Bayesian semiparametric modeling and inference with mixtures of symmetric distributions

Athanasios Kottas [1] and Gilbert W. Fellingham[2]

[1] Department of Applied Mathematics and Statistics, University of California, Santa Cruz, USA

[2] Department of Statistics, Brigham Young University, USA

Abstract: We propose a semiparametric modeling approach for mixtures of symmetric distributions. The mixture model is built from a common symmetric density with different components arising through different location parameters. This structure ensures identifiability for mixture components, which is a key feature of the model as it allows applications to settings where primary interest is inference for the subpopulations comprising the mixture. We focus on the two-component mixture setting and develop a Bayesian model using parametric priors for the location parameters and for the mixture proportion, and a nonparametric prior probability model, based on Dirichlet process mixtures, for the random symmetric density. We present an approach to inference using Markov chain Monte Carlo posterior simulation. The performance of the model is studied with a simulation experiment and through analysis of a rainfall precipitation data set as well as with data on eruptions of the Old Faithful geyser.

Key words: Dirichlet process prior; Identifiability; Markov chain Monte Carlo; Mixture deconvolution; Scale uniform mixtures.

[1]Address for correspondence: Athanasios Kottas, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064, USA. E-mail: thanos@ams.ucsc.edu

1 Introduction

In seeking to extend the scope of standard parametric families of distributions one is naturally led to mixture models. Continuous mixture models, arising through a parametric continuous family of mixing distributions, typically achieve increased heterogeneity but are still limited to unimodality and usually symmetry. Finite mixture distributions provide more flexible modeling, since with appropriate mixing, and sufficiently large number of mixture components, they can approximate a wide range of density shapes.

Hence, mixture models are commonly utilized to provide a flexible framework for modeling distributions with non-standard features. Under this setting, inferential interest focuses on the mixture distribution rather than the individual components that comprise the mixture. In this context, key Bayesian modeling approaches build on discrete mixtures with an unknown number of components or nonparametric mixture models based mainly on Dirichlet process (DP) priors (see, e.g., the respective reviews in Marin, Mengersen and Robert, 2005, and Müller and Quintana, 2004).

The methodology proposed in this paper is concerned with mixture deconvolution problems, which provide another important application area for finite mixture models. We consider univariate continuous distributions on the real line, for which the density of a general finite mixture model can be denoted by $\sum_{j=1}^{k} \pi_j f_j(\cdot)$, where the $f_j(\cdot)$ are densities on $\mathbb{R}$, and the mixture weights $\pi_j \geq 0$ with $\sum_{j=1}^{k} \pi_j = 1$. For mixture deconvolution settings, of primary interest is inference for the subpopulations $f_j(\cdot)$ comprising the mixture population model, which thus shifts the focus to mixtures with a specified, typically small, number of components, k, each one of which has a specific interpretation under the particular application. For instance, in epidemiological studies, the subpopulations of interest may correspond to non-diseased and diseased subjects (and thus $k = 2$), or the entire mixture may model an infected population with components corresponding to

k distinct stages of infection.

It is well known that the mixture components $f_j(\cdot)$ in the general mixture model $\sum_{j=1}^{k} \pi_j f_j(\cdot)$ are not identifiable without restrictions on the model parameters. Various forms of identifiability results are possible when the $f_j(\cdot)$ are assumed to correspond to specific parametric families of distributions (e.g., Titterington, Smith and Makov, 1985). Addressing the nonidentifiability issue is more challenging under a semiparametric mixture model setting where the $f_j(\cdot)$ are left unspecified under a classical nonparametric approach, or assigned nonparametric priors under Bayesian nonparametric modeling. Recent work of Bordes, Mottelet and Vandekerkhove (2006) and Hunter, Wang and Hettmansperger (2007) provides identifiability results for a more structured semiparametric version of the general finite mixture discussed above. Specifically, $\sum_{j=1}^{k} \pi_j f(\cdot - \mu_j)$, where $f(\cdot)$ is a density on $\mathbb{R}$, which is symmetric about 0, and the μ_j are component-specific location parameters. In particular, Bordes et al. (2006), and, independently, Hunter et al. (2007) proved identifiability when $k = 2$. Moreover, Hunter et al. (2007) established also identifiability for $k = 3$, although under complex constraints on the location parameters and mixture weights. In addition to the estimation approaches in Bordes et al. (2006) and Hunter et al. (2007), classical semiparametric estimation techniques have been proposed by Cruz-Medina and Hettmansperger (2004) and Bordes, Chauveau and Vandekerkhove (2007).

Our objective is to develop a fully inferential Bayesian modeling framework for such mixtures of location-shifted symmetric distributions. We focus on the two-component mixture setting, $\pi f(\cdot - \mu_1) + (1 - \pi) f(\cdot - \mu_2)$, and propose a Bayesian semiparametric probability model based on parametric priors for μ_1, μ_2 and π, and a nonparametric prior model for the symmetric density $f(\cdot)$, which is further taken to be unimodal. The additional assumption of unimodality for $f(\cdot)$ is natural for deconvolution problems, and

in our context, facilitates the choice of the prior for the nonparametric component of the mixture model. In particular, we employ a nonparametric scale uniform DP mixture prior for $f(\cdot)$ that supports the entire class of symmetric unimodal densities on $\mathbb{R}$. We argue that the combination of the theoretical identifiability results discussed above with flexible probabilistic modeling for $f(\cdot)$ provides a practically useful framework for inference in mixture deconvolution problems.

The outline of the paper is as follows. Section 2 presents the modeling framework, including approaches to prior specification and methods for posterior inference, with technical details on the latter deferred to the Appendix. Section 3 illustrates model performance through a simulation study and through analysis of data on eruptions of the Old Faithful geyser. We also consider comparison with a two-component normal mixture model, using a rainfall precipitation data set. Finally, Section 4 concludes with a summary and discussion of possible extensions.

2 Methodology

Section 2.1 presents the modeling approach. Prior specification is addressed in Section 2.2, and posterior inference is discussed in Section 2.3.

2.1 Semiparametric modeling for the two-component mixture

We develop an approach to modeling and inference under the following two-component mixture setting,

$$\pi f(y-\mu_1) + (1-\pi) f(y-\mu_2), \quad y \in \mathbb{R}, \tag{1}$$

where $f(\cdot)$ is a density on $\mathbb{R}$ assumed to be unimodal and symmetric about 0, π is the mixing proportion, and μ_1 and μ_2 are location parameters associated with the two mixture components. We build a semiparametric modeling framework working with parametric priors for π, μ_1 and μ_2, and a nonparametric prior model for $f(\cdot)$.

The foundation for constructing the probability model for $f(\cdot)$ is the representation of non-increasing densities on the positive real line as scale mixtures of uniform densities. Specifically, a density $h(\cdot)$ on $\mathbb{R}^+$ is non-increasing if and only if there exists a distribution function G on $\mathbb{R}^+$ such that $h(t) \equiv h(t;G) = \int \theta^{-1} 1_{[0,\theta)}(t) \mathrm{d}G(\theta)$, $t \in \mathbb{R}^+$ (see, e.g., Feller, 1971, p. 158.) Using this result, we have a representation for $f(\cdot)$ through a scale mixture of symmetric uniform densities, that is, $f(y) \equiv f(y;G) = \int 0.5\theta^{-1} 1_{(-\theta,\theta)}(y) \mathrm{d}G(\theta)$, $y \in \mathbb{R}$, where the mixing distribution G is supported on $\mathbb{R}^+$. This one-to-one mapping between f and G enables a nonparametric model for $f(\cdot)$ through a nonparametric prior for G. In particular, placing a DP prior on G, a scale uniform DP mixture emerges for $f(\cdot;G)$. We refer to Brunner and Lo (1989), Brunner (1992; 1995), and Hansen and Lauritzen (2002) for related work as well as Lavine and Mockus (1995), Kottas and Gelfand (2001), and Kottas and Krnjajić (2009) for DP-based modeling involving variations and extensions of the above representation leading to $f(\cdot;G)$.

As developed by Ferguson (1973), the DP serves as a prior model for random distributions (equivalently, distribution functions) G, which, in our context, have support on $\mathbb{R}^+$. The DP is defined in terms of two parameters, a parametric base distribution G_0 (the mean of the process) and a positive scalar parameter α, which can be interpreted as a precision parameter; larger values of α result in DP realizations G that are closer to G_0. We will write $G \sim \mathrm{DP}(\alpha, G_0)$ to indicate that a DP prior is used for the random distribution G. In fact, DP-based modeling typically utilizes mixtures of DPs (Antoniak, 1974), i.e., a more general version of the DP prior that involves hyperpriors for α and/or the

parameters of G_0. The most commonly used DP definition is its constructive definition (Sethuraman, 1994), which characterizes DP realizations as countable mixtures of point masses (and thus as random discrete distributions). Specifically, a random distribution G generated from a $\text{DP}(\alpha, G_0)$ is (almost surely) of the form

$$G(\cdot) = \sum_{\ell=1}^{\infty} w_\ell \, \delta_{\vartheta_\ell}(\cdot) \tag{2}$$

where $\delta_a(\cdot)$ denotes a point mass at a. The locations of the point masses, ϑ_ℓ, are i.i.d. realizations from G_0; the corresponding weights, w_ℓ, arise from a *stick-breaking* mechanism based on i.i.d. draws $\{\zeta_k : k = 1, 2, ...\}$ from a beta$(1, \alpha)$ distribution. In particular, $w_1 = \zeta_1$, and, for each $\ell = 2, 3, ..., \; w_\ell = \zeta_\ell \prod_{k=1}^{\ell-1}(1 - \zeta_k)$. Moreover, the sequences $\{\vartheta_\ell, \ell = 1, 2, \ldots\}$ and $\{\zeta_k : k = 1, 2, ...\}$ are independent.

Utilizing the representation theorem discussed above with a DP prior for the mixing distribution G, our proposed nonparametric prior probability model for the density $f(\cdot)$ in (1) is given by

$$f(y) \equiv f(y; G) = \int_{\mathbb{R}^+} u(y; \theta) \mathrm{d}G(\theta), \;\; y \in \mathbb{R}; \quad G \mid \alpha, \beta \sim \text{DP}(\alpha, G_0) \tag{3}$$

where $u(\cdot; \theta) = 0.5\theta^{-1} 1_{(-\theta, \theta)}(\cdot)$ denotes the density of the uniform distribution on $(-\theta, \theta)$. We take an inverse gamma distribution for G_0 with mean $\beta/(c-1)$ (provided $c > 1$). We work with fixed shape parameter c (with $c > 1$) and random scale parameter β with prior density denoted by $p(\beta)$. We also place a prior $p(\alpha)$ on the DP precision parameter α. (Note that we retain the more common "DP prior" terminology even though the prior for G is, in fact, a mixture of DPs.) The model is completed with priors $p(\mu_1)$, $p(\mu_2)$ and $p(\pi)$ for μ_1, μ_2 and π. Parameters μ_1, μ_2, π, α and β are assumed independent in their prior. We discuss prior choice and specification in Section 2.2.

Hence, the semiparametric model for the data $= \{y_i : i = 1, ..., n\}$ is given by

$$\begin{aligned} y_i \mid \pi, \mu_1, \mu_2, G &\overset{\text{ind.}}{\sim} \pi f(y_i - \mu_1; G) + (1-\pi) f(y_i - \mu_2; G), \;\; i = 1, ..., n \\ G \mid \alpha, \beta &\sim \text{DP}(\alpha, G_0(\cdot; \beta)) \\ \pi, \mu_1, \mu_2, \alpha, \beta &\sim p(\pi)p(\mu_1)p(\mu_2)p(\alpha)p(\beta) \end{aligned} \tag{4}$$

where $f(\cdot; G)$ is the scale uniform DP mixture defined in (3).

The model can be expressed hierarchically by introducing two sets of latent mixing parameters for the observables y_i, $i = 1, ..., n$. The first set of binary mixing parameters, say, $\{z_i : i = 1, ..., n\}$, can be used to break the two-component mixture $\pi f(y_i - \mu_1; G) + (1-\pi) f(y_i - \mu_2; G)$, so that the y_i, given z_i and μ_1, μ_2, G, are independent $f(y_i - \mu_{z_i}; G)$, and the z_i, given π, are i.i.d. with $\Pr(z_i = 1 \mid \pi) = \pi$ and $\Pr(z_i = 2 \mid \pi) = 1 - \pi$ (denoted below by $\Pr(z_i \mid \pi)$). The second set of mixing parameters, $\{\theta_i : i = 1, ..., n\}$, is added to break the DP mixture $f(y_i - \mu_{z_i}; G)$, where now the y_i, given z_i, μ_1, μ_2 and θ_i are independent $u(y_i - \mu_{z_i}; \theta_i)$, and the θ_i, given G, are i.i.d. G. Therefore, the full hierarchical model can be written as

$$\begin{aligned} y_i \mid \mu_1, \mu_2, \theta_i, z_i &\overset{\text{ind.}}{\sim} u(y_i - \mu_{z_i}; \theta_i), \;\; i = 1, ..., n \\ z_i \mid \pi &\overset{\text{ind.}}{\sim} \Pr(z_i \mid \pi), \;\; i = 1, ..., n \\ \theta_i \mid G &\overset{\text{i.i.d.}}{\sim} G, \;\; i = 1, ..., n \\ G \mid \alpha, \beta &\sim \text{DP}(\alpha, G_0(\cdot; \beta)) \\ \pi, \mu_1, \mu_2, \alpha, \beta &\sim p(\pi)p(\mu_1)p(\mu_2)p(\alpha)p(\beta). \end{aligned} \tag{5}$$

It is straightforward to show that if we integrate over the θ_i and then over the z_i in (5), we obtain (4).

2.2 Prior specification

To apply the model developed in Section 2.1, we work with $\mathrm{N}(a_k, b_k^2)$ priors (with standard deviation b_k) for μ_k, $k = 1, 2$, a beta(a_π, b_π) prior for π, a gamma(a_α, b_α) prior (with mean a_α/b_α) for α, and an exponential prior (with mean $1/b_\beta$) for β.

If prior information is available on the modes of the mixture distribution and their relative weights, it can be used to specify the priors for μ_1, μ_2 and π. Less informative prior specification is also possible, and indeed results in robust and accurate posterior inference, as illustrated with the data examples of Section 3. Specifically, we use the same normal prior for μ_1 and μ_2 with mean set equal to a guess at the center of the data, and variance defined through a guess at the data range. Note that such an approach may be too uninformative for a general parametric mixture, as shown with the rainfall data in Section 3.3, but worked well under model (4) for all the data sets we considered. Moreover, a uniform distribution on $(0, 1)$ is the natural noninformative prior for π.

To specify the priors for the DP hyperparameters α and β, based on weak prior information, we can use the role they play in the DP mixture prior model. The DP precision parameter α controls the number, $n^* \leq n$, of distinct DP mixture components, i.e., the number of distinct θ_i in (5) implied by the discrete random mixing distribution G (e.g., Antoniak, 1974; Escobar and West, 1995). For instance, for moderately large n, $\mathrm{E}(n^* \mid \alpha) \approx \alpha \log\{(\alpha + n)/\alpha\}$, which can be averaged over the gamma prior for α to obtain $\mathrm{E}(n^*)$. Hence, prior beliefs about n^* can be converted into particular prior choices for α. We note that applications of DP mixture models involve mainly location or location-scale mixing with unimodal kernels, in which case n^* would typically be small relative to n. In contrast to location DP mixtures, model (3) involves scale uniform mixing, which typically implies a large number of mixture components, i.e., large n^*, to model the underlying symmetric unimodal density $f(\cdot)$.

Regarding the scale parameter β of the inverse gamma distribution G_0, a practically useful approach to prior specification emerges by considering the limiting case of model (5) as $\alpha \to \infty$, which implies a distinct θ_i for each observation y_i (i.e., $n^* = n$). As discussed above, this is a natural limiting case of DP mixture model (3). Then, the θ_i are i.i.d. $G_0(\cdot;\beta)$, and thus $\mathrm{E}(\theta_i) = \mathrm{E}\{\mathrm{E}(\theta_i \mid \beta)\} = \mathrm{E}(\beta(c-1)^{-1}) = b_\beta^{-1}(c-1)^{-1}$. Therefore, $2b_\beta^{-1}(c-1)^{-1}$ can be used as a proxy for the range of the mixture kernel. Finally, based on the guess at the data range, say, R, we use $R/2$ as a guess at the range of each mixture component, and specify b_β through $2b_\beta^{-1}(c-1)^{-1} = R/2$. Recall that the shape parameter c is fixed; for all the data examples of Section 3, c is set to 2, yielding an inverse gamma distribution for G_0 with infinite variance, and mean β.

Finally, for any particular prior specification it is straightforward to estimate the associated prior predictive density, $p(y_0)$, which corresponds to model (5) for a single generic y_0. The corresponding mixing parameter θ_0 arises, given β, from the $G_0(\theta_0;\beta)$ distribution. Moreover, for the associated binary mixing parameter z_0, we have $\Pr(z_0 = 1 \mid \pi) = \pi$ and $\Pr(z_0 = 2 \mid \pi) = 1 - \pi$. Hence, marginalizing over z_0, $p(y_0)$ is given by

$$\int\int \{\pi u(y_0 - \mu_1;\theta_0) + (1-\pi)u(y_0 - \mu_2;\theta_0)\} g_0(\theta_0;\beta) p(\pi) p(\mu_1) p(\mu_2) p(\beta)\, \mathrm{d}\theta_0 \mathrm{d}\pi \mathrm{d}\mu_1 \mathrm{d}\mu_2 \mathrm{d}\beta,$$

where g_0 is the density of G_0. Note that $p(y_0)$ is defined through only the center G_0 of the DP prior, i.e., the DP precision parameter does not enter its expression. The prior predictive density reveals the combined effect of four of the five priors that need to be specified in order to implement inference under model (4). Moreover, as shown with the data examples of Sections 3.1 and 3.2, comparison of the prior predictive with the posterior predictive density (the latter developed in Section 2.3) indicates the amount of learning from the data for Bayesian density estimation under the semiparametric mixture model (4).

2.3 Posterior inference

To develop inference under the model of Section 2.1, we employ a Markov chain Monte Carlo (MCMC) algorithm for posterior sampling from the marginalized version of model (5) that results by integrating G over its DP prior,

$$\begin{aligned} y_i \mid \mu_1, \mu_2, \theta_i, z_i &\overset{\text{ind.}}{\sim} u(y_i - \mu_{z_i}; \theta_i), \;\; i = 1, ..., n \\ (\theta_1, ..., \theta_n) \mid \alpha, \beta &\sim p(\theta_1, ..., \theta_n \mid \alpha, \beta) \\ z_i \mid \pi &\overset{\text{ind.}}{\sim} \Pr(z_i \mid \pi), \;\; i = 1, ..., n \\ \pi, \mu_1, \mu_2, \alpha, \beta &\sim p(\pi)p(\mu_1)p(\mu_2)p(\alpha)p(\beta). \end{aligned} \tag{6}$$

The induced joint prior for the θ_i can be developed using the Pólya urn characterization of the DP (Blackwell and MacQueen, 1973). Specifically,

$$p(\theta_1, ..., \theta_n \mid \alpha, \beta) = g_0(\theta_1; \beta) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} g_0(\theta_i; \beta) + \frac{1}{\alpha + i - 1} \sum_{\ell=1}^{i-1} \delta_{\theta_\ell}(\theta_i) \right\}. \tag{7}$$

As the above expression indicates, the discreteness of the DP prior induces a partition of the vector of DP mixing parameters, $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, into n^* ($\leq n$) distinct components θ_j^*, $j = 1, ..., n^*$. The θ_j^* along with configuration indicators $(s_1, ..., s_n)$, defined by $s_i = j$ if and only if $\theta_i = \theta_j^*$, provide an equivalent representation for $\boldsymbol{\theta}$. Denote by $\boldsymbol{\psi}$ the full parameter vector consisting of $\boldsymbol{z} = (z_1, ..., z_n)$, $\boldsymbol{\theta}$, π, μ_1, μ_2, α, and β. The Appendix provides the details of the MCMC method for sampling from the posterior of $\boldsymbol{\psi}$. We note that key to the good performance of the MCMC algorithm are joint updates for each pair, (z_i, θ_i), of mixing parameters corresponding to the i-th observation.

The samples from $p(\boldsymbol{\psi} \mid \text{data})$ can be used to obtain posterior predictive inference for y_0, a generic *new* observation with corresponding latent mixing parameters z_0 and θ_0. Adding y_0, z_0, and θ_0 to the model structure in (5), and obtaining next the induced

marginalized version (which thus extends (6)), we have

$$p(y_0, z_0, \theta_0, \boldsymbol{\psi} \mid \text{data}) = u(y_0 - \mu_{z_0}; \theta_0) p(\theta_0 \mid \boldsymbol{\theta}, \alpha, \beta) \Pr(z_0 \mid \pi) p(\boldsymbol{\psi} \mid \text{data})$$

where the form of the conditional posterior, $p(\theta_0 \mid \boldsymbol{\theta}, \alpha, \beta)$, for θ_0 arises from the DP Pólya urn structure. Specifically,

$$p(\theta_0 \mid \boldsymbol{\theta}, \alpha, \beta) = \frac{\alpha}{\alpha + n} g_0(\theta_0; \beta) + \frac{1}{\alpha + n} \sum_{j=1}^{n^*} n_j \delta_{\theta_j^*}(\theta_0).$$

Therefore,

$$p(y_0 \mid \text{data}) = \int \int \sum_{z_0} u(y_0 - \mu_{z_0}; \theta_0) p(\theta_0 \mid \boldsymbol{\theta}, \alpha, \beta) \Pr(z_0 \mid \pi) p(\boldsymbol{\psi} \mid \text{data}) \, \mathrm{d}\theta_0 \mathrm{d}\boldsymbol{\psi}$$

or, evaluating the sum over z_0,

$$p(y_0 \mid \text{data}) = \int \int \{\pi u(y_0 - \mu_1; \theta_0) + (1 - \pi) u(y_0 - \mu_2; \theta_0)\} p(\theta_0 \mid \boldsymbol{\theta}, \alpha, \beta) p(\boldsymbol{\psi} \mid \text{data}) \, \mathrm{d}\theta_0 \mathrm{d}\boldsymbol{\psi}.$$

Based on this last expression, we can estimate the posterior predictive density as follows. For each posterior sample $\boldsymbol{\psi}_r$, $r = 1, ..., B$ (where B is the posterior sample size), we can draw θ_{0r} from $p(\theta_0 \mid \boldsymbol{\theta}_r, \alpha_r, \beta_r)$, and then y_{0r} from $\pi_r u(y_0 - \mu_{1r}; \theta_{0r}) + (1 - \pi_r) u(y_0 - \mu_{2r}; \theta_{0r})$. A (smoothed) histogram of the posterior predictive sample, $\{y_{0r} : r = 1, ..., B\}$, yields the posterior predictive density estimate.

The posterior predictive density is a point estimate for the two-component mixture density, in particular, it can be shown that $p(y_0 \mid \text{data}) = \pi \mathrm{E}(f(y_0 - \mu_1; G) \mid \text{data}) + (1 - \pi) \mathrm{E}(f(y_0 - \mu_2; G) \mid \text{data})$. Therefore, for any $x \in \mathbb{R}$, a posterior point estimate for

the underlying symmetric density $f(x) \equiv f(x; G)$ in (3) is given by

$$\mathrm{E}(f(x; G) \mid \text{data}) = \int \int u(x; \theta_0) p(\theta_0 \mid \boldsymbol{\theta}, \alpha, \beta) p(\boldsymbol{\psi} \mid \text{data}) \, \mathrm{d}\theta_0 \mathrm{d}\boldsymbol{\psi},$$

and this can be obtained similarly to the posterior predictive density estimate, using the posterior samples for $(\boldsymbol{\theta}, \alpha, \beta)$ and the additional draws from $p(\theta_0 \mid \boldsymbol{\theta}, \alpha, \beta)$.

Finally, we note that the approaches to posterior inference discussed above utilize only the posterior expectation of the random mixing distribution G. Although not needed for the data illustrations considered in Section 3, the posterior distribution for G can be sampled by augmenting each realization from $p(\boldsymbol{\psi} \mid \text{data})$ with an additional draw from the conditional posterior distribution for G, given $(\boldsymbol{\theta}, \alpha, \beta)$. This distribution is given by a DP with precision parameter $\alpha+n$ and base distribution $(\alpha+n)^{-1} \left[\alpha G_0(\cdot; \beta) + \sum_{i=1}^{n} \delta_{\theta_i}(\cdot)\right]$ (Antoniak, 1974). The draw from the conditional posterior distribution of G can be obtained using a truncation approximation to the DP based on its stick-breaking definition given in (2) (e.g., Gelfand and Kottas, 2002; Kottas, 2006).

3 Data illustrations

We first study in Section 3.1 the performance of the semiparametric model with simulated data. Next, Section 3.2 illustrates posterior predictive inference under the model with data on eruptions of the Old Faithful geyser. Finally, in Section 3.3, we consider comparison with a parametric mixture model, using a data set on rainfall precipitation.

3.1 Simulation study

The simulation study involved four data sets. In the first three cases, the data were drawn from a mixture of two normal distributions with means of -1 and 2, and standard

deviations of 1. The probability associated with the $N(-1, 1)$ component was 0.15 for the first data set, 0.25 for the second, and 0.35 for the third. To study results under an underlying symmetric density with heavier tails, we also considered a data set generated from a mixture of two t_2 distributions (t distributions with 2 degrees of freedom) with location parameters -1 and 2, common scale parameter 1, and probability 0.15 for the first component. The sample size was $n = 250$ for all four simulated data sets.

Regarding the priors for the model parameters, for all four data sets, we used a $N(0, 10^2)$ prior for both μ_1 and μ_2, a uniform prior for π, and a gamma$(50, 0.1)$ prior for α (with mean 500). The shape parameter c of the base distribution G_0 was set to 2, and β was assigned an exponential prior with mean 2, based on a data range of 8 for the approach of Section 2.2.

We used the MCMC posterior simulation method of Section 2.3 to obtain inference for the model parameters and to estimate the symmetric density and the posterior predictive density. For all the parameter estimates, convergence was tested using a criterion from Raftery and Lewis (1995). Specifically, we computed $I = (M + N)/N_{min}$, where M is the number of burn-in iterations, N is the number of iterations after burn-in, and N_{min} is the size of a pilot sample. M, N, and N_{min} were all calculated using the `gibbsit` program (available at `http://lib.stat.cmu.edu/general/gibbsit`). Raftery and Lewis (1995) appeal to empirical evidence to suggest that values of I greater than 5 are problematic. In general, this diagnostic measure can be used to indicate problems in the chain due to a bad starting value, high posterior correlation, or "stickiness" in the chain. In no instance did any I value in the analysis of the simulated data sets exceed 2. Regarding autocorrelations in the sampled chains for the model parameters, μ_1 and μ_2 were the more challenging parameters requiring about 200 iterations for the autocorrelations to drop to small values. This was also the case for the real data discussed in Sections 3.2 and 3.3.

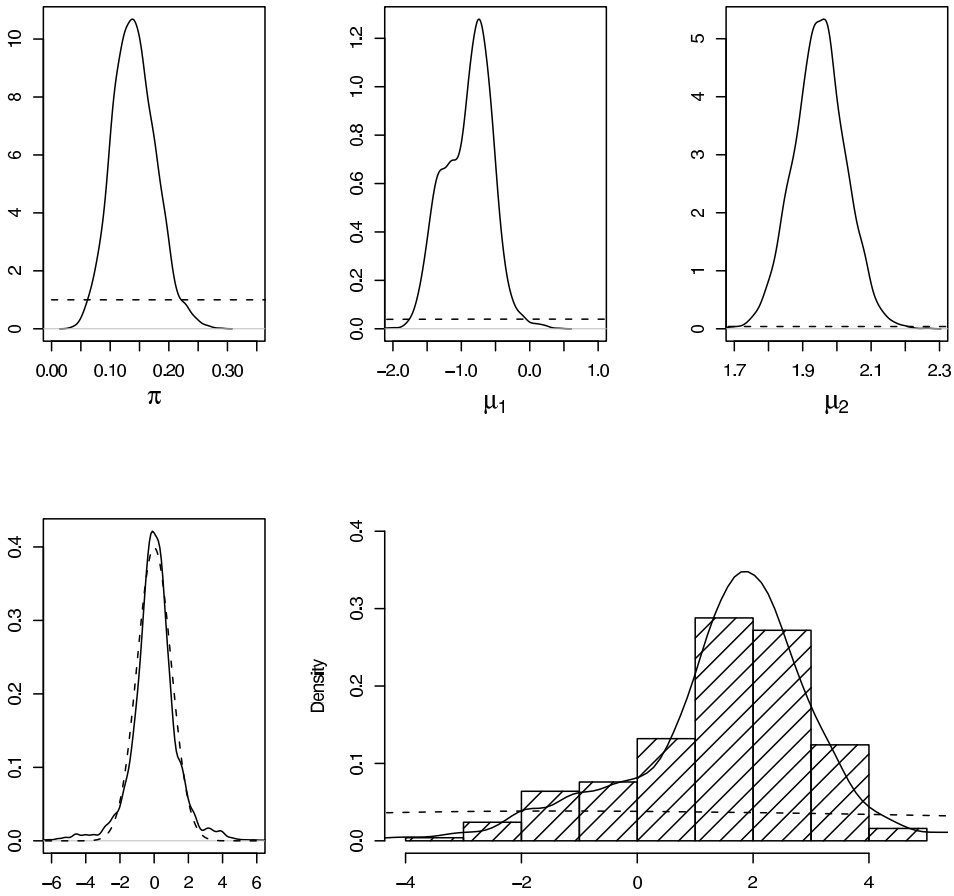

Figure 1: Simulated data from mixture of normals with true $\pi = 0.15$. The top row includes prior densities (dashed lines) and posterior densities (solid lines) for π, μ_1, and μ_2. The bottom left panel plots the symmetric density estimate (solid line) overlaid on the true underlying standard normal density (dashed line). The bottom right panel shows the prior and posterior predictive densities (dashed and solid lines, respectively) overlaid on the data histogram.

The results discussed below were based on 20000 burn-in iterations, followed by 600000 iterations taking output every 200^{th} to obtain the final posterior sample of size 3000.

Figures 1–4 show for each of the four simulated data sets the prior and posterior predictive density overlaid on the histogram of the raw data, as well as prior and posterior density plots for π, μ_1, and μ_2. Also plotted in each figure is the estimate of the underlying symmetric density overlaid on the corresponding density that was used to generate the data, that is, the $\mathrm{N}(0, 1)$ density for Figures 1–3 and the standard t_2 density for Figure 4.

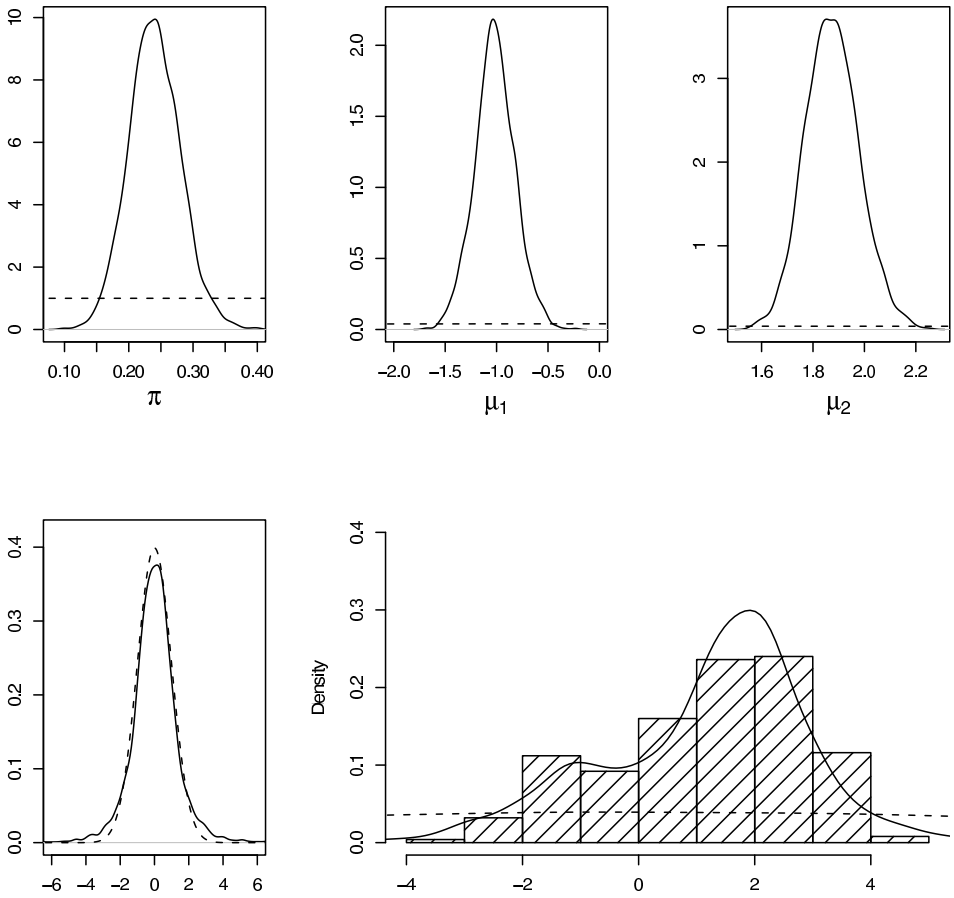

Figure 2: Simulated data from mixture of normals with true $\pi = 0.25$. The top row includes prior densities (dashed lines) and posterior densities (solid lines) for π, μ_1, and μ_2. The bottom left panel plots the symmetric density estimate (solid line) overlaid on the true underlying standard normal density (dashed line). The bottom right panel shows the prior and posterior predictive densities (dashed and solid lines, respectively) overlaid on the data histogram.

We note that in all simulation cases, the true values of μ_1, μ_2 and π are recovered successfully by their corresponding posteriors. In Figures 1–3, the posterior uncertainty for μ_1 and μ_2 changes with the true value of π, which should be anticipated, since the amount of information from the data for each mixture component depends on the mixing weight. In all four examples, the estimated posterior predictive density follows closely the shape of the data histogram, thus indicating that posterior predictive inference resulting from the model is accurate. Moreover, the symmetric density estimate recovers quite

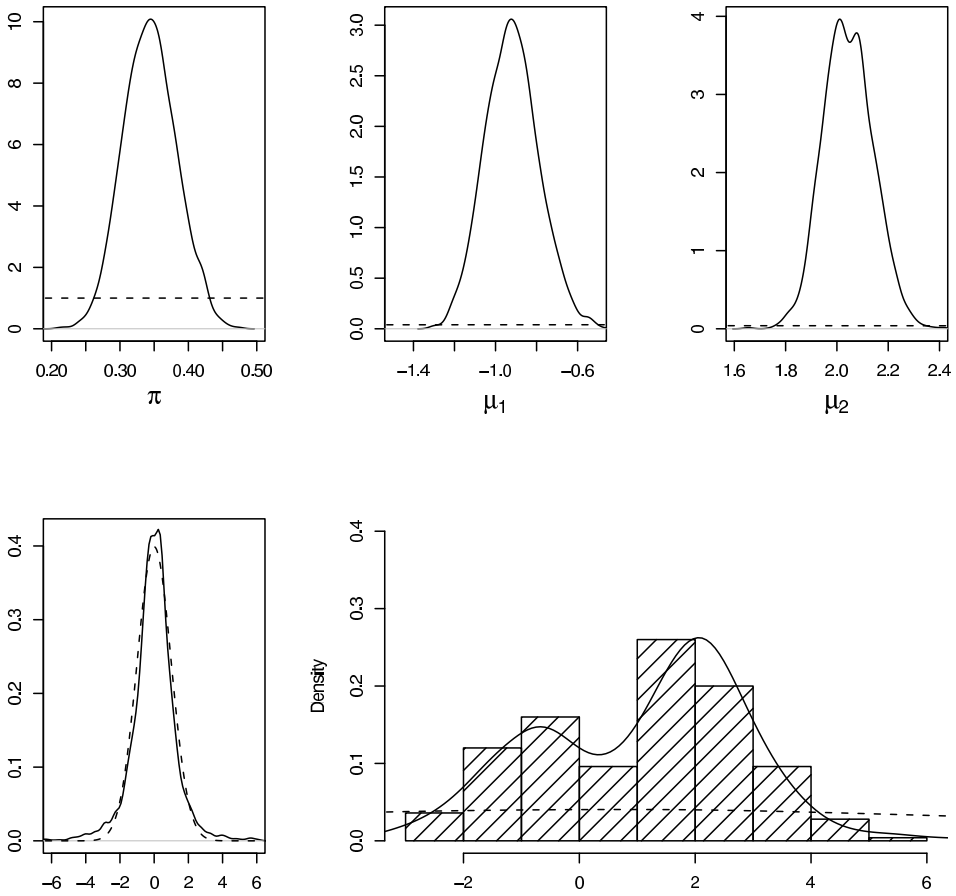

Figure 3: Simulated data from mixture of normals with true $\pi = 0.35$. The top row includes prior densities (dashed lines) and posterior densities (solid lines) for π, μ_1, and μ_2. The bottom left panel plots the symmetric density estimate (solid line) overlaid on the true underlying standard normal density (dashed line). The bottom right panel shows the prior and posterior predictive densities (dashed and solid lines, respectively) overlaid on the data histogram.

well the underlying symmetric density, especially, noting that in the bottom left panels of Figures 1–4 we are comparing the model posterior point estimate with the true standard normal or t_2 density from which a single data set was generated for each case of the simulation study. Finally, as is evident from Figures 1–4, the priors used for all simulated data sets were very uninformative. Hence, it is noteworthy that, at least based on this simulation experiment, the excellent performance of the model with regard to posterior inference and posterior predictive estimation does not rely on strong prior information,

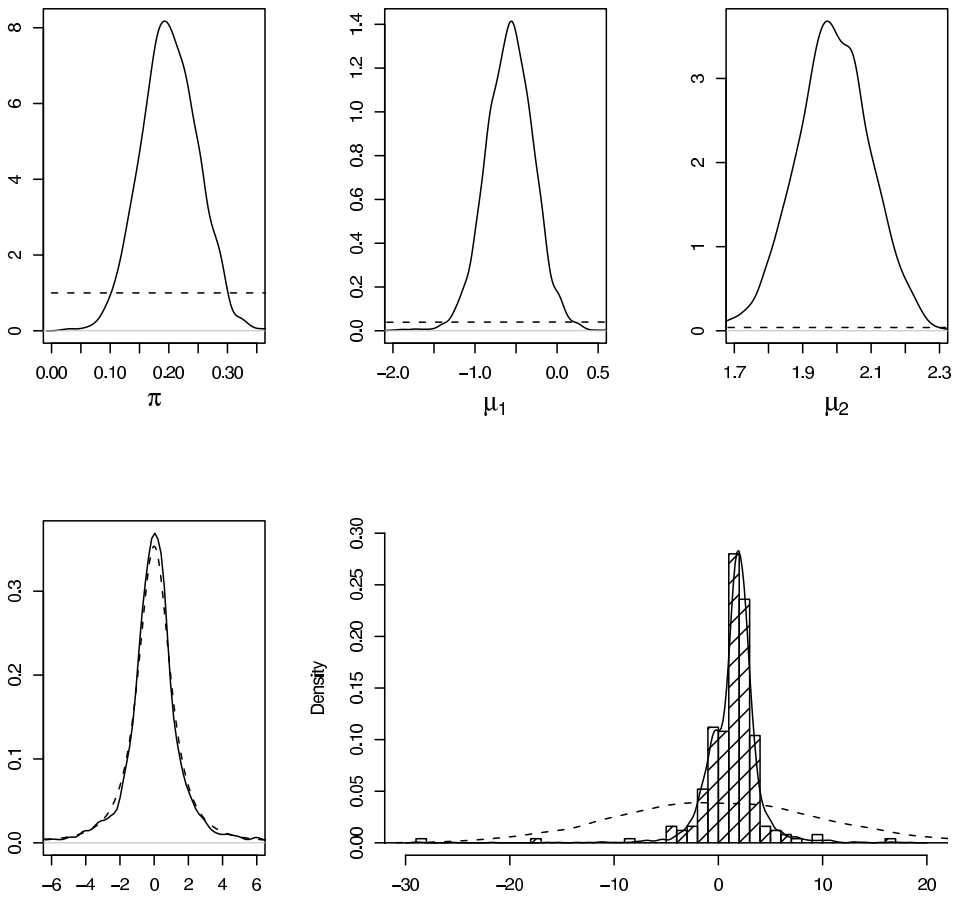

Figure 4: Simulated data from mixture of t_2 densities with true $\pi = 0.15$. The top row includes prior densities (dashed lines) and posterior densities (solid lines) for π, μ_1, and μ_2. The bottom left panel plots the symmetric density estimate (solid line) overlaid on the true underlying standard t_2 density (dashed line). The bottom right panel shows the prior and posterior predictive densities (dashed and solid lines, respectively) overlaid on the data histogram.

but is rather driven by the data.

3.2 Old Faithful eruptions data

We consider two data sets on eruptions of the Old Faithful geyser in Yellowstone National Park, USA (both included as part of the `datasets` R library). The first data set records the duration of the eruption in minutes, and the second the waiting time in minutes between eruptions. In both cases, the sample size is $n = 272$. Histograms of the eruption

time and waiting time data are plotted in Figures 5 and 6, respectively, indicating a bimodal shape for the underlying distributions.

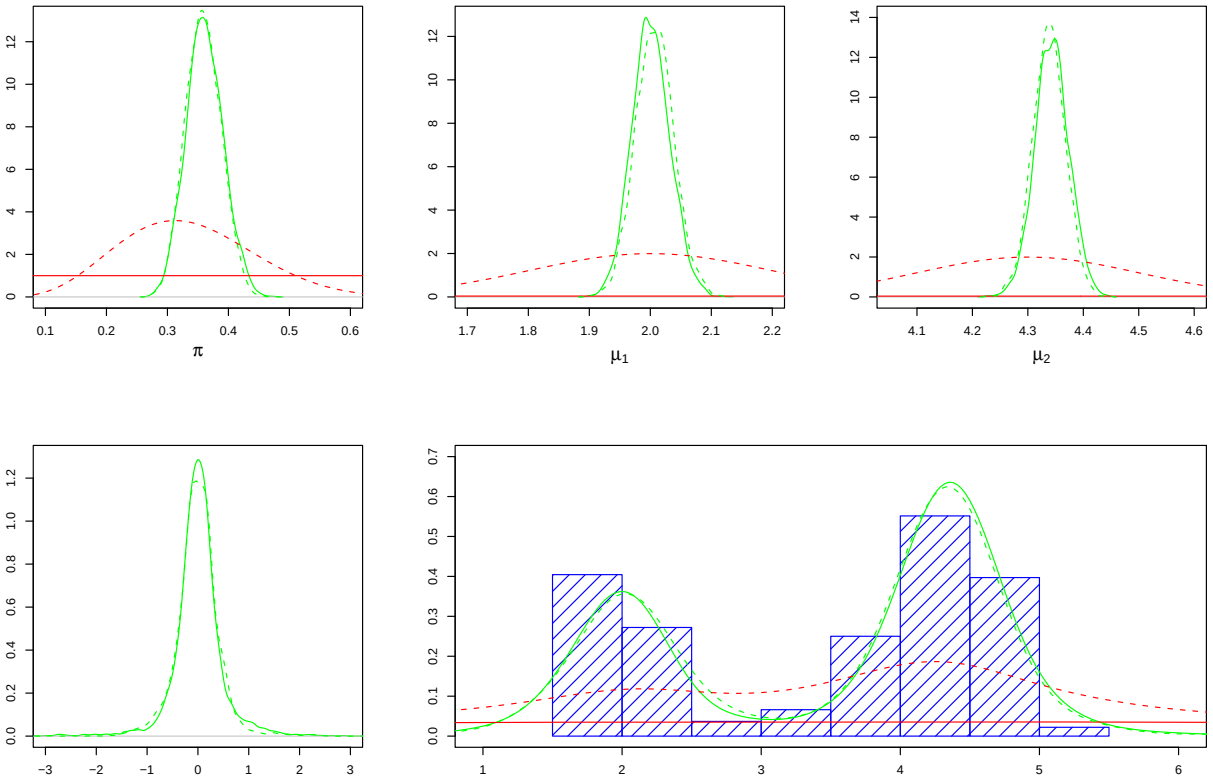

Figure 5: Old Faithful eruption time data. In all panels, the solid and dashed lines denote results under the diffuse and the more informative prior choice, respectively (see Section 3.2 for details). The top row includes prior densities (in red color) and posterior densities (in green) for π, μ_1, and μ_2. The bottom left panel plots the symmetric density estimates, and the bottom right panel shows the prior and posterior predictive densities (red and green lines, respectively) overlaid on the data histogram.

To further illustrate robustness of posterior inference to the prior choice, for both data sets, we obtained results under two distinct prior specifications, one inducing fairly diffuse priors, and one implying more informative priors. In all cases, the shape parameter c of the inverse gamma base distribution was set to 2, and thus the mean of G_0 is given by β. As in Section 3.1, we checked convergence of the MCMC algorithm using the criterion from Raftery and Lewis (1995). In no case did the I value exceed 4. All results reported below were based on a burn-in period of 20000 iterations, followed by 600000 iterations

Table 1: Old Faithful eruption time data. 95% posterior intevals for model parameters under the diffuse and the more informative prior choice discussed in Section 3.2.

Parameter	Diffuse Lower	Diffuse Upper	Informative Lower	Informative Upper
π	0.30	0.42	0.30	0.42
μ_1	1.94	2.06	1.95	2.07
μ_2	4.28	4.40	4.28	4.40
β	0.71	0.94	0.69	0.97
α	361	634	109	206

taking output every 200^{th} to obtain the 3000 posterior samples used for inference.

Focusing first on the the eruption time data, the diffuse set of prior specifications corresponds to a $\mathrm{N}(4, 10^2)$ prior for μ_1 and μ_2, a uniform prior for π, a gamma$(50, 0.1)$ prior for α (with mean 500), and an exponential prior for β with mean 10. For the more informative prior specification we used a $\mathrm{N}(2, 0.2^2)$ prior for μ_1, a $\mathrm{N}(4.3, 0.2^2)$ prior for μ_2, a beta$(6, 12)$ prior for π, a gamma$(40, 0.25)$ prior for α (with mean 160), and an exponential prior for β with mean 10. Figure 5 shows results for parameters π, μ_1 and μ_2, the underlying symmetric density, and the posterior predictive density, under both sets of priors. Moreover, Table 1 reports 95% posterior intervals for all model parameters. The robustness of the estimates for the symmetric density and the posterior predictive density, as well as of posterior inference results for the practically important model parameters π, μ_1 and μ_2 is particularly noteworthy. The prior sensitivity for the DP precision parameter α is typical with DP mixture models, since this DP prior hyperparameter is, in general, more difficult to inform from the data. However, this aspect of the DP mixture prior model does not affect substantially posterior predictive inference even with small sample sizes; in our example, note that the two posterior predictive densities (bottom right panel of Figure 5) are essentially indistinguishable.

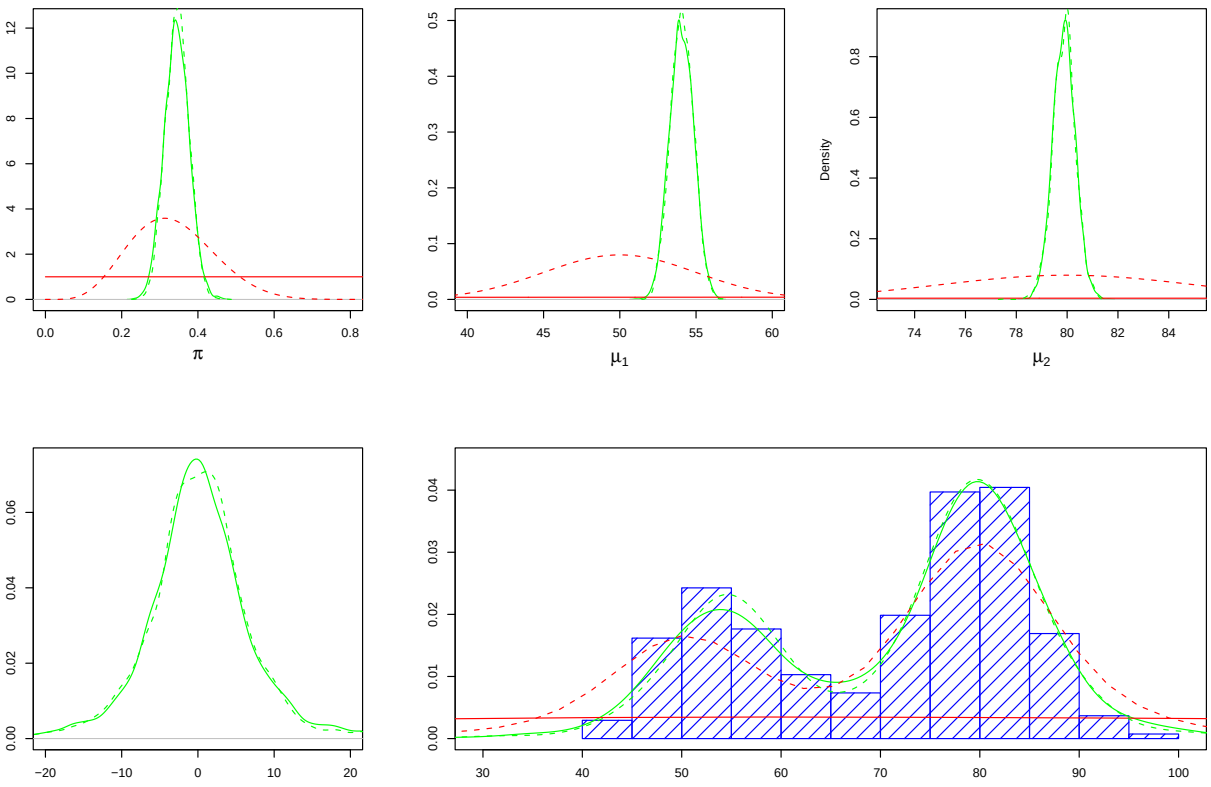

Figure 6: Old Faithful waiting time data. In all panels, the solid and dashed lines denote results under the diffuse and the more informative prior choice, respectively (see Section 3.2 for details). The top row includes prior densities (in red color) and posterior densities (in green) for π, μ_1, and μ_2. The bottom left panel plots the symmetric density estimates, and the bottom right panel shows the prior and posterior predictive densities (red and green lines, respectively) overlaid on the data histogram.

Turning to the waiting time data, the diffuse prior specification involved a $\mathrm{N}(70, 100^2)$ prior for μ_1 and μ_2, a uniform prior for π, a gamma$(50, 0.1)$ prior for α (with mean 500), and an exponential prior for β with mean 100. The more informative prior specification was based on a $\mathrm{N}(50, 5^2)$ prior for μ_1, a $\mathrm{N}(80, 5^2)$ prior for μ_2, a beta$(6, 12)$ prior for π, a gamma$(40, 0.25)$ prior for α (with mean 160), and an exponential prior for β with mean 14.3. Results under both prior choices for parameters π, μ_1 and μ_2, for the underlying symmetric density, and for the posterior predictive density are shown in Figure 6. Table 2 includes 95% posterior intervals for all model parameters, again, under both sets of priors. As with the eruption time data, posterior inference for all model parameters (other than

Table 2: Old Faithful waiting time data. 95% posterior intevals for model parameters under the diffuse and the more informative prior choice discussed in Section 3.2.

Parameter	Diffuse Lower	Diffuse Upper	Informative Lower	Informative Upper
π	0.28	0.41	0.29	0.40
μ_1	52.6	55.6	52.6	55.5
μ_2	79.0	80.7	79.0	80.7
β	12.2	16.5	11.7	16.6
α	358	634	104	203

precision parameter α) was robust to the very different prior specifications. This is also the case for the symmetric density estimates, and for the posterior predictive density estimates (bottom right panel of Figure 6) that capture well the bimodal density shape suggested by the data for the waiting time distribution.

3.3 Rainfall precipitation data

Here, we study the performance of the semiparametric DP mixture model with one more standard data set (available from the `reldist` R library), which records the average amount of rainfall precipitation in inches for each of $n = 70$ United States (and Puerto Rico) cities. The histogram of the data (included in Figure 8) suggests bimodality, albeit with components that are not as well separated as the ones for the Old Faithful data of Section 3.2.

The main objective with this example is to draw comparison with a parametric mixture model. To this end, we consider the two-component mixture of normals

$$\pi \mathrm{N}(y; \mu_1, \sigma^2) + (1 - \pi)\mathrm{N}(y; \mu_2, \sigma^2) \tag{8}$$

which is a special case of (1) with parametrically specified symmetric density $f(y)$ given by a $\mathrm{N}(y; 0, \sigma^2)$ density.

We center the comparison around label switching, a key challenge for Bayesian analysis of mixture models (e.g., Celeux, Hurn and Robert, 2000; Stephens, 2000; Jasra, Holmes and Stephens, 2005; Frühwirth-Schnatter, 2006). Label switching arises because the likelihood under a finite mixture distribution is invariant under permutations of the mixture model parameter vector. Hence, unless the prior distribution includes information that distinguishes between mixture components, the posterior distribution will be invariant to permutations of the labeling of the parameters. A problematic implication of label switching is that, for data sets that correspond to weakly separated mixture components, the standard MCMC algorithms will encounter the symmetry of the posterior distribution resulting in switching of the labels for the component specific parameters. But then, for instance, for a data set supporting a bimodal distribution, the posterior densities of μ_1 and μ_2 under model (8) will be bimodal and identical. Although this will not affect the posterior predictive density for the mixture distribution, it renders inference for the individual mixture components impractical.

Here, we compare inference under the DP mixture model and the normal mixture model in (8) without the use of any particular method to avoid label switching. Hence, we implemented posterior inference under the normal mixture model using the standard Gibbs sampling approach based on data augmentation with latent binary mixing parameters. We used a relatively informative inverse gamma prior for σ^2 with shape parameter 2 and mean 50 (results were similar under both less and more dispersed priors for σ^2). For the DP mixture model, we used a gamma$(50, 0.1)$ prior for α (with mean 500), and an exponential prior for β with mean 10 (as before, the shape parameter c of the base distribution G_0 was set to 2). For both models, we used a uniform prior for π, and three

different specifications for the independent normal priors for μ_1 and μ_2: a $N(10, 10^2)$ prior for μ_1 and a $N(50, 10^2)$ prior for μ_2 (Prior A); a $N(10, 30^2)$ prior for μ_1 and a $N(50, 30^2)$ prior for μ_2 (Prior B); and a $N(35, 30^2)$ prior for both μ_1 and μ_2 (Prior C).

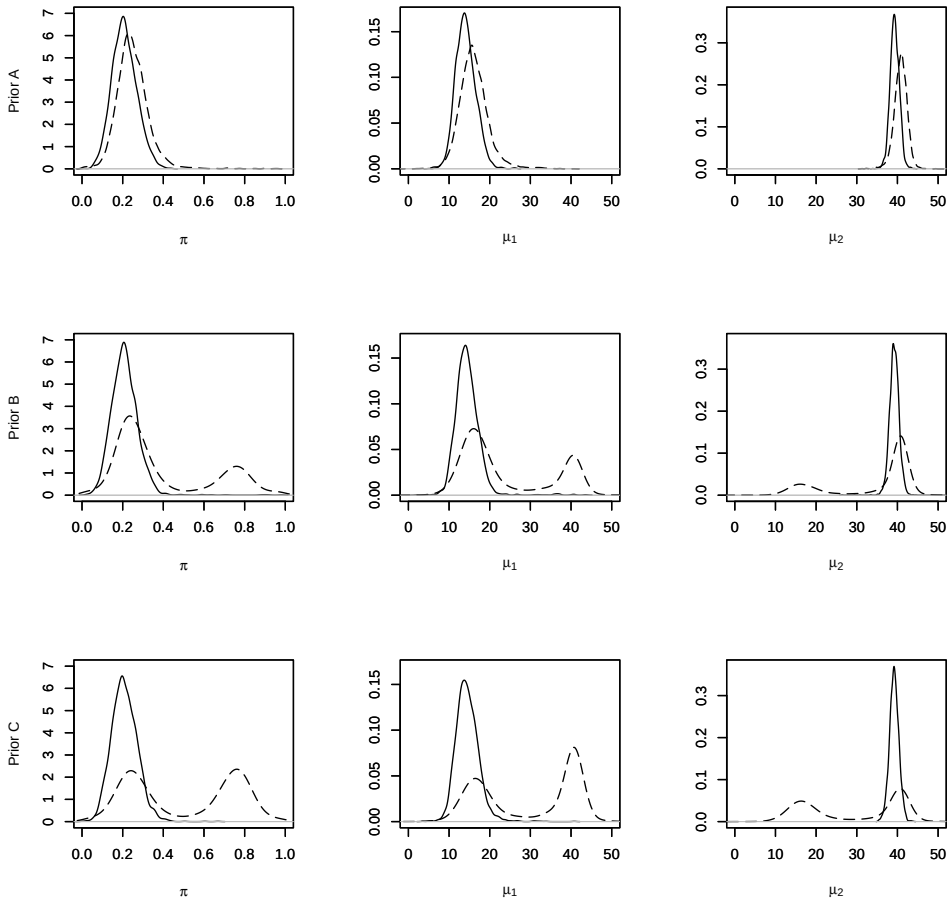

Figure 7: Rainfall precipitation data. Posterior densities for π, μ_1, and μ_2 under the DP mixture model (solid lines) and the normal mixture model (dashed lines). Each row corresponds to one of the three prior choices discussed in Section 3.3.

Under each of the two models, Figure 7 plots the posterior densities for μ_1, μ_2 and π for prior choices A, B and C. The capacity of the parametric mixture model to identify the two components depends on the prior choice for μ_1 and μ_2. The normal mixture successfully distinguishes μ_1 and μ_2 in their posterior distributions under informative

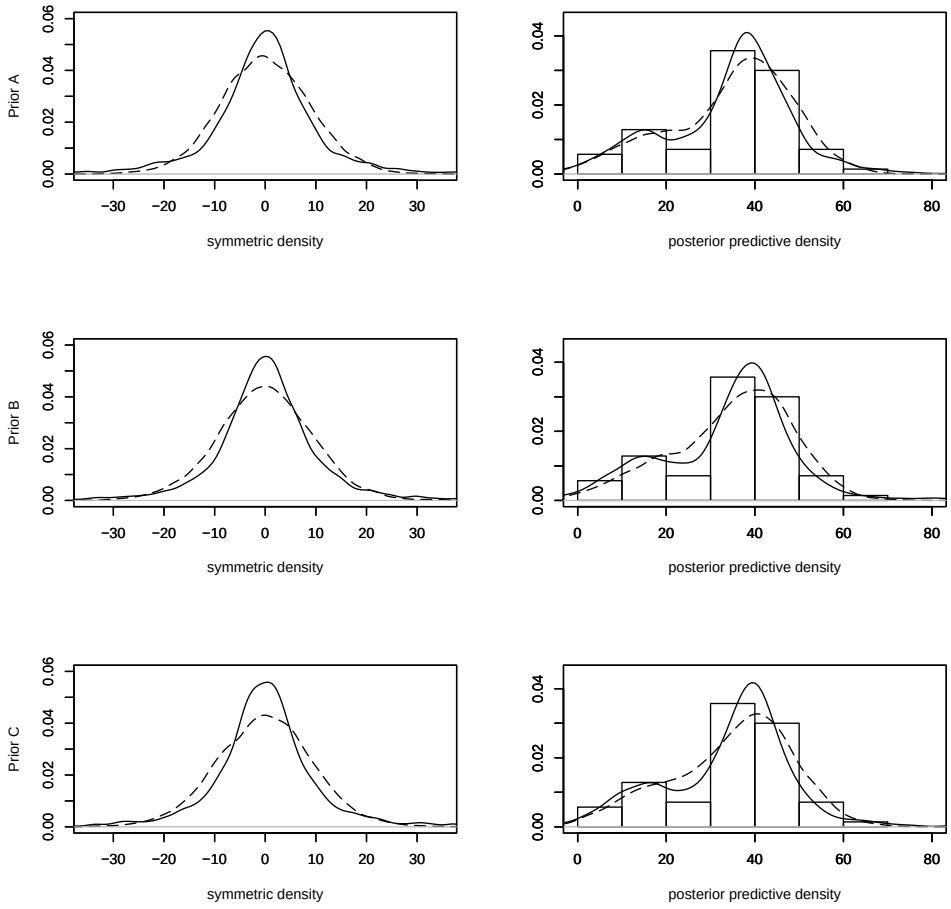

Figure 8: Rainfall precipitation data. Estimates for the symmetric density and the posterior predictive density (the latter overlaid on the data histogram) under the DP mixture model and the normal mixture model, denoted by the solid and dashed lines, respectively. Each row corresponds to one of the three prior choices discussed in Section 3.3.

prior A, but yields bimodal posteriors under the other priors. We note that assigning distinct informative prior means to μ_1 and μ_2, as in prior B, does not suffice to identify the parameters (in fact, prior B results in unbalanced label switching). The results for prior C correspond to typical label switching where the posterior densities for μ_1 and μ_2 are practically identical; this was also the case under a common $\mathrm{N}(35, 10^2)$ prior for μ_1 and μ_2 (results not shown). In contrast, the DP mixture model yields robust inference for μ_1, μ_2 and π under the three priors, with posterior densities that seem plausible for the

rainfall data. Prior C was the most challenging for the semiparametric model requiring careful tuning of the variances for the Metropolis-Hastings proposals for μ_1 and μ_2 in the MCMC algorithm. Increasing further the prior variance for μ_1 and μ_2 results in label switching under the DP mixture model.

The posterior predictive densities under both models capture the general shape of the rainfall data histogram (see Figure 8). However, under all three priors, the DP mixture model is more successful in capturing the first mode. This can be attributed to the shape of the underlying symmetric density; as shown on the left panels of Figure 8, the tail behavior of the symmetric density, as estimated by the DP mixture model, is evidently non-Gaussian.

We note that, when applied to the Old Faithful data of Section 3.2, the normal mixture model in (8) yields similar results with the DP mixture model, also without facing issues with label switching. Hence, the superior performance of the semiparametric mixture model with the rainfall data is likely due to a combination of the more challenging mixture structure for these data and the shape of the underlying symmetric density. Regardless, the ability of the DP mixture model to avoid label switching even under fairly uninformative prior specifications is encouraging with respect to its potential for application to mixture deconvolution problems.

4 Discussion

We have developed a Bayesian semiparametric modeling approach for mixtures of symmetric unimodal densities on the real line. The mixture is based on a common symmetric density, which defines all mixture components through distinct location parameters. This structure ensures identifiability of mixture components rendering the model an appealing choice for mixture deconvolution problems. We have argued for the utility of a non-

parametric prior model for the symmetric density that defines the structured mixture. The prior probability model is based on scale uniform Dirichlet process mixtures and it supports all unimodal symmetric (about 0) densities on the real line. Compared with existing estimation methods for the model, a distinguishing feature of our work is that it is based on a fully inferential probabilistic modeling framework. Moreover, the additional assumption of unimodality for the underlying symmetric density can be an asset for certain classes of mixture deconvolution problems. Finally, as illustrated with the rainfall precipitation data of Section 3.3, a promising feature of the semiparametric mixture model is that it is more robust to label switching than standard parametric mixture models.

For simpler exposition of the modeling approach, we opted to focus on the two-component mixture setting. However, it is straightforward to extend the Bayesian semiparametric modeling framework to mixtures of the form in (1) with more than two components, i.e., $\sum_{j=1}^{k} \pi_j f(y-\mu_j)$, $y \in \mathbb{R}$, for fixed $k \geq 2$, with $\pi_j \geq 0$ such that $\sum_{j=1}^{k} \pi_j = 1$, and with $f(\cdot)$ a density on $\mathbb{R}$, which is unimodal and symmetric about 0. This is because the nonparametric component of the model, i.e., the Dirichlet process mixture prior for $f(\cdot)$, remains the same. Moreover, the structure of the MCMC posterior simulation method would be similar, now requiring updates for the additional location parameters and mixture weights. Current work studies the practical utility of the three-component extension of (1) for mixture deconvolution problems from epidemiological research.

Acknowledgements

The authors wish to thank two reviewers and an Associate Editor for useful comments that led to an improved presentation of the material in the paper. The work of the first author was supported in part by the National Science Foundation under award DEB 0727543.

References

Antoniak, C.E. (1974), "Mixtures of Dirichlet Processes With Applications to Nonparametric Problems," *The Annals of Statistics*, 2, 1152-1174.

Blackwell, D., and MacQueen, J.B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353-355.

Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007), "A stochastic EM algorithm for a semiparametric mixture model," *Computational Statistics & Data Analysis*, 51, 5429-5443.

Bordes, L., Mottelet, S., and Vandekerkhove, P. (2006), "Semiparametric estimation of a two-component mixture model," *The Annals of Statistics*, 34, 1204-1232.

Brunner, L.J. (1992), "Bayesian nonparametric methods for data from a unimodal density," *Statistics and Probability Letters*, 14, 195-199.

Brunner, L.J. (1995), "Bayesian Linear Regression With Error Terms That Have Symmetric Unimodal Densities," *Journal of Nonparametric Statistics*, 4, 335-348.

Brunner, L.J., and Lo, A.Y. (1989), "Bayes methods for a symmetric unimodal density and its mode," *Annals of Statistics*, 17, 1550-1566.

Bush, C.A., and MacEachern, S.N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275-285.

Celeux, G., Hurn, M., and Robert, C.P. (2000), "Computational and inferential difficulties with mixture posterior distributions," *Journal of the American Statistical Association*, 95, 957-970.

Cruz-Medina, I.R., and Hettmansperger, T.P. (2004), "Nonparametric estimation in semi-parametric univariate mixture models," *Journal of Statistical Computation and Simulation*, 74, 513-524.

Escobar, M.D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577-588.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications, Vol II*, Second edition. New York: Wiley.

Ferguson, T.S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209-230.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer.

Gelfand, A.E., and Kottas, A. (2002), "A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 11, 289-305.

Hansen, M.B., and Lauritzen, S.L. (2002), "Nonparametric Bayes inference for concave distribution functions," *Statistica Neerlandica*, 56, 110-127.

Hunter, D.R., Wang, S., and Hettmansperger, T.P. (2007), "Inference for mixtures of symmetric distributions," *The Annals of Statistics*, 35, 224-251.

Jasra, A., Holmes, C.C., and Stephens, D.A. (2005), "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, 20, 50-67.

Kottas, A. (2006), "Nonparametric Bayesian survival analysis using mixtures of Weibull distributions," *Journal of Statistical Planning and Inference*, 136, 578-596.

Kottas, A., and Gelfand, A.E. (2001), "Bayesian semiparametric median regression modeling," *Journal of the American Statistical Association*, 96, 1458-1468.

Kottas, A., and Krnjajić, M. (2009), "Bayesian semiparametric modelling in quantile regression," *Scandinavian Journal of Statistics*, 36, 297-319.

Lavine, M., and Mockus, A. (1995), "A nonparametric Bayes method for isotonic regression," *Journal of Statistical Planning and Inference*, 46, 235-248.

Marin, J.M., Mengersen, K., and Robert, C.P. (2005), "Bayesian modelling and inference on mixtures of distributions," in *Bayesian Thinking: Modeling and Computation (Handbook of Statistics, vol. 25)*, Dey D.K. and Rao C.R. (eds). Amsterdam: Elsevier, pp. 459-508.

Müller, P., and Quintana, F.A. (2004), "Nonparametric Bayesian data analysis," *Statistical Science*, 19, 95-110.

Neal, R.M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9, 249-265.

Raftery, A.E., and Lewis, S.M. (1995), "Implementing MCMC." In *Markov Chain Monte Carlo in Practice* (Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. eds.), 115-130. Chapman & Hall, London.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639-650.

Stephens, M. (2000), "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society, Series B*, 62, 795-809.

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

Appendix: Posterior simulation method

Here, we present the MCMC posterior simulation method for the model of Section 2.1. Again, let $\boldsymbol{\psi}$ denote the full parameter vector comprising $\boldsymbol{z} = (z_1, ..., z_n)$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, π, μ_1, μ_2, α, and β. To sample from the posterior of $\boldsymbol{\psi}$, it is possible to use the standard Gibbs sampler (e.g., Escobar and West, 1995), which samples directly from the posterior full conditionals for the θ_i, or, in our case, the joint full conditional for (z_i, θ_i). However, the following method based on Metropolis-Hastings steps (in the spirit of the posterior simulation algorithms for DP mixtures considered in Neal, 2000) results in an easier to implement algorithm, which has proved to be sufficiently efficient in terms of mixing for all the data sets we considered.

First, note that the MCMC algorithm starting values for μ_1, μ_2 and the θ_i need to be chosen taking into account the restrictions of the uniform kernel of the DP mixture.

For each $i = 1, ..., n$, we update the DP mixing parameter, θ_i, and the mixing parameter for the two-component mixture, z_i, as a pair with a Metropolis-Hastings step. For each $i = 1, ..., n$, the joint posterior full conditional is given by

$$p(z_i, \theta_i \mid \mu_1, \mu_2, \{\theta_\ell : \ell \neq i\}, \pi, \alpha, \beta, \text{data}) \propto u(y_i - \mu_{z_i}; \theta_i) p(\theta_i \mid \{\theta_\ell : \ell \neq i\}, \alpha, \beta) \Pr(z_i \mid \pi)$$

where

$$p(\theta_i \mid \{\theta_\ell : \ell \neq i\}, \alpha, \beta) = \frac{\alpha}{\alpha + n - 1} g_0(\theta_i; \beta) + \frac{1}{\alpha + n - 1} \sum_{\ell \neq i} \delta_{\theta_\ell}(\theta_i)$$

is the prior full conditional for θ_i arising from (7). The Metropolis-Hastings update details are as follows.

- Let $(z_i^{(\text{old})}, \theta_i^{(\text{old})})$ be the current state of the chain. Repeat the following update R times ($R \geq 1$).

- Draw a candidate $(\tilde{z}_i, \tilde{\theta}_i)$ from the proposal distribution, which is given by the product of the prior full conditionals for z_i and θ_i. Hence, $\tilde{z}_i$ and $\tilde{\theta}_i$ are drawn independently, where $\tilde{z}_i = 1$ with probability π, and $\tilde{\theta}_i \sim p(\theta_i \mid \{\theta_\ell : \ell \neq i\}, \alpha, \beta)$.

- Set $(z_i, \theta_i) = (\tilde{z}_i, \tilde{\theta}_i)$ with probability $p = \min\left\{1, u(y_i - \mu_{\tilde{z}_i}; \tilde{\theta}_i)/u(y_i - \mu_{z_i^{(\text{old})}}; \theta_i^{(\text{old})})\right\}$, and $(z_i, \theta_i) = (z_i^{(\text{old})}, \theta_i^{(\text{old})})$ with probability $1 - p$.

Once all the θ_i are updated, we can compute: n^*, the number of distinct elements of vector $(\theta_1, ..., \theta_n)$; θ_j^*, $j = 1, ..., n^*$, the realizations of the distinct θ_i; the vector of configuration indicators $\boldsymbol{s} = (s_1, ..., s_n)$ such that $s_i = j$ if and only if $\theta_i = \theta_j^*$; and $n_j = |\{i : s_i = j\}|$, the size of the j-th distinct component. These posterior realizations are used in the updates for α and β as well as in sampling from the posterior predictive distribution. Moreover, given the currently imputed n^* and vector $\boldsymbol{s}$, we can re-sample the values for the distinct θ_j^* to improve mixing of the MCMC algorithm (Bush and MacEachern, 1996). Specifically, for each $j = 1, ..., n^*$, the posterior full conditional for θ_j^* is given by

$$\begin{aligned} p(\theta_j^* \mid \mu_1, \mu_2, \boldsymbol{z}, \boldsymbol{s}, \beta, \text{data}) &\propto g_0(\theta_j^*; \beta) \prod_{\{i:s_i=j\}} u(y_i - \mu_{z_i}; \theta_j^*) \\ &\propto \theta_j^{*-(c+n_j+1)} \exp(-\beta/\theta_j^*) 1(\theta_j^* > \max_{\{i:s_i=j\}} |y_i - \mu_{z_i}|), \end{aligned}$$

which is therefore a truncated inverse gamma distribution with shape parameter $c + n_j$ and scale parameter β, with the constraint over the interval $(\max_{\{i:s_i=j\}} |y_i - \mu_{z_i}|, \infty)$. After drawing from the full conditionals for all the θ_j^*, we update the values for the θ_i using their definition through the (re-sampled) θ_j^* and the vector $\boldsymbol{s}$.

With the z_i, $i = 1, ..., n$, updated, we obtain $m_\ell = |\{i : z_i = \ell\}|$, $\ell = 1, 2$ (with $m_1 + m_2 = n$). Then, $p(\pi \mid \boldsymbol{z}, \text{data}) \propto \pi^{a_\pi - 1}(1-\pi)^{b_\pi - 1}\pi^{m_1}(1-\pi)^{m_2}$, and thus the posterior full conditional for π is beta$(a_\pi + m_1, b_\pi + m_2)$.

The posterior full conditional for μ_1 is given by

$$\begin{aligned} p(\mu_1 \mid \boldsymbol{z}, \boldsymbol{\theta}, \text{data}) &\propto p(\mu_1) \prod_{\{i:z_i=1\}} u(y_i - \mu_1; \theta_i) \propto p(\mu_1) 1(\textstyle\bigcap_{\{i:z_i=1\}}(y_i - \theta_i < \mu_1 < y_i + \theta_i)) \\ &\propto p(\mu_1) 1(\max_{\{i:z_i=1\}}(y_i - \theta_i) < \mu_1 < \min_{\{i:z_i=1\}}(y_i + \theta_i)). \end{aligned}$$

Hence, with the $\mathrm{N}(a_1, b_1^2)$ prior for μ_1, the posterior full conditional is given by a $\mathrm{N}(a_1, b_1^2)$ distribution truncated over the interval $\left(\max_{\{i:z_i=1\}}(y_i - \theta_i), \min_{\{i:z_i=1\}}(y_i + \theta_i)\right)$. Similarly, the posterior full conditional for μ_2 is a $\mathrm{N}(a_2, b_2^2)$ distribution truncated over the interval $\left(\max_{\{i:z_i=2\}}(y_i - \theta_i), \min_{\{i:z_i=2\}}(y_i + \theta_i)\right)$. Therefore, μ_1 and μ_2 could be updated with Gibbs steps. A more numerically stable alternative involves the following Metropolis-Hastings updates for each μ_k, $k = 1, 2$:

- Let $\mu_k^{(\text{old})}$ be the current state of the chain.
- Draw a candidate $\tilde{\mu}_k$ from a normal proposal distribution with mean $\mu_k^{(\text{old})}$ and variance that is tuned to obtain appropriate acceptance rates. (For all the data sets we considered, acceptance rates for the μ_k, $k = 1, 2$, were between 20% and 40%.)
- Set $\mu_k = \tilde{\mu}_k$ with probability $q = \min\left\{1, p(\tilde{\mu}_k)1(\tilde{\mu}_k \in \mathcal{A}_k)/(p(\mu_k^{(\text{old})})1(\mu_k^{(\text{old})} \in \mathcal{A}_k))\right\}$, where $\mathcal{A}_k = \left(\max_{\{i:z_i=k\}}(y_i - \theta_i), \min_{\{i:z_i=k\}}(y_i + \theta_i)\right)$, and $\mu_k = \mu_k^{(\text{old})}$ with probability $1-q$.

Finally, regarding the DP prior hyperparameters, we use the augmentation technique of Escobar and West (1995) to update α. Moreover, the posterior full conditional for β is given by $p(\beta \mid \boldsymbol{\theta}, \text{data}) \propto p(\beta) \prod_{j=1}^{n^*} g_0(\theta_j^*; \beta)$ resulting in a gamma distribution with shape parameter $cn^* + 1$ and rate parameter $b_\beta + \sum_{j=1}^{n^*} \theta_j^{*-1}$.