Stat Comput (2011) 21: 93–105
DOI 10.1007/s11222-009-9150-y

Slice sampling mixture models

Maria Kalli · Jim E. Griffin · Stephen G. Walker

Received: 9 October 2008 / Accepted: 3 September 2009 / Published online: 19 September 2009

Abstract We propose a more efficient version of the slice sampler for Dirichlet process mixture models described by Walker (Commun. Stat., Simul. Comput. 36:45–54, 2007). This new sampler allows for the fitting of infinite mixture models with a wide-range of prior specifications. To illustrate this flexibility we consider priors defined through infinite sequences of independent positive random variables. Two applications are considered: density estimation using mixture models and hazard function estimation. In each case we show how the slice efficient sampler can be applied to make inference in the models. In the mixture case, two submodels are studied in detail. The first one assumes that the positive random variables are Gamma distributed and the second assumes that they are inverse-Gaussian distributed. Both priors have two hyperparameters and we consider their effect on the prior distribution of the number of occupied clusters in a sample. Extensive computational comparisons with alternative "conditional" simulation techniques for mixture models using the standard Dirichlet process prior and our new priors are made. The properties of the new priors are illustrated on a density estimation problem.

Keywords Dirichlet process · Markov chain Monte Carlo · Mixture model · Normalized weights · Slice sampler · Hazard function

M. Kalli
Centre for Health Services Studies, University of Kent, Canterbury, UK

J.E. Griffin (✉) · S.G. Walker
Institute of Mathematics, Statistics & Actuarial Science, University of Kent, Canterbury, UK
e-mail: jeg28@kent.ac.uk

1 Introduction

The well-known and widely used mixture of Dirichlet process (MDP) model was first introduced by Lo (1984). The MDP model, with Gaussian kernel, is given by

$$f_P(y) = \int \mathrm{N}(y; \mu, \sigma^2)\, \mathrm{d}P(\phi)$$

with $P \sim D(M, P_0)$. We write $P \sim D(M, P_0)$ to denote that P follows a Dirichlet process (Ferguson 1973) with parameters $M > 0$, the scale parameter, and P_0, a distribution on $\mathbb{R} \times \mathbb{R}_+$ where $\phi = (\mu, \sigma^2)$ with μ to represent the mean and σ^2 the variance of the normal component. Since the advent of Markov chain Monte Carlo methods within the mainstream statistics literature (Smith and Roberts 1993), and the specific application to the MDP model (Escobar 1988, 1994; Escobar and West 1995), the model has become one of the most popular in Bayesian nonparametrics since it is possible to integrate P from the posterior defined by this model.

Variations of the original algorithm of Escobar (1988) have been numerous; for example, MacEachern (1994), MacEachern and Müller (1998), Neal (2000). All of these algorithms rely on integrating out the random distribution function from the model, removing the infinite dimensional problem. These are usually referred to as "marginal" methods. Recent ideas have left the infinite dimensional distribution in the model and found ways of sampling a sufficient but finite number of variables at each iteration of a Markov chain with the correct stationary distribution. See Ishwaran and James (2000), who propose an approximate method, Papaspiliopoulos and Roberts (2008) and Walker (2007); the latter paper using slice sampling ideas. These define so-called "conditional" methods.

There has recently been interest in defining nonparametric priors for P that move beyond the Dirichlet process (see

e.g. Lijoi et al. 2007) in infinite mixture models. These alternative priors allow for more control over the prior cluster structure than is possible with the Dirichlet process. The availability of general computational methods allows the development of these priors without the need to develop specific computational methods on a case-by-case basis.

The purpose of this paper is: (1) to develop more efficient versions of the slice sampling algorithm for MDP models proposed by Walker (2007) and to extend it to more general nonparametric priors such as general stick-breaking processes and normalised weights priors, (2) to develop a new class of nonparametric priors for infinite mixture models by normalizing an infinite sequence of positive random variables, which will be termed a normalized weights prior and (3) to illustrate how slice sampling ideas can be applied to more general applications such as survival analysis. The lay-out of the paper is as follows. In Sect. 2 we describe the slice-efficient sampler for the MDP model. Section 3 describes the normalized weights prior and discusses constructing a slice sampler for infinite mixture models with this prior. Section 4 discusses an application of the normalized weights prior to modeling the hazard in survival analysis and Sect. 5 contains numerical illustrations. Finally, Sect. 6 contains conclusions and a discussion.

2 Slice-efficient samplers for the MDP

It is well-known that $P \sim D(M, P_0)$ has a stick-breaking representation (Sethuraman 1994) given by

$$P = \sum_{j=1}^{\infty} w_j \delta_{\phi_j},$$

where $\phi_1, \phi_2, \phi_3, \ldots$ are independent and identically distributed from P_0 and

$$w_1 = v_1, \qquad w_j = v_j \prod_{l<j} (1 - v_l)$$

with the (v_j) being independent and identically distributed from $\mathrm{Be}(1, M)$ where $\mathrm{Be}(a, b)$ represents the Beta distribution with parameters a and b. It is possible to integrate P from the posterior defined by the MDP model. However, the stick-breaking representation is essential to estimation via the non-marginal methods of Papaspiliopoulos and Roberts (2008) and Walker (2007). The idea is that we can write

$$f_{v,\mu,\sigma^2}(y) = \sum_{j=1}^{\infty} w_j \mathrm{N}(y; \mu_j, \sigma_j^2)$$

and the key is to find exactly which (finite number of) variables need to be sampled to produce a valid Markov chain with correct stationary distribution.

The details of the slice sampler algorithm are given in Walker (2007), but we briefly describe the basis for the algorithm here. Our starting point is the joint density

$$f_{v,\mu,\sigma^2}(y, u) = \sum_{j=1}^{\infty} \mathbf{1}(u < w_j) \mathrm{N}(y; \mu_j, \sigma_j^2).$$

Given u, the number of components is finite, the indices being $A_u = \{j : w_j > u\}$. One has

$$f_{v,\mu,\sigma^2}(y|u) = N_u^{-1} \sum_{j \in A_u} \mathrm{N}(y; \mu_j, \sigma_j^2),$$

and the size of A_u is $\sum_{j=1}^{\infty} \mathbf{1}(w_j > u)$ while $N_u = \sum_{j \in A_u} w_j$.

One can then introduce a further latent variable which indicates which of these finite number of components provides the observation to give the joint density

$$f_{v,\mu,\sigma^2}(y, u, d) = \mathbf{1}(u < w_d) \mathrm{N}(y; \mu_d, \sigma_d^2).$$

Hence, a complete likelihood function for (v, μ, σ^2) is available as a simple product of terms and crucially d is finite. Without u, d can take an infinite number of values which would make the implementation of a Markov chain Monte Carlo algorithm problematic.

The joint posterior distribution is proportional to

$$\prod_{i=1}^{n} \mathbf{1}(u_i < w_{d_i}) \mathrm{N}(y_i; \mu_{d_i}, \sigma_{d_i}^2)$$

and this allows a simple Gibbs sampling scheme for the posterior to be derived, which is given in Walker (2007). There are several problems with this algorithm. Firstly, it will often mix slowly due to the correlation between u and w. Secondly, updating u can lead to changes in the set $\bigcup_{i=1}^{n} A(u_i)$ which can lead to the simulation of more w's. As we shall see in the normalized weights section, simulating these values can involve some additional work. The first problem can be addressed by a suitable blocking structure for the Gibbs sampler, which was independently noted by Papaspiliopoulos (2008). The second problem can be addressed by a more general approach to slice sampling.

A general class of slice sampler can be defined by writing

$$f_{v,\mu,\sigma^2}(y, u, d) = \xi_d^{-1} \mathbf{1}(u < \xi_d) w_d \mathrm{N}(y; \mu_d, \sigma_d^2)$$

where $\xi_1, \xi_2, \xi_3, \ldots$ is any positive sequence. Previous implementations of slice sampler (see e.g. Walker 2007; Papaspiliopoulos 2008; Dunson 2008 and Yau et al. 2008) arise when $\xi_j = w_j$. Typically, the sequence will be a deterministic, decreasing sequence but a random sequence could also be considered. The choice of $\xi_1, \xi_2, \xi_3, \ldots$ is a delicate issue and any choice has to balance efficiency and computational

time. We find that mixing depends on the rate at which the ratio $r_i = E[w_i]/\xi_i$ increases with i. Faster rates of increase are associated with better mixing but longer running times, since the average size of A_u increases. We suggest increasing the rate of increase of r_i until the gains in mixing are counter-balanced by the longer running time. In our examples, we find that $r_i \propto (1.5)^i$ strikes a good balance.

The variables that need to be sampled at each sweep of a Gibbs sampler are

$$\{(\mu_j, \sigma_j^2, v_j), j = 1, 2, \ldots; (d_i, u_i),\ i = 1, \ldots, n\}$$

and the joint posterior distribution is proportional to

$$\prod_{i=1}^{n} \mathbf{1}(u_i < \xi_{d_i}) w_{d_i}/\xi_{d_i} \mathrm{N}(y_i; \mu_{d_i}, \sigma_{d_i}^2).$$

If ξ and v are conditionally independent then our Gibbs sampler is

1. $\pi(\mu_j, \sigma_j^2|\cdots) \propto p_0(\mu_j, \sigma_j^2) \prod_{d_i=j} \mathrm{N}(y_i; \mu_j, \sigma_j^2)$.
2. $\pi(v_j) \propto \mathrm{Be}(v_j; a_j, b_j)$, where

$$a_j = 1 + \sum_{i=1}^{n} \mathbf{1}(d_i = j)$$

and

$$b_j = M + \sum_{i=1}^{n} \mathbf{1}(d_i > j).$$

3. $\pi(u_i|\cdots) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$.
4. $\mathrm{P}(d_i = k|\cdots) \propto \mathbf{1}(k : \xi_k > u_i) w_k/\xi_k \mathrm{N}(y_i; \mu_k, \sigma_k^2)$.

This naturally defines a blocking scheme for u and v which are conditionally independent. If $\xi_j = w_j$ then we can also define a blocking scheme by jointly updating u and v which leads to the algorithm above.

Obviously, we can not sample all of the (μ_j, σ_j^2, v_j). But it is not required to in order to proceed with the chain. We only need to sample up to the integer N for which we have found all the appropriate k in order to do step 4 exactly. In fact it is easy to find the set of (k) required since it will be of the kind $\{1, \ldots, N\}$ where $N = \max_i\{N_i\}$ and N_i is the largest integer l for which $\xi_l > u_i$. This can often be found analytically for suitable choice of ξ_l. If we take $\xi_j = w_j$, as in Papaspiliopoulos (2008), it is sufficient to find an N_i such that $\sum_{k=1}^{N_i} w_k > 1 - u_i$ then it is not possible for any w_k, for $k > N_i$, to be greater than u_i. This search is more cumbersome since it can only be checked by simulation.

There are some important points to make here. First, it is a trivial extension to consider more general stick-breaking processes for which $v_j \sim \mathrm{Be}(\alpha_j, \beta_j)$ independently. Then, in this case, we would have

$$a_j = \alpha_j + \sum_{i=1}^{n} \mathbf{1}(d_i = j)$$

and

$$b_j = \beta_j + \sum_{i=1}^{n} \mathbf{1}(d_i > j).$$

This easy extension to more general priors is not a feature of alternative, marginal sampling algorithms. Secondly, the algorithm is remarkably simple to implement; all full conditionals are standard. Thirdly, further levels of a hierarchical model can be updated using the model marginalising over $u_1, u_2, \ldots, u_n$ and will have the same form as for the block sampler of Ishwaran and James (2001), with the exception of a random truncation point.

Later, for the illustrations and comparison with the other 'conditional' algorithms, namely the blocked Gibbs sampler (Ishwaran and James 2001) and the retrospective sampler (Papaspiliopoulos and Roberts 2008), we will consider three types of slice sampler. The original "slice" algorithm appearing in Walker (2007), a "dependent slice-efficient" with $\xi_j = w_j$ and an "independent slice-efficient" where $\xi_1, \xi_2, \xi_3, \ldots$ is a deterministic sequence.

The retrospective sampler (Papaspiliopoulos and Roberts 2008) is an alternative, conditional method which defines a Markov chain with the correct posterior for the infinite dimensional model. The difference between this approach and our slice sampling approach rests on the way that the allocation variables d_i are sampled. Our approach uses a slice variable to make the choice of d_i finite at each iteration of a Gibbs, whereas the retrospective sampler proposes a new value of d_i in a Metropolis-Hastings update where the proposal is cleverly chosen to define an efficient algorithm. From a computational point of view, the retrospective sampler involves the potential simulation of extra variables n time per iteration (once for every update of each d_i) whereas the slice sampler only generates extra variables once per iteration.

3 Mixtures based on normalized weights

3.1 Definition and properties

The slice sampling idea can be extended to mixture models with weights obtained via normalization. The Dirichlet process has been the dominant prior in nonparametrics but the definition of alternative nonparametric priors has been a recent area of interest. For example, Lijoi et al. (2007) define nonparametric priors through the normalization of the generalized Gamma process to define an NGG prior. The generalized Gamma process is a Lévy process and so has independent and identically distributed jumps. We discuss an alternative construction using the normalization of an infinite sequence of positive random variables. These play the

same role as the jumps in the construction of the NGG but are no longer identically distributed. We consider

$$f(y) = \sum_{j=1}^{\infty} w_j K(y; \phi_j)$$

where $w_j = \lambda_j/\Lambda$, $\Lambda = \sum_{j=1}^{\infty} \lambda_j$ and $\phi_1, \phi_2, \phi_3, \ldots$ are independent and identically distributed with distribution P_0. We will also use $\Lambda_m = \sum_{j=m+1}^{\infty} \lambda_j$. Here the $\lambda_1, \lambda_2, \lambda_3, \ldots$ are positive and will be assigned independent prior distributions, say $\lambda_j \sim \pi_j(\lambda_j)$. These must be constructed so as to ensure that $\Lambda < +\infty$ a.s. which is easy to achieve by ensuring that $\sum_j \mathrm{E}(\lambda_j) < +\infty$. Since $\mathrm{E}(\lambda_j)$ is our choice, we can obviously pick these so the sum is finite. If this sum is finite then it is easy to show that $\sum_{j=1}^{\infty} \lambda_j < +\infty$ a.s. We suggest defining specific priors by defining $\mathrm{E}(\lambda_j) = Mq_j$ where $M > 0$ and $q_1, q_2, q_3, \ldots$ are probabilities from a known probability distribution. For example, if they follow a geometric distribution, then

$$q_j = (1-\theta)\theta^{j-1}.$$

The parameter θ controls the rate at which $\mathrm{E}(\lambda_j)$ tends to zero. We have defined a nonparametric prior with two parameters θ and M. As we will see in the following examples, the choice of the distributions (π_j) controls the properties of the process. So we would not necessarily wish for the (q_j) to decay too slowly ensuring we do not put too much mass on large integers. Hyperpriors for (θ, M) can be assigned though we do not consider that here.

Example 1 (Gamma distribution) Here we take the (λ_j) to be independent gamma distributions, say $\lambda_j \sim \mathrm{Ga}(\gamma_j, 1)$. To ensure that $\Lambda < +\infty$ a.s. we take $\sum_{j=1}^{\infty} \gamma_j < +\infty$. Clearly, w_j has expectation q_j and variance $q_j(1-q_j)/(M+1)$, since marginally $w_j \sim \mathrm{Be}(Mq_j, M(1-q_j))$. We can interpret M as a mass parameter. We will refer to this model as an infinite Dirichlet prior since if we have a finite number of unnormalized weights $(\lambda_1, \lambda_2, \ldots, \lambda_N)$ then $(w_1, w_2, \ldots, w_N)$ would be Dirichlet distributed. In infinite mixture models, the prior distribution on the number of clusters from n observations is important. Figure 1 shows this distribution for $n = 30$. Larger values of θ for fixed M place more mass on larger numbers of clusters (as we would expect since the weights decay increasingly slowly with larger θ). The mass parameter M also plays an important role. Larger values of M lead to more dispersed distributions with a larger median value.

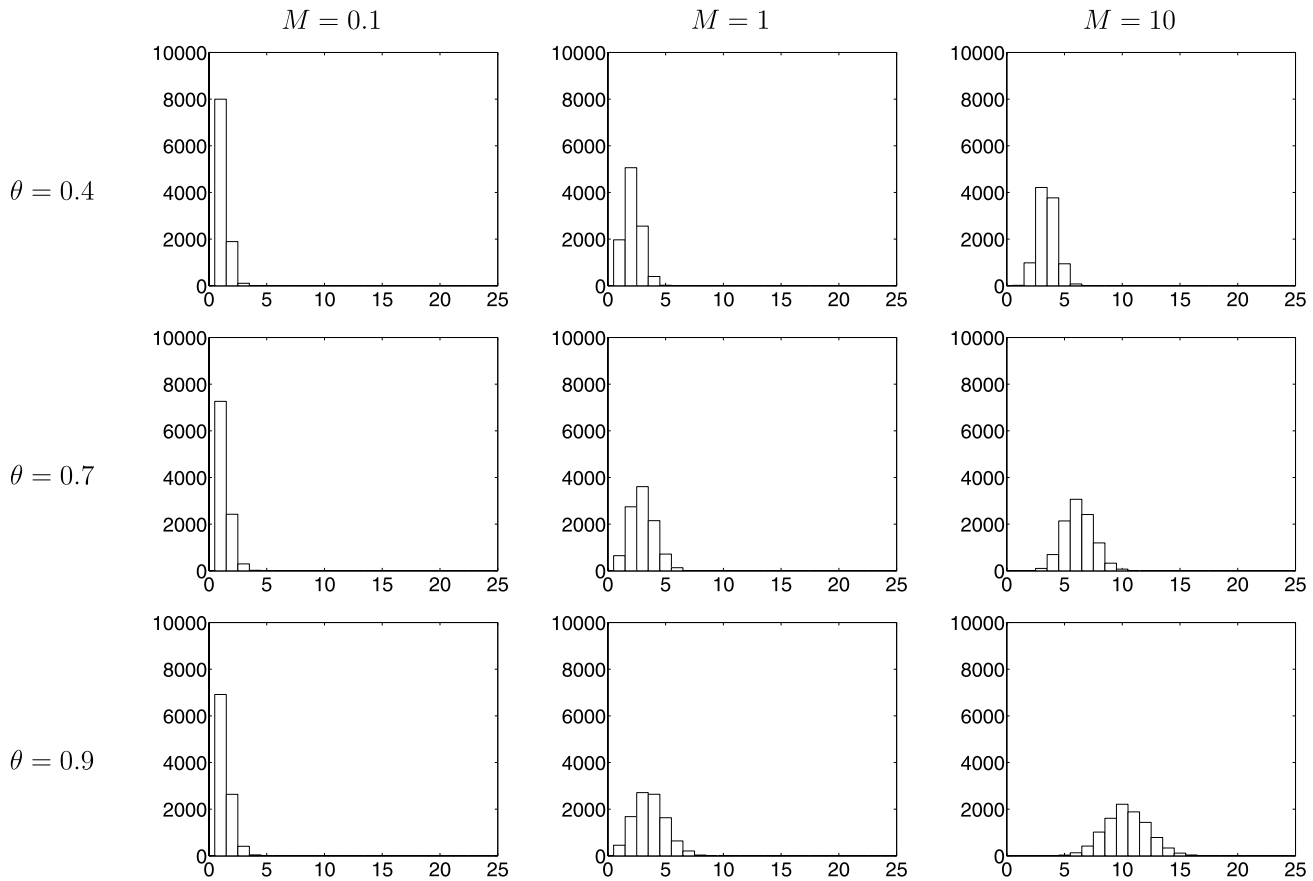

Fig. 1 Prior distribution of the number of clusters from 30 observations with the infinite Dirichlet prior

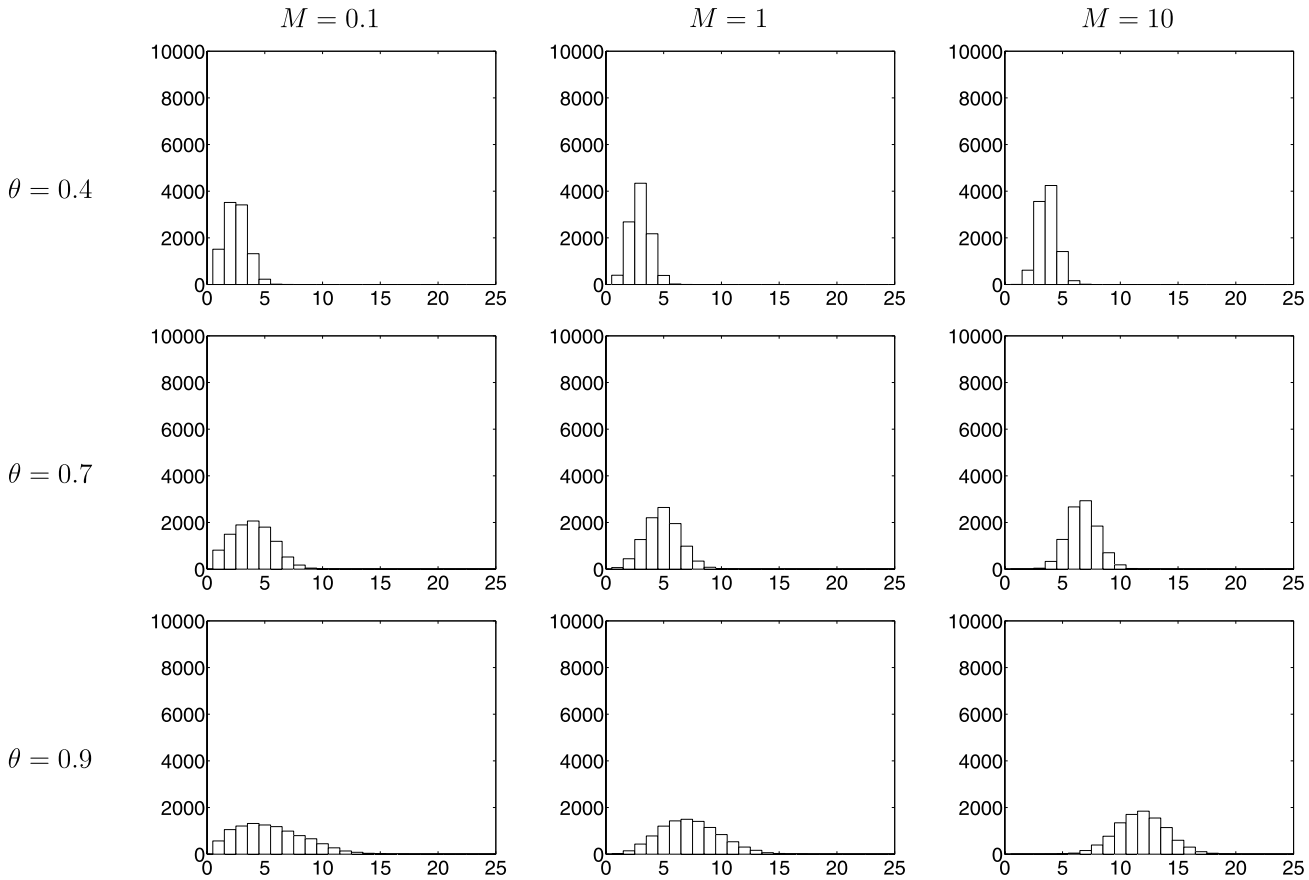

Fig. 2 Prior distribution of the number of clusters from 30 observations for the infinite normalized inverse-Gaussian prior

Stick-breaking priors were introduced to Bayesian nonparametrics by Ishwaran and James (2001). They are defined by two infinite vectors of parameters. Clearly, there is a need to develop priors within this class that have a few hyperparameters to allow easy prior specification. The Dirichlet process and Poisson-Dirichlet process are two such priors and the infinite Dirichlet prior represents another. The stick-breaking representation of the infinite Dirichlet prior takes $\alpha_j = Mq_j$ and $\beta_j = M(1 - \sum_{i=1}^{j} q_i)$.

Example 2 (Inverse-Gaussian distribution) The inverse-Gaussian distribution, IG(α, γ), has a density function given by

$$\pi(\lambda) = \frac{\alpha}{\sqrt{2\pi}}\lambda^{-3/2}\exp\left\{-\frac{1}{2}\left(\frac{\alpha^2}{\lambda} + \eta^2\lambda\right) + \eta\alpha\right\},$$

where α and η can be interpreted as a shape and a scale parameter, respectively. We take λ_j to follow independent IG$(\gamma_j, 1)$ distributions. Then $\Lambda_m = \sum_{j=m+1}^{\infty}\lambda_j$ is distributed as IG$(\sum_{j=m+1}^{\infty}\gamma_j, 1)$ and the normalization is well-defined if $\sum_{j=1}^{\infty}\gamma_j < +\infty$ which implies that Λ is almost surely finite. The finite dimensional normalized distribution $(\lambda_1/\Lambda, \lambda_2/\Lambda, \ldots, \lambda_m/\Lambda)$ has been studied by Lijoi et al. (2005) as the normalized inverse-Gaussian distribution. We again define $\gamma_j = Mq_j$ and it follows directly from their results that w_i has expectation q_i and variance $q_i(1-q_i)M^2\exp\{M\}\Gamma(-2, M)$. This prior will be referred to as the infinite normalized inverse-Gaussian prior. Figure 2 shows the prior distribution of the number of clusters in 30 observations. The effects of M and θ follow the same pattern as the infinite Dirichlet case discussed above. However, the effect of M is less marked for small M. In the infinite Dirichlet case for $M = 0.1$, the distributions are almost indistinguishable for different values of θ but in this case it is clear that the location of the distribution is increasing with θ.

3.2 Slice sampler

The model can be fitted using an extension of the slice sampler developed in Sect. 2. We will assume that the distribution of Λ_m has a known form for all m, which we will denote by $\pi_m^\star(\Lambda_m)$. We introduce the additional latent variable v, we consider the joint density

$$f(y, v, u, d) = \exp(-v\Lambda)\mathbf{1}(u < \xi_d)\lambda_d/\xi_d K(y; \phi_d).$$

Clearly the marginal density is

$$f(y,d) = \frac{\lambda_d}{\Lambda} K(y;\phi_d).$$

The likelihood function based on a sample of size n is given by

$$\prod_{i=1}^{n} \exp(-v_i\Lambda)\mathbf{1}(u_i < \xi_{d_i})\lambda_{d_i}/\xi_{d_i} K(y_i;\phi_{d_i}).$$

It is simpler to deal with the posterior replacing $v_1, v_2, \ldots, v_n$ by $v = \sum_{i=1}^{n} v_i$ which has the form

$$v^{n-1}\exp(-v\Lambda)\prod_{i=1}^{n}\mathbf{1}(u_i < \xi_{d_i})\lambda_{d_i}/\xi_{d_i} K(y_i;\phi_{d_i}).$$

We describe two algorithms: (1) a dependent slice-efficient sampler where $\xi_j = \lambda_j$ and (2) the independent slice-efficient sampler where ξ_j and λ_j are independent.

In the dependent slice-efficient sampler the full conditional distributions of u_i, v and ϕ_j are trivial. The distribution of d_i is also trivial. Complications arise when we try to find the number of λ_j's (and also ϕ_j's) to be sampled in order to implement the sampling of d_i. The non-trivial aspect of the algorithm is the sampling of the sufficient number of $\{\lambda_j\}$. We simulate $\lambda_1, \ldots, \lambda_m, \Lambda_m$ (where m is the number of atoms given in the previous iteration) in a block from their full conditional distribution which is proportional to

$$\exp\{-v\Lambda_m\}\pi^{\star}_m(\Lambda_m)\prod_{j=1}^{m}\exp\{-v\lambda_j\}\lambda_j^{n_j}\pi_j(\lambda_j),$$

where $n_j = \sum_{i=1}^{n}\mathbf{1}(d_i = j)$. We need to find the smallest value of m' for which $\Lambda_{m'} < \min_i\{u_i\}$ so that we can evaluate the full conditional distribution of d_i. This value can be found by sequentially simulating $[\lambda_j, \Lambda_j|\Lambda_{j-1}]$ for $j = m+1, \ldots, m'$. The conditional distribution of $[\lambda_j = x, \Lambda_j = \Lambda_{j-1} - x|\Lambda_{j-1}]$ is given by

$$f(x) \propto \pi_j(x)\pi^{\star}_j(\Lambda_{j-1} - x),$$
$$0 < x < \Lambda_{j-1}.$$

In some cases simulation from the distribution will be straightforward. If not, generic univariate simulation methods such as Adaptive Rejection Metropolis Sampling (Gilks et al. 1995) can be employed. The algorithm is

1. $\pi(\mu_j, \sigma_j^2|\cdots) \propto p_0(\mu_j,\sigma_j^2)\prod_{d_i=j}\mathrm{N}(y_i;\mu_j,\sigma_j^2)$.
2. $\pi(\lambda_j) \propto \lambda_j^{n_j}\exp\{-v\lambda_j\}$ and $\pi(\Lambda_m) \propto \exp\{-v\Lambda\} \times \pi^{\star}(\Lambda_m)$.
3. $\pi(u_i|\cdots) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$.
4. $\mathrm{P}(d_i = k|\cdots) \propto \mathbf{1}(k : \xi_k > u_i)w_k/\xi_k\mathrm{N}(y_i;\mu_k,\sigma_k^2)$.
5. v is Gamma distributed with shape parameter $n-1$ and mean $(n-1)/(\Lambda_m + \sum_{i=1}^{m}\lambda_i)$.

In these models, the dependent slice efficient sampler (and retrospective sampler) can be difficult to implement since we need to simulate λ_j conditional on Λ_{j-1}. The reason for conditioning is that we need Λ_j to update v and to check which components to include in each full conditional for d_i. In the independent slice-efficient sampler, we do not need Λ_m to find which elements to include in the full conditional of d_i and we can integrate Λ_m from the model and update v using this marginalized version of the posterior distribution. That is the full conditional distribution of v is proportional to

$$\mathrm{E}[\exp\{-v\Lambda_m\}]v^{n-1}\exp\left\{-v\sum_{j=1}^{m}\lambda_j\right\}.$$

Since $\mathrm{E}[\exp\{-v\Lambda_m\}]$ is the moment generating function of Λ_m, it's form will often be available analytically. An important advantage of this approach is that we can simulate λ_{m+1} from its unconditional distribution.

1. $\pi(\mu_j, \sigma_j^2|\cdots) \propto p_0(\mu_j,\sigma_j^2)\prod_{d_i=j}\mathrm{N}(y_i;\mu_j,\sigma_j^2)$.
2. $\pi(\lambda_j) \propto \lambda_j^{n_j}\exp\{-v\lambda_j\}$.
3. $\pi(u_i|\cdots) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$.
4. $\mathrm{P}(d_i = k|\cdots) \propto \mathbf{1}(k : \xi_k > u_i)w_k/\xi_k\mathrm{N}(y_i;\mu_k,\sigma_k^2)$.
5. $\pi(v) \propto \mathrm{E}[\exp\{-v\Lambda_m\}]v^{n-1}\exp\{-v\sum_{i=1}^{m}\lambda_i\}$ which is a univariate distribution and can be updated using standard methods.

We now consider a couple of examples.

Example 1 (Gamma distribution) The non-standard element of the sampler is simulating $\lambda_j|\Lambda_{j-1}$. In this case the simulation is simple since $\eta_j = \lambda_j/\Lambda_{j-1}|\Lambda_{j-1} \sim \mathrm{Be}(\gamma_j, \sum_{i=j+1}^{\infty}\gamma_i)$. We simulate $\eta_j \sim \mathrm{Be}(\gamma_j, \sum_{i=j+1}^{\infty}\gamma_i)$ and set $\lambda_j = \eta_j\Lambda_{j-1}$ and $\Lambda_j = (1-\eta_{j-1})\Lambda_{j-1}$.

Example 2 (Inverse-Gaussian distribution) The full conditional distribution of λ_j is given by

$$\pi(\lambda_j|\cdots) \propto \lambda_j^{n_j-3/2}\exp\left\{-\frac{1}{2}\left(\frac{\gamma_j^2}{\lambda_j} + (1+2v)\lambda_j\right)\right\},$$

where n_j is the number of observations allocated to component j. The full conditional distribution of Λ_m is proportional to

$$\Lambda_m^{-3/2}\exp\left\{-\frac{1}{2}\left(\frac{(\sum_{i=m+1}^{\infty}\gamma_i)^2}{\lambda_j} + (1+2v)\lambda_j\right)\right\}.$$

These are both generalized inverse-Gaussian distributions which can be simulated directly; see e.g. Devroye (1986).

We can simulate from $\lambda_j|\Lambda_{j-1}$ by defining $\lambda_j = \eta_j\Lambda_{j-1}$ and $\Lambda_j = (1-\eta_j)\Lambda_{j-1}$ where the density of η_j is given by

$$g(\eta_j) \propto \eta_j^{-3/2}(1-\eta_j)^{-3/2}$$

$$\times \exp\left\{-\frac{1}{2}\left[\frac{\gamma_j^2}{\Lambda_{j-1}\eta_j}+\frac{(\sum_{i=j+1}^{\infty}\gamma_i)^2}{\Lambda_{j-1}(1-\eta_j)}\right]\right\}.$$

Unlike the gamma case, this conditional distribution depends on Λ_{j-1}. The distribution of $\eta_j/(1-\eta_j)$ can be identified as a two-mixture of generalized inverse-Gaussian distributions and hence can be sampled easily (details are given in the Appendix).

4 Hazard functions

Another model which has a similar form of posterior to the normalized mixture models arises in the modeling of random hazard functions. Suppose we model the unknown hazard function $h(t)$, for $t>0$, using a set of known functions $\{h_k(t)\}_{k=1}^{\infty}$, via

$$h(t)=\sum_{k=1}^{\infty}\lambda_k h_k(t).$$

Here the $\{\lambda_k>0\}$ are the model parameters and can be assigned independent gamma prior distributions; say $\lambda_k\sim \mathrm{Ga}(a_k,b_k)$. Obviously we will need to select (a_k,b_k) to ensure that $h(t)<+\infty$ a.s. for all $t<+\infty$. The corresponding density function is given by

$$f(t)=\sum_{k=1}^{\infty}\lambda_k h_k(t)\exp\left\{-\sum_{k=1}^{\infty}\lambda_k H_k(t)\right\},$$

where H_k is the cumulative hazard corresponding to h_k.

So with observations $\{t_i\}_{i=1}^n$, the likelihood function is given by

$$l(\lambda|t)\propto\prod_{i=1}^{n}\left[\sum_{k=1}^{\infty}\lambda_k h_k(t_i)\exp\left\{-\sum_{k=1}^{\infty}\lambda_k H_k(t_i)\right\}\right].$$

Our approach is based on the introduction of a latent variable, say u, so that we consider the joint density with t given by

$$f(t,u)=\sum_{k=1}^{\infty}\mathbf{1}(u<\xi_k)\lambda_k/\xi_k h_k(t)\times\exp\left\{-\sum_{k=1}^{\infty}\lambda_k H_k(t)\right\}.$$

A further latent variable d picks out the mixture component from which (t,u) come,

$$f(t,u,d)=\mathbf{1}(u<\xi_d)\lambda_d/\xi_d h_d(t)\times\exp\left\{-\sum_{k=1}^{\infty}\lambda_k H_k(t)\right\}.$$

We will now introduce the key latent variables, one for each observation, and label them (u_i,d_i), into the likelihood, which is given by

$$l(\lambda|t,u,d)\propto\prod_{i=1}^{n}\mathbf{1}(u_i<\xi_{d_i})\lambda_{d_i}/\xi_{d_i}h_{d_i}(t_i)\times\exp\left\{-\sum_{k=1}^{\infty}\lambda_k H_k(t_i)\right\}.$$

The point is that the choice of d_i is finite. It is now clear that the sampling algorithm for this model is basically the same now as for the normalized case. We could take the λ_j to be gamma with parameters $a_j+\sum_{d_i=j}1$ and $b_j+\sum_{d_i=j}H_j(t_i)$ and we would first sample up to $M=\max_i d_i$. Then the u_i are from $\mathrm{Un}(0,\xi_{d_i})$. In order to sample the d_i we need to find all the ξ_j greater than u_i which is trivial to do.

5 Illustration and comparisons

In this section we carry out a comparison of the slice sampling algorithms with the retrospective sampler and the blocked Gibbs sampler (Ishwaran and James 2001) with the Dirichlet process and a comparison between the slice sampler algorithms and retrospective sampler for the normalized weights prior. We also consider inference for the normalized weights prior mixture model applied to the galaxy data and the hazard function model.

5.1 Algorithmic performance for mixture models

To monitor the performance of the algorithms we look at the convergence of two quantities:

- The number of clusters: at each iteration there are $j=1,\ldots,N$ clusters of the $i=1,\ldots,n$ data points with m_j being the size of the j cluster, so that $\sum_{j=1}^{N}m_j=n$.
- The deviance, D, of the estimated density, calculated as

$$D=-2\sum_{i=1}^{n}\log\left(\sum_j\frac{m_j}{n}K(y_i|\phi_j)\right).$$

These variables have been used in the previous comparison studies of Papaspiliopoulos and Roberts (2008), Green and Richardson (2001) and Neal (2000). Here D is one of the most common functionals used in comparing algorithms, because it is seen as a global function of all model parameters. Although we concentrate on this variable and study its algorithmic performance we are also concerned with the convergence of the number of clusters.

The efficiency of the algorithms is summarized by computing an estimate $\widehat{\tau}$ of the integrated autocorrelation time,

τ, for each of the variables. Integrated autocorrelation time is defined by Sokal (1997) as

$$\tau = \frac{1}{2} + \sum_{l=1}^{\infty} \rho_l,$$

where ρ_l is the autocorrelation at lag l. An estimate of τ has been used in Papaspiliopoulos and Roberts (2008), Green and Richardson (2001) and Neal (2000). Integrated autocorrelation time is of interest as it controls the statistical error in Monte Carlo measurements of a desired function f. To clarify this point, consider the Monte Carlo sample estimate, $\bar{f}$. The variance of $\bar{f}$ (Sokal 1997) is

$$\mathrm{Var}(\bar{f}) \approx \frac{1}{M} 2\tau \times V,$$

where V is the marginal variance of f and M is the number of iterations. This variance is a factor of 2τ larger than the variance when the samples are independent. Therefore a run of M iterations contains only $M/(2\tau)$ "effectively independent data points". This means that the algorithm with the smallest estimated value of τ will be the most efficient. The problem with the calculation of τ lies in accurately estimating the covariance between the states, which in turn is used to calculate the autocorrelation ρ_l. Sokal (1997) suggests the estimator

$$\widehat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1} \widehat{\rho_l}$$

for τ where $\widehat{\rho_l}$ is the estimated autocorrelation at lag l (obtained via MatLab) and C is a cut-off point. In our comparisons we define, as is commonly done,

$$C = \min\left\{ l : |\widehat{\rho_l}| < 2/\sqrt{M} \right\}.$$

Then C is the smallest lag for which we would not reject the null hypothesis $H_0 : \rho_l = 0$. A similar approach has also been used in Papaspiliopoulos (2008). According to Sokal (1997), this approach works well when a sufficient quantity of data is available which we can control by running the sampler for a sufficient number of iterations.

The algorithms are compared using the normal kernel $K(y|\phi)$ with components $\phi = (\mu, \sigma^2)$, and $P_0(\mu, \sigma^{-2}) = \mathrm{N}(\mu|\mu_0, \sigma_0^2) \times \mathrm{Ga}(\sigma^{-2}|\gamma, \beta)$. For comparison purposes we consider a real data set and two simulated data sets. The real data set is the galaxy data which consist of the velocities of 82 distant galaxies diverging from our own galaxy. This is the most commonly used data set in density estimation studies. The simulated data sets are based on the models used in Green and Richardson (2001) and consist of 100 draws from a bimodal and a leptokurtic mixture. The bimodal mixture assumes that $f(y_i) = 0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$ and the leptokurtic mixture assumes that $f(y_i) = 0.67N(0, 1) + 0.33N(0.3, 0.25^2)$. Both of these simulated data sets were used in the algorithm comparison study carried out in Papaspiliopoulos and Roberts (2008); since we are comparing our slice sampler with the retrospective sampler, we decided to use these simulated data sets.

The hyperparameters of P_0 are set according to Green and Richardson (2001). If R is the range of the data; then we take $\mu_0 = R/2$, $\sigma_0^2 = R$, $\gamma = 2$, and $\beta = 0.2R^2$. In the comparison of the estimates of the statistics used, we took the Monte Carlo sample size to be $N = 2{,}000{,}000$ for each algorithm, with the initial 10,000 used as a burn in period.

5.1.1 *Dirichlet process*

In these comparison the precision parameter of the Dirichlet Process is set at $M = 1$. We fit a class of independent slice-efficient samplers by defining $\xi_j = (1-\kappa)\kappa^{j-1}$. An interesting choice is $\kappa = 0.5$ which guarantees that $\xi_j = \mathrm{E}[w_j]$. However, we also consider alternative choices of κ. The blocked Gibbs sampler of Ishwaran and James (2001) approximates the infinite dimensional P by a finite version which is chosen to ensure that the discrepancy between the marginal likelihood of the data under the full model and finite version is less than ϵ. We consider the values $\epsilon = 10^{-6}$ and $\epsilon = 10^{-10}$.

Density estimates using the retrospective and dependent slice-efficient samplers are shown in Fig. 3. They show a strong agreement between the estimates using the two samplers as we would expect.

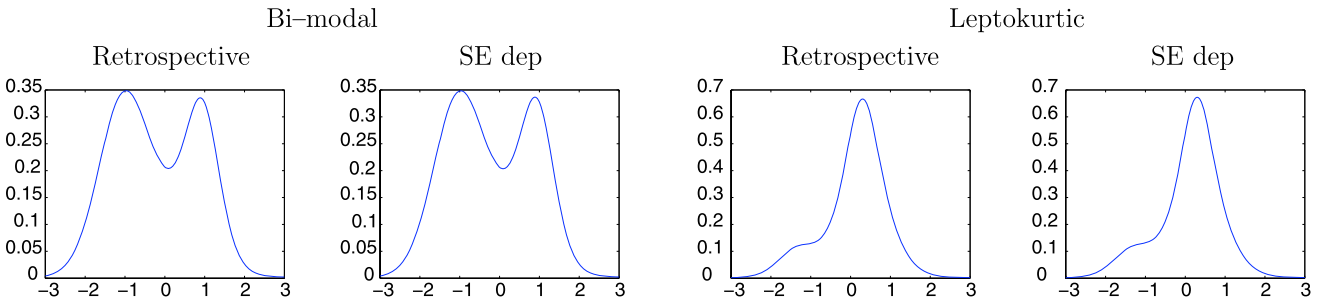

Fig. 3 Predictive densities for the two simulated data sets using the retrospective sampler and dependent slice-efficient sampler

Table 1 Estimates of the integrated autocorrelation times for the deviance (D) and for the number of clusters (K) with the three data sets with the Dirichlet process mixture model. The algorithms are: original slice sampler (Slice), dependent slice-efficient (SE dep) and independent slice-efficient (SE ind) with various values of κ

	Galaxy data		Leptokurtic data		Bimodal data	
	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D
Slice	62.56	20.04	313.52	238.18	334.50	108.72
SE dep	20.08	8.28	65.60	51.60	53.12	21.18
SE ind ($\kappa = 0.3$)	65.86	21.86	136.18	78.84	125.40	56.82
SE ind ($\kappa = 0.4$)	41.66	16.34	77.86	49.96	76.00	33.48
SE ind ($\kappa = 0.5$)	32.36	15.50	56.20	38.54	52.46	26.88
SE ind ($\kappa = 0.6$)	24.76	10.22	43.04	30.62	40.48	20.86
SE ind ($\kappa = 0.7$)	19.30	8.72	34.64	24.34	39.64	18.42
SE ind ($\kappa = 0.8$)	16.56	7.08	30.30	21.38	29.70	15.54
Retrospective	13.04	5.48	26.82	18.10	28.94	13.82
Block Gibbs ($\epsilon = 10^{-6}$)	13.18	6.06	25.36	18.40	26.54	13.02
Block Gibbs ($\epsilon = 10^{-10}$)	13.52	5.92	26.92	17.86	25.92	13.00

Table 2 Estimates of the integrated autocorrelation times for the deviance (D) and for the number of clusters (K) with the three data sets fitting the infinite Dirichlet distribution mixture model. The algorithms are: dependent slice-efficient sampler (SE dep), independent slice-efficient sampler (SE ind) and Retrospective sampler

	Galaxy data		Leptokurtic data		Bimodal data	
	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D
SE dep	50.50	23.92	230.22	158.90	127.88	34.06
SE ind ($\kappa = 0.5$)	98.60	16.20	99.24	64.04	279.32	83.18
SE ind ($\kappa = 0.6$)	94.26	11.58	81.44	51.58	86.00	17.36
SE ind ($\kappa = 0.7$)	76.90	7.10	71.82	37.96	63.04	12.72
SE ind ($\kappa = 0.8$)	43.34	7.14	59.48	33.72	50.02	10.36
Retrospective	53.74	13.66	96.14	57.76	87.60	16.78

Table 3 Estimates of the integrated autocorrelation times for the deviance (D) and for the number of clusters (K) with the three data sets fitting the infinite normalized inverse-Gaussian distribution mixture model. The algorithms are: dependent slice-efficient sampler (SE dep), independent slice-efficient sampler (SE ind) and Retrospective sampler

	Galaxy data		Leptokurtic data		Bimodal data	
	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D	$\widehat{\tau}$ for K	$\widehat{\tau}$ for D
SE dep	44.32	17.28	83.40	62.78	68.94	31.08
SE ind ($\kappa = 0.5$)	71.12	17.40	53.92	40.02	50.36	20.94
SE ind ($\kappa = 0.6$)	45.80	11.92	41.62	31.56	37.98	16.02
SE ind ($\kappa = 0.7$)	37.18	9.92	34.70	23.44	36.62	14.30
SE ind ($\kappa = 0.8$)	33.06	9.50	31.02	20.86	31.08	12.24
Retrospective	33.32	9.00	54.76	42.54	45.90	18.40

The integrated autocorrelation times for the algorithms for the three data sets are presented in Table 1. For each data set we find that the original slice sampler is the least efficient. The dependent slice-efficient sampler is much more efficient with a reduction in integrated autocorrelation time of 3 to 6 times. The efficiency of the independent slice-efficient sampler depends on the value of κ with the integrated autocorrelation time falling as κ increases. The value of κ makes a big difference with the integrated autocorrelation time when $\kappa = 0.3$ roughly 4 times the value when $\kappa = 0.8$ for all three data sets. The retrospective sampler is usually the most efficient but differences between the most

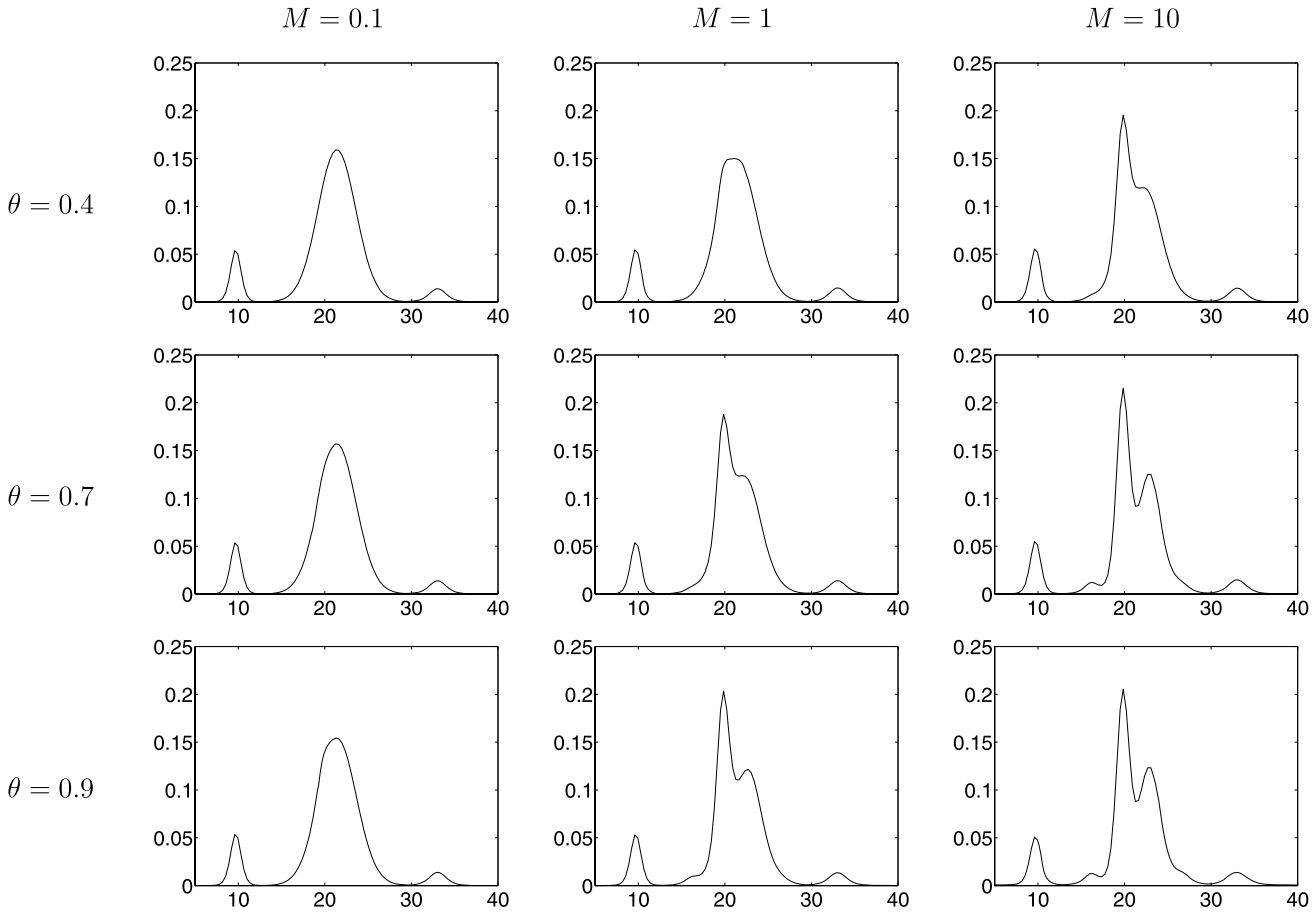

Fig. 4 Posterior mean density estimates for the galaxy data using the infinite Dirichlet prior with different values of M and θ

efficient slice samplers and these methods are very small. Even though the retrospective sampler performs marginally better, the slice-efficient sampler is easier to use as simulating the v and d is carried out in an easy way, as opposed to the complexity of the set up of the retrospective sampling steps.

5.1.2 Mixtures based on normalized weights

We reject the slice sampler and independent slice-efficient sampler with $\kappa < 0.5$ and concentrate on the other methods. We use the infinite Dirichlet (Table 2) and infinite normalized inverse-Gaussian (Table 3) mixture models with $M = 1$ and $\theta = 0.5$ on the three data sets. The relative performance of the samplers for the normalized weights prior is similar to the relative performance for the Dirichlet process prior. The retrospective sampler is usually more efficient than the dependent slice-efficient sampler with a small difference for the galaxy data set and a much larger difference for the two simulated data sets. The effect is also more pronounced for the infinite Dirichlet distribution prior rather than the infinite normalized inverse-Gaussian prior.

The integrated autocorrelation time for the independent slice-sampling methods depends on the choice of κ. In all case, the best independent slice-efficient samplers (for larger κ) outperforms the dependent slice-efficient and retrospective samplers. It seems that the difference in the algorithm is due to integrating Λ_m from the model. Standard MCMC theory would suggest that this will define more efficient samplers (which is supported by the results of Celeux et al. 2000 for finite mixture models).

5.2 Inference for mixture models with the normalized weights priors

The galaxy data has been a popular data set in Bayesian nonparametric modelling and we will illustrate the infinite Dirichlet and infinite normalized inverse-Gaussian priors on it. The posterior mean density estimates are shown in Fig. 4 for the infinite Dirichlet prior and Fig. 5 for the infinite normalized inverse-Gaussian prior. The hyperparameters of the prior distributions have a clear effect on the posterior mean estimates. Prior distributions that places more mass on a small number of components tend to find estimates

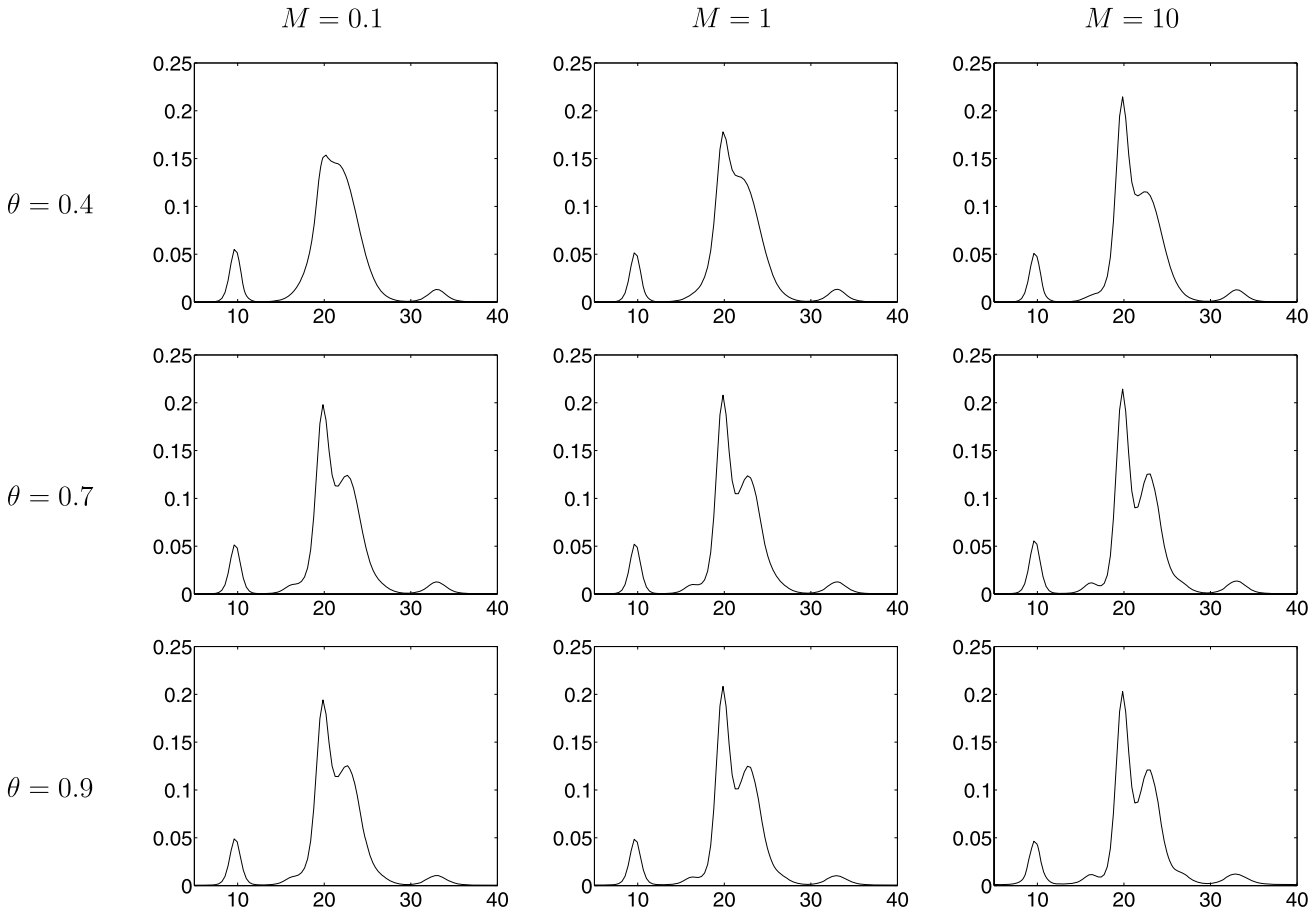

Fig. 5 Posterior mean density estimates for the galaxy data using the infinite normalized inverse-Gaussian prior with different values of M and θ

with three clear modes. As the prior mean number of components increases so do the number of modes in the estimate from four to five for the prior within each class that places most mass on a large number of components ($M = 10$ and $\phi = 0.9$). However, there are some clear differences between the two classes of prior. The effects of the two hyperparameters on the prior distribution of the number of non-empty components were more clearly distinguishable in the infinite normalized inverse-Gaussian prior than the infinite Dirichlet prior. In the infinite normalized inverse-Gaussian prior θ controls the mean number of non-empty components whereas M controls the dispersion around the mean. This property is carried forward to the posterior mean density and the number of modes in the posterior mean increases with θ. For example, when $M = 0.1$, there are three modes in the posterior mean if $\theta = 0.4$ whereas there are four when $\theta = 0.9$. Similarly, larger values of M are associated with larger variability in the prior mean and favour distributions which uses a larger number of components. This suggests that infinite normalized inverse-Gaussian distribution may be a more easily specified prior distribution than the infinite Dirichlet prior.

5.3 Inference for hazard functions

This example considers applying the Bayesian nonparametric model for the hazard function in Sect. 4. We choose the components of the hazard function $h_k(t)$ to have a Weibull form so that $h_k(t) = t^{\alpha_k}$, $\lambda_k \sim \text{Ga}(M(1-\theta)\theta^{k-1}, 1)$ and $\alpha_k \sim \text{Ga}(2, 1/2)$. The parameter $M = 10$ and $\theta = 0.7$ which places most of the prior mass on four to ten components. The data consists of remission times in weeks of leukemia patients. There were 21 observed times of which 12 were censored. Figure 6 shows the posterior median integrated hazard function with 95% pointwise credible interval and a Kaplan-Meier estimate. There is clear agreement between the two nonparametric estimates with the Bayesian model offering a "smoothed" version of the Kaplan-Meier estimate.

6 Conclusions and discussion

This paper has shown how mixture models based on random probability measures, of either the stick-breaking or normalized types, can be easily handled via the introduction of a

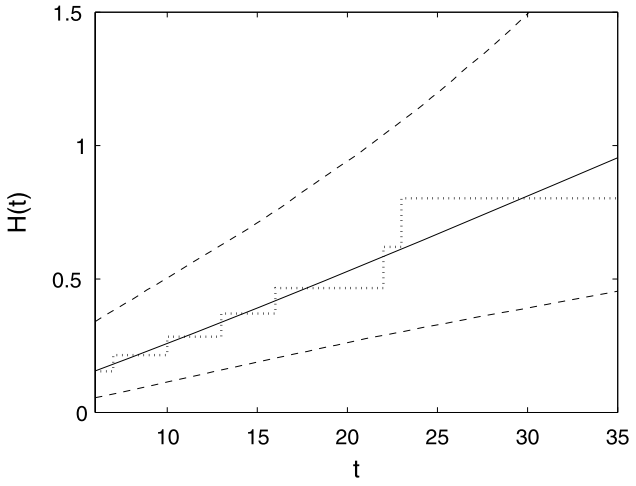

Fig. 6 Estimates of the integrated hazard function, $H(t)$, using the nonparametric prior showing the median (*solid line*) and 95% credible interval (*dashed line*) with Kaplan-Meier estimator (*dotted line*)

key latent variable which makes finite the number of mixtures. The more complicated of the two is the normalized type, which requires particular distributions of the unnormalized weights in order to be able to make the simulation algorithm work. Nevertheless, such distributions based on the gamma and inverse-Gaussian distributions are popular choices anyway.

Further ideas which need to be worked out include the case when we can generate weights which are decreasing. This for example would make the search for those $w_j > u$ are far simpler exercise and would lead to more efficient algorithms.

In conclusion, concerning performance of slice-efficient and retrospective samplers, we note that once running, both samplers are approximately the same in terms of efficiency and performance. However, the savings are in the pre-running work where setting up a slice sampler is far easier than setting up a retrospective sampler.

The slice sampler allows the Gibbs sampling step for a finite mixture model to be used at each iteration and introduce a method for updating the truncation point in each iteration. This allows standard methods for finite mixture models to be used directly. For example, Van Gael et al. (2008) fit an infinite hidden Markov model using the forward-backward sampler for finite hidden Markov model using the slice sampling idea. This would be difficult to implement in a retrospective framework since the truncation point changes when updating the allocations.

Acknowledgements We acknowledge the helpful comments of two referees and an associate editor.

Appendix

Simulation for the Inverse-Gaussian model We wish to simulate from the density $g(x_{j+1})$

$$g(x_{j+1}) \propto x_{j+1}^{-3/2}(1-x_{j+1})^{-3/2} \times \exp\left\{-\frac{1}{2}\left[\frac{\gamma_j^2}{\Lambda_j x_{j+1}} + \frac{(\sum_{i=j+1}^{\infty}\gamma_i)^2}{\Lambda_j(1-x_{j+1})}\right]\right\}.$$

The transformation $y_{j+1} = \frac{x_{j+1}}{1-x_{j+1}}$ has the density

$$g(y_{j+1}) \propto y_{j+1}^{-3/2}(1+y_{j+1}) \times \exp\left\{-\frac{1}{2}\left[\frac{\gamma_j^2}{\Lambda_j y_{j+1}} + \frac{(\sum_{i=j+1}^{\infty}\gamma_i)^2}{\Lambda_j} y_{j+1}\right]\right\},$$

which can be expressed as a mixture of two generalized inverse-Gaussian distributions

$$wGIG\left(-1/2, \gamma_j/\Lambda_j, \sum_{i=j+1}^{\infty}\gamma_i/\Lambda_j\right) + (1-w)GIG\left(1/2, \gamma_j/\Lambda_j, \sum_{i=j+1}^{\infty}\gamma_i/\Lambda_j\right)$$

where

$$w = \frac{\gamma_j}{\sum_{i=j+1}^{\infty}\gamma_i}$$

and $GIG(p,a,b)$ denotes a distribution with density

$$\frac{(b/a)^{p/2}}{2K_p(\sqrt{ab})} x^{(p-1)} \exp\{-(a/x+bx)/2\}$$

where K_ν denotes the modified Bessel function of the third kind with index ν.

References

Celeux, G., Hurn, M., Robert, C.P.: Computational and inferential difficulties with mixture posterior distributions. J. Am. Stat. Assoc. **95**, 957–970 (2000)

Devroye, L.: Non-Uniform Random Variate Generation. Springer, New York (1986)

Dunson, D.: Kernel local partition processes for functional data. Discussion paper 2008-26, Department of Statistical Science, Duke University (2008)

Escobar, M.D.: Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University (1988)

Escobar, M.D.: Estimating normal means with a Dirichlet process prior. J. Am. Stat. Assoc. **89**, 268–277 (1994)

Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**, 577–588 (1995)

Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)

Gilks, W.R., Best, N.G., Tan, K.K.C.: Adaptive rejection Metropolis sampling within Gibbs sampling. Appl. Stat. **44**, 455–472 (1995)

Green, P.J., Richardson, S.: Modelling heterogeneity with and without the Dirichlet process. Scand. J. Stat. **28**, 355–375 (2001)

Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**, 161–173 (2001)

Ishwaran, H., Zarepour, M.: Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. Biometrika **87**, 371–390 (2000)

Lijoi, A., Mena, R.H., Prünster, I.: Hierarchical mixture modeling with normalized inverse-Gaussian priors. J. Am. Stat. Assoc. **100**, 1278–1291 (2005)

Lijoi, A., Mena, R.H., Prüenster, I.: Controlling the reinforcement in Bayesian nonparametric mixture models. J. R. Stat. Soc. B **69**, 715–740 (2007)

Lo, A.Y.: On a class of Bayesian nonparametric estimates I. Density estimates. Ann. Stat. **12**, 351–357 (1984)

MacEachern, S.N.: Estimating normal means with a conjugate style Dirichlet process prior. Commun. Stat., Simul. Comput. **23**, 727–741 (1994)

MacEachern, S.N., Müller, P.: Estimating mixtures of Dirichlet process models. J. Comput. Graph. Stat. **7**, 223–238 (1998)

Neal, R.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**, 249–265 (2000)

Papaspiliopoulos, O.: A note on posterior sampling from Dirichlet mixture models. Preprint (2008)

Papaspiliopoulos, O., Roberts, G.O.: Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika **95**, 169–186 (2008)

Sethuraman, J.: A constructive definition of Dirichlet priors. Stat. Sin. **4**, 639–650 (1994)

Sokal, A.: Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms Functional Integration, Cargése, 1996. NATO Adv. Sci. Inst. Ser. B Phys., vol. 361, pp. 131–192. Plenum, New York (1997),

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

Van Gael, J., Saatchi, Y., Teh, Y.W., Ghahramani, Z.: Beam sampling for the infinite hidden Markov model. Technical Report: Engineering Department, University of Cambridge (2008)

Walker, S.G.: Sampling the Dirichlet mixture model with slices. Commun. Stat., Simul. Comput. **36**, 45–54 (2007)

Yau, C., Papaspiliopoulos, O., Roberts, G.O., Holmes, C.: Bayesian nonparametric hidden Markov models with application to the analysis of copy-number-variation in mammalian genomes. Technical Report, Man Institute, Oxford (2008)