On a loss-based prior for the number of components in mixture models

Clara Grazian , University of New South Wales & Università degli Studi “Gabriele d'Annunzio”

Cristiano Villa, University of Kent

Brunero Liseo, Sapienza Università di Roma

Abstract

We introduce a prior distribution for the number of components of a mixture model. The prior considers the worth of each possible mixture, measured by a loss function with two components: one measures the loss in information in choosing the “wrong” mixture and one the loss due to complexity.

Keywords: mixture models ; Bayesian inference ; default priors ; loss-based priors ; clustering

1 Introduction

This paper takes a novel look at the construction of a prior distribution for the number of components for finite mixture models. These models represent a flexible and rich way of modeling data, allowing to extend the collection of probability distributions that can be considered and used. Mixture models have been widely developed and researched upon for over a century. To name a few key contributions, we have Titterington et al. (1985), Neal (1992), McLachlan and Peel (2000), Marin et al. (2005), Frühwirth–Schnatter (2006) and the recently issued Celeux et al. (2019). Besides the general literature on mixture models, a wide range of applications have been discussed, including genetics and gene expression profiling (McLachlan et al., 2002; Yeung et al., 2001), economics and finance (Juárez and Steel, 2010; Dias et al., 2010), social sciences (Reynolds at al., 2000; Handcock at al., 2007) and more.

The basic idea of a mixture model is to assume that observations x are drawn from a density which is the result of a combination of components

$$x \sim \sum_{j=1}^{k} \omega_j f_j(\cdot \mid \theta_j), \tag{1}$$

where the form of f_j is known for each j, while the parameters θ_j and the weights ω_j are unknown and have to be estimated. In this work, we assume k to be unknown as well and, in accordance to the Bayesian framework, we assign a prior distribution to it. We think that a golden standard approach in how to identify the prior distribution for the parameter k is still an open problem in the literature and a comparative study among the available proposals to understand advantages and disadvantages of them is missing. In this paper, we aim at presenting such comparison among the main proposals, namely the uniform prior, the Poisson prior and the loss-based prior here proposed. Other methods to deal with an unknown k are the following. One way is based on model selection and consists in fitting mixtures with $k = 1, \dots, K$ (for a suitable K) and comparing the models through some index, such as the Bayesian information criterion; see, for example, Baudry et al. (2010) and Celeux et al. (2019) for an analysis of model choice approaches in this setting. Alternatively, one could set a large k and let the weights' posterior behaviour to identify which components are meaningful. This is known as an *overfitted mixture model* and the aim is to define a prior distribution which has a conservative property in reducing a posteriori the number of meaningful components (Rousseau and Mengersen, 2011); Grazian and Robert (2018) have discussed the same approach by using the Jeffreys prior for the mixture weights conditionally on the other parameters, while Malsiner–Walli et al. (2016) estimate the posterior distribution of the number of meaningful components by specifying a sparse Dirichlet prior on the component weights.

The other main line of research in the setting of mixture is using a nonparametric Dirichlet process prior, as in Antoniak (1974), where an infinite components mixture is assumed by construction and the number of clusters is inferred by implementing Monte Carlo Markov Chain (MCMC) algorithms, see Müller and Mitra (2013) for a recent survey. However, while these models seem to have good properties in terms of density estimation (see, for example, Ghosal and Van der Vaart (2017) for a thorough review of posterior asymptotics results), there is some suggestion that inference of the number of components is not consistent (Miller and Harrison, 2014) for a large class of nonparametric mixtures over a large variety of families of distributional components.

From the point of view of the implementation, several techniques have been proposed to deal with k through the use of a prior $P(k)$: see, for example, Richardson and Green (1997), Stephens (2000), Nobile and Fearnside (2007) and McCullagh and Yang (2008). A well-known and widely used method is the reversible-jump MCMC (Green, 1995) which, due to its non-trivial set up, has led to the search of alternatives. A recent and interesting one is proposed by Miller and Harrison (2018), where the model in (1) is written as

$$\begin{aligned}
&k \sim P(k), \\
&(\omega_1, \ldots, \omega_k) \sim \mathrm{Dir}(\gamma, \ldots, \gamma), \qquad Z_1, \ldots, Z_n \sim (\omega_1, \ldots, \omega_k), \\
&\theta_1, \ldots, \theta_k \sim H, \\
&x_i \sim f_{\theta_{Z_j}},
\end{aligned} \tag{2}$$

where $P(k)$ is the prior on the number of components defined over the set $\{1, 2, \ldots\}$, H is the prior base measure, both the Zs and the θs are conditionally independent and identically distributed and the Zs are latent variables describing the component membership. Miller and Harrison (2018) then show how the stick-breaking representation of the Dirichlet mixture model can be efficiently exploited in a finite mixture model as well. Here we follow this suggestion and both our simulation studies and real data analysis have been obtained by the use of the Jain-Neal split-merge samplers (Jain and Neal, 2004, 2007), as implemented by the above authors. Our paper is concentrated on the choice of a prior for k and, conditionally on k, we follow the model described by equations (2). Notwithstanding, it is interesting to note how our model can be reinterpreted, from a Bayesian nonparametric perspective, as a mixture of Pitman–Yor processes with a negative discount parameter, i.e. $-\gamma$; see for example Gnedin and Pitman (2005) and De Blasi at al. (2015). This observation is on the line of building a bridge from parametric to nonparametric Bayesian inference, already discussed and explored by Rousseau and Mengersen (2011) and Malsiner–Walli et al. (2016).

In terms of the determination of $P(k)$, which is the focus of this work, the literature is definitely gaunt. In particular, it appears that there is only one proposed prior for k with a non-informative flavour, that is $k \sim \mathrm{Poi}(1)$ (Nobile, 2005). Although other authors proposed to use a prior proportional to a Poisson distribution, see for example Phillips and Smith (1996) and Stephens (2000), only Nobile (2005) gave some theoretical justifications on how to choose the Poisson parameter when there is lack of prior knowledge about k. Another option, suitable when there is no sufficient prior information, would be to assign equal prior mass to every value of k; however, in the case one would like to consider, at least theoretically, the possibility of having an infinite support for the number of components, this last solution would not be viable or would need a truncation of the support which might influence inference. Alternatively, the geometric distribution depicts a possible representation of prior uncertainty (Miller and Harrison, 2018), although no discussion is reserved in setting the value of the parameter in a scenario of insufficient prior information for the number of components. Finally, Gnedin (2010) discusses a heavy-tailed prior for k in a nonparametric setting.

Although the illustrations we present here will refer to mixtures of univariate and multivariate normal densities, the loss-based prior for k we introduce does not depend on the form of the f_js, therefore it is suitable for any mixture. Throughout the paper we will adopt, for the weights and the component parameters, the priors proposed in Miller and Harrison (2018); this will not affect the analysis of the results and the comparisons among different priors for k.

2 Prior for the number of components

Let us consider the finite mixture distribution

$$g(x \mid k, \omega, \theta) = \sum_{j=1}^{k} \omega_j f_j(x_i \mid \theta_j), \qquad i = 1, \ldots, n, \tag{3}$$

for a set of observations $x_1, \ldots, x_n$, where $f_j(\cdot|\cdot)$ is the probability distribution of the j-th component, $\theta = (\theta_1, \ldots, \theta_k)$, θ_j is the (possibly vector-valued) parameter of f_j and $\omega = (\omega_1, \ldots, \omega_k)$ are the weights of the components, with $\omega_j > 0$ for $\forall j = 1, \cdots, k$ and $\sum_{j=1}^{k} \omega_j = 1$. In the following, we will focus on the prior distribution for k and we consider the mixture weights $\{\omega_j\}_{j=1}^{k}$ and the parameters of the mixture components $\{\theta\}_{j=1}^{k}$ as independent; although other possibilities are easy to implement.

For model (3) the prior can be specified as $\pi(k, \omega, \theta) = P(k)\pi(\omega \mid k)\pi(\theta \mid k)$. The aim of this paper is to define a prior for k, therefore the prior distributions for ω and θ will be chosen to be proper "standard" priors, minimally informative if necessary; see, for example, Richardson and Green (1997) or Miller and Harrison (2018). The posterior for k is then given by

$$P(k \mid x) \propto \int f(x \mid k, \omega, \theta) \times P(k)\pi(\omega \mid k)\pi(\theta \mid k)\, d\omega\, d\theta.$$

It is now fundamental to discuss the support of k. Although for practical purposes the range of values k can take is finite, $k = 1, 2, \ldots, K$, it may be appropriate to define a prior over $\mathbb{N}$. In fact, by truncating the support of k there may be possible distortions of the posterior around the boundary, affecting the inferential results. It has to be noted that this is needed when using a uniform prior, since the prior on k must be proper, as proved by Nobile (2005). It seems, therefore, more reasonable to use a proper prior defined on $\mathbb{N}$.

The posterior distribution on the number of components of a mixture is known to show inconsistency problems in the nonparametric setting related to the use of Pitman-Yor prior processes (Miller and Harrison, 2014). We believe these problems could be prevented by penalising larger values and, therefore, we propose to define the prior distribution with a loss-based approach. While a theoretical analysis of the properties of the prior distribution we propose is out of the scope of this work, we aim at showing through simulations that the posterior distribution can concentrate around the "true" number of components.

To obtain the loss-based prior on k, we build on Villa and Walker (2015) and Villa and Lee (2019). In particular, we define the prior on k by assigning a prior on the space of models determined by the mixtures with $k = 1, 2, \ldots$ components. The basic idea is that we can assign a *worth* to each mixture by objectively measuring what is lost if such mixture is removed from the space of models, and it is the true one. While in Villa and Walker (2015) the worth of a model was associated to a measure of loss in information only, in cases of mixture models it is sensible to include a component of loss due to the complexity of the model as well (Villa and Lee, 2019). Thus, the loss associated to a mixture with k components is formed by the cumulative loss $\text{Loss}(k) = \text{Loss}_I(k) + \text{Loss}_C(k)$, where $\text{Loss}_I(k)$ is the component measuring the loss in information and $\text{Loss}_C(k)$ is the component measuring the loss due to complexity.

The quantification of the worth comes from a result in Berk (1966) which states that, if the model is misspecified, the posterior distribution asymptotically tends to accumulate at the most similar model so to minimise the loss in information, where the similarity is measured by the Kullback–Leibler divergence. In general, if we consider mixture models $M_s = \left\{g_s(x|\tilde{\theta}_s), \pi_s(\tilde{\theta}_s)\right\}$, for $s = 1, \ldots, l$, where $\tilde{\theta}_s = (\omega_s, \theta_s)$ we have that

$$\text{Loss}_I(k) = \mathbb{E}_{\pi_s}\left\{\inf_{\tilde{\theta}_m, m \neq j} D_{KL}\Big(g_s(x|\tilde{\theta}_s) \| g_m(x|\tilde{\theta}_m)\Big)\right\}, \tag{4}$$

where the expectation is with respect to the prior on the parameters $\tilde{\theta}_s$ to reflect the uncertainty about their true value, and the infimum is obtained by considering θ_s as fixed. The above loss is linked to the prior mass by means of the *self-information* loss function (Merhav and Feder, 1998), which associates a loss to a probability statement, say $P(A)$, and it has the form $-\log P(A)$. As such, we can equate the self-information loss associated to a mixture with k components to the information loss related to its *worth*, given by $-\text{Loss}_I(k)$, obtaining

$$P(k) \propto \exp\left\{\text{Loss}_I(k)\right\}.$$

It is straightforward to see that the loss in (4) attains its minimum at zero. In fact, consider mixtures $g_k = \sum_{j=1}^{k} \omega_j f_j(x|\theta_j)$ and $g_{k'} = \sum_{j=1}^{k} \breve{\omega}_j f_j(x|\breve{\theta}_j) + \breve{\omega}_{k+1} f_{k+1}(x)$; the minimum, which is zero, is obtained when we set $\breve{\omega}_j = \omega_j$ and $\breve{\theta}_j = \theta_j$, for $j = 1, \ldots, k$; implying $\breve{\omega}_{k+1} = 0$. The same result applies for any $k' > k$; however, as we are seeking to the most "similar" model to the true one, setting $k' = k$ is sensible, as the perturbation in terms of overall uncertainty would be minimal. It is easy to see that if $k' < k$, then the Kullback–Leibler divergence will be larger than zero. In conclusion, the infimum in (4) is zero for every k, resulting in $\text{Loss}_I(k) = 0$ for every k.

To fully describe the *worth* of a mixture model it is also necessary to take into consideration its complexity. This is determined as follows. If we keep the mixture model with k components, the loss would be related to the number of parameters that have to be estimated, and therefore the number of components. So, the loss of keeping a mixture model increases with the number of components it contains, and we have $\mathrm{Loss}_C(k) = \mathrm{U}(\text{keep } k) = -c \cdot k$, where $\mathrm{U}(\cdot)$ is a utility function. As seen above, for mixture models, there is no loss in information in selecting the "wrong" mixture, as such the prior on k becomes

$$P(k) \propto \exp\left\{-c \cdot k\right\}, \tag{5}$$

where $c > 0$ is included as loss functions are defined up to a constant. Although the prior in (5) could be directly used, with the interpretation that c is a hyper-parameter which allows to control for sparsity, our recommendation is to reparametrise it by setting $p = \exp(-c)$ and assign a suitable prior to p. In particular, by having $p \sim \mathrm{Beta}(\alpha, \beta)$, the prior for k is a particular beta-negative-binomial, that is a beta-geometric distribution, when the support for k is infinite, as the following Theorem 2.1 (whose proof is in the Supplementary Material) shows. The complexity loss is set to be linear for simplicity, and other choices are also possible. However, choosing a linear loss seems reasonable when seeing it as a penalisation on the increasing number of parameters. Other loss functions can be considered, for example to take into account asymmetric penalisation between small and large numbers of components. Theorem 2.1 shows that a linear loss provides an elegant derivation of the prior distribution $P(k)$ which is analytically manageable, in comparison with other choices.

Theorem 2.1 *Consider the prior distribution for the number of components of a finite mixture model, as defined in* (5)*, where we set* $p = \exp\{-c\}$ *and* $k = 1, 2, \ldots$*. If we choose* $p \sim$ *Beta*(α, β)*, with* $\alpha, \beta > 0$*, then*

$$P(k|p) = p^{k-1}(1-p),$$

which is a geometric distribution with parameter $1-p$*, and*

$$P(k) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(k+\beta-1)\Gamma(\alpha+1)}{\Gamma(k+\alpha+\beta)}, \tag{6}$$

which is a beta-negative-binomial distribution where the number of failures before the experiment is stopped is equal to 1, and shape parameters α *and* β*.*

The prior in (6) is strictly positive on the whole support of k. This is a necessary condition (Nobile, 1994) to have consistency on the number of components. In addition, the prior in (6) is proper, which is another requirement to yield a proper posterior (Nobile, 2004) when the support is $k = \{1, 2, \ldots\}$. On this aspect, as the Jeffreys prior for a geometric distribution is improper, the prior for k will be improper as well. As such, a default choice for $P(k)$ should be chosen on different grounds. In particular, the default choice will not give any preference to particular values of p, and this can be achieved by setting $p \sim \text{Beta}(1, 1)$. The resulting prior is then a beta-negative-binomial with all parameter values equal to one which can be approximated by using a Stirling's approximation to the beta function as $P(k) = [k(k+1)]^{-1}$.

In a more general setting, the parameters α and β of the Beta prior on p can be used to reflect available prior information about the true number of components. The expectation and the variance of the prior in (6) are respectively

$$\mathbb{E}(k) = \mathbb{E}(\mathbb{E}\{k|p\}) = \mathbb{E}(p^{-1}) = \frac{\alpha + \beta - 1}{\alpha - 1}, \qquad \text{for } \alpha > 1, \tag{7}$$

$$\text{Var}(k) = \mathbb{E}(\text{Var}\{k|p\}) + \text{Var}(\mathbb{E}\{k|p\}) = \frac{\alpha\beta(\alpha + \beta - 1)}{(\alpha - 2)(\alpha - 1)^2}, \qquad \text{for } \alpha > 2. \tag{8}$$

From equation (7) we see that, as $\beta \to 0$, then $\mathbb{E}(k) \to 1$. So, for a given $\alpha > 1$, we have that the hyper-parameter β can be interpreted as the quantity controlling how many components in the mixture we want *a priori*. The choice of α, among values strictly larger than 2, allows to control the variance of the prior, i.e. how certain (or uncertain) *a priori* we are about the true value of k.

If the support for k is finite, say $k = \{1, 2, \ldots, K\}$, the prior for the number of components (with $p \sim \text{Beta}(\alpha, \beta)$) will have the form:

$$P(k) = \int_0^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{k+\alpha-2}(1-p)^{\beta} \frac{1}{1 - p^K} \, dp, \tag{9}$$

which does not have a closed form. Although the prior in (9) can be easily implemented in a Markov Chain Monte Carlo procedure, one has to be careful as its performance might depend on the choice of K. Besides this, the prior certainly yields a proper posterior for k and is consistent on the number of components.

3 Illustrations

To illustrate the performance of the loss-based prior we have run a simulation study (results are shown in the Appendix) and analysed two data sets. We have considered univariate and multivariate scenarios, comparing the proposed prior, under default settings, with current alternatives found in the literature.

Before describing the analysis and illustrate the results, the following clarifications have to be made. First, as the aim of this paper is to propose a novel prior distribution for the number of components, we do not discuss in detail the prior assigned to model weights and to the parameters of the components of the mixture. Second, for the same reason, we limit the examples to mixture of normal densities. In fact, keeping both model and priors relatively straightforward allows to better appreciate any difference in the priors. Finally, the computational algorithm implemented assumes that the maximum number of components in the mixture is 50, so that the uniform prior is defined over $k = \{1, \ldots, 50\}$; although the truncation is necessary for the uniform prior only, so to have a proper posterior, the choice of 50 is sufficiently large to not interfere with any of the analysis performed.

For the implementation, we have used the algorithms described in Miller and Harrison (2018).

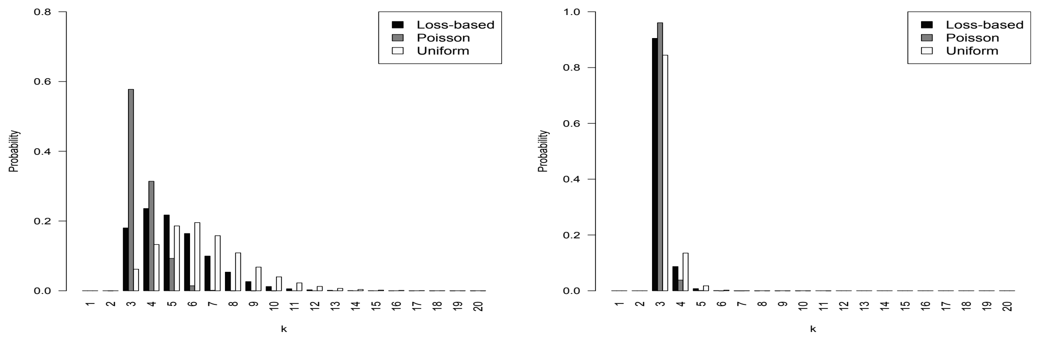
3.1 Real data sets

In this section we illustrate the performance of the prior by analysing two available data sets. The first dataset is the galaxy data set (Roeder, 1990), which is considered a benchmark for comparison in the univariate case. We also consider a multivariate case; in particular, the discriminating cancer subtypes using gene expression data set (Armstrong et al., 2001), which has $n = 72$ observations for $d = 1081$ variables.

3.1.1 The Galaxy Dataset

The galaxy data sets contains the velocities of 82 galaxies in the Corona Borealis region. Given that the focus here is on the prior for the number of component, we do not go beyond an already tested set up for the model. In particular, the model used in Richardson and Green (1997) where the components of the mixture are normal densities, i.e. $f_j(x) = N(x|\mu_j, \lambda_j^{-1}))$, with independent priors for the parameters, normal densities for the means ($\mu_j \sim N(\mu_0, \sigma_0^2)$) and gamma densities for the precision ($\lambda_j \sim \text{Gamma}(a, b)$). We also have $a = 2$, $b \sim \text{Gamma}(a_0, b_0)$, with $a_0 = 0.2$, while data-dependent priors are chosen for the remaining hyper-parameters: $\mu_0 = (\max x_i + \min x_i)/2$, $\sigma_0 = \max x_i - \min x_i$ and $b_0 = 10/\sigma_0^2$.

The evaluation of the number of components has proved to be delicate, for example, Roeder and Wasserman (1997) select three components with information criteria, Richardson and Green (1997) identify five or six components by using a uniform prior for k and implementing a reversible-jump MCMC method, Escobar and West (1995) propose an nonparametric approach based on Dirichlet processes which shows a posterior mode on seven components and Grazian and Robert (2018) present a conservative approach based on the use of the Jeffreys prior for the mixture weights where the posterior mass is concentrated around three non-zero weights.



(a) Galaxy dataset

(b) Gene dataset

Figure 1: Posterior distributions of the number of components.

The posteriors obtained by implementing the loss-based prior, the uniform prior and the Poisson($\lambda = 1$) prior are plotted in Figure (1a) and have modes, respectively, at $k = 4$, $k = 6$ and $k = 3$. There is no unanimous agreement in the number of components in the literature and this is supported by the results in Figure (1a), which shows estimates of k comparable to what has been already identified. However, while the posterior 95% credible intervals obtained with the loss-based prior and the uniform prior, $[3, 9]$ and $[3, 12]$ respectively, are sensible, the interval for the Poisson(1) appears to be quite narrow $[3, 5]$, excluding values of k previously estimated in the literature. It seems that the loss-based prior provides an intermediate posterior distribution, between the one deriving from the Poisson prior, which is very peaked around 3, and the one deriving from the uniform prior, which gives non-negligible posterior mass to large values as 12.

3.1.2 Gene expression data

Mixture modelling is becoming popular in genomics to identify clusters based on how much a gene is expressed in different tissues. For example, identification of cancer types based on Gaussian mixture models has been proposed in Yeung et al. (2001), McLachlan et al. (2002) and Medvedovic et al. (2004) among others, with deSouto et al. (2008) showing that Gaussian mixture models exhibit the best performance among seven clustering methods on 35 datasets, given that the true number of components is known. This is a problem of primary interest in health sciences, since, once a particular cancer type is identified, it is possible to offer patient-specific treatments As stated before, Gaussian mixture models may represent an essential tool in this setting, however it is necessary to identify the right number of groups (deSouto et al., 2008). In practice, this could be difficult, therefore the availability of a method to perform inference on the number of components which clearly states the assumption and the *a priori* knowledge is essential.

Following Miller and Harrison (2018), we analyze a dataset collected by Armstrong et al. (2001) for a study of leukemia subtypes, measuring gene expression levels in 72 patients. The goal of our analysis is showing if it is evident from the data a third type of leukemia, beyond the standard ones (acute lymphoblastic leukemia and myelogenous leukemia), as proposed by the authors. The analysis of this dataset has given results very similar for the three priors. As Figure (1b) shows, the posterior distributions for k do not differ much, which is supported by the posterior mode, $k = 3$, and posterior 95% credible intervals, $[3, 4]$, in all cases. It is obvious that the amount on information about k in the data is sufficiently strong to dominate any of the used priors.

4 Conclusions

We see that, in a setting of limited information, the prior chosen for the number of components influences the posterior distribution; in particular, the uniform prior, which is often used as a default prior, does not seem to be conservative. In terms of inference, some level of conservativeness should be preferred, given the fact that the complexity of the inferential problem explodes with the number of meaningful components. On the other hand, the Poisson(1) prior seems to be too conservative, so that the true value may not even included in the posterior credible interval. In an informative context, the proposed loss-based prior allows to include information on both centrality and variability of the uncertainty about k. This possibility appears to lack in currently used options, such as the Poisson or the geometric prior, where only one piece of information can be included. Analysis on both real and simulated data shows that the loss-based prior represents a good compromise between having a prior which excessively penalises for complexity (Poisson(1)) and the uniform prior which suffers from theoretical and implementation weaknesses. We think that this work may also offer an important contribution in an applied context: mixture models offer a flexible tool to analyse non-standard data, however identifying the correct number of components is essential for a good fit. We believe it is important to use a prior distribution for the number of components which both can represent particular assumptions on the model and shows a property of conservativeness to better interpret and estimate the model, as the one derived here.

Acknoledgements

The work of the first author has partially been developed under the PRIN2015 supported-project: Environmental processes and human activities: capturing their interactions via statistical methods (EPHAStat), funded by MIUR (Italian Ministry of Education, University and Scientific Research) (20154X8K23-SH3).

We would like to thank the Editor, the Associate Editor and two anonymous referees for their valuable comments. We would also like to thank Jeff Miller (Harvard T.H. Chan School of Public Health) for the useful discussion about the implementation of the algorithm developed in Miller and Harrison (2018). The codes used in this work are a slight modification of the Julia codes available at the website `https://github.com/jwmi/BayesianMixtures.jl/`.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics* 1152-1174.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. (2001). MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia. *Nature Genetics* **30**, 41–47.

Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining mixture components for clustering. *Journal of computational and graphical statistics* **19**(2), 332–353.

Berk, R.H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. of Math. Statist.* **37**, 51–58

Celeux, G., Frühwirth–Schnatter, S. and Robert, C.P. (2019). *Handbook of Mixture Analysis.* Boca Raton, FL: CRC Press.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I. and Ruggero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet Process? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 212–229.

de Souto,M. C., Costa, I. G., deAraujo, D. S., Ludermir, T. B., and Schliep, A. (2008). Clustering Cancer Gene Expression Data: A Comparative Study. *BMC Bioinformatics* **9**, 497.

Dias, J.G., Vermunt, J.K. and Ramos, S. (2010). Mixture hidden Markov models in finance research. A. Fink et al., (eds.), *Advances in Data Analysis, Data Handling and Business Intelligence, Studies in Classification, Data Analysis, and Knowledge Organization* Springer-Verlag Berlin Heidelberg

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Frankel, A. and Kamenica, E. (2018) Quantifying information and uncertainty. Proceedings. University of Chicago.

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

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference.* Cambridge University Press, Cambridge.

Gnedin, A. (2010). A species sampling model with finitely many types. *Electron. Commun. Probab.* **15**, 79–88

Gnedin, A., and Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauch. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **325**, 83–102

Grazian, C., and Robert, C. P. (2018). Jeffreys priors for mixture estimation: Properties and alternatives. *Computational Statistics & Data Analysis* **121**, 149–163

Green, AP.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732

Handcock, M.S., Raftery, A.E. and Tantrum, J.M. (2007). Model-based clustering for social networks. *J. R. Stat. Soc. A* **170**, 301–354

Jain, S., and Neal, R. M. (2004). A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics* **13**, 158–182

Jain, S., and Neal, R. M. (2007). Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model. *Bayesian Analysis* **2**, 445–472

Juárez, M.A. and Steel, M.F.J. (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *J. Bus. Econ. Stat.* **28**, 52–66

Malsiner–Walli, G., Frühwirth–Schnatter, S., and Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26**, 303–324

McCullagh, P. and Yang, J. (2008). How many clusters? *Bayesian Analysis* **3**, 101–120

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models.* New York: J. Wiley

McLachlan, G.J., Bean, R.W. and Peel, D. (2002). A mixture-model based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422

Medvedovic,M., Yeung, K. Y., and Bumgarner, R. E. (2004). Bayesian Mixture Model Based Clustering of Replicated-Microarray Data *Bioinformatics* **20**, 1222–1232.

Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Trans. Inf. Theory* **44**, 2124–2147

Marin, J.M., Mengersen, K.L. and Robert, C.P. (2005). Bayesian Modelling and Inference on Mixtures of Distributions. D. Dey and C.R. Rao (eds.), *Handbook of Statistics, Vol. 25, pp. 459–507* Elsevier.

Miller, J.W. and Harrison, M.T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research* **15**(1), 3333-3370.

Miller, J.W. and Harrison, M.T. (2018). Mixture models with a prior on the number of components. *J. American Stats. Assoc.* **113**, 340–356.

Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference–why and how. *Bayesian analysis* **8**(2).

Neal, R.M. (1992). *Bayesian Mixture Modeling,* in Maximum Entropy and Bayesian Methods *eds. C.R. Smith, G.J. Erickson and P.O. Neudorfer.* New York: Springer

Nobile, A. (1994). *Bayesian Analysis of Finite Mixture Distributions.* Ph.D. thesis, Pittsburgh, PA: Department of Statistics, Carneige Mellon University.

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *Ann. of Statist.* **32**, 2044–2073

Nobile, A. (2005). Bayesian finite mixtures: a note on prior specification and posterior computation. *Technical Report* Department of Statistics, University of Glasgow

Nobile, A. and Fearnside, A.T. (2007). Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Statistics and Computing* **17**, 147–162

Phillips, D.B and Smith, A.F.M. (1996). Bayesian model comparison via jump diffusions. Gilks W.R., Richardson S. and Spiegelhalter D.J. (eds.), *Markov Chain Monte Carlo in Practice* Chapman & Hall, London

Reynolds, D.A., Quatieri, T.F. and Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. *Data Signal Processing* **10**, 19–41

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures With an Unknown Number of Components. *J. R. Stat. Soc. B* **59**, 731–792

Roeder, K. (1990). Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies. *J. American Stats. Assoc.* **85**, 617–624

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

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Ann. of Statist.* **28**, 40–74

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

Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science* **185**, 1124–1131.

Villa, C. and Lee, J.E. (2019). A loss-based prior for variable selection in linear regression methods. *Bayesian Analysis*, Advance Publication, doi:10.1214/19-BA1162.

Villa, C. and Walker, S.G. (2015). An objective Bayesian criterion to determine model prior probabilities. *Scandinavian Journal of Statistics* **42**, 947–966

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **98**, 894–902.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. (2001). Model-based clustering and data transformation for gene expression data. *Bioinformatics* **17**, 977–987