Quantifying the uncertainty of partitions for infinite mixture models

Aurore Lavigne[a], Silvia Liverani[b]

[a] *Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, Lille, F-59000, France*
[b] *School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK*

Abstract

Bayesian clustering models, such as Dirichlet process mixture models (DPMMs), are sophisticated flexible models. They induce a posterior distribution on the set of all partitions of a set of observations. Analysing this posterior distribution is of great interest, but it comes with several challenges. First of all, the number of partitions is overwhelmingly large even for moderate values of the number of observations. Consequently the sample space of the posterior distribution of the partitions is not explored well by MCMC samplers. Second, due to the complexity of representing the uncertainty of partitions, usually only maximum a posteriori estimates of the posterior distribution of partitions are provided and discussed in the literature. In this paper we propose a numerical and graphical method for quantifying the uncertainty of the clusters of a given partition of the data and we suggest how this tool can be used to learn about the partition uncertainty.

Keywords: Dirichlet process mixture model, Bayesian methods, clustering, uncertainty

1. Introduction

An infinite mixture model is a mixture model with potentially infinitely many mixture components. One example of such models is the Dirichlet process mixture model (DPMM), an extremely popular model, used for density estimation, prediction or clustering, which can estimate the number of components. This is a flexible method which allows for many types of data (e.g. continuous, count, categorical, survival) and can also allow for joint modelling of covariates and response variables. These models are widely used in many fields, including machine learning, genetics and epidemiology.

Clustering aims at grouping individuals according to their degree of similarity. We expect that variability is small intra-cluster but high extra-clusters. In model based clustering, each component of a mixture is assigned to a cluster. Data from a same cluster are assumed to follow a same parametric probability density, leading to the rule one component, one cluster. However, clusters may not arise from a single parametric distribution and the mixture may fail to retrieve the clusters, over-estimating the number of clusters (Baudry et al., 2010; Hennig, 2010). Bayesian mixture models are more flexible, since we consider a probability distribution on the space of all possible partitions. In the case of DPMM the mixture distribution is assumed to be engendered by a Dirichlet process. As a result, observations are not thought to be generated by a unique mixture, but they could be generated by a large set of mixture distributions. In this context, the clustering task is still debated. We could consider the partition derived from only one mixture, as it is done when the maximum a posteriori is researched (Fritsch et al., 2009), but we choose not to ignore all of the other possible mixtures. In this case, in practice, clustering is about finding a consensus from a large set of data partitions. The consensus partition obtained is not necessarily a member of the initial set, and it cannot be related to a given mixture.

In order to relativise the importance of the estimate partition, it is important to get information about its uncertainty. Posterior similarity matrices are the tools generally used to assess partition uncertainty. Based on the MCMC draws, the posterior similarity matrix gives for each pair of subject their probability to be in a same cluster. However this matrix has the drawback of not giving an index of confidence for one partition. Wade et al. (2018) develop the foundation of a mathematical framework to compute credible balls in the partition space, using the variation of information as a measure of distance on the partition space. However,

(1) the upper and lower bounds are generally not unique, (2) in practice, even for very small datasets, the number of partitions expected within the 95% credible ball is so large that an analyst cannot consider all of the them and (3) estimates are given among the MCMC draws. Considering only partitions explored in the MCMC draws is a big limitation of the method, as the most desirable partition is often outside the MCMC sample (Hastie et al., 2015; Jing et al., 2022). In practice, the 'consensus' partitions are usually obtained by postprocessing the MCMC draws, for example applying partitioning around medoids on the dissimilarity matrix (Liverani et al., 2015).

Finite mixture models are not subjected to all of these pitfalls, because first the space of partition is limited by the fixed number of clusters and second, constraints are generally added in application to make the different clusters identifiable accross the MCMC draws. As a consequence, it is possible to evaluate the posterior distribution of a given cluster in a partition. It is then possible to evaluate the uncertainty of the partition, by computing for each individual its posterior probability of belonging to a cluster.

In this paper we propose to take advantage of the two approaches: finite mixtures models and DPMM. We supply a simple and efficient method for quantifying the uncertainty of any partition in a DPMM, as it can be done for finite mixture models. Our method is based on the predictive distribution, a recommended choice for model checking. The uncertainty can be evaluated for any partitions including those that do not belong to the explored MCMC sample.

This paper is organised as follows. In Section 2 we review DPMMs and methods to identifying the consensus partition. In Section 3 we propose our tools for quantifying the uncertainty of a given partition. In Section 4 we discuss how the methods proposed perform on a dataset.

2. Bayesian clustering with Dirichlet process mixture models

The DPMM is a Bayesian clustering model, in which a Dirichlet process (DP) is assigned as latent distribution on the parameters of the observation distribution. We use the stick breaking representation of the DP by Sethuraman (1994), in which the infinite mixture is explicit. The (possibly multivariate) data $\boldsymbol{D}_n = (D_1, D_2, \ldots, D_n)$ follow an infinite mixture distribution, where component c of the mixture is a parametric density of the form $f_c(\cdot) = f(\cdot|\Theta_c)$ parametrised by some component specific parameter Θ_c, so

$$\begin{aligned} D_i|\Theta_1, \Theta_2, \cdots &\sim \sum_{c=1}^{\infty} \psi_c f(D_i|\Theta_c) \text{ i.i.d. for } i = 1, 2, \ldots, n, \\ \Theta_c &\sim P_{\Theta_0} \text{ i.i.d. for } c \in \mathbb{Z}^+, \text{ and } \psi_c = V_c \prod_{l<c}(1 - V_l) \text{ for } c \in \mathbb{Z}^+ \setminus \{1\} \end{aligned} \tag{1}$$

with Θ_c independent of V_c for $c \in \mathbb{Z}^+$, with $\psi_1 = V_1$, and $V_c \sim \text{Beta}(1, \alpha)$ i.i.d. for $c \in \mathbb{Z}^+$. By introducing the latent allocation variable Z_i of observation D_i, the first line of the DPMM in Eq. (1) can be replaced by $D_i|Z_i, \Theta_1, \Theta_2, \cdots \sim f(D_i|\Theta_{Z_i})$ i.i.d. for $i = 1, 2, \ldots, n$, and $\mathbf{P}(Z_i = c) = \psi_c$ for $i = 1, 2, \ldots, n$ and for $c \in \mathbb{Z}^+$. Bayesian posterior samples for the latent allocation variables can be effectively obtained from the model above. However, due to the categorical nature of the clustering variables and the lack of scalable algorithms, it is not immediately clear how one can appropriately summarise the output of partitions from this model. There are several methods available in the literature for selecting a single clustering estimate $\boldsymbol{Z}^*$ for an unknown number of clusters. Among them, many aim to retrieve the maximum a posteriori (Fritsch et al., 2009). As Liverani et al. (2015), we prefer to use Partitioning Around Medoids (PAM) to identify the partition $\boldsymbol{Z}$. This method is very effective and robust in our experience. It consists of processing the similarity matrix, S, through a deterministic clustering procedure where an optimal number of clusters can be chosen by maximizing an associated clustering score. Our proposal below to quantify the partition uncertainty is not influenced by the method used to derive the partition estimate.

3. Quantifying the uncertainty of a consensus partition

Finite mixture models are an attractive alternative to deterministic methods such as k-means because they supply a probabilistic framework for dealing with partition uncertainty. In these models, the density of an observation y is defined as a weighted sum of standard densities which represent the clusters of a partition and can overlap: $f(y) = \sum_{c=1}^{K} \omega_c f_c(y)$. The posterior probability that an observation y belongs to cluster c, is thus readily obtained by the Bayes theorem and is given by $P(\text{“}y \text{ belongs to cluster } c\text{”}) = w_c f_c(y)/f(y)$. However, when an infinite mixture model is considered, this method can not be readily applied, because of the infinite number of potential clusters and label switching. Here, we propose to write the predictive distribution of a DPMM, i.e. an infinite mixture model as a finite mixture, in order to take advantage of the Bayes' theorem. In the following section we present how we split the predictive distribution in k^* components and how they are connected to the k^* clusters of a given partition. Then, the results of this method will be visualised in an uncertainty table, as this is the most effective way to process the results.

3.1. The posterior predictive distribution as a finite mixture of distributions

Escobar and West (1995) discuss the posterior predictive distribution for DPMMs. They show that given a partition $\boldsymbol{Z}$, the predictive distribution of a new observation D_{n+1} is a mixture of marginal densities, one for each cluster formed by the partition $\boldsymbol{Z}$ and one for a potential new cluster. The predictive distribution of a DPMM is obtained by integrating over the space of partitions,

$$P(D_{n+1}|\mathbf{D}_n) = \frac{\alpha}{\alpha+n} f_0(D_{n+1}) + \frac{1}{\alpha+n} \sum_{\boldsymbol{Z}\in\mathcal{Z}} \sum_{c:n_c(\boldsymbol{Z})>0} n_c f(D_{n+1}|\{D_i : Z_i = c\}) p(\boldsymbol{Z}|\mathbf{D}_n) \tag{2}$$

where $\mathcal{Z}$ denotes the space of partitions of $\mathbf{D}_n$, $f_0(D_{n+1}) = \int f(D_{n+1}|\tilde{\Theta}_{n+1}) P_{\Theta_0}(\tilde{\Theta}_{n+1}) d\tilde{\Theta}_{n+1}$, and $f(D_{n+1}|\{D_i : Z_i = c\})$ is the predictive distribution for cluster c of partition $\boldsymbol{Z}$, $f(D_{n+1}|\{D_i : Z_i = c\})$ can be written as $\int f(D_{n+1}|\Theta_c) p(\Theta_c|\{D_i : Z_i = c\}) d\Theta_c$.

The posterior predictive can be rewritten, using the power set of $\boldsymbol{D}_n$, $\mathcal{P}(\boldsymbol{D}_n)$. The order within $\boldsymbol{D}_n$ does not influence the posterior predictive because the DPMM considered here is exchangeable. We now consider $\boldsymbol{D}_n$ as a set. A partition $\boldsymbol{Z}$ is thus a collection of some of the 2^n elements of $\mathcal{P}(\boldsymbol{D}_n)$ denoted $\boldsymbol{S}_j$. $\boldsymbol{Z} = \{\boldsymbol{S}_{j_1}, \cdots, \boldsymbol{S}_{j_{k_{\boldsymbol{Z}}}}\}$, where elements are non empty, pairwise disjoint and covering $\boldsymbol{D}_n$. Without loss of generality we set $\boldsymbol{S}_1 = \emptyset$ and $f_0(D_{n+1}) = f(D_{n+1}|\boldsymbol{S}_1)$. We show that

$$P(D_{n+1}|\mathbf{D}_n) \quad = \quad \frac{1}{\alpha+n} \sum_{j=1}^{2^n} \omega_j n_j f(D_{n+1}|\boldsymbol{S}_j), \tag{3}$$

where $\omega_j = \sum_{\{\boldsymbol{Z}\in\mathcal{Z}:\boldsymbol{S}_j\in\boldsymbol{Z}\}} p(\boldsymbol{Z}|\mathbf{D}_n)$ is the posterior probability that subset $\boldsymbol{S}_j$ is a cluster of the partition. For the empty set, we set $n_1\omega_1 = \alpha$, so that $\sum_{j=1}^{2^n} n_j\omega_j = \alpha + n$. The details of this proof are available in Appendix A.

3.2. The predictive distribution as a finite mixture

Our aim is to formulate the predictive posterior distribution as a finite mixture model linked to partition $\boldsymbol{Z}^*$, which is the consensus partition identified by postprocessing. The partition $\boldsymbol{Z}^*$ has k clusters, $\boldsymbol{S}_1^*, \boldsymbol{S}_2^*, \cdots, \boldsymbol{S}_k^*$. Therefore, our ideal finite mixture model will be a sum of k components with weights proportional to n_l, the number of observations in cluster l. Each component of the finite mixture will be represented by $\tilde{f}_l(\cdot)$ for cluster l, as follows,

$$P(D_{n+1}|\mathbf{D}_n) = \sum_{l=1}^{k} \frac{n_l}{n} \tilde{f}_l(D_{n+1}). \tag{4}$$

Because of the form of predictive distribution in Equation (3), the components of the finite mixture should be of the form

$$\tilde{f}_l(D_{n+1}) = \sum_{j=1}^{2^n} \frac{n_j \omega_j}{\alpha + n} \beta_j^l f(D_{n+1}|\boldsymbol{S}_j) \tag{5}$$

where β_j^l represents the part of the marginal distribution $f(D_{n+1}|\boldsymbol{S}_j)$ relating to component l in the finite mixture in Equation (4). Of course, $\tilde{f}_l(D_{n+1})$ should be a density distribution, constraining the sum $\sum_{j=1}^{2^n} n_j \omega_j \beta_j^l$ to be equal to $\alpha + n$, for any l. Also, the mixture in Equation (4) should be equal to Equation (3), meaning that for each subset j, $\sum_{l=1}^{k} \frac{n_l}{n} \beta_j^l = 1$. Therefore, our problem can be reduced to finding the β_j^l which satisfy these two sequences of constraints, and such that each component $\tilde{f}_l(D_{n+1})$ is as close as possible to the predictive distribution of data in cluster $\boldsymbol{S}_l^*$.

We propose a solution based on the simple empirical principle that the more $\boldsymbol{S}_j$ has data in common with $\boldsymbol{S}_l^*$, the more $f(.|\boldsymbol{S}_j)$ is close to the predictive distribution of $\boldsymbol{S}_l^*$. This principle leads us to propose this simple rule of allocation based on proportionality. The part of $f(.|\boldsymbol{S}_j)$ allocated to cluster l will be proportional to n_{jl}, the number of data both in $\boldsymbol{S}_j$ and $\boldsymbol{S}_l^*$. From these rules, a solution that satisfies all above mentioned constraints follows, $\beta_j^l = (n n_{lj})/(n_j n_l)$ if $j \leq 2$ and $\beta_j^l = 1$ if $j = 1$. The proof is given in Appendix B.

3.3. Estimation using MCMC draws

Consider the consensus partition $\boldsymbol{Z}^*$ and its k clusters. Using MCMC draws, we propose a Monte Carlo estimate of $\tilde{f}_l$. From a sequence of parameters $(\boldsymbol{Z}^t, \Theta_1^t, \Theta_2^t, \cdots)$ for $t \in \{1, \cdots, T\}$ drawn from the posterior distribution, and a sequence of draws, Θ_0^t, in the baseline distribution, we approximate $\tilde{f}_l(D_{n+1})$ using the following Monte-Carlo estimate

$$(\alpha + n)\hat{\tilde{f}}_l(D_{n+1}) = \frac{1}{T} \sum_{t=1}^{T} \left(\alpha f(D_{n+1}|\Theta_0^t) + \frac{n}{n_l} \sum_{c:n(\boldsymbol{Z}^t)>0} n_{lc} f(D_{n+1}|\Theta_c^t) \right). \tag{6}$$

This estimate is based on the posterior draws, readily available when MCMC inference is done. Its computation demands at each step, the evaluation of a small number of functions in about n points. If the integrals are in closed form, it is then possible to use the following estimate, with smaller variance

$$(\alpha + n)\hat{\tilde{f}}_l(D_{n+1}) = \alpha f_0(D_{n+1}) + \frac{n}{n_l} \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{c:n(\boldsymbol{Z}^t)>0} n_{lc} f(D_{n+1}|\{D_i : Z_i^t = c\}) \right). \tag{7}$$

Finally, we estimate p_i^l as the posterior probability that data point i belongs to cluster l of the $\boldsymbol{Z}^*$ partition as: $\hat{p}_i^l = n_l \hat{\tilde{f}}_l(D_i) / \sum_{j=1}^{k^*} n_j \hat{\tilde{f}}_j(D_i)$ We propose to calculate these probabilities for each data of $\boldsymbol{D}_n$ and for each cluster of $\boldsymbol{Z}^*$ and to visualize them graphically. In Appendix C we provide an application of this to mixtures of Gaussian distributions.

4. Application to velocity galaxy data

We consider the velocity galaxy dataset proposed by Roeder (1990). See Appendix D for a detailed description of the dataset and our implementation.

The predictive density and the densities $\tilde{f}_l(D_{n+1})$ are represented on the left hand side of Fig. 1. From the representation of the mixture of densities, the distinction between cluster 3 and 4 is unclear: some element of cluster 4 have a higher probability to belong to cluster 3. A clear representation of the proposed methodology is presented on the right hand side of Fig. 1, where the $\hat{p}_i^l$ are displayed for a graphical analysis of the uncertainty of the selected partition. In this figure we note that cluster 1, 2 and 5 are well defined, because they are composed of elements that have a high probability to belong to the cluster that they are

allocated to and low probability to belong to a different cluster. However, doubts may exist on the choice of clusters 3 and 4, as observations that have been allocated to them have high probability of belonging to either cluster. The graphical representation of cluster uncertainty was critically helpful to quantify this uncertainty, and it can scale to higher dimensions.

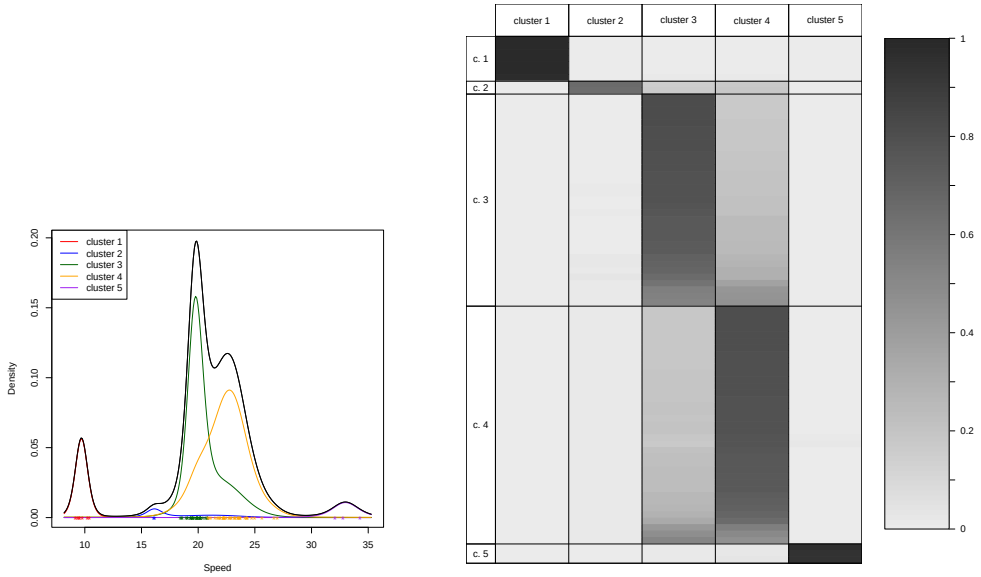


Figure 1: The plot on the left represents the predictive density (black), and densities $\tilde{f}_l(D_{n+1})$ of the mixture of Eq. (C.2). The consensus partition recognizes 5 clusters, data are represented by a star on the x-axis, they are colored according to their cluster allocation. The plot on the right hand side is a graphical representation of $\hat{p}_i^l$ for a visual analysis of cluster uncertainty. In this matrix, each row of the matrix corresponds to an observation in the original dataset. Each column represents a cluster in the partition that has been identified by PAM as a consensus. The observations are ordered according first to their cluster allocation in the consensus partition, and then according to their probability of belonging to each cluster.

5. Discussion

In this paper we have provided the mathematical justification that underpins a much needed uncertainty quantification tool for any given partition. As far as we know, we are the first to write the posterior predictive distribution for a DPMM as a finite mixture. The uncertainty that we have computed can be visualised and we have demonstrated this to be a powerful tool to learn about the partitions and the uncertainty of each cluster. In practice, if this tool identifies significant uncertainty, it may trigger a further exploration of the partition space or a more in-depth analysis of the observations at the boundaries between clusters.

Even though we have focused on Bayesian DPMM clustering in this paper, our work can be extended to all Bayesian models which provide a sample of the posterior distribution of the partitions, and this is the focus of our future work.

References

Baudry, J.-P., A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo (2010). Combining mixture components for clustering. Journal of Computational and Graphical Statistics 19(2), 332–353.

Coker, E., S. Liverani, J. K. Ghosh, M. Jerrett, B. Beckerman, A. Li, B. Ritz, and J. Molitor (2016). Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. Environment International 91, 1–13.

Coker, E., S. Liverani, J. G. Su, and J. Molitor (2018). Multi-pollutant modeling through examination of susceptible subpopulations using profile regression. Current Environmental Health Reports 5(1), 59–69.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures . Journal of the American Statistical Association 90(430), 577– 588.

Fraser, D. and M. S. Haq (1969). Structural probability and prediction for the multivariate model. Journal of the Royal Statistical Society: Series B (Methodological) 31(2), 317–331.

Fritsch, A., K. Ickstadt, et al. (2009). Improved criteria for clustering based on the posterior similarity matrix. Bayesian analysis 4(2), 367–391.

Hastie, D. I., S. Liverani, L. Azizi, S. Richardson, and I. Stücker (2013). A semi-parametric approach to estimate risk functions associated with multidimensional exposure profiles: application to smoking and lung cancer. BMC Medical Research Methodology 13, 129.

Hastie, D. I., S. Liverani, and S. Richardson (2015). Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. Statistics and Computing 25(5), 1023–1037.

Hennig, C. (2010). Methods for merging gaussian mixture components. Advances in data analysis and classification 4(1), 3–34.

Jing, W., M. Papathomas, and S. Liverani (2022). Variance matrix priors for Dirichlet process mixture models with Gaussian kernels. arXiv preprint arXiv:2202.03946.

Lavigne, A., A. Freni-Sterrantino, D. Fecht, S. Liverani, M. Blangiardo, K. De Hoogh, J. Molitor, and A. L. Hansell (2020). A spatial joint analysis of metal constituents of ambient particulate matter and mortality in england. Environmental Epidemiology 4(4), 0.

Liu, X., S. Liverani, K. J. Smith, and K. Yu (2020). Modeling tails for collinear data with outliers in the english longitudinal study of ageing: Quantile profile regression. Biometrical Journal 62(4), 916–931.

Liverani, S., D. I. Hastie, L. Azizi, M. Papathomas, and S. Richardson (2015). PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. Journal of Statistical Software 64(7), 1–30.

Liverani, S., A. Lavigne, and M. Blangiardo (2016). Modelling collinear and spatially correlated data. Spatial and Spatio-temporal Epidemiology 18, 63–73.

Liverani, S., L. Leigh, I. L. Hudson, and J. E. Byles (2021). Clustering method for censored and collinear survival data. Computational Statistics 36(1), 35–60.

Mattei, F., S. Liverani, F. Guida, M. Matrat, S. Cenée, L. Azizi, G. Menvielle, M. Sanchez, C. Pilorget, B. Lapôtre-Ledoux, et al. (2016). Multidimensional analysis of the effect of occupational exposure to organic solvents on lung cancer risk: the icare study. Occupational and environmental medicine 73(6), 368–377.

Molitor, J., I. J. Brown, Q. Chan, M. Papathomas, S. Liverani, N. Molitor, S. Richardson, L. Van Horn, M. L. Daviglus, A. Dyer, J. Stamler, P. Elliott, and I. R. Group (2014). Blood pressure differences associated With optimal macronutrient intake trial for heart health (OMNIHEART)–like diet compared with a typical American diet. Hypertension 64(6), 1198–1204.

Pirani, M., N. Best, M. Blangiardo, S. Liverani, R. W. Atkinson, and G. W. Fuller (2015). Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. Environment International 79, 56–64.

Ricciardi, F., S. Liverani, and G. Baio (2022). Dirichlet process mixture models for regression discontinuity designs. Statistical Methods for Medical Research 0(0), 0.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. Journal of the American Statistical Association 85(411), 617–624.

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

Wade, S., Z. Ghahramani, et al. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). Bayesian Analysis 13(2), 559–626.

Appendix A. The posterior predictive distribution as a finite mixture of distributions

Escobar and West (1995) discuss the posterior predictive distribution for Dirichlet process mixture models. They show that given a partition $\boldsymbol{Z}$, the predictive distribution of a new observation D_{n+1} is given by

$$\begin{aligned} P(D_{n+1}|\boldsymbol{Z},\mathbf{D}_n) &= \frac{\alpha}{\alpha+n}\int f(D_{n+1}|\tilde{\Theta}_{n+1})P_{\Theta_0}(\tilde{\Theta}_{n+1})d\tilde{\Theta}_{n+1} \\ &\quad + \sum_{c:n_c>0}\frac{n_c}{\alpha+n}\int f(D_{n+1}|\Theta_c)p(\Theta_c|\{D_i:Z_i=c\})d\Theta_c \end{aligned} \tag{A.1}$$

where n_c is the number of individuals belonging to cluster c. Note that the predictive distribution is a mixture of marginal densities. In the first term of the sum above, the parameter $\tilde{\Theta}_{n+1}$ is sampled from the baseline distribution $P_{\Theta_0}(\cdot)$, whereas in the remaining terms of the sum, parameters Θ_c are sampled from their posterior distribution given prior $P_{\Theta_0}(\cdot)$ and data $\{D_i : Z_i = c\}$. This implies that a new observation D_{n+1} could be assigned to one of the clusters defined by observations $\mathbf{D}_n = (D_1, \ldots, D_n)$ or to a new cluster. For simplicity, in the following, we denote $f_0(\cdot)$ the first component of the mixture in Equation (A.1) and we note that it does not depend on the partition $\boldsymbol{Z}$.

We obtain predictive distribution by integrating Equation (A.1) over the space of partitions, so that

$$P(D_{n+1}|\mathbf{D}_n) = \int_{\mathcal{Z}} P(D_{n+1}|\boldsymbol{Z},\mathbf{D}_n)p(\boldsymbol{Z}|\mathbf{D}_n)d\boldsymbol{Z} \tag{A.2}$$

$$\begin{aligned} &= \frac{\alpha}{\alpha+n}f_0(D_{n+1}) \\ &\quad + \frac{1}{\alpha+n}\sum_{\boldsymbol{Z}\in\mathcal{Z}}\sum_{c:n_c(\boldsymbol{Z})>0} n_c f(D_{n+1}|\{D_i:Z_i=c\})p(\boldsymbol{Z}|\mathbf{D}_n) \end{aligned} \tag{A.3}$$

where $\mathcal{Z}$ denotes the space of partitions of $\mathbf{D}_n$, and $f(D_{n+1}|\boldsymbol{Z},\{D_i : Z_i = c\})$ is the predictive distribution for cluster c of partition $\boldsymbol{Z}$,

$$f(D_{n+1}|\boldsymbol{Z},\{D_i:Z_i=c\}) = \int f(D_{n+1}|\Theta_c)p(\Theta_c|\{D_i:Z_i=c\})d\Theta_c. \tag{A.4}$$

Equation (A.4) is in closed form if the baseline distribution $P_{\Theta_0}(.)$ is a conjugate prior for the likelihood $f(.|\Theta_c)$. Therefore, the predictive distribution in Eq. (A.3) can be interpreted as a mixture of parametric densities, weighted by the marginal posterior partition probability $p(\boldsymbol{Z}|\mathbf{D}_n)$.

Then we introduce the 2^n subsets of $\boldsymbol{D}_n$, denoted $\boldsymbol{S}_j$ with $j = 1, 2, \ldots, 2^n$. We recall that any partition $\boldsymbol{Z}$ is a set of these subsets $\boldsymbol{S}_j$, $\boldsymbol{Z} = \{\boldsymbol{S}_{j_1}, \cdots, \boldsymbol{S}_{j_{k_Z}}\}$ such that

$$\boldsymbol{D}_n = \boldsymbol{S}_{j_1} \cup \cdots \cup \boldsymbol{S}_{j_{k_Z}}$$

$$|\boldsymbol{S}_{j_l}| > 0 \quad \text{for all} \quad l = 1, 2, \ldots, k_Z \quad \text{and}$$

$$\boldsymbol{S}_{j_c} \cap \boldsymbol{S}_{j'_c} = \emptyset$$

for any pair of subsets $(\boldsymbol{S}_{j_c}, \boldsymbol{S}_{j'_c})$ for $j_c = 1, 2, \ldots, k_Z$ and $j_{c'} = 1, 2, \ldots, k_Z$. Without loss of generality we set $\boldsymbol{S}_1 = \emptyset$ and $f_0(D_{n+1}) = f(D_{n+1}|\boldsymbol{S}_1)$. We then rewrite Equation (A.3) using the subsets $\boldsymbol{S}_j$ and then

switch the summations between partitions and subsets.

$$\begin{aligned} P(D_{n+1}|\mathbf{D}_n) &= \frac{\alpha}{\alpha+n} f_0(D_{n+1}) \\ &\quad + \frac{1}{\alpha+n} \sum_{\boldsymbol{Z}\in\mathcal{Z}} \sum_{c=1}^{k_Z} n_{j_c} f(D_{n+1}|\boldsymbol{S}_{j_c}) p(\boldsymbol{Z}|\mathbf{D}_n) \qquad \text{(A.5)} \\ &= \frac{\alpha}{\alpha+n} f(D_{n+1}|\boldsymbol{S}_1) \\ &\quad + \frac{1}{\alpha+n} \sum_{j=2}^{2^n} \sum_{\{\boldsymbol{Z}\in\mathcal{Z}:\boldsymbol{S}_j\in\boldsymbol{Z}\}} p(\boldsymbol{Z}|\mathbf{D}_n) n_j f(D_{n+1}|\boldsymbol{S}_j) \qquad \text{(A.6)} \end{aligned}$$

For a non-empty subset j, we set $\omega_j = \sum_{\{\boldsymbol{Z}\in\mathcal{Z}:\boldsymbol{S}_j\in\boldsymbol{Z}\}} p(\boldsymbol{Z}|\mathbf{D}_n)$, the posterior probability that subset j is sampled over the space of partitions. For the empty set, we set $n_1\omega_1 = \alpha$. We notice that $\sum_{j=1}^{2^n} n_j\omega_j = \alpha+n$ because $\sum_{\boldsymbol{Z}\in\mathcal{Z}} p(\boldsymbol{Z}|\mathbf{D}_n) = 1$. We have

$$P(D_{n+1}|\mathbf{D}_n) = \frac{1}{\alpha+n} \sum_{j=1}^{2^n} \omega_j n_j f(D_{n+1}|\boldsymbol{S}_j). \qquad \text{(A.7)}$$

With this formulation, we have shown that the predictive distribution is a mixture of posterior parametric distributions given each subset. The predictive distribution belongs then to the space generated by the functions $f(\cdot|\boldsymbol{S}_j)$.

Appendix B. The predictive distribution as a finite mixture

Our aim is to formulate the predictive posterior distribution as a finite mixture model linked to partition $\boldsymbol{Z}^*$, which is the consensus partition identified by postprocessing. The partition $\boldsymbol{Z}^*$ has k clusters, $S_1^*, S_2^*, \cdots, S_k^*$. Therefore, our ideal finite mixture model will be a sum of k components with weights propotional to n_l, the number of observations in cluster l. Each component of the finite mixture will be represented by $\tilde{f}_l(\cdot)$ for cluster l, as follows,

$$P(D_{n+1}|\mathbf{D}_n) = \sum_{l=1}^{k} \frac{n_l}{n} \tilde{f}_l(D_{n+1}). \qquad \text{(B.1)}$$

Because of the form of predictive distribution in Equation (A.7), the components of the finite mixture should be of the form

$$\tilde{f}_l(D_{n+1}) = \sum_{j=1}^{2^n} \frac{n_j\omega_j}{\alpha+n} \beta_j^l f(D_{n+1}|\boldsymbol{S}_j) \qquad \text{(B.2)}$$

where β_j^l represents the part of the marginal distribution $f(D_{n+1}|\boldsymbol{S}_j)$ relating to component l in the finite mixture in Equation (B.1). Of course, $\tilde{f}_l(D_{n+1})$ should be a density distribution, constraining the sum $\sum_{j=1}^{2^n} n_j\omega_j\beta_j^l$ to be equal to $\alpha+n$. Also, the mixture in Equation (B.1) should be equal to Equation (A.7), meaning that for each subset j, $\sum_{l=1}^{k} \frac{n_l}{n}\beta_j^l = 1$. Therefore, our problem can be reduced to finding the β_j^l which satisfy these two sequences of constraints,

$$\begin{aligned} \sum_{j=1}^{2^n} n_j\omega_j\beta_j^l &= \alpha+n \quad \text{for all } l \in \{1,2,\cdots,k\} \qquad \text{(B.3)} \\ \sum_{l=1}^{k} \frac{n_l}{n}\beta_j^l &= 1 \quad \text{for all } j \in \{1,2,\cdots,2^n\} \qquad \text{(B.4)} \end{aligned}$$

and such that each component $\tilde{f}_l(D_{n+1})$ is as close as possible to the predictive distribution of data in cluster $\boldsymbol{S}_l^*$.

We propose a solution based on the simple empirical principle that the more $\boldsymbol{S}_j$ has data in common with $\boldsymbol{S}_l^*$, the more $f(.|\boldsymbol{S}_j)$ is close to the predictive distribution of $\boldsymbol{S}_l^*$. This principle leads us to propose this simple rule of allocation based on proportionality. The part of $f(.|\boldsymbol{S}_j)$ allocated to cluster l will be proportional to n_{jl}, the number of data both in $\boldsymbol{S}_j$ and $\boldsymbol{S}_l^*$. From these rules, a solution that satisfies all constraints in Equations (B.3) and (B.4) follows.

$$\begin{array}{l} \beta_j^l = \frac{n n_{lj}}{n_j n_l} \quad \text{if } j \leq 2 \\ \beta_j^l = 1 \quad \text{if } j = 1. \end{array} \tag{B.5}$$

For the constraints in Equation (B.3) we have

$$\begin{aligned} \sum_{j=1}^{2^n} n_j \omega_j \beta_j^l &= n_1 \omega_1 \beta_1^l + \sum_{j=2}^{2^n} n_j \omega_j \frac{n_{lj} n}{n_l n_j} = \alpha + \frac{n}{n_l} \sum_{j=2}^{2^n} n_{lj} \omega_j \\ &= \alpha + \frac{n}{n_l} \sum_{j=2}^{2^n} \sum_{\boldsymbol{Z} \in \{Z \in \mathcal{Z} : \boldsymbol{S}_j \in Z\}} p(\boldsymbol{Z}|\mathbf{D}_n) n_{lj} \\ &= \alpha + \frac{n}{n_l} \sum_{\boldsymbol{Z} \in \mathcal{Z}} p(\boldsymbol{Z}|\mathbf{D}_n) \sum_{j=2}^{2^n} n_{lj} \mathbf{1}_{S_j \in Z} = \alpha + \frac{n}{n_l} n_l = \alpha + n. \end{aligned}$$

For the constraints in Equation (B.4) we have, if $j \neq 1$,

$$\sum_{l=1}^{k} \frac{n_l}{n} \beta_j^l = \sum_{l=1}^{k} \frac{n_l}{n} \frac{n}{n_l} \frac{n_{lj}}{n_j} = \sum_{l=1}^{k} \frac{n_{lj}}{n_j} = 1 \tag{B.6}$$

and, if $j = 1$,

$$\sum_{l=1}^{k} \frac{n_l}{n} \beta_j^l = \sum_{l=1}^{k} \frac{n_l}{n} = 1.$$

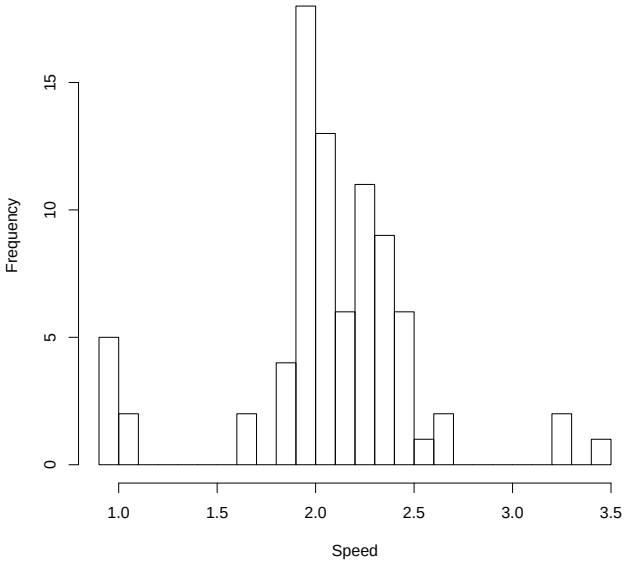
Appendix C. Application to mixtures of Gaussian distributions

Perhaps the most common model to be implemented under the DPMM framework is the Gaussian mixture model, where $\boldsymbol{D}_n = \boldsymbol{Y}_n$ for some continuous multidimensional data Y_i of dimension p, and Y_i follows a mixture of Gaussian distributions. Under this setting for each cluster c, the cluster specific parameters are given by $\Theta_c = (\mu_c, \Sigma_c)$, where $\mu_c \in \mathbb{R}^p$ is a mean vector and $\Sigma_c \in \mathbb{R}^{p \times p}$ is a covariance matrix. Under this setting

$$p(Y_i|Z_i, \Theta_{Z_i}) = f(Y_i|\mu_{Z_i}, \Sigma_{Z_i}) = (2\pi)^{-\frac{p}{2}} |\Sigma_{Z_i}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (Y_i - \mu_{Z_i})^\top \Sigma_{Z_i}^{-1} (Y_i - \mu_{Z_i}) \right\}. \tag{C.1}$$

The normal inverse Wishart prior is a convenient prior choice for (μ_c, Σ_c) due to its conjugacy with the multivariate Gaussian distribution, facilitating Gibbs updates. It is parametrised with μ_0, ν_0, κ_0, R_0 ($NIW(\mu_0, \nu_0, \kappa_0, R_0)$), and is such that $\mu_c \sim \text{Normal}(\mu_0, (1/\nu_0)\Sigma_c)$ and $\Sigma_c \sim \text{InvWishart}(R_0, \kappa_0)$, for each c.

If the Normal inverse Wishart prior for parameters μ_0, ν_0, κ_0, R_0 ($NIW(\mu_0, \nu_0, \kappa_0, R_0)$) is chosen, then $f_0(\cdot)$ is the density of a multivariate Student distribution with mean μ_0, covariance matrix $\Psi = \frac{1+\nu_0}{(\kappa_0 - p + 1)\nu_0} R_0$ and degree of freedom $\kappa_0 - p + 1$, denoted $\mathcal{T}_{\kappa_0 - p + 1}(\mu_0, \Psi)$ with $Y_i \in \mathbb{R}^p$ (Fraser and Haq, 1969). We use this



h

Figure D.2: Histogram of galaxy velocity data.

widely used model for illustrative purposes but see Jing et al. (2022) for a warning on the use of these prior distributions for highly dimensional data.

Thus we can write $\tilde{f}_l(D_{n+1})$ as

$$\begin{aligned}\tilde{f}_l(D_{n+1}) &= \frac{\alpha}{\alpha+n}\mathcal{T}_{\kappa_0-p+1}(\mu_0,\Psi) \\ &\quad + \frac{1}{\alpha+n}\sum_{\boldsymbol{Z}\in\mathcal{Z}}\sum_{c:n_c(\boldsymbol{Z})>0} n_{cl}(\boldsymbol{Z}) f(D_{n+1}|\{D_i : Z_i = c\}) p(\boldsymbol{Z}|\mathbf{D}_n).\end{aligned}$$

In the same way, $f(D_{n+1}|\boldsymbol{Z}, \{D_i : Z_i = c\}) \propto \int f(D_{n+1}|\Theta_c) p(\Theta_c|\boldsymbol{Z}, \{D_i : Z_i = c\}) d\Theta_c$ is the density of a p multivariate Student distribution $\mathcal{T}_{\kappa_0+n_c-p+1}(\mu_c, \Sigma_c)$, with

$$\mu_c = \frac{\nu_0}{\nu_0+n_c}\mu_0 + \frac{n_c}{\nu_0+n_c}\bar{y}_l \tag{C.2}$$

$$\Sigma_c = \frac{\nu_0+n_c+1}{(\nu_0+n_c)(\kappa_0+n_c-p+1)}(R_0 + S_c + \frac{\nu_0 n_c}{\nu_0+n_c}(\mu_0-\bar{D}_l)(\mu_0-\bar{D}_l)') \tag{C.3}$$

$$S_c = \sum_{i:Z_i=c}(D_i-\bar{D}_c)(D_i-\bar{D}_c)'. \tag{C.4}$$

Appendix D. Application to velocity galaxy data

We consider the velocity galaxy dataset proposed by Roeder (1990) and largely used in the literature for comparing different clustering methods. It contains the velocities of $n = 82$ galaxies from a redshift survey in the Corona Borealis region (Fig. D.2). The distribution of the data appears clearly multimodal.

Escobar and West (1995) apply a DPMM to these data for density estimation. We use the same model to illustrate our method. Observation D_i for $i = 1, \ldots, n$ are modelled with the DPMM. Conditionally on

the partition, observations are supposed Normally distributed,

$$D_i|Z_i, \mathbf{\Theta} \sim \mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i})$$

with $\Theta_c = (\mu_c, \Sigma_c)$. We chose of the conjugated normal inverse Gamma distribution ($NIW(\mu_0, \nu_0, \alpha, \beta)$) as baseline distribution. We use the hyperparameters as proposed by Escobar and West (1995), i.e. $\mu_0 = 0$, $\nu_0 = 0,001$, $\alpha = 2$ and $\beta = 1$. For the illustration purposes, we fix $\alpha = 4$. The model is fit with the PReMiuM package (see below), with one run of 300,000 iterations of a Gibbs algorithm. The 'consensus' partition $\boldsymbol{Z}^*$ is obtained applying PAM on the dissimilarity matrix.

The plot on the left hand side of Fig. 1 shows that each component $\tilde{f}_l$ is not Gaussian. Indeed, each component is a mixture of the 2^n marginal densities. This point suggests that mixture components derived from a clustering using DPMM should not be thought of as the parametric component $f(.|\Theta_c)$ of the model. Only if each data of cluster j has a posterior probability 1 to be sampled from this cluster, each component corresponds to a cluster. The finite mixture model components $\tilde{f}_l$ are themselves a mixture of densities. This phenomenon has been discussed by Baudry et al. (2010) and Hennig (2010). Clusters are supposed to represent subpopulations, and the associated distribution is not necessarily Gaussian. As a consequence, several mixture components can account for a single cluster. With this decomposition of the predictive distribution, we show that using DPMM for clustering, implicitly results in multi-components clusters.

All the functions discussed in this paper are available in the R package PReMiuM, a package developed by Liverani et al. (2015) for profile regression, a method that has a wide range of applications, such as spatial modelling (Liverani et al., 2016; Lavigne et al., 2020) and epidemiology (Hastie et al., 2013; Molitor et al., 2014; Pirani et al., 2015; Mattei et al., 2016; Coker et al., 2016, 2018; Liu et al., 2020; Liverani et al., 2021; Ricciardi et al., 2022). The functions used to create the figures and plots in this paper are open source and available on Github at `https://github.com/silvialiverani/partitionuncertainty`