zeta^2\sigma_{u-1}^2}\right\}\exp\left(-\alpha/2\sigma_u^2\right)\mathrm{d}\mu_u\mathrm{d}\sigma_u.$$

Assume in addition that $\alpha>0$ because this is the case with u small enough. If $n_u=0$, the bound is then

$$C\int_{\sigma_u\leq\sigma_{u-1}}\sigma_u^{-1+\gamma}\exp\left[-\delta\frac{(\overline{x}_v-\mu_{u-1})^2}{2\sigma_{u-1}^2}\right]\exp\left(-\alpha/2\sigma_u^2\right)\mathrm{d}\sigma_u$$
$$\leq C\exp\left(-\alpha/4\sigma_{u-1}^2\right)\exp\left[-\delta\frac{(\overline{x}_v-\mu_{u-1})^2}{2\sigma_{u-1}^2}\right]\sigma_{u-1}^{\gamma+\epsilon}\times\int_0^{\infty}\omega^{(\epsilon-2)/2}\exp\left(-\alpha\omega/4\right)\mathrm{d}\omega,$$

where $\epsilon>0$ is arbitrarily small. (In the special case $\alpha=0$, $\gamma\geq 1$ and ϵ can be removed from the above equation.) Therefore,

$$\sigma_{u-1}^{\gamma+\epsilon}\exp\left[-\delta\frac{(\overline{x}_v-\mu_{u-1})^2}{2\sigma_{u-1}^2}\right]\exp\left(-\alpha/4\sigma_{u-1}^2\right)$$

is the bound for the next step. If $n_u=1$, a similar development leads to

$$\sigma_{u-1}^{\gamma+\epsilon}\exp\left[-\frac{(\overline{x}_u-\mu_{u-1})^2}{2\zeta^2\sigma_{u-1}^2}\right]\exp\left(-\alpha/4\sigma_{u-1}^2\right)$$

(and δ is thus eventually positive). When $n_u>1$, the bound is

$$C\int_0^{\sigma_{u-1}}\sigma_u^{-n_u-1+\gamma}\exp\left[-\frac{(\overline{x}_u-\mu_{u-1})^2}{2\zeta^2\sigma_{u-1}^2}\right]\exp\left(-s_u/2\sigma_u^2\right)\times\exp\left(-\alpha/2\sigma_u^2\right)\mathrm{d}\sigma_u$$
$$\leq C\exp\left(-s_u/4\sigma_{u-1}^2\right)\exp\left[-\frac{(\overline{x}_u-\mu_{u-1})^2}{2\zeta^2\sigma_{u-1}^2}\right]\times\int_{\sigma_{u-1}^{-2}}^{\infty}\omega^{((n_u-\gamma)/2)-1}\exp\left(-s_u\omega/4\right)\mathrm{d}\omega,$$

which implies that $\alpha=s_u/2>0$ for $n_u>1$. Therefore, depending on n_u, the bound is either $C\sigma_{u-1}$ when $n_u=\gamma=2$, or C otherwise, leading to a new value of γ smaller than the previous one. Finally, consider the last integral in (μ_1,σ_1),

$$\int\sigma_1^{-n_1-3+\epsilon}\exp\left[-\frac{n_1(\overline{x}_1-\mu_1)^2+s_1}{2\sigma_1^2}-\frac{(\overline{x}_v-\mu_1)^2}{2\zeta^2\sigma_1^2}\right]\times\exp\left(-\alpha/\sigma_1^2\right)\mathrm{d}\mu_1\mathrm{d}\sigma_1,$$

where v is the index of the last non-empty component. When $n_1=0$, this integral is bounded by

$$C\int_0^{\infty}\sigma_1^{-2+\epsilon}\exp\left(-\alpha/\sigma_1^2\right)\mathrm{d}\sigma_1 = C\int_0^{\infty}\omega^{(-1-\epsilon)/2}\exp\left(-\alpha\omega\right)\mathrm{d}\omega,$$

which is indeed finite for $\epsilon>1$. The other cases lead to the same conclusion.

Poisson case: The finiteness of the marginal distribution of $(x_1,\ldots,x_n)$ follows from the development of the k^n terms of the likelihood. (The prior on $\mathbb{P}$ being proper, there is no problem with integrating over the p_{uv}s.) When integrating

$$\frac{\lambda_k}{\lambda_1^{1/2}}\prod_{u=1}^{k}\lambda_u^{s_u-1}\exp(-n_u\lambda_u)\;\mathbb{I}_{\lambda_k\leq\cdots\leq\lambda_1},$$

we start with λ_1, obtaining the upper bound

$$\int_{\lambda_2}^{\infty} \lambda_1^{s_1-3/2} \exp\Big(-n_1\lambda_1\Big)\mathrm{d}\lambda_1 \leq \begin{cases} C\lambda_2^{-1/2} & \text{if } n_1 = 0, \\ C\lambda_2^{-(s_1-(1/2))^+-\epsilon} \exp\Big(-n_1\lambda_2/2\Big) & \text{if } n_1 > 0, \end{cases}$$

where $(s_1 - (1/2))^+ = \max(s_1 - (1/2), 0)$ and $\epsilon > 0$ is arbitrary. By recursion, the bound on the integral in λ_u is of the form $C\lambda_u^{-\delta}\exp(-\alpha\lambda_u)$, where $\delta \geq 0$ decreases for $s_u > 0$ or slightly increases otherwise. Indeed,

$$\int_{\lambda_{u+1}}^{\infty} \lambda_u^{-\delta+s_u-1} \exp[-(\alpha+n_u)\lambda_u]\mathrm{d}\lambda_u \leq \exp(-(\alpha+n_u)\lambda_{u+1}/2)\lambda_{u+1}^{-(\delta-s_u)^+-\epsilon}$$

gives such a bound with δ equal to zero if $s_u > \delta$ and to $\delta - s_u + \epsilon$ otherwise. Therefore, for u large enough, δ is arbitrarily close to zero if some s_vs are different from zero $(v < u)$. In this case, the final integral involves

$$\int_0^{\infty} \lambda_k^{s_k-\delta} \exp[-(\alpha+n_k)\lambda_k]\mathrm{d}\lambda_k,$$

which is indeed finite for $\delta < 1$. Note that, if $s_1 = \cdots = s_k = 0$, the overall integral is

$$\begin{aligned} &\int_{\lambda_k\leq\cdots\leq\lambda_1} \lambda_1^{-3/2}\lambda_2^{-1}\cdots\lambda_{k-1}^{-1} \exp\Big(-\sum_u n_u\lambda_u\Big)\mathrm{d}\lambda_1\cdots\mathrm{d}\lambda_k \\ &\quad\leq \int_{\lambda_k\leq\cdots\leq\lambda_1} \lambda_1^{-3/2}\lambda_2^{-1}\cdots\lambda_{k-1}^{-1} \\ &\qquad\times \exp\Big(-\sum_u n_u\lambda_k\Big)\mathrm{d}\lambda_1\cdots\mathrm{d}\lambda_k \\ &\quad= C\int_0^{\infty} \lambda_k^{1/2}\exp\Big(-\sum_u n_u\lambda_k\Big)\mathrm{d}\lambda_k < \infty. \end{aligned}$$

Appendix 2: Proof that the penalized likelihood is truly bounded

Normal case: The modified prior leads to the penalized likelihood

$$\sum_{(n_u,\bar{x}_u,s_u)} \prod_{u=2}^{k} \sigma_u^{-n_u} \exp\left[-\frac{n_u(\bar{x}_u-\mu_u)^2+s_u}{2\sigma_u^2}\right] \exp\left[-\frac{(\mu_u-\mu_{u-1})^2}{2\zeta^2\sigma_{u-1}^2}\right] \times \sigma_1^{-n_1-2k+1} \exp\left[-\frac{n_1(\bar{x}_1-\mu_1)^2+s_1}{2\sigma_1^2}\right] \sigma_k \mathbb{I}_{\sigma_k\leq\cdots\leq\sigma_1},$$

where the sum is taken over all the different partitions. For $1 < u < k$, the term in (μ_u, σ_u) is

$$\sigma_u^{-n_u} \exp\left[-\frac{n_u(\bar{x}_u-\mu_u)^2+s_u}{2\sigma_u^2}\right] \exp\left[-\frac{(\mu_u-\mu_{u-1})^2}{2\zeta^2\sigma_{u-1}^2}\right] \times \mathbb{I}_{\sigma_{u+1}\leq\sigma_u\leq\sigma_{u-1}} \leq \sigma_u^{-n_u}\exp\big[-s_u/2\sigma_u^2\big] \qquad \text{(A2)}$$

when $n_u \geq 2$, which is clearly finite. If $n_u = 1$, Equation A2 is bounded by σ_u^{-1} when $\mu_u = \mu_{u+1} = \bar{x}_u$. If $n_{u+1} = 0$, the bound does not change. Otherwise, Equation A2 is then bounded by

$$\sigma_u^{-1}\sigma_{u+1}^{-n_{u+1}} \exp\left[-\frac{n_{u+1}(\bar{x}_{u+1}-\bar{x}_u)^2+s_{u+1}}{2\sigma_{u+1}^2}\right] \leq \sigma_{u+1}^{-n_{u+1}-1} \exp\left[-\frac{(\bar{x}_{u+1}-\bar{x}_u)^2}{2\sigma_{u+1}^2}\right],$$

which is finite because $\bar{x}_{u+1} \neq \bar{x}_u$. In the first case, we iterate upwards on indices above u until we reach either a non-empty component, or attain $u = k$ and $n_k = 0$, the bound on Equation A2 being then

$$\sigma_u^{-1}\sigma_k \mathbb{I}_{\sigma_k\leq\sigma_u} \leq 1.$$

When $n_k = 0$, we can invert the reasoning to bound the likelihood when σ_k gets to infinity by descending the component indices until $n_u > 0$. The case $u = 1$ can be processed similarly because

$$\sigma_1^{-n_1-2k+1} \exp\left[-\frac{n_1(\bar{x}_1-\mu_1)^2+s_1}{2\sigma_1^2}\right] \exp\left[-\frac{(\mu_2-\mu_1)^2}{2\zeta^2\sigma_1}\right]$$

is bounded for $n_1 > 1$. When $n_1 = 1$, the control can be found in the first non-empty component, v, because $\bar{x}_1 = \mu_1 = \cdots = \mu_v$ provides the upper bound

$$\sigma_1^{-2k}\sigma_v^{-n_v} \exp\left[-\frac{(\bar{x}_v-\bar{x}_1)^2}{2\sigma_v^2}\right] \leq \sigma_v^{-n_v-2k} \exp\left[-\frac{(\bar{x}_v-\bar{x}_1)^2}{2\sigma_v^2}\right],$$

which is bounded. The same control can be obtained for $n_1 = 0$. Note that the choice

$$\pi(\sigma_1,\ldots,\sigma_k) = \sigma_k/\sigma_1^{2k-1}$$

is necessary in the sense that negative powers for the σ_us or a smaller power for σ_k would lead to an unbounded penalized likelihood (consider the special case $n_2 = \cdots = n_k = 0$). Appendix 3 details why the power of σ_1 has to be $2k - 1$.

Poisson case: The penalized likelihood is the sum of terms like

$$\prod_{u=1}^{k} \lambda_u^{s_u-1}\exp(-n_u\lambda_u)\frac{\lambda_k}{\lambda_1^{3/2}}\mathbb{I}_{\lambda_k\leq\cdots\leq\lambda_1};$$

where $s_u \geq 0$, $n_u \geq 0$ $(1 \leq u \leq k)$. If we use the scale parameterization in τ_i, the penalized likelihood is

$$\begin{aligned} &\tau_1^{\sum_{u=1}^{k} s_u-1/2}\cdots\tau_k^{s_k}\exp(-n_1\tau_1)\cdots\exp(-n_k\tau_1\cdots\tau_k) \\ &\quad\leq \tau_1^{\sum_{u=1}^{k} s_u-1/2}\exp\Big(-\sum_{u=1}^{k} n_u\tau_1\cdots\tau_k\Big); \end{aligned}$$

because the τ_ks are smaller than one. Therefore this penalised likelihood is bounded near $\tau_1 = 0$ when $\sum_{u=1}^{k} s_u = \sum_{i=1}^{n} x_i > 0$. If $\sum_{u=1}^{k} s_u = 0$, we need to change the prior on τ_1 to $\pi(\tau_1) = 1$.

Appendix 3: Proof that the modified posterior is still proper

Normal case: The proof is similar to that of Appendix 1. The powers of the σ_us are modified from $\sigma_u^{-n_u-2}$ to $\sigma_u^{-n_u}$ for $1 < u < k$. The main difference from Appendix 1 is that components with no allocated observation ($n_u = 0$) lead to an increase of the power of σ_u. To be more precise, the bound at step u is related to

$$\sigma_u^{\delta} \exp\left[-\frac{(\bar{x}_v - \mu_u)^2}{2\zeta^2\sigma_u^2}\right] \exp(-\alpha/\sigma_u^2),$$

with $\alpha > 0$ if there exists $v > u$ such that $n_v > 1$. By integrating out (μ_u, σ_u), we get

$$\begin{aligned} C \int_{\sigma_u \le \sigma_{u-1}} \sigma_u^{\delta - n_u} \exp\left[-\frac{(\bar{x}_v - \mu_u)^2}{2\zeta^2\sigma_u^2} - \frac{(\mu_u - \mu_{u-1})^2}{2\zeta^2\sigma_{u-1}^2}\right] \\ \times \exp(-\alpha/\sigma_u^2) \mathrm{d}\mu_u \mathrm{d}\sigma_u \\ \le C \exp\left[-\frac{(\bar{x}_v - \mu_{u-1})^2}{2\zeta^2\sigma_{u-1}^2}\right] \exp(-\alpha/2\sigma_{u-1}^2) \sigma_{u-1}^{(\delta+2-n_u)^+ + \epsilon}, \end{aligned}$$

where $(\delta + 2 - n_u)^+ = \max(\delta + 2 - n_u, 0)$ and $\epsilon > 0$ is arbitrarily small. Therefore, $n_u = 0$ produces an increase in the power of σ_u from δ to $\delta + 2 + \epsilon$. The worst possible case among the different allocations is therefore $n_k = n$ and $n_1 = \cdots = n_{k-1} = 0$, since the successive integrations from $u = k$ to $u = 2$ lead to the bound

$$\begin{aligned} C \int \sigma_1^{2(k-2)+\epsilon-\eta} \exp\left[-\frac{(\bar{x}_k - \mu_1)^2}{2\zeta^2\sigma_1^2}\right] \exp\left(-\alpha/\sigma_1^2\right) \mathrm{d}\mu_1 \mathrm{d}\sigma_1 \\ = C \int_0^{\infty} \omega^{[(\eta+2-\epsilon-2k)/2]-1} \exp(-\alpha\omega) \mathrm{d}\omega, \end{aligned}$$

which is finite for $\eta > 2k - 2$. This therefore imposes a lower limit of $\eta = 3$ for $k = 2$. The corresponding conditional distributions of the σ_us in Equation 8 are modified into

$$\begin{aligned} \sigma_1^{-2} &\sim \mathcal{G}a\left[\frac{n_1 + 2k - 4}{2}, \frac{n_1(\bar{x}_1 - \mu_1)^2 + s_1^2 + (\mu_2 - \mu_1)^2\zeta^{-2}}{2}\right] \\ &\quad \times \mathbb{I}_{(\sigma_2, \infty)}(\sigma_1), \\ \sigma_u^{-2} &\sim \mathcal{G}a\left[\frac{n_u - 3}{2}, \frac{n_u(\bar{x}_u - \mu_u)^2 + s_u^2 + (\mu_{u+1} - \mu_u)^2\zeta^{-2}}{2}\right] \\ &\quad \times \mathbb{I}_{(\sigma_{u+1}, \sigma_{u-1})}(\sigma_u), \\ \sigma_k^{-2} &\sim \mathcal{G}a\left[\frac{n_k - 3}{2}, \frac{n_k(\bar{x}_k - \mu_k)^2 + s_k^2}{2}\right] \mathbb{I}_{(0, \sigma_{k-1})}(\sigma_k), \end{aligned}$$

while the conditional distributions of the μ_us remain the same.

Poisson case: Because the prior has only to be modified when $\sum_u s_u = 0$, we consider this special case. Recursive reasoning similar to the proof in Appendix 1 leads to the inequality

$$\begin{aligned} \int \prod_{u=1}^{k-1} \lambda_u^{-1} \exp(-n_u\lambda_u) \exp(-n_k\lambda_k) \mathbb{I}_{\lambda_k \le \cdots \le \lambda_1} \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_k \\ \le C \int_{\lambda_k < \lambda_{k-1}} \lambda_{k-1}^{-1+\epsilon} \exp(-\alpha\lambda_{k-1} - n_k\lambda_k) \mathrm{d}\lambda_{k-1} \mathrm{d}\lambda_k \\ \le C \int_0^{\infty} \lambda_k^{2\epsilon} \exp(-\alpha\lambda_k/2) \mathrm{d}\lambda_k, \end{aligned}$$

with $\alpha > 0, \epsilon > 0$. The integral, I, is indeed finite.