Non-informative reparameterisations for location-scale mixtures*

Kaniav Kamary
Université Paris-Dauphine, CEREMADE

Kate Lee
Auckland University of Technology, New Zealand

Christian P. Robert
Université Paris-Dauphine, CEREMADE, Dept. of Statistics, University of Warwick, and CREST, Paris

Abstract. While mixtures of Gaussian distributions have been studied for more than a century (Pearson, 1894), the construction of a reference Bayesian analysis of those models still remains unsolved, with a general prohibition of the usage of improper priors (Frühwirth-Schnatter, 2006) due to the ill-posed nature of such statistical objects. This difficulty is usually bypassed by an empirical Bayes resolution (Richardson and Green, 1997). By creating a new parameterisation cantered on the mean and variance of the mixture distribution itself, we are able to develop here a genuine non-informative prior for Gaussian mixtures with an arbitrary number of components. We demonstrate that the posterior distribution associated with this prior is almost surely proper and provide MCMC implementations that exhibit the expected exchangeability. While we only study here the Gaussian case, extension to other classes of location-scale mixtures is straightforward.

Key words and phrases: Noninformative prior, improper prior, Mixture of distributions, Bayesian analysis, Dirichlet prior, exchangeability, plane-sphere intersection, polar coordinates.

1. INTRODUCTION

A mixture density is traditionally represented as a weighted average of densities from standard families, i.e.,

$$f(x|\boldsymbol{\theta}, \mathbf{p}) = \sum_{i=1}^{k} p_i f(x|\theta_i) \qquad \sum_{i=1}^{k} p_i = 1\,. \tag{1}$$

Each component of the mixture is characterised by a component-wise parameter θ_i and the weights p_i of those components translate the importance of each of those components in the model.

*Kaniav Kamary, CEREMADE, Université Paris-Dauphine, 75775 Paris cedex 16, France, kamary@ceremade.dauphine.fr, Kate Lee, Auckland University of Technology, New Zealand, jeong.lee@aut.ac.nz, Christian P. Robert, CEREMADE, Université Paris-Dauphine, 75775 Paris cedex 16 France xian@ceremade.dauphine.fr. Research partly supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2012–2015 grant ANR-11-BS01-0010 "Calibration" and by a 2010–2015 senior chair grant of Institut Universitaire de France. The authors are grateful to Robert Kohn for his helpful comments.

This particular representation gives a separate meaning to each component through its parameter θ_i, even though there is a well-known lack of identifiability in such models, due to the invariance of the sum by permutation of the indices. This issue relates to the equally well-known "label switching" phenomenon in the Bayesian approach to the model, which pertains both to inference and to simulation of the corresponding posterior (Celeux et al., 2000; Stephens, 2000; Frühwirth-Schnatter, 2001; Frühwirth-Schnatter, 2004; Jasra et al., 2005). From this Bayesian viewpoint, the choice of the prior distribution on the component parameters is quite open, the only constraint being that the corresponding posterior is proper (Diebolt and Robert, 1994; Frühwirth-Schnatter, 2004). Diebolt and Robert (1994) and Wasserman (1999) discussed the alternative approach of *imposing* proper posteriors on improper priors by banning almost empty components from the likelihood function. While consistent, this approach induces dependence between the observations, higher computational costs and is not handling overfitting very well. It has therefore seen little following.

The prior distribution on the weights p_i is equally open for choice, but a standard version is a Dirichlet distribution with common hyperparameter a, $\text{Dir}(a, \ldots, a)$. Recently, Rousseau and Mengersen (2011) demonstrated that the choice of this hyperparameter a relates to the inference on the total number of components, namely that a small enough value of a manages to handle over-fitted mixtures in a convergent manner. In a Bayesian non-parametric modelling, Griffin (2010) showed that the prior on the weights may have a higher impact when inferring about the number of components, relative to the prior on the component-specific parameters. As indicated above, the prior distribution on the θ_i's has received less attention and conjugate choices are most standard, since they facilitate simulation via Gibbs samplers (Diebolt and Robert, 1990; Escobar and West, 1995; Richardson and Green, 1997) if not estimation, since posterior moments remain unavailable in closed form. In addition, Richardson and Green (1997) among others proposed data-based priors that derive some hyperparameters as functions of the data, towards an automatic scaling of such priors. An R package, bayesm (Rossi and McCulloch, 2010) incorporates some of those ideas. In the case when $\theta_i = (\mu_i, \sigma_i)$ is a location-scale parameter, Mengersen and Robert (1996) proposed a reparameterisation of (1) that express each component as a local perturbation of the previous one, namely $(i > 1)$

$$\mu_i = \mu_{i-1} + \sigma_{i-1}\delta_i\,, \quad \sigma_i = \tau_i\sigma_{i-1}\,,\ \tau_i < 1\,,$$

with μ_1 and σ_1 being the reference values. Based on this reparameterisation, Robert and Titterington (1998) established that a particular improper prior on (μ_1, σ_1) still leads to a proper prior. We propose here to modify further this reparameterisation towards using the global mean and global variance of the mixture distribution as reference location and scale, respectively. This modification has foundational consequences in terms of using improper and non-informative priors over mixtures, in sharp contrast with the existing literature (see, e.g. Diebolt and Robert, 1993, 1994; O'Hagan, 1994; Wasserman, 1999).

Bayesian computing for mixtures covers a wide variety of proposals, starting with the introduction of the Gibbs sampler (Diebolt and Robert, 1990; Gelman and King, 1990; Escobar and West, 1995), some concerned with approximations (Roeder, 1990; Wasserman, 1999) and MCMC features (Richardson and Green, 1997; Celeux et al., 2000; Casella et al., 2002), and others with asymptotic justifications, in particular when over-fitting mixtures (Rousseau and Mengersen, 2011; Kamary et al., 2014), but most attempting to overcome the methodological hurdles in estimating mixture model (Chib, 1995; Neal, 1999; Berkhof et al., 2003; Marin et al., 2005; Frühwirth-Schnatter, 2006; Lee et al., 2009; Mengersen et al., 2011).

In this paper, we introduce and study the global mean-variance reparameterisation (Section 2), which main consequence is to constrain all other parameters to a compact space. We study several possible parameterisations of that kind and demonstrate that the improper Jeffreys-like prior associated with them is proper. In Section 3, we propose some MCMC implementation to estimate the parameters of the mixture, discussing label switching (Section 3.2) and its

resolution by tempering. Extensions to non-Gaussian mixtures are briefly discussed in Section 6.

2. MIXTURE REPRESENTATION

2.1 Mean-variance reparameterisation

Let us first recall how both mean and variance of a mixture distribution can be represented in terms of the mean and variance parameters of the component of the mixture:

Lemma 1 *If μ_i and σ_i^2 denote the mean and variance of the distribution with density $f(\cdot|\theta_i)$, respectively, the mean of the mixture distribution (1) is given by*

$$\mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X] = \sum_{i=1}^{k} p_i \mu_i$$

and its variance by

$$\text{var}_{\boldsymbol{\theta},\mathbf{p}}(X) = \sum_{i=1}^{k} p_i \sigma_i^2 + \sum_{i=1}^{k} p_i (\mu_i^2 - \mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X]^2)$$

Proof: The population mean given by

$$\mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X] = \sum_{i=1}^{k} p_i \mathbb{E}_{f(\cdot|\theta_i)}[X] = \sum_{i=1}^{k} p_i \mu_i$$

where $\mathbb{E}_{f(\cdot|\theta_i)}[X]$ is the expected value component i. Similarly, the population variance is given by

$$\text{var}_{\boldsymbol{\theta},\mathbf{p}}(X) = \sum_{i=1}^{k} p_i \mathbb{E}_{f(\cdot|\theta_i)}[X^2] - \mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X]^2 = \sum_{i=1}^{k} p_i (\sigma_i^2 + \mu_i^2) - \mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X]^2 \,,$$

which concludes the proof □

For any location-scale mixture, we then propose a reparameterisation of the mixture model that starts by scaling all parameters in terms of its global mean μ and global variance σ^2. For instance, we can switch to the representation

$$\mu_i = \mu + \sigma\alpha_i \quad \text{and} \quad \sigma_i = \sigma\tau_i \tag{2}$$

of the component-wise parameters, where $\tau_i > 0$ and $\alpha_i \in \mathbb{R}$. This is formally equivalent to the reparameterisation of Mengersen and Robert (1996), except that they put no special meaning on the global mean and variance parameters. Once the global mean and variance are set, this imposes natural constraints on the other parameters of the model. For instance, setting the global variance to σ^2 implies that $(\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k)$ belongs to a specific ellipse conditional on the weights and σ^2, by virtue of Lemma 1.

Considering the α_i's and the τ_i's in (2) as the new parameters of the components, the following result states that the global mean and variance parameters are the sole freely varying parameters. In other words, once both the global mean and variance are set, there exists a parameterisation such that all remaining parameters of a mixture distribution are restricted to a compact set, which is most helpful in selecting a non-informative prior distribution.

Lemma 2 *The parameters α_i and τ_i in (2) are constrained by*

$$\sum_{i=1}^{k} p_i \alpha_i = 0 \quad \textit{and} \quad \sum_{i=1}^{k} p_i \tau_i^2 + \sum_{i=1}^{k} p_i \alpha_i^2 = 1 \,.$$

Proof: The result is a trivial consequence of Lemma 1. The population mean is

$$\mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X] = \sum_{i=1}^{k} p_i\mu_i = \sum_{i=1}^{k} p_i(\mu+\sigma\alpha_i) = \mu + \sum_{i=1}^{k} p_i\alpha_i = \mu$$

and the first constraint follows. The population variance is

$$\begin{aligned}
\text{var}_{\boldsymbol{\theta},\mathbf{p}}(X) &= \sum_{i=1}^{k} p_i\sigma_i^2 + \sum_{i=1}^{k} p_i(\mu_i^2 - \mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X]^2) \\
&= \sum_{i=1}^{k} p_i\sigma^2\tau_i^2 + \sum_{i=1}^{k} p_ip_i(\mu^2 + 2\sigma\mu\alpha_i + \sigma^2\alpha_i^2 - \mu^2) \\
&= \sum_{i=1}^{k} p_i\sigma^2\tau_i^2 + \sum_{i=1}^{k} p_i\sigma^2\alpha_i^2 = \sigma^2
\end{aligned}$$

The last equation simplifies to the second constraint above. □

2.2 Reference priors

The constraints in Lemma 2 define a set of values of $(p_1,\ldots,p_k,\alpha,\ldots,\alpha,\tau,\ldots,\tau)$ that is obviously compact. From a Bayesian perspective, this allows for the call to uniform and other non-informative proper priors, conditional on (μ,σ). Furthermore, since (μ,σ) is a location-scale parameter, we may invoke Jeffreys (1939) to use the Jeffreys prior $\pi(\mu,\sigma) = 1/\sigma$ on this parameter, even though this is not the genuine Jeffreys prior for the mixture model (Grazian and Robert, 2015). In the same spirit as Robert and Titterington (1998) who established properness of the posterior distribution derived by Mengersen and Robert (1996), we now establish that this choice of prior produces a proper posterior distribution for a minimal sample size of two.

Theorem 1 *The posterior distribution associated with the prior $\pi(\mu,\sigma) = 1/\sigma$ and with the likelihood derived from (1) is proper when the components $f(\cdot|\mu,\sigma)$ are Gaussian densities, provided (a) proper distributions are used on the other parameters and (b) there are at least two observations in the sample.*

Proof: When $n = 1$, it is easy to show that the Jeffreys posterior is not proper. The marginal likelihood is then

$$\begin{aligned}
M_k(x_1) &= \sum_{i=1}^{k} \int p_i f(x_1|\mu+\sigma\alpha_i,\sigma^2\tau_i^2)\pi(\mu,\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mu,\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
&= \sum_{i=1}^{k} \int \left\{ \int \frac{p_i}{\sqrt{2\pi}\sigma^2\tau_i} \exp\left(\frac{-(x_1-\mu-\sigma\alpha_i)^2}{2\tau_i^2\sigma^2}\right) \mathrm{d}(\mu,\sigma) \right\} \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
&= \sum_{i=1}^{k} \int \left\{ \int_0^\infty \frac{p_i}{\sigma}\,\mathrm{d}\sigma \right\} \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})
\end{aligned}$$

The integral against σ is then not defined.

For two data-points, $x_1, x_2 \sim \sum_{i=1}^{k} p_i f(\mu+\sigma\alpha_i,\sigma^2\tau_i^2)$, the associated marginal likelihood is

$$\begin{aligned}
M_k(x_1,x_2) &= \int \prod_{j=1}^{2} \left\{ \sum_{i=1}^{k} p_i f(x_j|\mu+\sigma\alpha_i,\sigma^2\tau_i^2) \right\} \pi(\mu,\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mu,\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \int p_ip_j f(x_1|\mu+\sigma\alpha_i,\sigma^2\tau_i^2) f(x_2|\mu+\sigma\alpha_j,\sigma^2\tau_j^2)\pi(\mu,\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mu,\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\ .
\end{aligned}$$

If all those k^2 integrals are proper, the Jeffrey posterior distribution is proper. An arbitrary integral ($1 \le i, j \le k$) in this sum leads to

$$\begin{aligned}
&\int p_i p_j f(x_1|\mu+\sigma\alpha_i, \sigma^2\tau_i^2) f(x_2|\mu+\sigma\alpha_j, \sigma^2\tau_j^2) \pi(\mu,\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mu,\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
&= \int \left\{ \int \frac{p_i p_j}{2\pi\sigma^3\tau_i\tau_j} \exp\left[\frac{-(x_1-\mu-\sigma\alpha_i)^2}{2\tau_i^2\sigma^2} + \frac{-(x_2-\mu-\sigma\alpha_j)^2}{2\tau_j^2\sigma^2} \right] \mathrm{d}(\mu,\sigma) \right\} \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
&= \int \left\{ \int_0^\infty \frac{p_i p_j}{\sqrt{2\pi}\sigma^2\sqrt{\tau_i^2+\tau_j^2}} \exp\left[\frac{-1}{2(\tau_i^2+\tau_j^2)} \left(\frac{1}{\sigma^2}(x_1-x_2)^2 + \frac{2}{\sigma}(x_1-x_2)(\alpha_i-\alpha_j) \right.\right.\right. \\
&\qquad \left.\left.\left. +(\alpha_i-\alpha_j)^2 \right) \right] \mathrm{d}\sigma \right\} \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\sigma,\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,.
\end{aligned}$$

Substituting $\sigma = 1/z$, the above is integrated with respect to z, leading to

$$\begin{aligned}
&\int \left\{ \int_0^\infty \frac{p_i p_j}{\sqrt{2\pi}\sqrt{\tau_i^2+\tau_j^2}} \exp\left(\frac{-1}{2(\tau_i^2+\tau_j^2)} \left(z^2(x_1-x_2)^2 + 2z(x_1-x_2)(\alpha_i-\alpha_j) \right.\right.\right. \\
&\qquad \left.\left.\left. +(\alpha_i-\alpha_j)^2 \right) \right) \mathrm{d}z \right\} \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
&= \int \left\{ \int_0^\infty \frac{p_i p_j}{\sqrt{2\pi}\sqrt{\tau_i^2+\tau_j^2}} \exp\left(\frac{-(x_1-x_2)^2}{2(\tau_i^2+\tau_j^2)} \Big(z + \frac{\alpha_i-\alpha_j}{x_1-x_2} \Big)^2 \right) \mathrm{d}z \right\} \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,\mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
&= \int \frac{p_i p_j}{|x_1-x_2|} \Phi\left(-\frac{\alpha_i-\alpha_j}{x_1-x_2} \frac{|x_1-x_2|}{\sqrt{\tau_i^2+\tau_j^2}} \right) \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\;\mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,,
\end{aligned}$$

where Φ is the cumulative distribution function of the standardised Normal distribution. Given that the prior is proper on all remaining parameters of the mixture and that the integrand is bounded by $1/|x_1-x_2|$, it integrates against the remaining components of $\boldsymbol{\theta}$.

Let us now consider the case $n \ge 3$. Since the posterior $\pi(\boldsymbol{\theta}|x_1,x_2)$ is proper, it constitutes a proper prior when considering only the observations $x_3,\ldots,x_n$. Therefore, the posterior is almost everywhere proper. □

2.3 Further reparameterisations

Before proposing relevant priors, let us note that the constraints in Lemma 2 suggest a new reparameterisation (among many possible ones): this reparameterisations uses the weights p_i in the definition of the component parameters, as to achieve a more generic constraint. The component location and scale parameters in (2) can indeed be reparameterised as

$$\alpha_i = \sigma\gamma_i/\sqrt{p_i} \quad \text{and} \quad \tau_i = \sigma\eta_i/\sqrt{p_i}\,,$$

leading to the mixture representation

$$f(x|\boldsymbol{\theta},\mathbf{p}) = \sum_{i=1}^k p_i f(x|\mu+\sigma\gamma_i/\sqrt{p_i}, \sigma\eta_i/\sqrt{p_i})\,, \quad \eta_i > 0\,, \tag{3}$$

Given $(p_1,\cdots,p_k)$, these new parameters are constrained by

$$\sum_{i=1}^k \sqrt{p_i}\gamma_i = 0 \quad \text{and} \quad \sum_{i=1}^k (\eta_i^2+\gamma_i^2) = 1\,,$$

which means that $(\gamma_1, \ldots, \eta_k)$ belongs to an hypersphere of $\mathbb{R}^{2k}$ intersected with an hyperplane of this space.

Given these constraints, further simplifications via new reparameterisations can be contemplated, as for instance separating mean and variance parameters in (3) by introducing a radius φ such that

$$\sum_{i=1}^{k} \gamma_i^2 = \varphi^2 \quad \text{and} \quad \sum_{i=1}^{k} \eta_i^2 = 1 - \varphi^2 \,.$$

This choice naturally leads to a hierarchical prior where, e.g., φ^2 and $(p_1, \ldots, p_k)$ are distributed from a $\mathcal{B}e(a_1, a_2)$ and a $\mathcal{D}ir(\alpha_0, \ldots, \alpha_0)$ distributions, respectively, while the vectors $(\gamma_1, \ldots, \gamma_k)$ and $(\eta_1, \ldots, \eta_k)$ are uniformly distributed on the spheres of radius φ and $\sqrt{1-\varphi^2}$, respectively, under the additional linear constraint $\sum_{i=1}^{k} \sqrt{p_i}\gamma_i = 0$.

We now describe how this reparameterisation leads to a practical construction of the constrained parameter space, for an arbitrary number of components k.

2.3.1 Spherical coordinate representation of the γ's. The vector $(\gamma_1, \ldots, \gamma_k)$ belongs both to the hypersphere of radius φ and to the hyperplane orthogonal to $(\sqrt{p_1}, \ldots, \sqrt{p_k})$. Therefore, $(\gamma_1, \ldots, \gamma_k)$ can be expressed in terms of spherical coordinates within that hyperplane. Namely, if $(F_1, \ldots, F_{k-1})$ denotes an orthonormal basis of the hyperplane, $(\gamma_1, \ldots, \gamma_k)$ can be written as

$$(\gamma_1, \ldots, \gamma_k) = \varphi \cos(\varpi_1) F_1 + \varphi \sin(\varpi_1) \cos(\varpi_2) F_2 + \ldots + \varphi \sin(\varpi_1) \cdots \sin(\varpi_{k-2}) F_{k-1}$$

with the angles $\varpi_1, \ldots, \varpi_{k-3}$ in $[0, \pi]$ and ϖ_{k-2} in $[0, 2\pi]$. The s-th orthonormal base F_s can be derived from the k-dimensional orthogonal vectors $\widetilde{F}_s$ where

$$\widetilde{F}_{1,j} = \begin{cases} -\sqrt{p_2}, & j = 1 \\ \sqrt{p_1}, & j = 2 \\ 0, & j > 2 \end{cases}$$

and the s-th vector is given by

$$\widetilde{F}_{s,j} = \begin{cases} -(p_j p_{s+1})^{1/2} \Big/ \left(\sum_{l=1}^{s} p_l\right)^{1/2}, & s > 1,\ j \leq s \\ \left(\sum_{l=1}^{s} p_l\right)^{1/2}, & s > 1,\ j = s+1 \\ 0, & s > 1,\ j > s+1 \end{cases}$$

Note the special case of $k = 2$ since the angle ϖ_1 is then missing. In this special case, the mixture location parameter is defined by $(\gamma_1, \gamma_2) = \varphi F_1$ and φ takes both positive and negative values. In the general setting, the parameter vector $(\gamma_1 \cdots, \gamma_k)$ is a transform of $(\varphi^2, p_1, \cdots, p_k, \varpi_1, \cdots, \varpi_{k-2})$. A natural reference prior for ϖ is made of uniforms, $\varpi_1, \cdots, \varpi_{k-3} \sim \mathcal{U}[0, \pi]$ and $\varpi_{k-2} \sim \mathcal{U}[0, 2\pi]$, although other choices are obviously possible and should be explored to test the sensitivity to the prior.

2.3.2 Dual representation of the η_i's. For the component variance parameters, the vector $(\eta_1, \cdots, \eta_k)$ belongs to the k-dimension sphere of radius $\sqrt{1-\varphi^2}$. A natural prior is then a Dirichlet distribution with common hyperparameter a,

$$\pi(\eta_1^2, \cdots, \eta_k^2, \varphi^2) = \mathrm{Dir}(\alpha, \cdots, \alpha)$$

If k is small enough, $(\eta_1, \cdots, \eta_k)$ can then be simulated from the corresponding posterior with no computational challenge. However, as k increases, sampling may become more delicate and

benefits from a similar spherical reparameterisation. In this approach, the vector $(\eta_1, \cdots, \eta_k)$ is rewritten through spherical coordinates with angle components $(\xi_1, \cdots, \xi_{k-1})$,

$$\eta_i = \begin{cases} \sqrt{1-\varphi^2}\cos(\xi_i)\,, & i = 1 \\ \sqrt{1-\varphi^2}\prod_{j=1}^{i-1}\sin(\xi_j)\cos(\xi_i)\,, & 1 < i < k \\ \sqrt{1-\varphi^2}\prod_{j=1}^{i-1}\sin(\xi_j)\,, & i = k \end{cases}$$

Unlike ϖ, the support for all angles $\xi_1, \cdots, \xi_{k-1}$ is limited to $[0, \pi/2]$, due to the positivity requirement on the η_i's. In this case, a reference prior on the angles is

$$(\xi_1, \cdots, \xi_{k-1}) \sim \mathcal{U}([0, \pi/2]^{k-1})\,,$$

while again other choices are possible.

3. MCMC IMPLICATIONS

3.1 The Metropolis-within-Gibbs sampler

Given the reparameterisations introduced in Section 2, different MCMC implementations are possible and we investigate in this section some of these. To this effect, we distinguish between two cases: (i) only $(\mu_1, \cdots, \mu_k)$ is expressed in spherical coordinates; and (ii) both the μ_i's and the σ_i's are associated with spherical coordinates.

Although the target density is similar to the target explored by early Gibbs samplers in Diebolt and Robert (1990) and Gelman and King (1990), simulating directly the new parameters implies managing constrained parameter spaces. The hierarchical nature of the parameterisation also leads us to consider a block Gibbs sampler that coincides with this hierarchy. Since the corresponding full conditional posteriors are not in closed form, a Metropolis-within-Gibbs sampler is implemented here with random walk proposals. In this approach, the scales of the proposal distributions are automatically calibrated towards optimal acceptance rates (Roberts et al., 1997; Roberts and Rosenthal, 2001, 2009; Rosenthal, 2011). Convergence of a simulated chain is assessed based on the rudimentary convergence monitoring technique of Gelman and Rubin (1992). The description of the algorithm is provided by the pseudo-code version in Figure 1. Note that the Metropolis-within-Gibbs version does not rely on latent variables and complete likelihood as in Tanner and Wong (1987) and Diebolt and Robert (1990). Following the adaptive MCMC method in Section 3 of Roberts and Rosenthal (2009), we derive the optimal scales associated with proposal densities, based on 10 batches with size 50. The scales ϵ are identified by a subscript with the corresponding parameter.

For the reparameterisation (i), all steps are the same except that steps 2.5 and 2.7 are combined together and that $((\varphi^2)^{(t)}, (\eta_1^2)^{(t)}, \ldots, (\eta_k^2)^{(t)})$ is updated in the same manner. One potential proposal density is a Dirichlet distribution,

$$((\varphi^2)', (\eta_1^2)', \ldots, (\eta_k^2)') \sim \mathrm{Dir}((\varphi^2)^{(t-1)}\epsilon, (\eta_1^2)^{(t-1)}\epsilon, \ldots, (\eta_k^2)^{(t-1)}\epsilon)\,.$$

Alternative proposal densities will be discussed later along with simulation studies in Section 4.

3.2 Removing and detecting label switching

The standard parameterisation of mixture models contains weights $\{p_i\}_{i=1}^k$ and component-wise parameters $\{\theta_i\}_{i=1}^k$ as shown in (1). The likelihood function is invariant under permutations of the component indices. If an exchangeable prior is chosen on weights and component-wise parameters, the posterior density reproduces the likelihood invariance and component labels are

Metropolis-within-Gibbs algorithm for reparameterised mixture model

1 Generate initial values $(\mu^{(0)}, \sigma^{(0)}, \mathbf{p}^{(0)}, \varphi^{(0)}, \xi_1^{(0)}, \ldots, \xi_{k-1}^{(0)}, \varpi_1^{(0)}, \ldots, \varpi_{k-2}^{(0)})$.

2 For $t = 1, \ldots, T$, the update of $(\mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t)}, \varphi^{(t)}, \xi_1^{(t)}, \ldots, \xi_{k-1}^{(t)}, \varpi_1^{(t)}, \ldots, \varpi_{k-2}^{(t)})$ follows;

- **2.1** Generate a proposal $\mu' \sim \mathcal{N}(\mu^{(t-1)}, \epsilon_\mu)$ and update $\mu^{(t)}$ against $\pi(\cdot|\mathbf{x}, \sigma^{(t-1)}, \mathbf{p}^{(t-1)}, \varphi^{(t-1)}, \boldsymbol{\xi}^{(t-1)}, \boldsymbol{\varpi}^{(t-1)})$.
- **2.2** Generate a proposal $\log(\sigma)' \sim \mathcal{N}(\log(\sigma^{(t-1)}), \epsilon_\sigma)$ and update $\sigma^{(t)}$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \mathbf{p}^{(t-1)}, \varphi^{(t-1)}, \boldsymbol{\xi}^{(t-1)}, \boldsymbol{\varpi}^{(t-1)})$.
- **2.3** Generate proposals $\xi_i' \sim \mathcal{U}[0, \pi/2]$, $i = 1, \cdots, k-1$, and update $(\xi_1^{(t)}, \ldots, \xi_{k-1}^{(t)})$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t-1)}, \varphi^{(t-1)}, \boldsymbol{\varpi}^{(t-1)})$.
- **2.4** Generate proposals $\varpi_i' \sim \mathcal{U}[0, \pi]$, $i = 1, \cdots, k-3$, and $\varpi_{k-2}' \sim \mathcal{U}[0, 2\pi]$. Update $(\varpi_1^{(t)}, \ldots, \varpi_{k-2}^{(t)})$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t-1)}, \varphi^{(t-1)}, \boldsymbol{\xi}^{(t)})$.
- **2.5** Generate a proposal $(\varphi^2)' \sim \mathcal{B}eta((\varphi^2)^{(t)}\epsilon_\varphi + 1, (1 - (\varphi^2)^{(t)})\epsilon_\varphi + 1)$ and update $\varphi^{(t)}$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t-1)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\varpi}^{(t)})$.
- **2.6** Generate a proposal $\mathbf{p}' \sim \mathrm{Dir}(p_1^{(t-1)}\epsilon_p + 1, \ldots, p_k^{(t-1)}\epsilon_p + 1)$, and update $\mathbf{p}^{(t)}$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \varphi^{(t)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\varpi}^{(t)})$.
- **2.7** Generate proposals $\xi_i' \sim U[\xi_i^{(t)} - \epsilon_\xi, \xi_i^{(t)} + \epsilon_\xi]$, $i = 1, \cdots, k-1$, and update $(\xi_1^{(t)}, \ldots, \xi_{k-1}^{(t)})$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t)}, \varphi^{(t)}, \boldsymbol{\varpi}^{(t)})$.
- **2.8** Generate proposals $\varpi_i' \sim U[\varpi_i^{(t)} - \epsilon_\varpi, \varpi_i^{(t)} + \epsilon_\varpi]$, $i = 1, \cdots, k-2$, and update $(\varpi_1^{(t)}, \ldots, \varpi_{k-2}^{(t)})$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t)}, \varphi^{(t)}, \boldsymbol{\xi}^{(t)})$.

Fig 1: Pseudo-code representation of the Metropolis-within-Gibbs algorithm used in this paper for the reparameterisation (ii) based on two sets of spherical coordinates. For simplicity's sake, we denote $\mathbf{p}^{(t)} = (p_1^{(t)}, \ldots, p_k^{(t)})$, $\mathbf{x} = (x_1, \ldots, x_n)$, $\boldsymbol{\xi}^{(t)} = (\xi_1^{(t)}, \ldots, \xi_{k-1}^{(t)})$ and $\boldsymbol{\varpi}^{(t)} = (\varpi_1^{(t)}, \ldots, \varpi_{k-2}^{(t)})$.

not identifiable. This phenomenon is called *label switching* and is well-studied in the literature (Celeux et al., 2000; Stephens, 2000; Frühwirth-Schnatter, 2001; Frühwirth-Schnatter, 2004; Jasra et al., 2005). This means that the posterior distribution consists of $k!$ symmetric modes and a Markov chain with such target distribution is expected to explore all of them. However, a chain often fails and rather ends up exploring a particular mode.

In our reparameterisation of Gaussian mixture models, each component mean and variance are functions of angular and radius parameters with weights. The mapping between both parameterisations is a one-to-one map conditional on the weights. In other words, there are unique component-wise means and variances given particular values for angular and radius parameters and weights. Although the new parameterisation is not exchangeable, due to the choice of the orthogonal basis, adopting an exchangeable prior on the weights (e.g., a Dirichlet distribution with a common parameter) and uniform priors on all angular parameters leads to an exchangeable posterior on the natural parameters of the mixture. Therefore, label switching should also occur with this prior modelling.

When an MCMC chain manages to jump between modes, the inference on each of the mixture components becomes harder (Geweke, 2007). To get component-specific inference and to give a meaning to each component, various relabelling methods have been proposed in the literature (see, e.g., Frühwirth-Schnatter, 2004). A first available alternative is to reorder labels so that the mixture weights are in increasing order (Frühwirth-Schnatter, 2001). A second alternative method proposed by, e.g., Lee et al. (2009) is that labels are reordered towards producing the shortest distance between the current posterior sample and the (or a) maximum posterior probability (MAP) estimate.

Let us denote by h the map from our reparameterisation to the standard parameterisation of (1), i.e.,

$$(\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k, \mathbf{p}) = h(\mathbf{p}, \boldsymbol{\theta}) \,,$$

with its inverse h^{-1} available as well. We also denote by $\mathfrak{S}_k$ the set of permutations of $\{1, \ldots, k\}$. Then, given an MCMC sample $\{\mathbf{p}^{(t)}, \boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$, the above relabelling technique procedure follows;

1. Reparameterise the MCMC sample $\{\mathbf{p}^{(t)}, \boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$ into component-wise means and standard deviations via the function h, resulting in $\{\mu_1^{(t)}, \ldots, \mu_k^{(t)}, \sigma_1^{(t)}, \ldots, \sigma_k^{(t)}, \mathbf{p}^{(t)}\}_{t=1}^{T}$.
2. Find the MAP estimate by computing the posterior values of the sample; denote the solution as $(\mu_1^*, \ldots, \mu_k^*, \sigma_1^*, \ldots, \sigma_k^*, \mathbf{p}^*)$.
3. Reorder $(\mu_1^{(t)}, \ldots, \mu_k^{(t)}, \sigma_1^{(t)}, \ldots, \sigma_k^{(t)}, \mathbf{p}^{(t)})$ as

$$(\widetilde{\mu}_1^{(t)}, \ldots, \widetilde{\mu}_k^{(t)}, \widetilde{\sigma}_1^{(t)}, \ldots, \widetilde{\sigma}_k^{(t)}, \widetilde{\mathbf{p}}^{(t)}) = \delta_j(\mu_1^{(t)}, \ldots, \mu_k^{(t)}, \sigma_1^{(t)}, \ldots, \sigma_k^{(t)}, \mathbf{p}^{(t)})$$

where $\delta_j = \arg\min_{\delta \in \mathfrak{S}_k} \|\delta(\mu_1^{(t)}, \ldots, \mu_k^{(t)}, \sigma_1^{(t)}, \ldots, \sigma_k^{(t)}, \mathbf{p}^{(t)}) - (\mu_1^*, \ldots, \mu_k^*, \sigma_1^*, \ldots, \sigma_k^*, \mathbf{p}^*)\|$.

The resulting permutation is then denoted $\lambda^{(t)} \in \mathfrak{S}_k$. Label switching occurrences in an MCMC sequence can be monitored via the changes in the sequence $\lambda^{(1)}, \ldots, \lambda^{(T)}$. If the chain fails to switch modes, the sequence is likely to remain at the same permutation. On the opposite, if a chain moves between some of the $k!$ symmetric posterior modes, the $\lambda^{(t)}$'s are expected to vary.

We proceed here by a simulation studies section and all algorithms used in this section are publicly available within the R package Ultimixt (Kamary and Lee, 2015). The package Ultimixt contains functions that implement adaptive determination of optimal scales and convergence monitoring based on Gelman and Rubin (1992) criterion. In addition, Ultimixt includes functions that summarise the simulations and compute point estimates of each parameter, such as posterior mean and median. It also produces an estimated mixture density in numerical and graphical formats. The output further includes graphical representations of the generated parameter samples. For the potentially unimodal parameters μ, σ and φ, averaging and calculating

the median over the generated chains directly returns valid point estimators, as those parameters are not subjected to label switching. For the other parameters (component weights, means and variances), since label switching is a possible issue, we need to postprocess the MCMC draws as discussed earlier, by first relabelling these simulations. We then derive point estimates by clustering over the parameter space, using k-mean clustering (Hastie et al., 2001).

4. SIMULATION STUDIES

In this section, we examine the performances of the above Metropolis-within-Gibbs algorithm, when applied to both reparameterisations defined above. We also consider the special case $k = 2$ in Section 4.1. All simulations were conducted using the package Ultimixt (Kamary and Lee, 2015).

4.1 The case $k = 2$

In this specific case, we do not have to simulate any angle. Two straightforward proposals are compared over simulation experiments. One is based on Beta and Dirichlet proposals:

$$p^* \sim \mathcal{B}eta(p^{(t)}\epsilon_p, (1-p^{(t)})\epsilon_p)\,, \quad (\varphi^{2^*}, \eta_1^{2^*}, \eta_2^{2^*}) \sim \mathcal{D}ir(\varphi^{2^{(t)}}\epsilon, \eta_1^{2^{(t)}}\epsilon, \eta_2^{2^{(t)}}\epsilon)$$

(this will be called Proposal 1) and another one is based on Gaussian random walks:

$$\begin{aligned}
\log(p^*/(1-p^*)) &\sim \mathcal{N}(\log(p^{(t)}/(1-p^{(t)})), \epsilon_p) \\
(\vartheta_1^*, \vartheta_2^*)^T &\sim \mathcal{N}(\chi_2^{(t)}, \epsilon_\vartheta I_2) \quad \text{with} \\
(\varphi^{2^*}, \eta_1^{2^*}, \eta_2^{2^*}) &= (\exp(\vartheta_1^*)/\bar{\vartheta}^*, \exp(\vartheta_2^*)/\bar{\vartheta}^*, 1/\bar{\vartheta}^*)\,, \\
\chi_2^{(t)} &= (\log(\varphi^{2^{(t)}}/\eta_2^{2^{(t)}}), \log(\eta_1^{2^{(t)}}/\eta_2^{2^{(t)}}) \\
\text{and} \quad \bar{\vartheta}^* &= 1 + \exp(\vartheta_1^*) + \exp(\vartheta_2^*)
\end{aligned}$$

(which will be called Proposal 2). The global parameters are proposed using Normal and Inverse-Gamma proposals

$$\mu^* \sim \mathcal{N}(\bar{x}, \epsilon_\mu) \quad \text{and} \quad \sigma^{2^*} \sim \mathcal{IG}((n+1)/2, (n-1)\bar{\sigma}^2/2)$$

where $\bar{x}$ and $\bar{\sigma}^2$ are sample mean and variance respectively. We present below some analyses and also explain how MCMC methods can be used to fit the reparameterised mixture distribution.

Example 4.1 In this experiment, a dataset of size 50 is simulated from the mixture $0.65\mathcal{N}(-8, 2) + 0.35\mathcal{N}(-0.5, 1)$, which implies that while the true value of $(\varphi, \eta_1, \eta_2)$ is $(0.91, 0.16, 0.38)$. Figure 2 illustrates the performances of a Metropolis-within-Gibbs algorithm based on Proposal 1. It shows the outcomes of 10 parallel chains, each started randomly from different starting values. The estimated densities are almost indistinguishable among the different chains and they all converge to a neighbourhood of the true values. The chains are well-mixed and the sampler output covers the entire sample space in this case.

We also run the Metropolis-within-Gibbs algorithm based on Proposal 2 using the same simulated dataset for comparison purposes. As shown in Figure 3, the outputs for both proposals are quite similar but Proposal 1 produces more symmetric chains on $p, \varphi, \eta_1, \eta_2$, thus suggesting higher mixing abilities.

The scales of the various proposals are determined by aiming at Roberts et al. (1997) goal of an average acceptance rate of either 0.44 or 0.234 depending on the dimension of the simulated parameter. As shown in Table 1, an adaptive Metropolis-within-Gibbs strategy manages to recover acceptance rates close to optimal values. ◀

Having exposed how our sampler behaves we now discuss a second example, in which we briefly outline how this method may behave for a benchmark dataset with a slightly larger sample size.

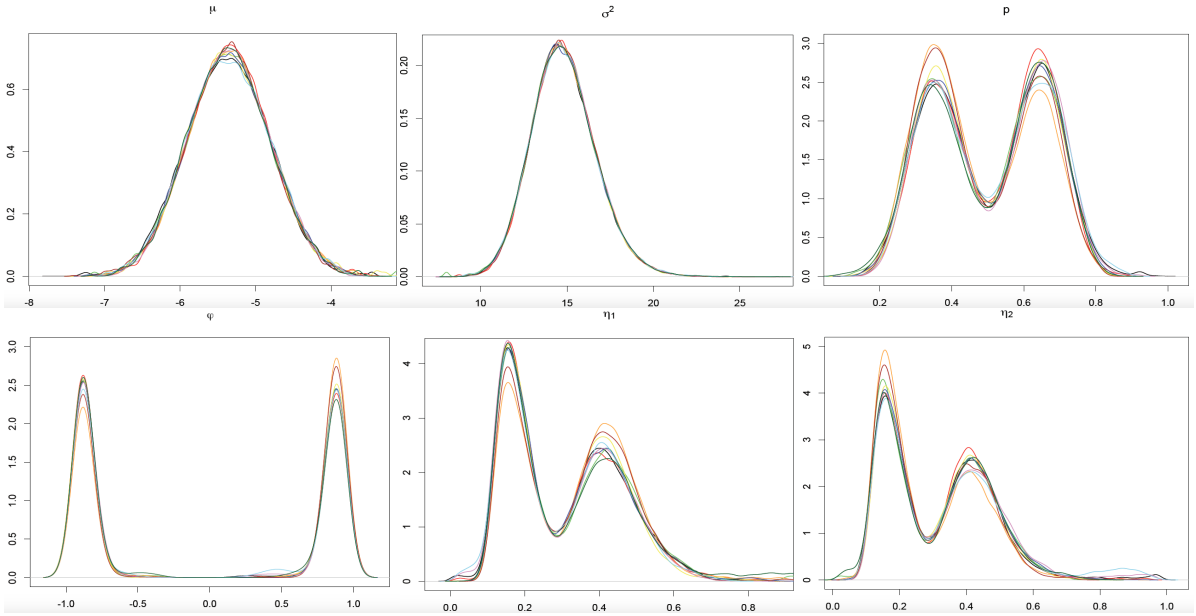

Fig 2: **Example 4.1:** Kernel estimates of the posterior densities of the parameters μ, σ, p, φ, η_i, based on 10 parallel MCMC chains for Proposal 1 and $2\,10^5$ iterations, based on a single simulated sample of size 50. The true value of $(\varphi, \eta_1, \eta_2)$ is $(0.91, 0.16, 0.38)$.

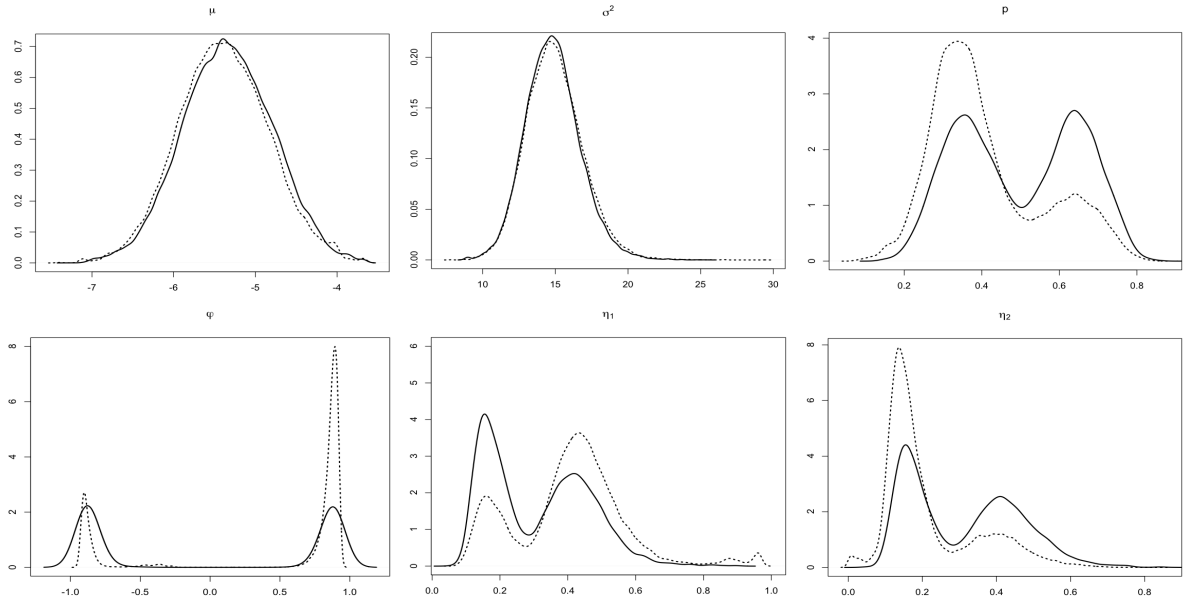

Fig 3: **Example 4.1:** Comparison between MCMC samples from our Metropolis-within-Gibbs algorithm using Proposal 1 (*solid line*) or Proposal 2 (*dashed line*), with $90,000$ iterations and the same sample as in Figure 2. The true value of $(\varphi, \eta_1, \eta_2)$ is $(0.91, 0.16, 0.38)$.

Example 4.2 We now analyse the benchmark Old Faithful dataset, available from R, using the 272 observations of eruption times and a mixture with two components. The empirical mean and variance of the observations are $(3.49, 1.30)$.

When using Proposal 1, the optimal scales $\epsilon_\mu, \epsilon_p, \epsilon$ after $50,000$ burn-in iterations are 0.07, 501.1, 802.19, respectively. The posterior distributions of the generated samples shown in Figure 4 demonstrate a strong concentration of (μ, σ^2) near the empirical mean and variance. Trace plots for the other parameters indicate a high dependence between successive iterations.There

Proposal 1	ar_μ	ar_σ	ar_p	$ar_{\varphi,\eta}$	ϵ_μ	ϵ_p	ϵ
	0.40	0.47	0.45	0.24	0.56	77.06	99.94
Proposal 2	ar_μ	ar_σ	ar_p	$ar_{\varphi,\eta}$	ϵ_μ	ϵ_p	ϵ_ϑ
	0.38	0.46	0.45	0.27	0.55	0.29	0.35

TABLE 1

Example 4.1: *Acceptance rate (ar) and corresponding proposal scale (ϵ) when the adaptive Metropolis-within-Gibbs sampler is used.*

is a strong indication that the chain gets trapped into a single mode of the posterior density. In Section 5, we reanalyse this dataset when using parallel tempering. ◀

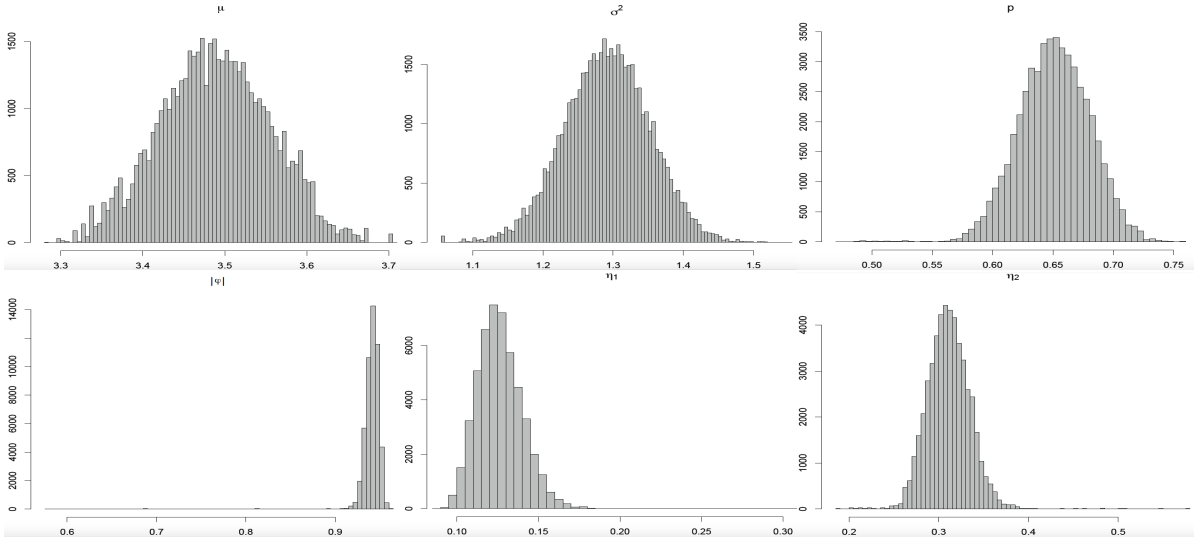

Fig 4: **Old Faithful dataset (Example 4.2):** Posterior distributions of the parameters of a two-component mixture distribution based on $50,000$ MCMC iterations.

4.2 The general case

We now consider the general case of estimating a reparameterised mixture for any k when the variance vector $(\eta_1^2, \ldots, \eta_k^2)$ also has the spherical coordinate system as represented in Section 2.3.

Example 4.3 We simulated 50 data points from the mixture

$$0.27\mathcal{N}(-4.5, 1) + 0.4\mathcal{N}(10, 1) + 0.33\mathcal{N}(3, 1)\,.$$

Running our adaptive Metropolis-within-Gibbs algorithm shows that the simulated samples are quite close to the true values. However, the sampler has apparently visited only one of the posterior modes. This lack of label switching helps us in producing point estimates directly from this MCMC output (Geweke, 2007) but this also shows an incomplete convergence of the MCMC sampler (Celeux et al., 2000). When considering the new parameters of this mixture, the single ϖ plays a significant role in the lack of label switching since transforming ϖ to $\pi - \varpi$ swaps first and second components.

If we restrict the proposal on ϖ to step 2.4 of the Metropolis-within-Gibbs algorithm, namely using only a uniform $\mathcal{U}(0, 2\pi)$ distribution, Figure 5 shows that the MCMC chains of the p_i's are both well-mixed and exhibiting strong exchangeability. However, the corresponding acceptance rate is quite low at 0.051.

If we consider in addition the random walk proposal of Step 2.8 on ϖ, namely a $\mathcal{U}(\varpi^{(t)} - \epsilon_\varpi, \varpi^{(t)} + \epsilon_\varpi)$ distribution, this step clearly improves performances, as illustrated in Figure 6,

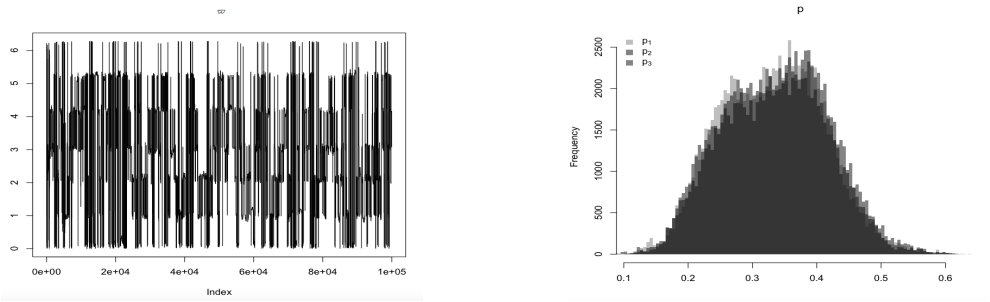

Fig 5: **Example 4.3:** *(Left)* Evolution of the sequence $(\varpi^{(t)})$ and *(Right)* histograms of the simulated weights based on 10^5 iterations of an adaptive Metropolis-within-Gibbs algorithm with independent proposal on ϖ.

with acceptance rates all close to 0.234 and 0.44. Almost perfect label switching occurs in this case.

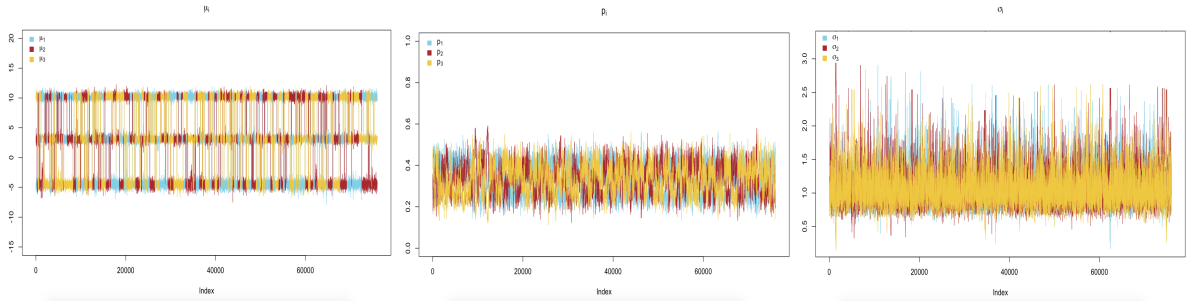

Fig 6: **Example 4.3:** Traces of the last $70,000$ simulations from the posterior distributions of the component means, standard deviations and weights, involving an additional random walk proposal on ϖ, based on 10^5 iterations.

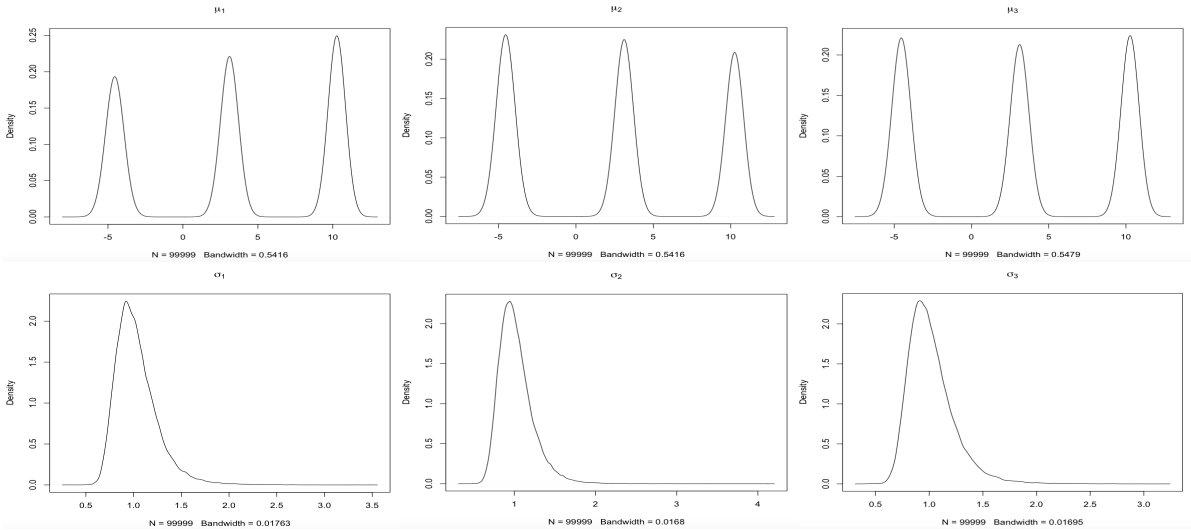

Fig 7: **Example 4.3:** Estimated marginal posterior densities of component means and standard deviations, based on 10^5 MCMC iterations.

The marginal posterior distributions of the means and standard deviations are shown in Figure 7. They are almost indistinguishable due to label switching. Point estimates are once

	Angular & component-wise parameters					
	k-means clustering			MAP estimate		
	ϖ	ξ_1	ξ_2	ϖ	ξ_1	ξ_2
Median	3.54	0.97	0.73	3.32	0.94	0.83
Mean	3.53	0.98	0.72	3.45	0.94	0.82
	p_1	p_2	p_3	p_1	p_2	p_3
Median	0.40	0.27	0.33	0.41	0.27	0.33
Mean	0.41	0.27	0.33	0.41	0.27	0.33
	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
Median	10.27	-4.55	3.11	10.27	-4.55	3.11
Mean	10.27	-4.54	3.12	10.26	-4.45	3.11
	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3
Median	0.93	1.04	1.01	0.93	1.04	1.03
Mean	0.95	1.08	1.05	0.95	1.07	1.05

	Global parameters		
	μ	σ	φ
Median	3.98	6.03	0.98
Mean	3.98	6.02	0.99

Proposal scales					
ϵ_μ	ϵ_σ	ϵ_p	ϵ_φ	ϵ_ϖ	ϵ_ξ
0.33	0.06	190	160	0.09	0.39
Acceptance rates					
ar_μ	ar_σ	ar_p	ar_φ	ar_ϖ	ar_ξ
0.22	0.34	0.23	0.43	0.42	0.22

TABLE 2

Example *4.3: Point estimators of the parameters of a mixture of 3 components, proposal scales and corresponding acceptance rates.*

more produced by relabelling and k-mean clustering, to be compared with the MAP estimates automatically deduced from the simulation output. Those estimate are shown on the left and right sides of Table 2, respectively. Estimates computed by both methods are almost identical and all parameters are close to the true values.

However, Bayesian inference for parameters related to individual components of the mixture using averaging over posterior draws is not possible in this case since the posterior means of the component specific parameters such as $p, \mu_i, \sigma_i; i = 1, 2, 3$ are the same for all components. We therefore revert to both methods of k-means clustering algorithm presented at the beginning of this section and removing label switching based on the distance between posterior sample and MAP estimate which are shown in left and right sides of Table 2, respectively. Bayesian estimations computed by both methods are almost identical and all parameters of the mixture distributions are accurately estimated in comparison with those of the true model with the acceptance rates of the proposal distributions of the Metropolis-within-Gibbs very close to the optimal ones.

Example 4.4 We now consider an 8 component mixture,

$$0.08\mathcal{N}(0, 0.8) + 0.12\mathcal{N}(1.5, 1.1) + 0.2\mathcal{N}(3, 0.9) + 0.1\mathcal{N}(5, 1.2) + 0.15\mathcal{N}(7.5, 2) + 0.1\mathcal{N}(9, 1.3) + 0.13\mathcal{N}(10.2, 0.7) + 0.12\mathcal{N}(11.5, 1.1)\,,$$

from which we simulated 20 samples of size 250. Calibration of the random walks is achieved after 10^4 for almost all samples.

When computing point estimates of the natural parameters of the components, we obtain the maximum errors of 0.08 and 0.11 for μ and σ, respectively. The average absolute error over the 20 samples is quite low. Furthermore, when comparing the true and estimated mixtures, we can resort to the Kullback-Leibler divergence. For the 20 simulated samples, the maximum value is 0.02, which means an information loss of at most 2%. If we consider the upper bound introduced by Sayyareh (2011) on Kullback-Leibler divergence, the obtained values indicates a good similarity between P_{true} and $P_{estimated}$ and illustrates the consistency of the estimates resulting from our Metropolis-within-Gibbs algorithm. ◀

Example 4.5 When an MCMC chain converges to a very small value for at least one component weight p_i, this may lead to an extremely large mean or large variance in the corresponding component. This happens partly because there is hardly any information from the data for this component and partly because the new parameters are functions of $1/\sqrt{p_i}$. We may thus face extreme points in the simplex parameter spaces. This phenomenon is illustrated with the *Galaxy dataset*, a constant benchmark for mixture estimation (Roeder, 1990; Richardson and Green,

1997), when we impose $k = 6$ components. The MCMC sample is again summarised by k-means clustering and MAP estimates, as presented in Section 3.2. The resulting means, medians and 95% credible intervals of the parameters of the mixture components are displayed in Table 3. Unsurprisingly, global mean and standard deviation are quite similar to the empirical estimates. Table 3 also displays estimates based on the Gibbs sampler of bayesm (Rossi and McCulloch, 2010) and on the EM algorithms of mixtools (Benaglia et al., 2009), with our approach being produced by Ultimixt (Kamary and Lee, 2015).

Obtaining very close estimations for two component means μ_i, as $\mu_1 = 19.59$ and $\mu_5 = 19.93$, and $\mu_2 = 21.97$ and $\mu_6 = 22$ and $\mu_4 = 22.21$ for bayesm, and $\mu_1 = 24.27$ and $\mu_6 = 24.26$ for mixtools, signals that overfitting occurs: there are more components than supported by the data. With our analysis, overfitting is handled in a different way: the mean of one or more component weights is close to zero. For instance, we obtained estimates of p_1 very close to zero, inducing estimates for μ_1 and σ_1 of 61.59 and 32.23 for μ_1 and σ_1 (obtained by k-means clustering) and of 67.26 and 20.53 (using MAP estimates), as shown in Table 3. If we examine the MCMC sequences in detail, the minimum simulated value for the first component weight and the corresponding first component mean and standard deviation are $1.045\,10^{-6}$, 449.25 and 284.34, respectively. Such extreme values are produced because of the extremely small weight. However, such large values have no impact on the resulting estimate of the mixture itself. This is clearly exhibited in Figure 8 for the *Galaxy dataset*, which shows that extreme values have no effect on the predictive density plots due to the small weights. Using our modelling, the resulting density estimate is remarkably smooth when considering that the number of observations is 82 and a number of components equal to 6.

If we repeat running the algorithm on the *Galaxy dataset* for $50,000$ iterations and a smaller number of components, for instance $k = 4$, summary and model fit statistics are provided in Table 3. In this case, extreme values do not occur and the predictive density plots show that a four component model fits the data equally well as displayed in Figure 9. The posterior estimates of the component parameters computed by three methods (k-means clustering, MAP, and EM estimates) are almost similar, while the Gibbs sampler results from bayesm yield two very close estimates of component means, $\mu_2 = 21.05$ and $\mu_4 = 20.90$ in this case.

The common priors for the standard parameters are

$$\mu_i \sim N(\bar{\mu}, 10\sigma_R)\,, \quad \sigma_R^2 \sim \text{IW}(\nu, 3) \quad \text{and} \quad (p_1, \ldots, p_k) \sim \mathcal{D}ir(\alpha_0, \ldots, \alpha_0)$$

where $\text{IW}(\nu, 3)$ is the Inverse-Wishart distribution with the scale parameter of 3 and the degrees of freedom of ν. Unknown hyperparameters $\bar{\mu}$, σ_R, α_0 and ν are given by bayesm from the empirical estimation of data and, the comparison of the proposed priors and the prior obtained from bayesm are graphically presented in Figure 10.

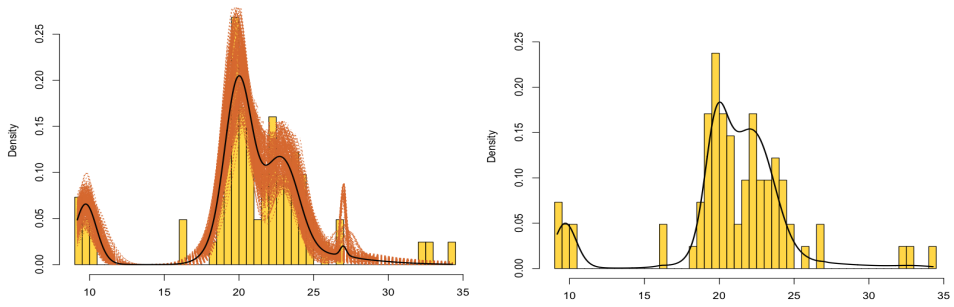

Fig 8: **Galaxy dataset:** (*Left*) Representation of 500 MCMC iterations as mixture distributions with the overlaid average curve for $k = 6$ components (*dark line*); (*Right*) mixture density estimate based on $15,000$ MCMC iterations for $k = 6$ components.

	6 components, $k = 6$						**4 components,** $k = 4$			
	k-means clustering						**k-means clustering**			
	p_1	p_2	p_3	p_4	p_5	p_6	p_1	p_2	p_3	p_4
Median	0.01	0.08	0.13	0.43	0.05	0.24	0.56	0.27	0.06	0.10
Mean	0.02	0.06	0.14	0.46	0.05	0.24	0.58	0.25	0.06	0.11
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1	μ_2	μ_3	μ_4
Median	25.95	9.72	22.06	19.83	32.71	22.87	20.19	21.52	32.79	9.72
Mean	61.59	9.725	22.09	19.84	32.70	22.93	20.27	21.48	33.29	9.73
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1	σ_2	σ_3	σ_4
Median	4.53	4.91	1.91	0.52	2.86	0.65	0.52	1.62	3.00	1.05
Mean	32.23	4.61	2.41	0.58	4.23	1.10	0.57	2.08	3.66	3.44
	MAP estimate						**MAP estimate**			
	p_1	p_2	p_3	p_4	p_5	p_6	p_1	p_2	p_3	p_4
Median	0.04	0.09	0.13	0.37	0.10	0.15	0.32	0.46	0.08	0.08
Mean	0.05	0.09	0.10	0.39	0.14	0.22	0.34	0.43	0.13	0.09
2.5%	$< 10^{-5}$	< 0.01	< 0.01	< 0.01	$< 10^{-3}$	< 0.01	0.04	0.02	0.01	0.04
97.5%	0.2.1	0.13	0.69	0.39	0.56	0.68	0.87	0.82	0.51	0.15
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1	μ_2	μ_3	μ_4
Median	30.96	9.70	21.75	19.73	20.61	23.12	19.84	22.17	28.23	9.71
Mean	67.26	8.18	21.58	18.73	20.84	24.33	19.83	22.34	29.03	9.50
2.5%	22.87	-9.28	19.60	9.68	12.83	21.29	17.59	20.14	22.27	9.17
97.5%	606.16	10.21	23.44	20.47	25.69	33.07	21.47	26.87	36.20	10.21
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1	σ_2	σ_3	σ_4
Median	4.82	0.54	1.76	0.60	3.41	1.73	0.69	2.22	3.22	0.53
Mean	20.53	2.05	2.06	0.73	15.59	2.34	0.96	3.23	4.15	0.91
2.5%	0.79	0.30	0.31	0.19	0.41	0.17	0.29	0.87	0.68	0.29
97.5%	198.23	17.28	7.63	2.13	35.95	7.62	2.44	9.62	10.57	1.34
	Gibbs sampler (bayesm)						**Gibbs sampler** (bayesm)			
	p_1	p_2	p_3	p_4	p_5	p_6	p_1	p_2	p_3	p_4
	0.17	0.09	0.14	0.23	0.19	0.19	0.33	0.31	0.18	0.18
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1	μ_2	μ_3	μ_4
	19.59	21.97	20.83	22.21	19.93	22.00	20.53	21.05	21.75	20.90
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1	σ_2	σ_3	σ_4
	0.35	0.23	0.22	0.24	0.26	0.31	0.22	0.19	0.21	0.27
	EM estimate (mixtools)						**EM estimate** (mixtools)			
	p_1	p_2	p_3	p_4	p_5	p_6	p_1	p_2	p_3	p_4
	0.04	0.08	0.17	0.41	0.09	0.20	0.52	0.33	0.04	0.11
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1	μ_2	μ_3	μ_4
	24.27	9.71	22.33	19.88	33.04	24.26	19.72	22.72	33.04	10.14
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1	σ_2	σ_3	σ_4
	0.08	0.42	0.44	.70	0.19	8.33	0.62	1.77	0.92	2.73

TABLE 3

Galaxy dataset*: Estimates of the parameters of a mixture of 6 and 4 components.*

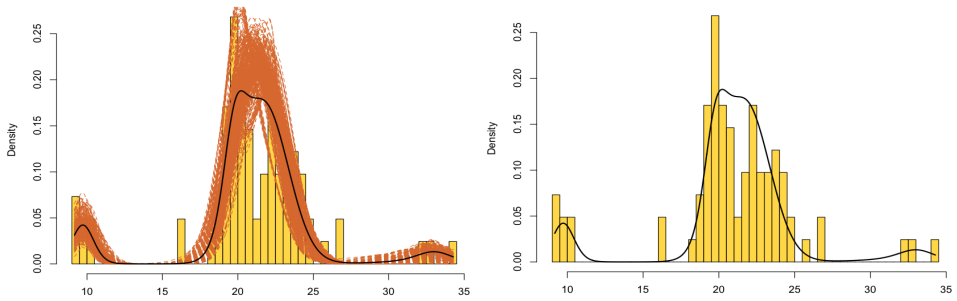

Fig 9: **Galaxy dataset:** (*left*) Representation of 500 Metropolis-within-Gibbs iterations for the mixture estimation and the overlay curve (*dark line*) obtained by averaging over iterations; (*right*) The mixture density estimate to histogram of dataset computed by averaging over $50,000$ MCMC iterations.

It is seen that the proposed prior is more dispersed for μ_1 and p_1 and is very skewed toward 0 for σ_1 with long tail. When $k = 6$, bayesm yields a more concentrated prior for p to accommodate all components and the proposed prior becomes dispersed to give flexible support on component-wise location and scale.

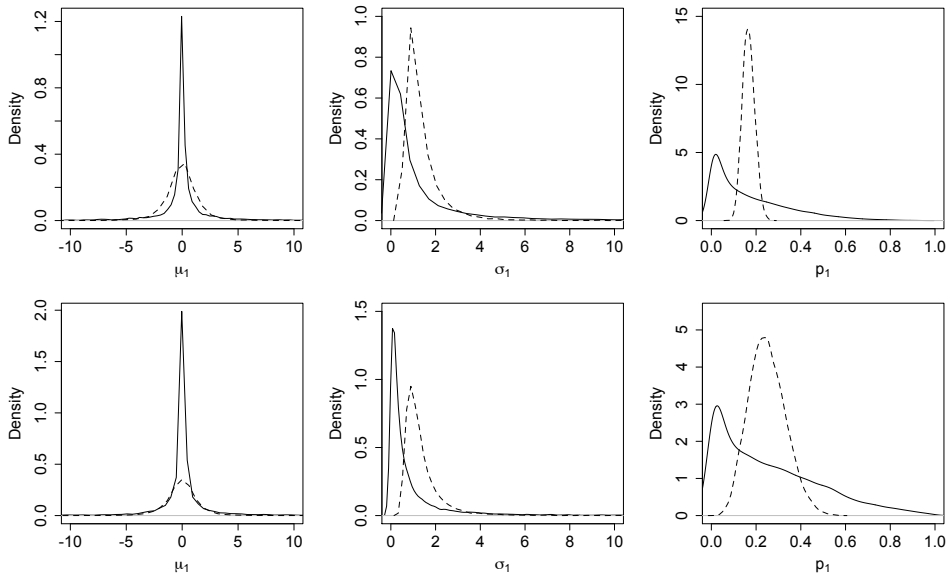

Fig 10: **Galaxy dataset:** Empirical prior densities based on 10^4 samples for μ_1, σ_1 and p_1 when (*Top*) $k = 6$ and (*Bottom*) $k = 4$. For the proposed prior (*solid lines*), the priors induced are $\pi(\mu_1) \propto \pi(\sigma\gamma_1/\sqrt{p_1})$ and $\pi(\sigma_1) \propto \pi(\sigma\eta_1/\sqrt{p_1})$. For the prior by bayesm (*dashed lines*), hyperparameters are $\alpha_0 = 5$ for $k = 4$ and $\alpha_0 = 25$ for $k = 6$ while $\bar{\mu} = 0$ and $\nu = 3$.

5. PARALLEL TEMPERING

In Example 4.2 we have seen that for the Old Faithful dataset, the multimodality of the mixture model is not reproduced in the MCMC output, which means the adaptive Metropolis-within-Gibbs sampler cannot escape one of the modes. In this case, parallel tempering may be used (Marinari and Parisi, 1992; Neal, 1996). This method allows for better mixing in multimodal target distributions, when using straightforward Metropolis-Hastings algorithms fail

(Miasojedow et al., 2013). It is indeed designed to overcome low probability regions between modal areas. Given the posterior density $f(\theta|x)$, we define tempered versions $f_\beta(\theta|x) \propto f(\theta|x)^\beta$, where $0 \leq \beta \leq 1$ is the inverse temperature and $\beta = 1$ corresponds to the original target distribution (Geyer, 1991). The tempered MCMC algorithm then runs a basic MCMC algorithm on a range of tempered distribution and, at each iteration, the current samples are considered for potential exchanges between adjacent temperatures, with a Metropolis–Hastings acceptance probability

$$\alpha_h = \min\left(1, \frac{f_{\beta_{h-1}}(\theta_h^{(t)}) f_{\beta_h}(\theta_{h-1}^{(t)})}{f_{\beta_{h-1}}(\theta_{h-1}^{(t)}) f_{\beta_h}(\theta_h^{(t)})}\right),$$

as the chances of accepting a swap are higher for nearby temperatures. Proposal scales are calibrated by adaptive MCMC method and is used for all tempered versions of the target. Temperatures are chosen of the form 2^j $(j = 1, \ldots)$ and the sequence is determined according to the degree of symmetry in the distribution of the p_i's or when the minimum acceptance rate for swaps between adjacent temperatures is larger than a default threshold.

Example 5.1 Considering again the Old Faithful benchmark, we set this symmetry threshold to .1 and this acceptance threshold to 0.3. Using the same proposals as in Example 4.2 and $N_{sim} = 50,000$, the algorithm selects 4 temperatures, thus equal to $1, 2, 4, 8$. Figure 11 demonstrates that the parallel tempering sampler visits all modes in the posterior distribution and that the mixing of the chains is greatly improved. ◀

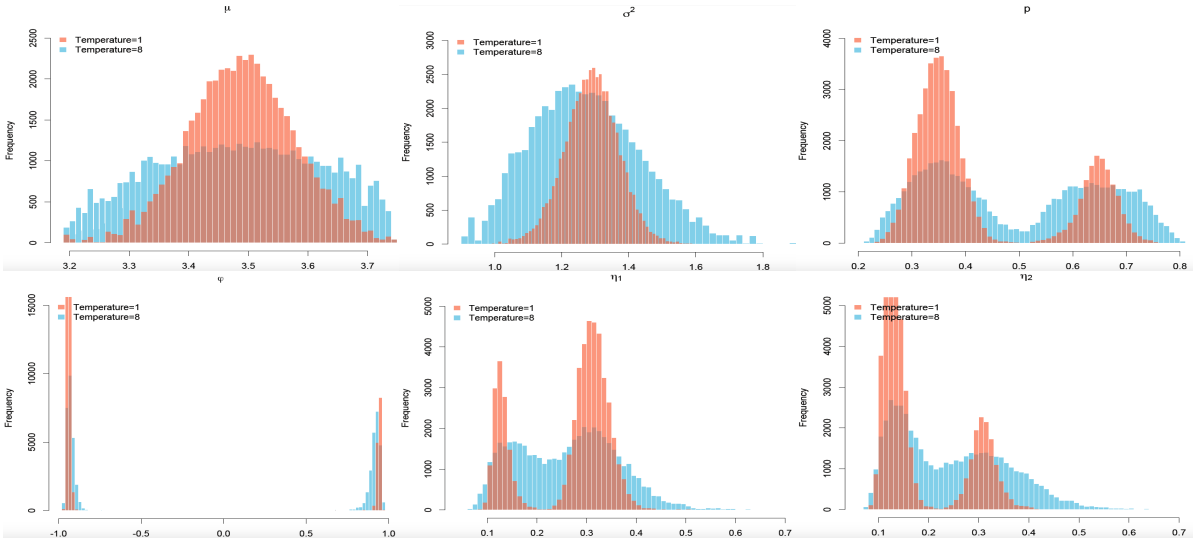

Fig 11: **Faithful dataset 4.2:** Posterior distribution of the mixture distribution parameters and comparison between the lowest and highest temperatures (target distribution and $f(x|\theta)^{1/8}$) of parallel tempering outputs based on $50,000$ iterations.

Example 5.2 We now implement parallel tempering for a mixture of $k = 3$ components applied to a benchmark dataset from Marin and Robert (2007). This dataset is derived from an image of a car license plate, and made of 2625 observations. In Marin and Robert (2007), a lack of label switching is observed when using a Gibbs sampler. Once again, this means each component can be estimated by its mean and standard deviation. The sample size is larger here and more likely to mixing problems. This is clearly exhibited in the six top plots of Figure 12 where the estimates provided for the three components are quite distinct. When implementing parallel tempering, the temperature increase stops when when all acceptance rates of swaps are above .4, meaning for this dataset 7 temperatures ranging from 1 to 64.

The six bottom plots of Figure 12 show that parallel tempering immensely improves the swaps between the posterior modes. The sample of ϖ's produced by parallel tempering visits a much larger region in $(0, 2\pi)$, when compared with the highly peaked output of the original MCMC output.

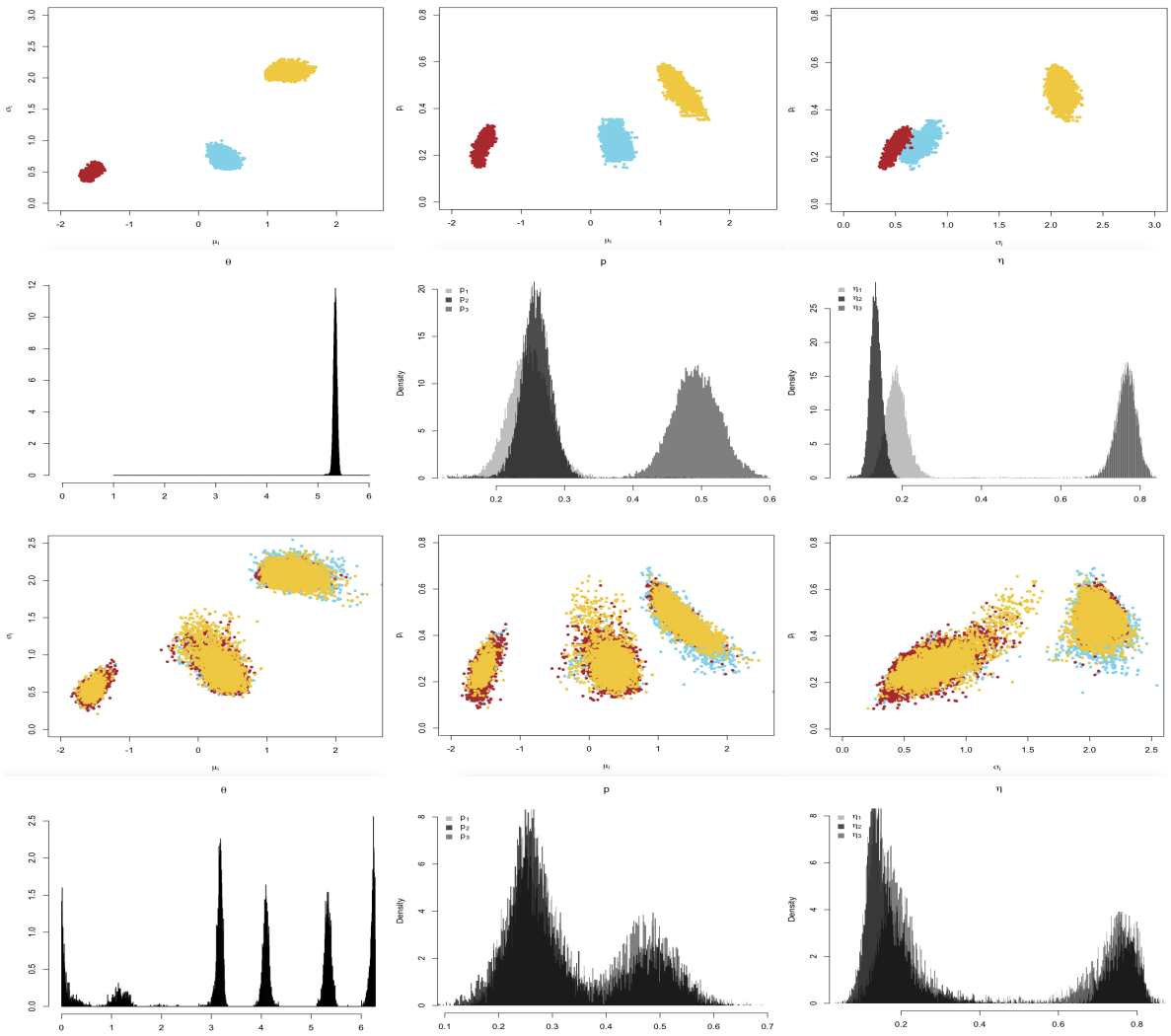

Fig 12: **Licence dataset (Example 5.2):** Comparison between Metropolis-within-Gibbs and parallel tempering outputs: The distributions of the samples of 10^4 last points and corresponding 2×2 plots.

The histograms in Figure 12 show that the posterior on p and η are now close to identical for each component. Two-dimensional plots also highlight this correct label switching behaviour, which demonstrates better mixing and convergence of the produced chain. ◀

6. CONCLUSION

This paper has introduced a new parametrisation for mixtures of location-scale models. By constraining the parameters in terms of the global mean and global variance of the mixture, i.e., by recognising the location-scale nature of such mixtures, it has been shown that the remaining parameters can be expressed as varying within a compact set. Therefore, it is possible to use a well-defined uniform prior on these parameters (as well as any proper prior) and we established that an improper prior of Jeffreys' type on the global mean and global variance returns a proper posterior distribution when handling at least two observations from the mixture. While the no-

tion of *non-informative* or *objective* prior is open to interpretations and sometimes controversies, we believe we have defined in this paper what can be considered as the first reference prior for mixture models.

We have demonstrated that relatively standard simulation algorithms are able to handle this new parametrisation and that they can manage the computing issues connected with label switching. In case of poor switching, we also established that parallel tempering can be easily implemented. As exhibited in the associated Ultimixt package, relabelling techniques are readily available.

While the extension to non-Gaussian cases with location-scale parameterisation (and beyond) is conceptually straightforward, considering this parameterisation in higher dimensions is delicate in terms of the covariance matrix. Indeed, even though we can easily set the global variance of the mixture as a parameter, reparameterising the component variances against this reference matrix remains an open question that we have not yet explored.

REFERENCES

Benaglia, T., D. Chauveau, D. Hunter, and D. Young (2009). mixtools: An r package for analyzing finite mixture models. *J. Statistical Software 32*(6), 1–29.

Berkhof, J., I. van Mechelen, and A. Gelman (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica 13*, 423–442.

Casella, G., K. Mengersen, C. Robert, and D. Titterington (2002). Perfect slice samplers for mixtures of distributions. *J. Royal Statist. Society Series B 64(4)*, 777–790.

Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *J. American Statist. Assoc. 95(3)*, 957–979.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. American Statist. Assoc. 90*, 1313–1321.

Diebolt, J. and C. Robert (1990). Estimation des paramètres d'un mélange par échantillonnage bayésien. *Notes aux Comptes–Rendus de l'Académie des Sciences I 311*, 653–658.

Diebolt, J. and C. Robert (1993). Discussion of "Bayesian computations via the Gibbs sampler" by A.F.M. Smith and G.O. Roberts. *J. Royal Statist. Society Series B 55*, 71–72.

Diebolt, J. and C. Robert (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Society Series B 56*, 363–375.

Escobar, M. and M. West (1995). Bayesian prediction and density estimation. *J. American Statist. Assoc. 90*, 577–588.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. American Statist. Assoc. 96*, 194–209.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal 7*(1), 143–167.

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

Gelman, A. and G. King (1990). Estimating the electoral consequences of legislative redistricting. *J. American Statist. Assoc. 85*(410), 274–282.

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Science 7*, 457–511.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works. *Comput. Statist. Data Analysis 51*(7), 3529–3550.

Geyer, C. J. (1991). Markov Chain Monte Carlo maximum likelihood. *Computing Science and Statistics 23*, 156–163.

Grazian, C. and C. Robert (2015). *Jeffreys priors for mixture estimation*, Volume 126, pp. 37–48. Springer Verlag.

Griffin, J. E. (2010). Default priors for density estimation with mixture models. *Bayesian Analysis 5*(1), 45–64.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Jasra, A., C. Holmes, and D. Stephens (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Science 20*(1), 50–67.

Jeffreys, H. (1939). *Theory of Probability* (First ed.). Oxford: The Clarendon Press.

Kamary, K. and K. Lee (2015). *Ultimixt: Bayesian Analysis of a Non-Informative Parametrization for Gaussian Mixture Distributions*. R package version 2.0.

Kamary, K., K. Mengersen, C. Robert, and J. Rousseau (2014). Testing hypotheses as a mixture estimation model. *arXiv:1214.4436*.

Lee, K., J.-M. Marin, K. Mengersen, and C. Robert (2009). Bayesian inference on mixtures of distributions. In

N. N. Sastry, M. Delampady, and B. Rajeev (Eds.), *Perspectives in Mathematical Sciences I: Probability and Statistics*, pp. 165–202. Singapore: World Scientific.

Marin, J. and C. Robert (2007). *Bayesian Core*. Springer-Verlag, New York.

Marin, J.-M., K. Mengersen, and C. Robert (2005). Bayesian modelling and inference on mixtures of distributions. In C. Rao and D. Dey (Eds.), *Handbook of Statistics*, Volume 25, pp. 459–507. Springer-Verlag, New York.

Marinari, E. and G. Parisi (1992). Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett. 19*, 451–458.

Mengersen, K. and C. Robert (1996). Testing for mixtures: A Bayesian entropic approach (with discussion). In J. Berger, J. Bernardo, A. Dawid, D. Lindley, and A. Smith (Eds.), *Bayesian Statistics 5*, pp. 255–276. Oxford University Press, Oxford.

Mengersen, K., C. Robert, and D. Titterington (2011). *Mixtures: Estimation and Applications*. John Wiley.

Miasojedow, B., E. Moulines, and M. Vihola (2013). An adaptive parallel tempering algorithm. *J. Comput. Graphical Statist. 22*(3), 649–664.

Neal, R. (1999). Erroneous results in "Marginal likelihood from the Gibbs output". Technical report, University of Toronto.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing 6*(4), 353–366.

O'Hagan, A. (1994). *Bayesian Inference*. Number 2B in Kendall's Advanced Theory of Statistics. New York: Chapman and Hall.

Richardson, S. and P. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B 59*, 731–792.

Robert, C. and M. Titterington (1998). Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing 8*(2), 145–158.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Applied Probability 7*(1), 110–120.

Roberts, G. O. and S. J. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Science 16*(4), 351–367.

Roberts, G. O. and S. J. Rosenthal (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist. 18*(2), 349 –367.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. American Statist. Assoc. 85*, 617–624.

Rosenthal, S. J. (2011). Optimal proposal distributions and adaptive MCMC. In G. J. S. Brooks, A. Gelman and X. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, pp. 93–112. CRC Press.

Rossi, P. and R. McCulloch (2010). Bayesm: Bayesian inference for marketing/micro-econometrics. *R package version 2*, 357–365.

Rousseau, J. and J. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. Royal Statist. Society Series B 73*(5), 689–710.

Sayyareh, A. (2011). A new upper bound for Kullback-Leibler divergence. *Applied Mathematical Sciences 5*(67), 3303–3317.

Stephens, M. (2000). Dealing with label switching in mixture models. *J. Royal Statist. Society Series B 62(4)*, 795–809.

Tanner, M. and W. Wong (1987). The calculation of posterior distributions by data augmentation. *J. American Statist. Assoc. 82*, 528–550.

Wasserman, L. (1999). Asymptotic inference for mixture models by using data-dependent priors. *J. Royal Statist. Society Series B 61*(1), 159–180.