Chapter 4
Jeffreys' Priors for Mixture Estimation

Clara Grazian and Christian P. Robert

Abstract Mixture models may be a useful and flexible tool to describe data with a complicated structure, for instance characterized by multimodality or asymmetry. The literature about Bayesian analysis of mixture models is huge, nevertheless an "objective" Bayesian approach for these models is not widespread, because it is a well-established fact that one needs to be careful in using improper prior distributions, since the posterior distribution may not be proper, yet noninformative priors are often improper. In this work, a preliminary analysis based on the use of a dependent Jeffreys' prior in the setting of mixture models will be presented. The Jeffreys' prior which assumes the parameters of a Gaussian mixture model is shown to be improper and the conditional Jeffreys' prior for each group of parameters is studied. The Jeffreys' prior for the complete set of parameters is then used to approximate the derived posterior distribution via a Metropolis–Hastings algorithm and the behavior of the simulated chains is investigated to reach evidence in favor of the properness of the posterior distribution.

Key words: Improper priors, Mixture of distributions, Monte Carlo methods, Noninformative priors

C. Grazian (✉)
CEREMADE, Université Paris-Dauphine, Paris, France

Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy
e-mail: grazian@ceremade.dauphine.fr

C.P. Robert
CEREMADE, Université Paris-Dauphine, Paris, France

Department of Statistics, University of Warwick, Coventry, UK
e-mail: xian@ceremade.dauphine.fr

S. Frühwirth-Schnatter et al. (eds.), *Bayesian Statistics from Methods to Models and Applications*, Springer Proceedings in Mathematics & Statistics 126,
DOI 10.1007/978-3-319-16238-6_4

4.1 Introduction

The probability density function of the random variable $\mathbf{x}$ of a mixture model is given as follows:

$$g(\mathbf{x} \mid \psi) = \sum_{i=1}^{K} w_i f_i(\mathbf{x} \mid \theta_i), \tag{4.1}$$

where $\mathbf{x}$ is a random variable with probability density function $g(\cdot)$ which depends on a parameter vector $\psi = (\theta_1, \cdots, \theta_K, w_1, \cdots, w_K)$, where $w_i \in (0,1)$ and $\sum_{i=1}^{K} w_i = 1$, K is the number of components and θ_i is the vector of parameters of the ith component, whose behavior is described by the density function $f_i(\cdot)$.

In this setting, the maximum likelihood estimation may be problematic, even in the simple case of Gaussian mixture models, as shown in [2]. For a comprehensive review, see [11]. In a Bayesian setting, [4] and [6] suggest to be careful when using improper priors, in particular because it is always possible that the sample does not include observations for one or more components, thus the data are not informative about those particular components. Avoiding improper priors is not necessary, however the properness of the posterior distribution has to be proven and works exist which show that independent improper priors on the parameters of a mixture model lead to improper posteriors, except for the case where one component has no observation in the sample is prohibited (as in [3]). Some proposals of "objective priors" in the setting of mixture models are the partially proper priors in [3, 7] and [10], the data dependent prior in [12] and the weakly informative prior in [8], which may or may not be data-dependent.

In this work, we want to analyze the posterior distribution for the parameters of a mixture model with a finite number of components when the Jeffreys' definition of a noninformative prior (see [5]) is applied. In particular, we want to assess the convergence of the Markov chain derived from an MCMC approximation of the posterior distribution when using the Jeffreys' prior for the parameters of a Gaussian mixture model, even when the prior for some parameters is improper conditional on the others.

The outline of the paper is as follows: in Sect. 4.2, the Jeffreys' prior is presented and an explanation about the reason why improper priors have to be used with care in the setting of mixture models is given; then, the Jeffreys' prior for the weights of a general mixture model conditional on the other parameters is presented in Sect. 4.2.1 and the Jeffreys' priors for the means and the standard deviations when every other parameter is known are presented in Sect. 4.2.2. Section 4.3 describes the algorithms used to implement simulations. Section 4.4 shows the results for the posterior distributions obtained when using a dependent Jeffreys' prior for all the parameters of a Gaussian mixture model based on simulations, in particular including an example for a three-component Gaussian mixture model; finally, Sect. 4.5 concludes the paper with a discussion.

4.2 Jeffreys' Priors for Mixture Models

We recall that Jeffreys' prior was introduced by [5] as a default prior based on the Fisher information matrix $I(\theta)$ as

$$\pi^J(\theta) \propto |I(\theta)|^{\frac{1}{2}}, \tag{4.2}$$

whenever the latter is well defined. In most settings, Jeffreys' priors are improper, which may explain their conspicuous absence in the domain of mixture estimations, since the latter prohibits the use of most improper priors by allowing any subset of components to go empty. That is, the likelihood of a mixture model can always be decomposed into a sum over all possible partitions of the data with 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 carries information about the parameters of a specific component. In particular, consider independent improper priors

$$\pi(\theta_1, \cdots, \theta_K) \propto \prod_{j=1}^{K} \pi^*(\theta_j), \tag{4.3}$$

such that $\int \pi^*(\theta_k) d\theta_k = \infty \ \forall\ k \in \{1, \cdots, K\}$. Mixture models are an example of latent variable models, where the density function may be rewritten in an augmented version as

$$g(\mathbf{x}; \psi) = \sum_{S \in \mathscr{S}_K} f_j(\mathbf{x}; S, \theta_j) \prod_{j=1}^{K} \pi^*(\theta_j) \pi(S \mid \mathbf{w}) \pi(\mathbf{w}), \tag{4.4}$$

where the summation runs over the set $\mathscr{S}_K$ of all the K^N possible classifications S. Then, if there is an empty component (let's say the jth), i.e. a component with no observation in the sample, the complete-data likelihood does not carry information about that particular component and the posterior distribution for it will depend only on the prior and will have an infinite integral, if the prior is improper:

$$\int \prod_{i:S_i=j} g(x_i; \theta_j) \pi^*(\theta_j) d\theta_j \propto \int \pi^*(\theta_j) d\theta_j = \infty. \tag{4.5}$$

Another obvious reason for the absence of Jeffreys' priors is a computational one, namely the closed-form derivation of the Fisher information matrix is almost inevitably impossible. The reason are integrals which cannot be analytically computed having the form

$$-\int_{\mathscr{X}} \frac{\partial^2 \log \left[\sum_{k=1}^{K} w_k f(\mathbf{x}|\theta_k)\right]}{\partial \theta_i \partial \theta_j} \left[\sum_{k=1}^{K} w_k f(\mathbf{x}|\theta_k)\right] \mathrm{d}x. \tag{4.6}$$

4.2.1 Jeffreys' Prior for the Weights of a Mixture Model

Consider a two-component mixture model with known parameters of the component distributions. The Jeffreys' prior for the weights is just a function of only one parameter because of the constraint on the sum of the weights:

$$\pi^J(w_1) \propto \sqrt{\int_{\mathscr{X}} \frac{\left(f(\mathbf{x};\theta_1) - f(\mathbf{x};\theta_2)\right)^2}{w_1 f(\mathbf{x};\theta_1) + w_2 f(\mathbf{x};\theta_2)} d\mathbf{x}} \tag{4.7}$$

$$\leq \sqrt{\int_{\mathscr{X}} \frac{\left(f(\mathbf{x};\theta_1) - f(\mathbf{x};\theta_2)\right)^2}{w_1 f(\mathbf{x};\theta_1)} d\mathbf{x}} \tag{4.8}$$

$$= \sqrt{\frac{1}{w_1} c_1}, \tag{4.9}$$

where c_1 is a positive constant and $\mathscr{X}$ is the support of the random variable $\mathbf{x}$ which is modelled as a mixture. The resulting prior may be easily generalized to the case of K components for which the generic element of the Fisher information matrix is

$$\int_{\mathscr{X}} \frac{(f(\mathbf{x};\theta_i) - f(\mathbf{x};\theta_K))(f(\mathbf{x};\theta_j) - f(\mathbf{x};\theta_K))}{\sum_{k=1}^K w_k f(\mathbf{x};\theta_k)} d\mathbf{x}, \tag{4.10}$$

where $i \in \{1, \cdots, K-1\}$ and $j \in \{1, \cdots, K-1\}$. As shown above, this prior is proper and it is easy to see that it is convex by studying its second derivative (in the general case of K components, it can be shown that the prior is still proper because all the integrals in the Fisher information matrix are finite and the marginals are still convex). The form of the prior depends on the type of components. For an approximation of the prior and of the derived posterior distribution based on a sample of 100 observations for a particular choice of parameters, see Fig. 4.1. We have compared this approximation to the ones obtained by fixing the parameters at different values: the Jeffreys' prior for the weights of a mixture model is more symmetric as the components are more similar in terms of variance and it is more concentrated around the extreme values of the support as the means are more distant.

The Jeffreys' prior for the weights of a mixture model could be approximated by a beta distribution, which represents the traditional marginal distribution for the weights of a mixture model (since in the literature the Dirichlet distribution is, in general, the default choice as proper prior distribution): after having obtained a sample via Metropolis–Hastings algorithm which approximates the Jeffreys' prior (which is not known in closed form), the parameters of the approximating beta distribution may be estimated from the sample with the method of moments. Figure 4.2 shows this approximation for the weight of the first component of a Gaussian mixture model for an increasing number of components, each having the same standard deviation. It is evident from the figure that the (marginal) Jeffreys' distribution and its beta approximation tend to be more and more concentrated

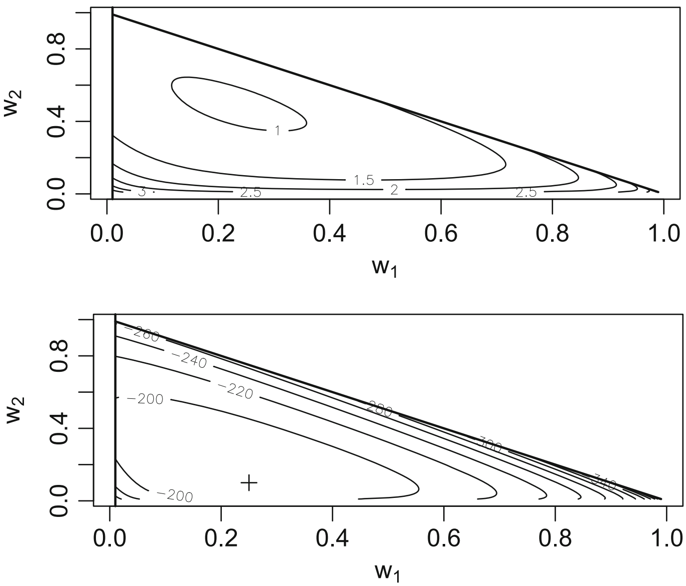

Fig. 4.1 Approximations of the (conditional) prior (*top*) and derived posterior (*bottom*) distributions for the weights of the three-component Gaussian mixture model $0.25 \cdot N(-1,1) + 0.10 \cdot N(0,5) + 0.65 \cdot N(2,0.5)$

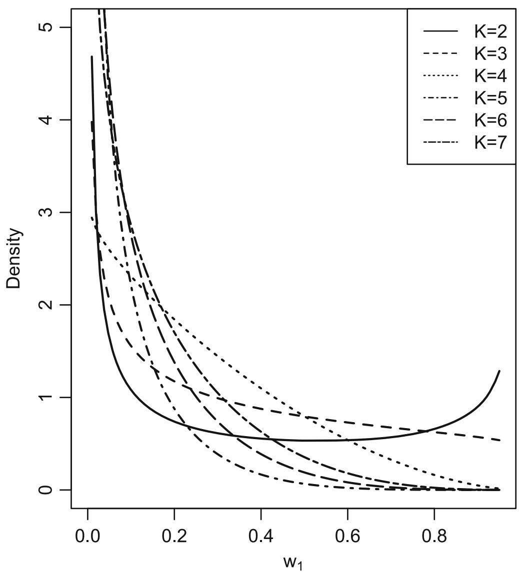

Fig. 4.2 Beta approximations of the (conditional) prior distributions for the weight of the first component of a Gaussian mixture model with an increasing number K of components (with a fixed standard deviation equal to 1 for all the components and location parameters chosen as the first K elements of $\{-9,9,0,-6,6,-3,3\}$)

around 0 as the number of components increases. Both the variance and the mean of the beta approximations tend to stabilize around values close to 0, while it is not evident that there is a particular behavior for the parameters of the beta distribution, which could be smaller or greater than 1.

4.2.2 Jeffreys' Prior for the Means and the Standard Deviations of a Gaussian Mixture Model

Consider a two-component Gaussian mixture model. The conditional Jeffreys' prior for the mean parameters depends on the following derivatives:

$$\frac{\partial^2 \log f}{\partial \mu_i^2} = \left\{ \frac{w_i N(\mu_i, \sigma_i) \left[\left(\frac{x-\mu_i}{\sigma_i^2} \right)^2 - \frac{1}{\sigma_i^2} \right]}{w_1 N(\mu_1, \sigma_1) + w_2 N(\mu_2, \sigma_2)} \right\} - \left\{ \frac{w_i N(\mu_i, \sigma_i) \exp\left(-\frac{1}{2} \left(\frac{x-\mu_i}{\sigma_i} \right) \right) \frac{x-\mu_i}{\sigma_i^2}}{w_1 N(\mu_1, \sigma_1) + w_2 N(\mu_2, \sigma_2)} \right\}^2, \tag{4.11}$$

for $i \in \{1,2\}$ and

$$\frac{\partial^2 \log f}{\partial \mu_1 \partial \mu_2} = -\frac{w_1 N(\mu_1, \sigma_1) \frac{x-\mu_1}{\sigma_1^2} \cdot w_2 N(\mu_2, \sigma_2) \frac{x-\mu_2}{\sigma_2^2}}{w_1 N(\mu_1, \sigma_1) + w_2 N(\mu_2, \sigma_2)}. \tag{4.12}$$

With a simple change of variable $y = x - \mu_i$ for some choice of $i \in \{1,2\}$, it is easy to see that each element of the Fisher information matrix depends only on the difference between the means but not on μ_i alone, therefore it is flat with respect to each μ_i. The generalization to K components is obvious.

An approximation of the prior and derived posterior distribution based on a sample of 100 observations of the means of a two-component Gaussian mixture model is shown in Fig. 4.3. When only the means are unknown, it is evident that the prior is constant on the difference between the means and is increasing with the difference of the means. On the contrary, the posterior distribution is concentrated around the true values and shows classical label switching. Again, Fig. 4.3 has been compared to similar approximations obtained by fixing the parameters at different values: the conditional Jeffreys' prior for the means of a Gaussian mixture model is more symmetric, as the standard deviations become more similar and the choice of the weights seems to influence only the approximation of the posterior distribution, where the density is more concentrated around the mean linked to the highest weight.

As an additional proof that the conditional Jeffreys' prior for the means of a Gaussian mixture model is improper, consider that, if the location or the scale parameters of a mixture model are unknown, this makes the model a location or a scale model, for which the Jeffreys' prior is improper in the location and the log-scale parameters, respectively (see [9] for details). In this case also the conditional Jeffreys' prior for the standard deviations when all the other parameters are considered known is improper.

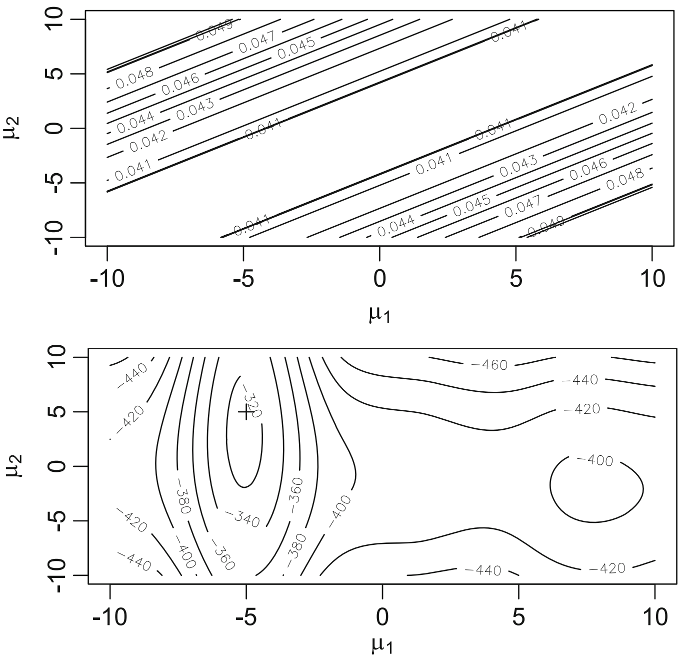

Fig. 4.3 Approximations of the (conditional) prior (*top*) and derived posterior (*bottom*) distributions for the means of the two-component Gaussian mixture model $0.5 \cdot N(-5,1) + 0.5 \cdot N(5,10)$

4.3 Implementation

Each element of the Fisher information matrix is an integral of the form presented in Eq. (4.2) which has to be approximated. We have applied both numerical integration and Monte Carlo integration and the results show that, in general, numerical integration obtained via Gauss–Kronrod quadrature yields more stable results. Nevertheless, when one or more proposed values for the standard deviations or the weights are too small, the approximations tend to be very dependent on the bounds used for numerical integration (usually chosen to omit a negligible part of the density). In this case Monte Carlo integration seems to yield more stable approximations and thus is applied by us. However, in these situations the approximation could lead to a negative determinant of the Fisher information matrix, even if it was very small in absolute value (of order 10^{-25} or even smaller). In this case, we have chosen to recompute the approximation until we get a positive number.

The computing expense due to deriving the Jeffreys' prior for a set of parameter values is $O(d^2)$, where d is the total number of (independent) parameters. A way to accelerate the Metropolis–Hastings algorithm used to approximate the posterior distribution derived from the Jeffreys' prior is the Delayed Acceptance algorithm proposed by Banterle et al. [1] (Algorithm 1).

Algorithm 1 Delayed Acceptance algorithm

Choose the initial values w^0, μ^0, σ^0
for i in $1:N$ **do**
  Propose $w^{prop}, \mu^{prop}, \sigma^{prop} \sim K(\cdot,\cdot,\cdot|w^{(i-1)}, \mu^{(i-1)}, \sigma^{(i-1)})$
  Simulate $u_1 \sim Unif(0,1)$ and $u_2 \sim Unif(0,1)$
  if $u_1 < \frac{l(w^{prop},\mu^{prop},\sigma^{prop})}{l(w^{(i-1)},\mu^{(i-1)},\sigma^{(i-1)})} \frac{K(w^{(i-1)},\mu^{(i-1)},\sigma^{(i-1)}|w^{prop},\mu^{prop},\sigma^{prop})}{K(w^{prop},\mu^{prop},\sigma^{prop}|w^{(i-1)},\mu^{(i-1)},\sigma^{(i-1)})}$ **then**
    if $u_2 < \frac{\pi^J(w^{prop},\mu^{prop},\sigma^{prop})}{\pi^J(w^{(i-1)},\mu^{(i-1)},\sigma^{(i-1)})}$ **then** Set $(w^{(i)}, \mu^{(i)}, \sigma^{(i)}) = (w^{prop}, \mu^{prop}, \sigma^{prop})$
    else $(w^{(i)}, \mu^{(i)}, \sigma^{(i)}) = (w^{(i-1)}, \mu^{(i-1)}, \sigma^{(i-1)})$
    end if
  else Set $(w^{(i)}, \mu^{(i)}, \sigma^{(i)}) = (w^{(i-1)}, \mu^{(i-1)}, \sigma^{(i-1)})$
  end if
end for

This exact algorithm allows to compute the Jeffreys' prior (the more expensive part of the posterior distribution) only when a first accept/reject step depending on the likelihood ratio (less costly) leads to acceptance and reduces the computational time by about 80 % (from an average of about 113 h with standard Metropolis–Hastings algorithm to an average of about 32 h with the Delayed Acceptance version) for 10^6 simulations for a three-component Gaussian mixture model with all parameters unknown, with a decrease of the acceptance rates from about 35 % with the standard Metropolis algorithm to about 20 % with the Delayed Acceptance. In combination with a reduction in the acceptance rate, the Delayed Acceptance version of the Metropolis–Hastings algorithm also induces a reduction of the effective sample size of about 35 %. We have used an algorithm that is adaptive during the burn-in period such that it leads to an acceptance rate above 20 % and below 40 %.

The Delayed Acceptance algorithm is an ideal solution for this setting: the likelihood is cheap to evaluate while the prior distribution is not only demanding but also non-informative. Therefore, it should have a limited influence with respect to the data when computing the posterior distribution and thus an early rejection due only to the likelihood ratio should not worsen the MCMC performances. Nevertheless, attention must be paid when applying the algorithm; since the prior distribution is improper, when the likelihood function is concentrated near regions of the parameter space where the prior distribution diverges (for examples in the case of Gaussian mixture models if the likelihood function is concentrated not far from values of standard deviations near 0), even if the first step based on the likelihood alone accepts a move, the first derivative of the prior distribution in that point may be too high in absolute value to allow the acceptance of the move and, therefore, the chain may be stuck or accept the proposed value only if the move is towards regions of even higher prior density.

A solution to this problem may be seen in splitting the likelihood ratio and using a part of it (relative to a small set of observations) jointly with the prior ratio in the second step. How many observations one has to consider in this splitting depends on the problem at hand and on the total number of observations.

4.4 The Posterior Distribution for a Mixture Model when Jeffreys' Prior is Used

It is a well-established fact in the literature that using independent improper priors for mixture models leads to improper posterior distributions, in particular if one of the components of the mixture model is not represented in the observed sample (i.e., there are no observations relative to at least one of the components of the mixture). One may use improper priors in mixture models by introducing some form of dependence between the components, as shown in [7]. Actually, the Jeffreys' prior does that by considering the Fisher information matrix. Checking for properness of the posterior distribution is unfeasible in an analytic way and the outcome of the classical Metropolis–Hastings algorithm and the version described in Sect. 4.3 targeting the posterior distribution derived from using the Jeffreys' prior has to be exploited in order to collect evidence that the posterior distribution is proper, even if the results which will be presented are not a conclusive proof of that.

4.4.1 Output of the MCMC Algorithm

Through extensive simulation, we have seen that for a sufficiently big sample size (at least equal to 20 for a three-component Gaussian mixture model) the MCMC chain never diverged. In particular, when the sample size decreases by up to ten data points, the component with the lowest weight is usually not identified and the chains may even get stuck, in particular because values of standard deviations close to 0 are accepted. The uncertainty on the posterior estimates depends on how variable that particular component is and how big the corresponding weight is. The acceptance rate of the MCMC algorithm is around 20 % for high sample sizes and it increases as the sample size decreases, unless the chain gets stuck (this happens with very low sample sizes).

4.4.2 An Example

Experiments with simulated data have been performed with different numbers of Gaussian components with different location and scale parameters generating the data and with different weights. The results are always similar, except for the fact that the uncertainty on the Bayesian estimates of the parameters increases as the components are closer, in terms of location. In particular, the Bayesian estimates of the components with the highest variances and/or the lowest weights are more variable.

Figures 4.4, 4.5, and 4.6 show the trace plots and the histograms of the MCMC chains approximating the marginal posterior distributions of the parameters of the three-component Gaussian mixture model

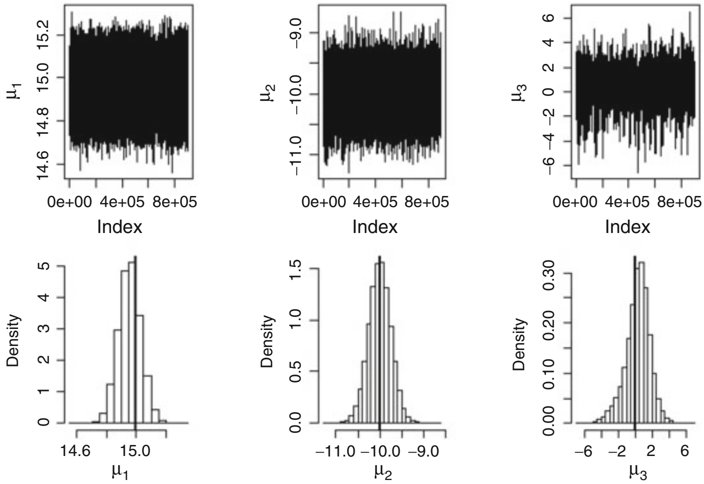

Fig. 4.4 Marginal posterior distributions (chains obtained via Metropolis–Hastings algorithm) for the means of a three-component Gaussian mixture model $0.65 \cdot N(15,0.5) + 0.25 \cdot N(-10,1) + 0.10 \cdot N(0,5)$

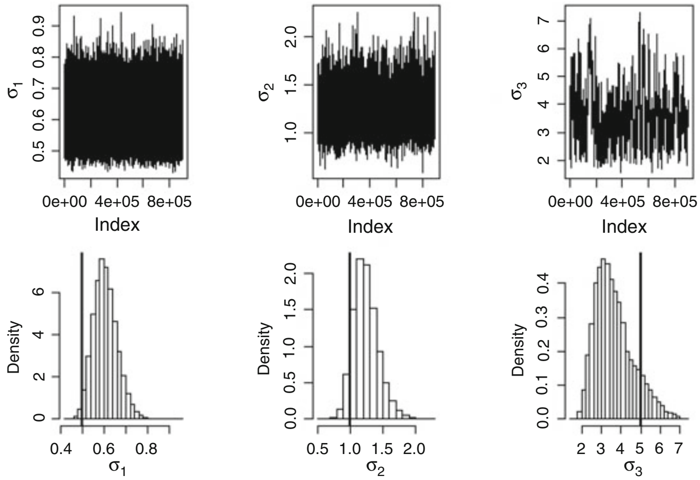

Fig. 4.5 Marginal posterior distributions for the standard deviations of the same model as in Fig. 4.4

$$0.65 \cdot N(15,0.5) + 0.25 \cdot N(-10,1) + 0.10 \cdot N(0,5) \tag{4.13}$$

and for a sample size equal to 100. They show that the chains have reached convergence, with a higher uncertainty when estimating the mean and the standard deviation of the third component, being the one with the highest variability and the smallest weight.

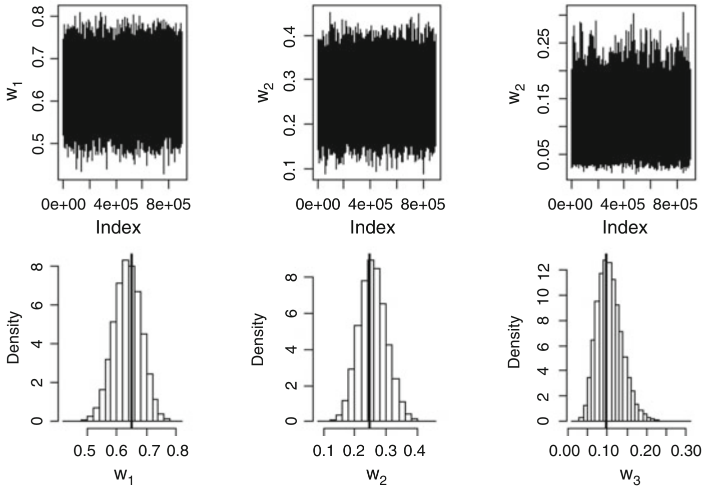

Fig. 4.6 Marginal posterior distributions for the weights of the same model as in Fig. 4.4

4.5 Discussion

The aim of this work is to propose an objective Bayesian analysis for the mixture model setting. The literature on mixture models shows that attention must be paid when using improper priors with these models and a generally accepted default solution does not exist. Nonetheless, we try to introduce a new objective Bayesian approach to handling mixture models by applying a Jeffreys' prior which considers the parameters of the dependent model. The prior has been shown to be improper, nevertheless extensive simulation studies suggest that the posterior could be proper, at least for a sufficiently high number of observations.

There are two important drawbacks when using the Jeffreys' prior for mixture models. First, the prior depends on integrals which have to be approximated. Possible solutions to this problem have been investigated, but they are beyond the scope of this paper. One may refer to [1] for an algorithm which may reduce the computational time of approximating the posterior distribution. Another solution is to reparameterize the model, as proposed by Mengersen and Robert [7], and exploit its independence features to reduce the dimension of the matrix to approximate. Second, the posterior distribution cannot be managed in an analytic way. Nevertheless there is no assurance that a distribution is proper if the Markov chain simulated via MCMC and used to approximate the posterior seems to converge. Future work will be aimed at the study of the relationship between the prior distribution and the likelihood function (in particular, the tails of the two functions) that makes the posterior proper in the setting of mixture models.

For the moment a widely accepted objective Bayesian approach in the setting of mixture models does not exist. This work could be seen as a way to understand if the Jeffreys' prior could represent a reasonable alternative to existing solutions. However, further research is needed, in particular to prove if and in which cases the posterior distribution derived from the Jeffreys' prior is proper and to generalize the Jeffreys' prior to models with non-Gaussian components or with a non-fixed number of components.

References

[1] Banterle, M., Grazian, C., Robert, C.P.: Accelerating Metropolis-Hastings algorithms: delayed acceptance with prefetching. arXiv:1406.2660 (2014)
[2] Basford, K.E., McLachlan, G.J.: Likelihood estimation with normal mixture models. Appl. Stat. **34**(3), 282–289 (1985)
[3] Diebolt, J. and Robert, C.P.: Estimation of finite mixture distributions through Bayesian sampling. J. R. Stat. Soc. Ser. B Stat Methodol. **56**(2), 363–375 (1994)
[4] Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes, 1st edn. Springer, New York (2006)
[5] Jeffreys', H.: Theory of Probability, 1st edn. The Clarendon Press, Oxford (1939)
[6] McLachlan, G.J., Peel, D.: Finite Mixture Models, 1st edn. Wiley, Newark (2000)
[7] Mengersen, K., Robert, C.: Testing for mixtures: a Bayesian entropic approach (with discussion). In: Berger, J., Bernardo, J., Dawid, A., Lindley, D., Smith, A. (eds.) Bayesian Statistics, vol. 5, pp. 255–276. Oxford University Press, Oxford (1996)
[8] Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components. J. R. Stat. Soc. Ser. B Stat Methodol. **59**(4), 731–792 (1997)
[9] Robert, C.: The Bayesian Choice, 2nd ed. Springer, New York (2001)
[10] Roeder, K., Wasserman, L.: Practical Bayesian density estimation using mixtures of normals. J. Am. Stat. Assoc. **92**(439), 894–902 (1997)
[11] Titterington, D.M., Smith, A.F.M., Makov, U.E.: Statistical Analysis of Finite Mixture Distributions, vol. 7. Wiley, Newark (1985)
[12] Wasserman, L.: Asymptotic inference for mixture models using data-dependent priors. J. R. Stat. Soc. Ser. B Stat Methodol. **62**(1), 159–180 (2000)