Jeffreys priors for mixture estimation: properties and alternatives

Clara Grazian*

University of Oxford

Christian P. Robert**

Université Paris-Dauphine, University of Warwick and CREST

Abstract

While Jeffreys priors usually are well-defined for the parameters of mixtures of distributions, they are not available in closed form. Furthermore, they often are improper priors. Hence, they have never been used to draw inference on the mixture parameters. This paper studies the implementation and the properties of Jeffreys priors in several mixture settings and shows that the associated posterior distributions most often are improper. Nevertheless, the Jeffreys prior for the mixture weights conditionally on the parameters of the mixture components will be shown to have the property of conservativeness with respect to the number of components, in case of overfitted mixture and it can be therefore used as a default priors in this context.

Keywords: *Noninformative prior; mixture of distributions; Bayesian analysis; improper prior*

1. Introduction

Bayesian inference in mixtures of distributions has been studied quite extensively in the literature. See, e.g., [20] and [10] for book-long references and [19]

*Corresponding Author: clara.grazian@ndm.ox.ac.uk, Nuffield Department of Medicine (Level 5), Room 7400, Headley Way, OX3 9DU, Oxford, UK. Tel: 0044 (0)1865 220 257.
**xian@ceremade.dauphine.fr

for one among many surveys. From a Bayesian perspective, one of the several difficulties with this type of distribution,

$$\sum_{\ell=1}^{k} p_\ell \, f_\ell(x|\theta_\ell)\,, \quad \sum_{\ell=1}^{k} p_\ell = 1\,, \tag{1}$$

is that its ill-defined nature (non-identifiability, multimodality, unbounded likelihood, etc.) leads to restrictive prior modelling since most improper priors are not acceptable. This is due in particular to the feature that a sample from (1) may contain no subset from one of the k components $f(\cdot|\theta_\ell)$ (see. e.g., 38). Albeit the probability of such an event is decreasing quickly to zero as the sample size grows, it nonetheless prevents the use of independent improper priors, unless such events are prohibited [7]. Similarly, the exchangeable nature of the components often induces both multimodality in the posterior distribution and convergence difficulties as exemplified by the *label switching* phenomenon that is now quite well-documented [5, 37, 15, 10, 11, 23]. This feature is characterized by a lack of symmetry in the outcome of a Monte Carlo Markov chain (MCMC) algorithm, in that the posterior density is exchangeable in the components of the mixture but the MCMC sample does not exhibit this symmetry. In addition, most MCMC samplers do not concentrate around a single mode of the posterior density, partly exploring several modes, which makes the construction of Bayes estimators of the components much harder.

When specifying a prior over the parameters of (1), it is therefore quite delicate to produce a manageable and sensible non-informative version and some have argued against using non-informative priors in this setting (for example, [20] argues that it is impossible to obtain proper posterior distributions from fully noninformative priors), on the basis that mixture models are ill-defined objects that require informative priors to give a meaning to the notion of a component of (1). For instance, the distance between two components needs to be bounded from below to avoid repeating the same component indefinitely. Alternatively, the components all need to be informed by the data, as exemplified in [7] who imposed a completion scheme (i.e., a joint model on both parame-

ters and latent variables) such that all components were allocated at least two observations, thereby ensuring that the (truncated) posterior was well-defined. [39] proved ten years later that this truncation led to consistent estimators and moreover that only this type of priors could produce consistency. While the constraint on the allocations is not fully compatible with the i.i.d. representation of a mixture model, it naturally expresses a modelling requirement that all components have a meaning in terms of the data, namely that all components genuinely contributed to generating a part of the data. This translates as a form of weak prior information on how much one trusts the model and how meaningful each component is on its own (by opposition with the possibility of adding meaningless artificial extra-components with almost zero weights or almost identical parameters).

While we do not seek Jeffreys priors as the ultimate prior modelling for non-informative settings, being altogether convinced of the lack of unique reference priors [27, 30], we think it is nonetheless worthwhile to study the performances of those priors in the setting of mixtures in order to determine if indeed they can provide a version of reference priors and if they are at least well-defined in such settings. We will show that only in very specific situations the Jeffreys prior provides reasonable inference.

In Section 2 we provide a formal characterisation of properness of the posterior distribution for the parameters of a mixture model, in particular with Gaussian components, when a Jeffreys prior is used for them. In Section 3 we will analyze the properness of the Jeffreys prior and of the related posterior distribution: only when the weights of the components (which are defined in a compact space) are the only unknown parameters it turns out that the Jeffreys prior (and so the relative posterior) is proper; on the other hand, when the other parameters are unknown, the Jeffreys prior will be proved to be improper and in only one situation it provides a proper posterior distribution. In Section 4 we present a way to realize a noninformative analysis of mixture models, in particular we propose to use the Jeffreys prior as a default prior in case of overfitted mixtures and introduce improper priors for at least some parameters.

The default proposal of Section 4 will be tested on several simulation studies in Section 5 and several real examples in Section 6, on both well known datasets in the mixture literature and a new dataset. Section 7 concludes the paper.

2. Jeffreys priors for mixture models

We recall that the Jeffreys prior was introduced by [16] as a default prior based on the Fisher information matrix

$$\pi^{\mathrm{J}}(\theta) \propto |I(\theta)|^{1/2} = \left| -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}} \log g(X;\theta) \right] \right|^{1/2}, \tag{2}$$

whenever the later is well-defined; $I(\cdot)$ stands for the expected Fisher information matrix and the symbol $|\cdot|$ denotes the determinant. Although the prior is endowed with some frequentist properties like matching and asymptotic minimal information [27, Chapter 3], it does not constitute the ultimate answer to the selection of prior distributions in non-informative settings and there exist many alternatives such as reference priors [2], maximum entropy priors [26], matching priors [12], and other proposals [18]. In most settings Jeffreys priors are improper, which may explain for their conspicuous absence in the domain of mixture estimation, since the latter prohibits the use of independent improper priors by allowing any subset of components to go "empty" with positive probability. That is, the likelihood of a mixture model can always be decomposed as a sum over all possible partitions of the data into k groups at most, where k is the number of components of the mixture. This means that there are terms in this sum where no observation from the sample brings any amount of information about the parameters of a specific component.

Approximations of the Jeffreys prior in the setting of mixtures can be found, e.g., in [9], where the authors revert to independent Jeffreys priors on the components of the mixture. This induces the same negative side-effect as with other independent priors, namely an impossibility to handle improper priors. [36] provides a closed-form expression for the Jeffreys prior for a location-scale mixture

with two components. The family of distributions considered in [36] is

$$\frac{2\epsilon}{\sigma_1} f\left(\frac{x-\mu}{\sigma_1}\right) \mathbb{I}_{x<\mu} + \frac{2(1-\epsilon)}{\sigma_2} f\left(\frac{x-\mu}{\sigma_2}\right) \mathbb{I}_{x>\mu}$$

(which thus hardly qualifies as a mixture, due to the orthogonality in the supports of both components that allows to identify which component each observation is issued from). The factor 2 in the fraction is due to the assumption of symmetry around zero for the density f. For this specific model, if we impose that the weight ϵ is a function of the variance parameters, $\epsilon = {}^{\sigma_1}/_{\sigma_1+\sigma_2}$, the Jeffreys prior is given by

$$\pi(\mu, \sigma_1, \sigma_2) \propto {}^{1}/_{\sigma_1\sigma_2\{\sigma_1+\sigma_2\}}.$$

However, in this setting, [36] demonstrates that the posterior associated with the (regular) Jeffreys prior is improper, hence not relevant for conducting inference. [36] also considers alternatives to the genuine Jeffreys prior, either by reducing the range or even the number of parameters, or by building a product of conditional priors. They further consider so-called non-objective priors that are only relevant to the specific case of the above mixture.

Another obvious explanation for the absence of Jeffreys priors is computational, namely the closed-form derivation of the Fisher information matrix is analytically unavailable. The reason is that the generic $[j,h]$-th element, with $j, h \in \{1, \cdots, k\}$, of the Fisher information matrix for mixture models is an integral of the form

$$-\int_{\mathcal{X}} \frac{\partial^2 \log\left[\sum_{\ell=1}^{k} p_\ell \, f_\ell(x|\theta_\ell)\right]}{\partial\theta_j \partial\theta_h} \left[\sum_{\ell=1}^{k} p_\ell \, f_\ell(x|\theta_\ell)\right]^{-1} dx \tag{3}$$

(in the special case of component densities with a univariate parameter) which cannot be computed analytically. Since these are unidimensional integrals, we derive an approximation of the elements of the Fisher information matrix based on Riemann sums. The resulting computational expense is of order $\mathrm{O}(b^2)$ if b is the total number of (independent) parameters. Since the elements of the information matrix usually are ratios between the component densities and the

mixture density, there may be difficulties with non-probabilistic methods of integration.

3. Characterization of the Jeffreys priors for mixture models and respective posteriors

Unsurprisingly, most Jeffreys priors associated with mixture models are improper, the exception being when only the weights of the mixture are unknown, as already demonstrated in [3].

We will characterize properness and improperness of Jeffreys priors and derived posteriors, when some or all of the parameters of distributions from location-scale families are unknown. These results are analytically established; the behavior of the Jeffreys prior and of the deriving posterior has also been studied through simulations, with sufficiently large Monte Carlo experiments (see Section 5). The following results are often presented for Gaussian mixture models, anyway, the Jeffreys prior has a behavior common to all the location-scale families; therefore the results may be generalized to any location-scale family.

3.1. Weights of mixture unknown

A representation of the Jeffreys prior and the derived posterior distribution for the weights of a three-component mixture model is given in Figure 1: the prior distribution is much more concentrated around extreme values in the support, i.e., it is a prior distribution conservative in the number of important components.

Lemma 3.1. *When the weights p_i are the only unknown parameters in* (1)*, the corresponding Jeffreys prior is proper.*

Proof. The generic element of the Fisher information matrix $I(p)$ of the mixture model (1) when the weights are the only unknown parameters is (for $j, h = \{1, \ldots, k-1\}$)

$$\int_{\mathcal{X}} \frac{(f_j(x) - f_k(x))(f_h(x) - f_k(x))}{\sum_{\ell=1}^{k} p_\ell f_\ell(x)} \mathrm{d}x \tag{4}$$

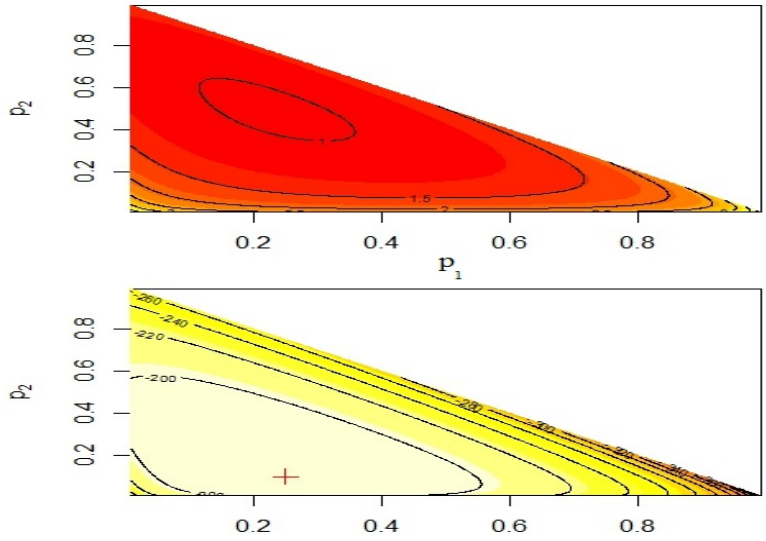

Figure 1: Approximations (on a grid of values) of the Jeffreys prior (on the log-scale) when only the weights of a Gaussian mixture model with 3 components are unknown (on the top) and of the derived posterior distribution (with known means equal to -1, 0 and 2 respectively and known standard devitations equal to 1, 5 and 0.5 respectively). The red cross represents the true values.

when we consider the parametrization in $(p_1, \ldots, p_{k-1})$, with

$$p_k = 1 - p_1 - \cdots - p_{k-1}\,.$$

Consider now a data augmented model, where a latent variable describing the allocations of each observation to the particular component is introduced. In other words, a latent variable z_i is considered such that $z_i = (0 \cdots 1 \cdots 0)$, where $z_{i\ell} = 1$ in the ℓ-th position of the vector if x_i has been generated from the ℓ-th components, for $i = 1, \cdots, n$ where n is the sample size and $\ell = 1, \cdots, k$. Therefore, $z = (z_1, \ldots, z_n)$ is a multinomial variable for k possible outcomes such that

$$\begin{aligned} g(x, z|\theta, p) &= g(x|z, \theta, p)g(z|\theta, p) = \prod_{i=1}^{n} g(x_i|z_i, \theta, p)g(z_i|\theta, p) \\ &= \prod_{i=1}^{n}\prod_{\ell=1}^{k} \left[f_\ell(x_i|\theta_\ell)p_\ell\right]^{\mathbb{I}_{[z_{i,\ell}]=1}} = \prod_{\ell=1}^{k}\left[\prod_{i:z_{i,\ell}=1} f_l(x_i|\theta_\ell)\right]\left[\prod_{\ell=1}^{k} p_\ell^{n_\ell}\right] \quad (5) \end{aligned}$$

where $\mathbb{I}_{[z_{i,\ell}=1]}$ is the indicator function that $z_{i,\ell} = 1$ and n_ℓ is the number of allocations to the ℓ-th component. For an extensive review of the techniques of data augmentation in the case of mixture models one may refer to [10].

Equation (6) shows that the likelihood function is separable for θ and p and that the second part is multinomial. Therefore, when looking for the Jeffreys prior for the weights of a complete (data-augmented) mixture model, the elements of the Fisher information matrix are

$$\begin{aligned} -\mathbb{E}\left[\frac{\partial^2}{\partial p_\ell^2}\log g(x, z|\theta, p)\right] &= -\frac{n_\ell n p_\ell}{p_\ell^2} = \frac{c}{p_\ell} \\ -\mathbb{E}\left[\frac{\partial^2}{\partial p_\ell \partial p_j}\log g(x, z|\theta, p)\right] &= 0 \end{aligned}$$

leading to the usual Jeffreys prior associated to the multinomial model, a Dirichlet distribution $\mathcal{D}ir(\frac{1}{2}, \cdots, \frac{1}{2})$.

The above only applies to the artificial case when the allocations z_i are known. When they are unknown, it is easy to see that the log-likelihood function becomes

$$\log g(x|\theta, p) = \log g(x, z|\theta, p) - \sum_{i=1}^{n}\sum_{\ell=1}^{k} \mathbb{I}_{[z_{i,\ell}=1]} \log p(z_{i,\ell} = 1|x_i, \theta, p) \quad (6)$$

where the second term on the right side of the equation represents the loss of information compared to the data-augmented likelihood function. Define the expected Fisher information matrix for model (6) (when only the weights are unknown) as $I^{data-aug}(p,\theta)$. Therefore, since the difference between both matrices is positive definite, this implies that

$$\begin{aligned}\det(I(p)) &\leq \det(I^{data-aug}(p)) \\ [\det(I(p))]^{1/2} &\leq \left[\det(I^{data-aug}(p))\right]^{1/2} \\ \pi_J(p) &\leq \pi_J^{data-aug}(p)\end{aligned}$$

This results shows that the Jeffreys prior on the weights of a mixture model when allocations are unknown is proper since bounded by the Jeffreys prior $\mathcal{D}ir(\frac{1}{2},\cdots,\frac{1}{2})$ for the complete model.

As a particular case, when all the mixands converge to the same distribution, each of the elements of the form (4) tends to

$$\int_{\mathcal{X}} \frac{(f_j(x)-f_k(x))(f_\ell(x)-f_k(x))}{f_j(x)}\mathrm{d}x$$

which does not depend on p. Therefore, in this case, the determinant of the deriving Fisher information matrix is constant in $p=(p_1,\cdots,p_k)$ and the resulting Jeffreys prior is uniform on the k-dimensional simplex.

□

We note that this result is a generalization to a k-component mixture of the prior derived in [3] for $k=2$ (however, these authors derive the reference prior for the limiting cases when all the components have pairwise disjoint supports and when all the components converge to the same distribution). This reasoning led [3] to conclude that the usual $\mathcal{D}(\lambda_1,\ldots,\lambda_k)$ Dirichlet prior with $\lambda_\ell \in [1/2,1]$ for $\forall \ell = 1,\cdots,k$ seems to be a reasonable approximation. They also prove that the Jeffreys prior for the weights p_ℓ is convex, with an argument based on the sign of the second derivative.

It is important to stress that, in a mixture model setting, it is usual to saturate the model when the number of components is not surely known a

priori and consider a large number of components k. The main difficulty in this setting is non-identifiability, in particular the rate of estimation for the satured model is much slower than the standard $1/\sqrt{n}$. [35] have studied the effect of a prior distribution on the weights of a general mixture on regularizing the posterior distribution, i.e. consistency to a single configuration of the reduced parameter space. This is achievable with a prior which allows to empty the extra-components or to merge the existing ones. In particular, [35] propose a Dirichlet prior distribution, with parameters $\lambda_1, \cdots, \lambda_k$ smaller than $r/2$ (where r is the dimension of θ_ℓ) to empty the extra-components or larger than $r/2$ to merge the extra-components. However, the choice of λ_j $(j = 1, \cdots, k)$ is quite influential for finite sample sizes. The configuration studied in the proof of Lemma 3.1 is compatible with the Dirichlet configuration of the prior proposed by [35]. This is an important property of the Jeffreys prior, since it makes the prior conservative in the number of the components. Namely, one can asymptotically identify the components that are artificially added to the model but have no meaning for the data. Moreover, it offers an automatic choice, on the contrary of the Dirichlet prior where the hyper-parameters have to been chosen.

The shape of the Jeffreys prior for the weights of a mixture model depends on the type of the components: see Appendix A of the Supplementary Material for a discussion. The marginal Jeffreys prior for the weight of one component is more concentrated around one if that component is more informative in terms of Fisher information matrix: for example, if we consider a two-component mixture model with a Gaussian and a Student t component, the Jeffreys prior for the weights will be more symmetric as the number of degrees of freedom of the Student t increases.

3.2. Weights, location and scale parameters of a mixture model unknown

In this Section we will consider mixtures of location-scale distributions. If the components of the mixture model (1) are distributions from a location-scale family and the location or scale parameters of the mixture components

are unknown, this turns the mixture itself into a location-scale model:

$$p_1 f_1(x|\mu,\tau) + \sum_{\ell=2}^{k} p_\ell f_\ell(\frac{a_\ell + x}{b_\ell}|\mu,\tau,a_\ell,b_\ell). \tag{7}$$

As a result, model (1) may be reparametrized following [21], in the case of Gaussian components

$$p\mathcal{N}(\mu,\tau^2) + (1-p)\mathcal{N}(\mu+\tau\delta,\tau^2\sigma^2) \tag{8}$$

namely using a reference location μ and a reference scale τ (which may be, for instance, the location and scale of a specific component). Equation (8) may be generalized to the case of k components as

$$\begin{aligned} &p\mathcal{N}(\mu,\tau^2) + \sum_{\ell=1}^{k-2}(1-p)(1-q_1)\cdots(1-q_{\ell-1})q_\ell \\ &\quad \cdot \mathcal{N}(\mu+\tau\theta_1+\cdots+\tau\cdots\sigma_{\ell-1}\theta_\ell, \tau^2\sigma_1^2\cdots\sigma_\ell^2) + \\ &+ (1-p)(1-q_1)\cdots(1-q_{k-2}) \\ &\quad \cdot \mathcal{N}(\mu+\tau\theta_1+\cdots+\tau\cdots\sigma_{k-2}\theta_{k-1}, \tau^2\sigma_1^2\cdots\sigma_{k-1}^2). \end{aligned}$$

Since the mixture model is a location-scale model, the Jeffreys prior is as in the following Lemma (see also Robert [27, Chapter 3]).

Lemma 3.2. *When the parameters of a location-scale mixture model are unknown, the Jeffreys prior is improper, constant in μ and powered as $\tau^{-d/2}$, where d is the total number of unknown parameters of the components (i.e. excluding the weights).*

An new version of the proof, never presented before, is available in Appendix B of the Supplementary Material, while the characterization of the Jeffreys prior for δ is given in Appendix C.

We now derive analytical characterizations of the posterior distributions associated with the Jeffreys priors for mixture models.

Consider, first, the case where only the location parameters of a mixture model are unknown.

There is a substantial difference between the cases where $k = 2$ or $k > 2$.

Lemma 3.3. *When $k = 2$, the posterior distribution derived from the Jeffreys prior when only the location parameters of model* (7) *are unknown is proper.*

The complete proof of lemma 3.3 is given in Appendix D of the Supplementary Material. Here it is worth noticing that the properness of the posterior distribution in the context of Lemma 3.3 depends on the representation of the mixture model as a location-scale distribution, where the second component is defined with respect to a reference component: if we focus the attention on the part of the likelihood depending only on the second component, even if the prior is constant with respect to the difference between the location parameters δ as $\delta \to \pm\infty$, the likelihood depends on δ as $\exp(-\frac{n-1}{2}\delta^2)$ and therefore the behavior of the posterior distribution is convergent.

Figure 2 shows an approximation of the Jeffreys prior for the location parameters of a two-component Gaussian mixture model on a grid of values and confirms that the prior is constant on the difference between the means and takes higher and higher values as the difference between them increases, while the posterior distribution, even if showing the classical multimodal nature [5], seems to concentrate around the true modes. It also appears to be perfectly symmetric because the other parameters (weights and standard deviations) have been fixed as identical.

The same proof cannot be extended to the general case of k components, because the location parameters are defined as several distances from the reference location parameter: if we again focus the attention on the part of the likelihood depending on the second component, the integral with respect to δ_2 converges, however the prior is constant with respect to any other δ_j $(j = 3, \cdots, k)$ as $\delta_j \to \pm\infty$ and the integral does not converge with respect to the other differences. Then the following Lemma holds (the formal proof is available in Appendix E).

Lemma 3.4. *When $k > 2$, the posterior distribution derived from the Jeffreys prior is improper when only the location parameters of model* (7) *are unknown.*

This result confirms the idea that each part of the likelihood gives infor-

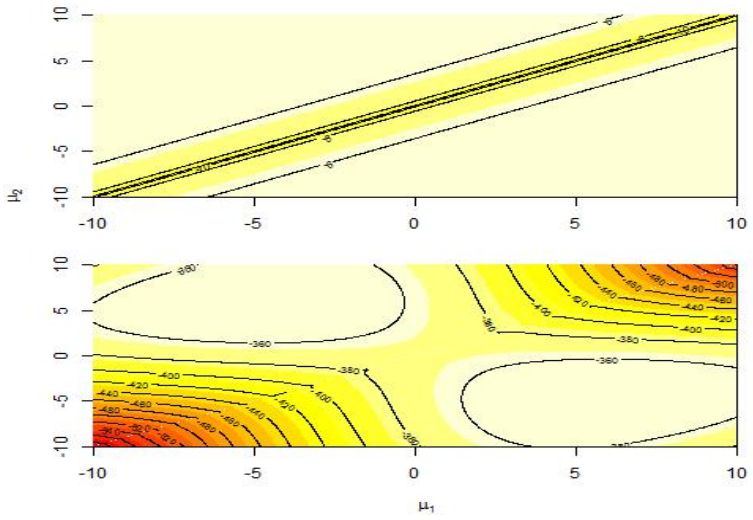

Figure 2: Approximations (on a grid of values) of the Jeffreys prior (on the log-scale) when only the means of a Gaussian mixture model with two components are unknown above and of the derived posterior distribution (with known weights both equal to 0.5 and known standard deviations both equal to 5) below.

mation about at most the difference between the locations of the respective components and the reference location, but not on the locations of the other components.

We can now consider the case where all the parameters of (7) are unknown.

Theorem 3.1. *The posterior distribution of the parameters of a mixture model with location-scale components derived from the Jeffreys prior when all parameters of model* (7) *are unknown is improper.*

The proof is available in Appendix F of the Supplementary Material.

4. A noninformative alternative to Jeffreys prior

The information brought by the Jeffreys prior or lack thereof does not seem to be enough to conduct inference in the case of mixture models. The computation of the determinant creates a dependence between the elements of the Fisher information matrix in the definition of the prior distribution which makes it difficult to find and justify moderate modifications of this prior that would lead to a proper posterior distribution. For example, using a proper prior for part of the scale parameters and the Jeffreys prior conditionally on them does not avoid impropriety, as it is shown Appendix G of the Supplementary Material.

The literature covers attempts to define priors that add a small amount of information that is sufficient to conduct the statistical analysis without overwhelming the information contained in the data. Some of these are related to the computational issues in estimating the parameters of mixture models, as in the approach of [4], who finds a way to use perfect slice sampler by focusing on components in the exponential family and conjugate priors. A characteristic example is given by [25], who proposes weakly informative priors, which are data-dependent (or empirical Bayes) and are represented by flat normal priors over an interval corresponding to the range of the data. Nevertheless, since mixture models belong to the class of ill-posed problems, the influence of a proper prior over the resulting inference is difficult to assess.

Another solution found in [21] proceeds through the reparametrization (8) and introduces a reference component that allows for improper priors. This approach then envisions the other parameters as departures from the reference and ties them together by considering each parameter θ_ℓ as a perturbation of the parameter of the previous component $\theta_{\ell-1}$. This perspective is justified by the argument that the $(\ell-1)$-th component may not be informative enough to absorb all the variability in the data. For instance, a three-component mixture model gets rewritten as

$$p\mathcal{N}(\mu,\tau^2) + (1-p)q\mathcal{N}(\mu+\tau\theta,\tau^2\sigma_1^2) \\ + (1-p)(1-q)\mathcal{N}(\mu+\tau\theta+\tau\sigma\epsilon,\tau^2\sigma_1^2\sigma_2^2)$$

where one can impose the constraint $1 \geq \sigma_1 \geq \sigma_2$ for identifiability reasons. Under this representation, it is possible to use an improper prior on the global location-scale parameter (μ,τ), while proper priors must be applied to the remaining parameters. This reparametrization has been used also for exponential components by [13] and Poisson components by [32]. Moreover, [34] proposes a Markov prior which follows the same reasoning of dependence between the parameters for Gaussian components, where each parameter is again a perturbation of the parameter of the previous component $\theta_{\ell-1}$. [17] also proposes a reparametrization of location-scale mixtures based on invariance that allows for weakly informative priors.

On one hand, this representation suggests to define a global location-scale parameter in a more implicit way, via a hierarchical model that considers more levels in the analysis and choose noninformative priors at the last level in the hierarchy.

On the other hand, we believe that an essential feature of a default prior is that it should let the analysis be able to identify the correct number of meaningful components, in particular in the standard case where an overfitted mixture is assumed because the a priori information on the number of components is weak.

We thus propose a prior scenario which combines both the hierarchical rep-

resentation and the conservativeness property in terms of components.

More precisely, consider the Gaussian mixture model (1)

$$g(x|\boldsymbol{\theta}) = \sum_{\ell=1}^{k} p_i \mathcal{N}(x|\mu_\ell, \sigma_\ell). \tag{9}$$

The parameters of each component may be considered as related in some way; for example, the observations induce a reasonable range, which makes it highly improbable to face very different means in the above Gaussian mixture model. A similar argument may be used for the standard deviations.

Therefore, at the second level of the hierarchical model, we may write

$$\begin{aligned} \mu_\ell &\overset{iid}{\sim} \mathcal{N}(\mu_0, \zeta_0) \\ \sigma_\ell &\overset{iid}{\sim} \frac{1}{2}\mathcal{U}(0, \zeta_0) + \frac{1}{2}\frac{1}{\mathcal{U}(0, \zeta_0)} \\ p|\mu, \sigma &\sim \pi^J(p|\mu, \sigma) \end{aligned} \tag{10}$$

which indicates that the location parameters vary between components, but are likely to be close, and that the scale parameters may be smaller or larger than ζ_0; we have decided to define both μ_ℓ and σ_ℓ as depending on hyperparameter ζ_0 without loss of generality, as one may notice by analysing mean and variance of the random variables; this representation allows the application of the MCMC scheme proposed in [31] which allows a better mixing of the chains. The mixture weights are given the prior distribution $\pi^J(p|\mu, \sigma)$ which is the Jeffreys prior for the weights, conditional on the location and scale parameters, given in Section 3.1; this choice makes use of the conservative property of the Jeffreys prior for the weights which is essential in the case of miss-specification of the number of components.

At the third level of the hierarchical model, the prior may be noninformative:

$$\pi(\mu_0, \zeta_0) \propto \frac{1}{\zeta_0}. \tag{11}$$

As in [21] the parameters in the mixture model are considered tied together; on the other hand, this feature is not obtained via a constrained representation of the mixture model itself, but via a hierarchy in the definition of the model and the parameters.

Theorem 4.1. *The posterior distribution derived from the hierarchical representation of the Gaussian mixture model associated with* (9)*,* (10) *and* (11) *is proper.*

The proof of Theorem 4.1 is available in Appendix H of the Supplementary Material.

As a side remark, even if Theorem 4.1 is stated for Gaussian mixture models, it may be extended to other location-scale distributions. Section 6 will present an example with log-normal components, Section 6.1 with Gumbel components. However it cannot be generalized to any location-scale distribution.

This hierarchical version of the mixture model presents some advantages; in particular, the Jeffreys prior used for the weights is conservative in terms of number of components in the case of misspecification. We remind that when the number of components is not known, it is usual in practice to fix a model with a high number of components (if one wants to avoid a nonparametric analysis), therefore it is essential that the posterior distribution gives hints on the right k. This feature of the Jeffreys prior allow the experimenter to do so in a noninformative way. More precisely, this hierchical prior respect the Assumption 5 of [35].

5. Simulation Study

In this Section we present the results of several simulations studies we conduct to support the theoretical results presented so far. The results of additional simulations are given in Appendix G and H of the Supplementary Material.

As a remark, integrals of the form (3) need to be approximated, as mentioned in Section 2. There are numerical issue here. We decided to use Riemann sums

(with 550 points) when the component standard deviations are sufficiently large, as they produce stable results, and Monte Carlo integration (with sample sizes of 1500) when they are small. In the latter case, the variability of MCMC results seems to decrease as σ_i approaches 0. See the Supplementary Material for a detailed description of these computational issues.

We can analyse the property of conservativeness in overfitted mixtures through simulations, by using the hierarchical prior proposed in Section 4. We consider a very simple example to illustrate this theoretical result. Suppose we want to fit a two-component Gaussian mixture model with weights p and $1-p$ and parameters unknown to a sample of data $\mathbf{x} = \{x_1, \cdots, x_n\}$ generated from a standard normal distribution $\mathcal{N}(0,1)$. We computed the posterior distribution for $M = 20$ replications of samples of size $n = (50, 100, 500, 1000, 5000, 10000)$. Figure 3 shows that the posterior means of p increases to 1 as n increases.

We have also considered a more complicated situation, where we want to fit a model with an increasing number of components ($k = (2, 3, 4, 5)$) to a data set $\mathbf{x} = \{x_1, \cdots, x_n\}$ generated from a two-component mixture model

$$0.5\mathcal{N}(-3, 1) + 0.5\mathcal{N}(3, 1). \tag{12}$$

Figures 4 and 5 show the boxplots for the posterior means of the weights obtained through $M = 20$ replications of the experiment, with a correct ($k = 2$) or a misspecified ($k = (3, 4, 5)$) model. It is clear that as the number of components increases, the additional weights are estimated by smaller and smaller values as the sample size increases. It is evident that the variability of the estimates (in repetitions of the experiment) is smaller when an exact number of components is assumed; however, in every case, the Bayesian analysis based on the Jeffreys prior is able to identify the right number of components. The higher variability in estimating the weights is reflected in the fact that, as the number of components increases, the estimated (and the predictive) densities are less and less smooth, nevertheless this feature is mitigated as the sample size increases, see Appendix H in the Supplementary Material.

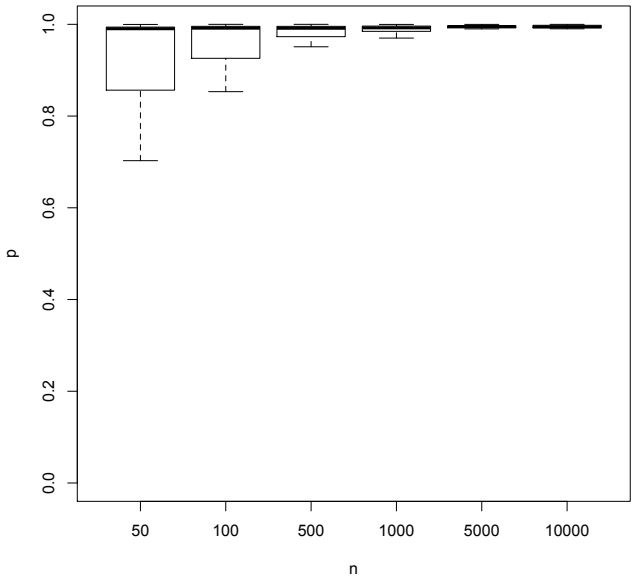

Figure 3: Boxplots of posterior means of the largest weight p, with the hierarchical prior on the parameters, in particular a conditional Jeffreys prior on the weights, for sample sizes $n = 50, 100, 500, 1000, 5000, 10000$.

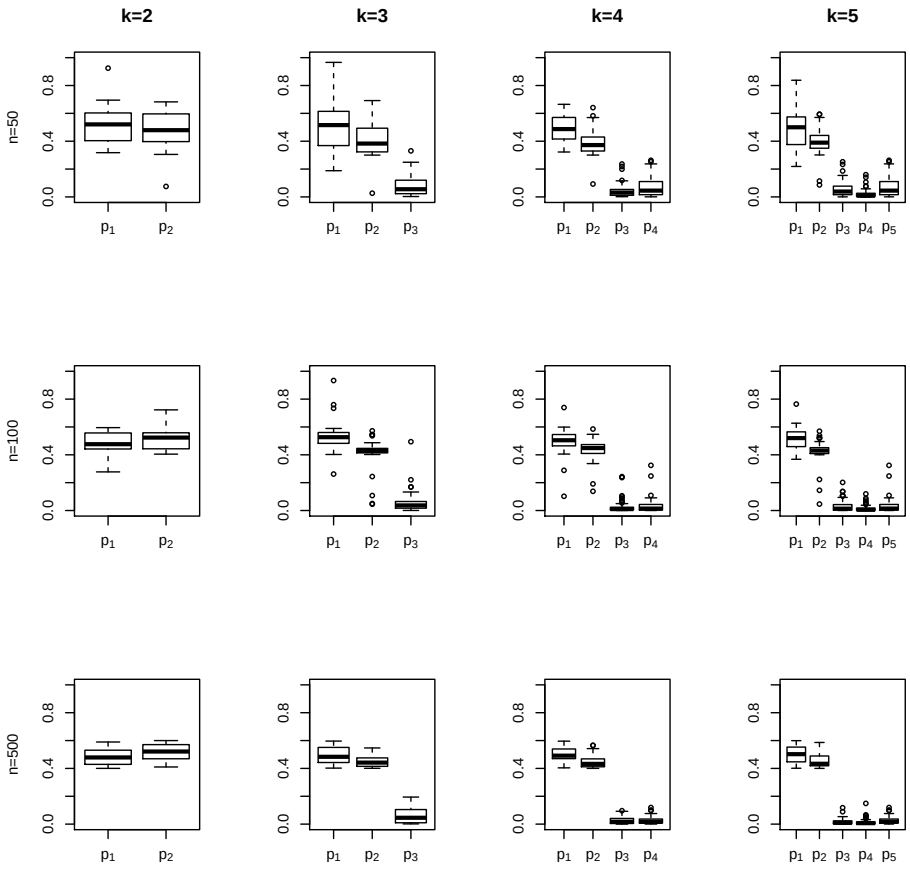

Figure 4: Boxplots of posterior means of the weights $\mathbf{p}$, with the hierarchical prior on the parameters, in particular a conditional Jeffreys prior on the weights, for sample sizes $n = (50, 100, 500)$ and with models with $k = (2, 3, 4, 5)$ components.

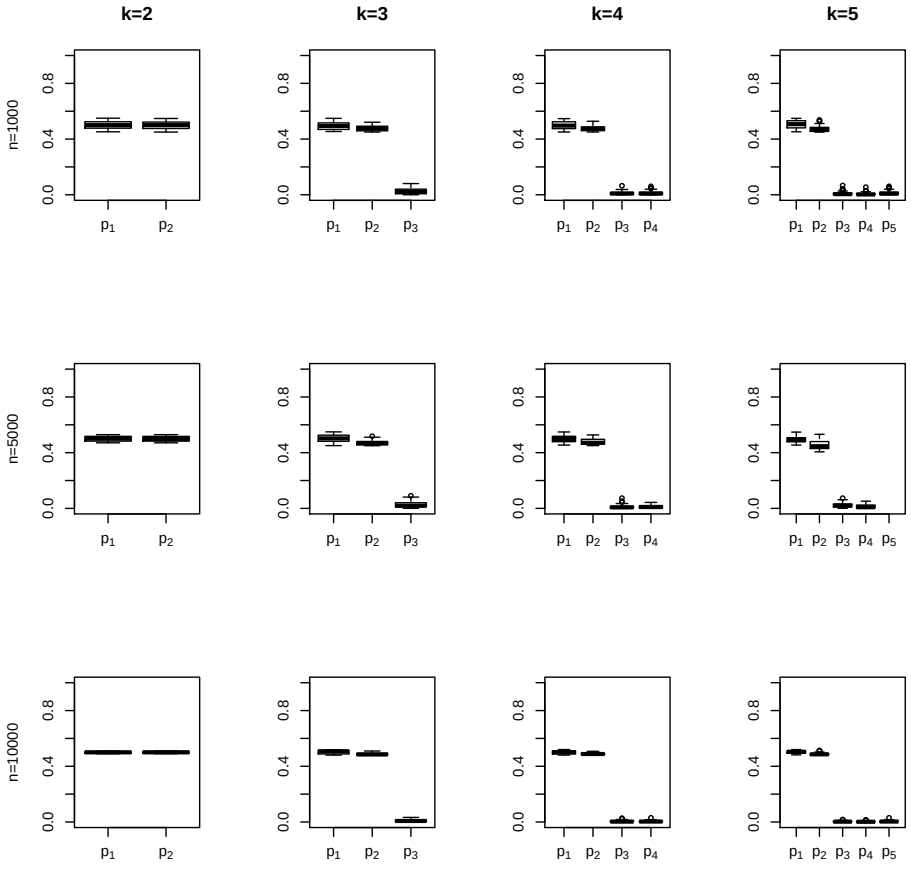

Figure 5: As in Figure 4, for sample sizes $n = (1000, 5000, 10000)$.

6. Illustrations

In this Section we will analyse the performance of the approach proposed in Section 4 in three datasets so well-known in the literature of mixture models that they can be taken as benchmarks and in a new dataset we propose here for the first time. In order to better present this new dataset, the analysis of it is presented separately.

The first dataset contains data about the velocity (in km per second) of 82 galaxies in the Corona Borealis region. The goal of this analysis is to understand the number of stellar populations, in order to support a particular theory of the formation of the Galaxy. The Galaxy dataset has been investigated by several authors, including [25], [24], [8] and [33] among others.

The galaxies velocities are considered as random variables distributed according to a mixture of k normal distributions. The evaluation of the number of components has proved to be delicate, with estimates from 3 in [34] to 5 in [25] and 7 in [8].

We have assumed a ten-component mixture model and check whether or not the hierarchical approach that uses the conditional Jeffreys prior on the weights of the mixture model manages to identify a smaller number of significant components. The results are available in Figure 6 and Table 1. The algorithm identifies 5 components with weights larger than zero, which is a result along the line of [25] and more conservative than [8], which confirms the Jeffreys prior's feature of being conservative in the number of the components. Credible intervals also show that the parameters of the components with marginal posterior distributions for the weights not concentrated around zero are estimated with lower uncertainty.

The second dataset is related to a population study to validate caffeine as a probe drug to establish the genetic status of rapid acetylators and slow acetylators [1]: many drugs, including caffeine, are metabolyzed by a polymorphic enzyme (EC 2.3.1.5) in humans and the white population is divided into two groups of slow acetylators and rapid acetylators. Caffeine is considered an inter-

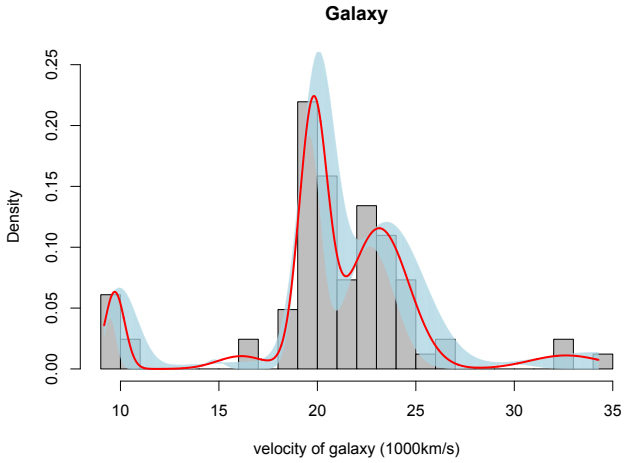

Figure 6: Predictive distribution of the galaxy dataset: the red line represent the estimation of the density, the shadow blue area represents the credible intervals in 10^5 simulations by assuming a ten-component mixture model.

esting drug to study the phenotype of people, because it is regularly consumed by a large amount of the population. Several population studies have been conducted, some of them reporting a bimodality, some others a trimodality. We focus on the study presented by [1], involving 245 unrelated patients and computing the molar ratio between two metabolites of caffeine, AFMU and 1X, both measured in urine 4 to 6 hours after ingestion of 200 mg of caffeine.

We have again assumed a ten-component mixture model and checked whether or not the hierarchical approach which uses the conditional Jeffreys prior on the weights of the mixture model is able to identify a smaller number of significant components.

The results are available in Figure 7 and Table 1. The algorithm identifies two components with weights clearly larger than zero and two other components with very small weights. [1] identify a bimodal density, while [25] consider highly likely a 3-5 component mixture. The Jeffreys prior allows to concentrate the analysis on mainly two subgroups and it suggests that Gaussian components may be inappropriate in this setting: by looking to the location of the com-

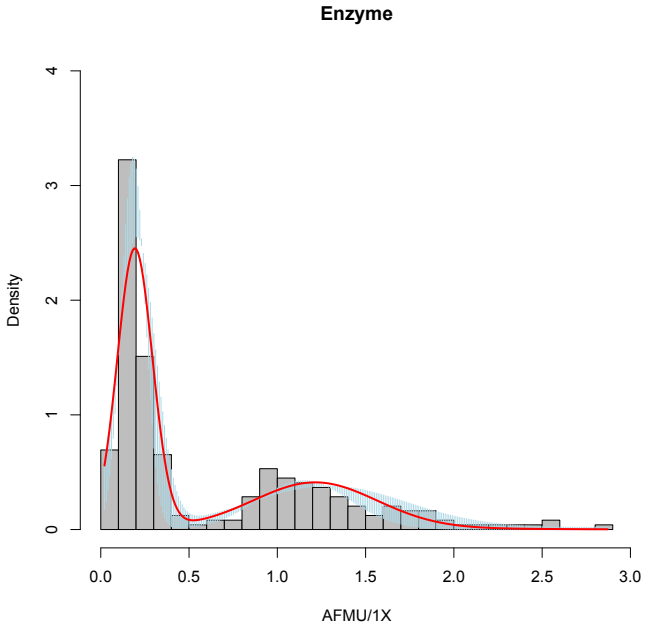

Figure 7: Predictive distribution of the enzyme dataset: the red line represent the estimation of the density, the shadow blue area represents the credible intervals in 10^5 simulations by assuming a ten-component mixture model.

ponents with small weights, it may be more adequate to consider asymmetric distributions.

Our third dataset is related to measuring the acid neutralizing capacity (ANC) (in log-scale) of a sample of 155 lakes in north-central Wisconsin, to determine the number of lakes that have been affected by acidic deposition [6]: the ANC measures the capability of a lake to neutralize acid, i.e. low values may indicate a problem for the lake's biological diversity.

The results are available in Figure 8 and Table 1. The algorithm identifies two components with significant weights and two other components with very small weights. [6] assume a bimodal density, while [25] consider highly likely a 3-5 component model. The Jeffreys prior again allows to concentrate the analysis on two main subgroups and suggests to investigate the importance of other two

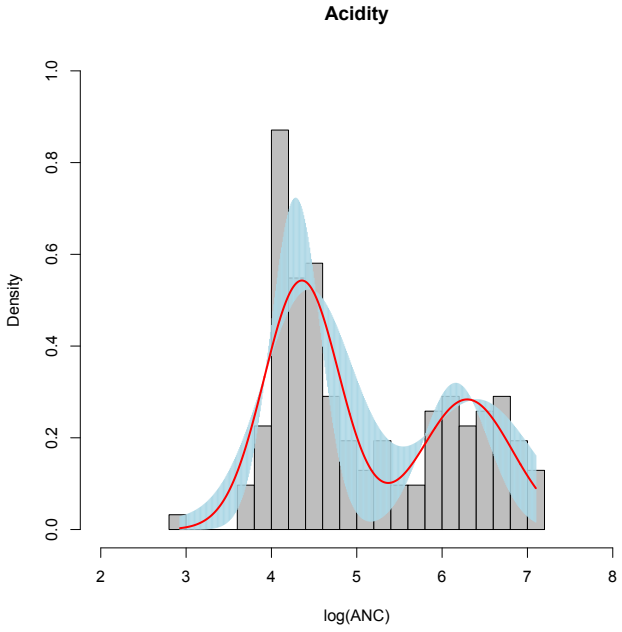

Figure 8: Predictive distribution of the acidity dataset: the red line represent the estimation of the density, the shadow blue area represents the credible intervals in 10^5 simulations by assuming a ten-component mixture model.

components and possibly the goodness-of-fit of the log-normal distribution in this setting.

6.1. Network dataset

A recent trend in computer network systems is the deployment of network functions in software [22]. The so-called "software dataplanes" are emerging as an alternative to traditional hardware switched and routers, reducing costs and enhancing programmability.

The monitoring of IP packets is, among all possible network functions, one of the most suitable for a software deployment. However, the monitoring has a huge cost in terms of consumed CPU (processing) time by packet. The main reason for this is that each incoming packet triggers the retrieval, from a large hash-table, of all the information related to the packet flow (i.e. the packet's family).

Table 1: Posterior means for the weights, the means and the standard deviations of a ten-component mixture model, assumed for the galaxy, the enzyme and the acidity datasets (the first number in brackets is the posterior mean and the second is the posterior standard deviation). We have decided to not shown the estimated location and scale parameters when the weights are concentrated around zero.

Dataset:	**galaxy**	**enzyme**	**acidity**
p_1	0.437	0.606	0.601
	(23.139, 1.507)	(0.193, 0.090)	(4.356,0.442)
p_2	0.390	0.343	0.378
	(19.790, 0.715)	(1.216, 0.348)	(6.294, 0.531)
p_3	0.080	0.021	0.003
	(9.709, 0.503)	(0.915, 1.174)	(0.083, 0.802)
p_4	0.056	0.018	0.003
	(32.630, 1.842)	(1.176, 0.702)	(0.125, 0.589)
p_5	0.037	0.000	0.000
	(16.138,1.226)	-	-
$\sum_{\ell=6}^{10} p_\ell$	0.000	0.000	0.000

This operation is generally called flow-entry retrieval. The time required for the flow-entry retrieval (retrieval time) mainly depends on whether such information is available in one of the processor caches (e.g. L1, L2, L3) or in memory.

The dataset used in this analysis consists of generated samples of retrieval time, each with 10^6 times, under two different set-ups. In the first one, the flow-entry has been forced to reside in fast processor caches ("hit"). In the second one, all flow-entries have been forced to reside in the server RAM (memory), which results in a slower flow-entry retrieval ("miss").

Both samples show a heavy tail, due to possible hash collisions on the table, as well as additional delays introduced by measuring the retrieval time at a nanosecond timescale. In the case of "miss", another reason for the heavy tail can be identified with the virtual/physical memory mapping, which can inflate the retrieval time in some cases.

The goal of a realistic analysis is to infer the proportion of reported times which may be considered from the "hit" distribution and the proportion of times which may be considered from the "miss" distribution, i.e. to derive what is the percentage of packets for which the flow-entry was in the cache and the percentage of packets for which the flow-entry was in memory.

However, a first simulation is generally used to test the procedure. The interest of the analysis will be in the region of the space where the two distributions are overlapping, therefore the interest is not in the external tails, which may, nonetheless, affect inference. Therefore, a preliminary analysis may be conducted in order to understand if a part of the future observations may be discarded from the analysis. In this particular case, the conservative property of the Jeffreys prior may be used in order to understand how much important are the tails of each distribution and to identify the right models to use. For instance, a comparison between a Gaussian mixture model and a mixture model with Gumbel components may be run: if in both cases the analysis run with a Jeffreys prior for the mixture weights identifies more than two (assumed) distributions of interest, this may be a suggestion that the observations allocated to the external components (not the "hit" or the "miss" ones) may be discarded,

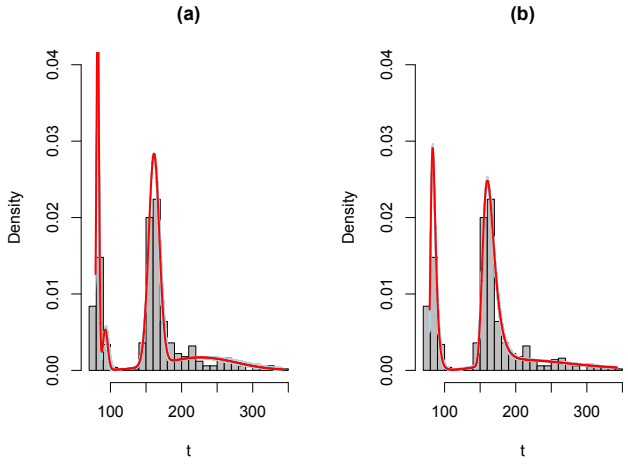

Figure 9: Predictive distribution of the network dataset: the red line represent the estimation of the density, the shadow blue area (very concentrated around the red lines) represents the credible intervals in 10^5 simulations by assuming a ten-component mixture model, with Gaussian components on the left and with Gumbel components on the right.

providing inference on the proportion of observations to discard as well.

Figure 9 and Table 2 show the results of this analysis: adopting a Jeffreys prior for the mixture weights when assuming Gumbel components allows to better estimate the first component and to describe the asymmetry observed in the data as an asymmetry in the first component instead of an additional component. Nevertheless it is not sufficient to identify the observations in the right tail of the second component as part of its tail, since the algorithm identifies a third component located in that part of the space.

In this setting, the Jeffreys prior allows to i) identify a miss-specification of the model assumptions (the approximated Bayes factor of the mixture of Gumbel components against the mixture of normal components is 2.10) and ii) identify which part of the observations to discard from further studies.

Table 2: Posterior means for the weights, the means and the standard deviations of a ten-component mixture model, assumed for the network dataset (credible intervals of level 0.95 in brackets).

Gaussian comp.		
p	μ	σ
0.214	224.318	50.271
(0.180,0.249)	*(222.657,233.842)*	*(45.483,55.265)*
0.519	161.645	7.497
(0.474,0.568)	*(160.216,161.882)*	*(6.830,8.212)*
0.221	82.847	1.888
(0.188,0.257)	*(81.057,82.270)*	*(1.666,2.135)*
0.046	92.826	3.474
(0.029,0.064)	*(91.710 ,93.700)*	*(2.698,4.388*
$\sum_{\ell=5}^{10} p_\ell = 0.000$		

Gumbel comp.		
p	μ	σ
0.214	213.512	59.080
(0.183,0.251)	*(213.446,213.846)*	*(53.526,64.667)*
0.520	160.164	7.959
(0.479,0.562)	*(160.113,160.482)*	*(7.465,8.482)*
0.265	83.260	3.348
(0.219,0.302)	*(83.251,83.270)*	*(3.005,3.753)*
$\sum_{\ell=4}^{10} p_\ell = 0.000$		

7. Conclusion

This thorough analysis of the Jeffreys priors in the setting of mixtures with location-scale components shows that mixture distributions deserve the qualification of an ill-posed problem with regard to the production of non-informative priors. Indeed, we have shown that most configurations for Bayesian inference in this framework do not allow for the standard Jeffreys prior to be taken as a reference. While this is not the first occurrence where Jeffreys priors cannot be used as reference priors, we have shown that the Jeffreys prior for the mixture weights has the important property to be conservative in the number of components, with a configuration compatible with the results of [35].This is a general feature of the Jeffreys prior for the mixture weights, which is independent from the shape of the distributions composing the mixture.

Nevertheless, we have decided to study its behavior in the specific case of components from location-scale families. We have proposed a hierarchical representation of the mixture model, which allow for improper priors at the highest level of the hierarchy and assumes the Jeffreys prior for the mixture weights in the second level, conditional on prior distributions for the location and scale parameters along the line of [21].

Through several examples, both on simulated and real datasets, we have shown that this representation seems to be more conservative on the number of components than other non or weakly informative prior distributions for mixture models available in the literature. In particular, it seems to be able to recognize the meaningful components, which is an essential property for a noninformative prior for mixture model: in fact, in an objective setting, it is essential to consider the possibility to have assumed a wrong number of components. In this sense, the Jeffreys prior for the mixture weights may be used to identify the meaningful components and possible miss-specifications of either the number or the distributional family of the components.

As a note aside, we have mainly analyzed mixture of Gaussian distributions in this paper, with extensions of the theoretical results to the other distributions

of the location-scale family. Nevertheless, the possible difficulties deriving from the use of distributions different from the Gaussian are not considered here and will be the focus of future research. In particular, all likelihoods poorly specified and ill-behaved cases are more likely to meet difficulties. However, the Jeffreys prior is known as a regularization prior that does not necessarily reflect prior beliefs, but in combination with the likelihood function yields posteriors with desirable properties; see [14] for a detailed review of ill-behaved posterior cases and the role of the Jeffreys prior in those cases.

Acknowledgements and Notes

The code used for the Gaussian mixture models is available online at the following link: `https://github.com/cgrazian/Jeffreys_mixtures`.

The Authors want to thank Gioacchino Tangari, from the Department of Electronic and Electrical Engineering, University College London, for having provided the simulations of Section 6.1.

References

[1] Bechtel Y.C., Bonaiti-Pellie, C., Poisson, N., Magnette, J. and Bechtel, P.R. (1993). A population and family study Nacetyltransferase using caffeine urinary metabolites. *Clinical Pharmacology & Therapeutics*, **54(2)** 134–141.

[2] Berger, J., Bernardo, J. and D., S. (2009). Natural induction: An objective Bayesian approach. *Rev. Acad. Sci. Madrid*, **A 103** 125–159. (With discussion).

[3] Bernardo, J. and Giròn, F. (1988). A Bayesian analysis of simple mixture problems. In *Bayesian Statistics 3* (J. Bernardo, M. DeGroot, D. Lindley and A. Smith, eds.). Oxford University Press, Oxford, 67–78.

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

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

[6] Crawford, S.L., DeGroot, M.H., Kadane, J.B. and Small, M.J. (1992). Modeling Lake-Chemistry Distributions: Approximate Bayesian Methods for Estimating a Finite-Mixture Model *Technometrics*, **34(4)** 441–453.

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

[8] Escobar, M.D., and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, **90**(430) 577–588.

[9] Figueiredo, M. and Jain, A. (2002). Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **24** 381–396.

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

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

[12] Ghosh, M., Carlin, B. P. and Srivastiva, M. S. (1995). Probability matching priors for linear calibration. *TEST*, **4** 333–357.

[13] Gruet, M., Philippe, A. and Robert, C. (1999). MCMC control spreadsheets for exponential mixture estimation. *J. Comput. Graph. Statist.*, **8** 298–317.

[14] Hoogerheide, L.F., and van Dijk, H.K. (2008). Possibly Ill-behaved Posteriors in Econometric Models.

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

[16] Jeffreys, H. (1939). *Theory of Probability*. 1st ed. The Clarendon Press, Oxford.

[17] Kamary, K., Lee, J.E. and Robert, C.P. (2007). Weakly informative reparameterisations for location-scale mixtures. *pre-print, arXiv:1601.01178v2*.

[18] Kass, R. and Wasserman, L. (1996). Formal rules of selecting prior distributions: a review and annotated bibliography. *J. American Statist. Assoc.*, **91** 343–1370.

[19] Lee, K., Marin, J.-M., Mengersen, K. and Robert, C. (2009). Bayesian inference on mixtures of distributions. In *Perspectives in Mathematical Sciences I: Probability and Statistics* (N. N. Sastry, M. Delampady and B. Rajeev, eds.). World Scientific, Singapore, 165–202.

[20] MacLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley, New York.

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

[22] Nunes, B.A., Mendonca, M., Nguyen, X.N., Obraczka, K. and Turletti,, T. (2014). A survey of software-defined networking: Past, present, and future of programmable networks *IEEE Communications Surveys & Tutorials* **16**(3) 1617–1634.

[23] Puolamäki, K. and Kaski, S. (2009). Bayesian solutions to the label switching problem. In *Advances in Intelligent Data Analysis VIII* (N. Adams, C. Robardet, A. Siebes and J.-F. Boulicaut, eds.), vol. 5772 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 381–392.

[24] Raftery, A.E. (1996). Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.), London: Chapman and Hall, pp. 163–188.

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

[26] Rissanen, J. (2012). *Optimal Estimation of Parameters*. Cambridge University Press.

[27] Robert, C. (2001a). *The Bayesian Choice*. 2nd ed. Springer-Verlag, New York.

[28] Robert, C. (2001b). *The Bayesian Choice*. 2nd ed. Springer-Verlag, New York.

[29] Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. 2nd ed. Springer-Verlag, New York.

[30] Robert, C., Chopin, N. and Rousseau, J. (2009). Theory of Probability revisited (with discussion). *Statist. Science*, **24(2)** 141–172 and 191–194.

[31] Robert, C. and Mengersen, K. (1999). Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Comput. Statist. Data Analysis*, **29** 325–343.

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

[33] Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. American Statist. Assoc.*, **85**(411) 617–624.

[34] Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. American Statist. Assoc.*, **92** 894–902.

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

[36] Rubio, F. and Steel, M. (2014). Inference in two-piece location-scale models with Jeffreys priors. *Bayesian Analysis*, **9** 1–22.

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

[38] Titterington, D., Smith, A. and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.

[39] Wasserman, L. (2000). Asymptotic inference for mixture models using data dependent priors. *J. Royal Statist. Society Series B*, **62** 159–180.

Supplementary Material for online publication only

Click here to download Supplementary Material for online publication only: GR17_CSDA_SuppMat.pdf