UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"

Bachelor of Science in Economics, Management and Computer Science

Bayesian Learning: Clustering with the Dirichlet Process Mixture Model

Supervisor:

Prof. **ANTONIO LIJOI**

Candidate:

LUIGI NOTO

3075187

Academic Year 2020 - 2021

Contents

Introduction

The purpose of this final paper is to explore Bayesian nonparametrics and its application to clustering, by means of the Dirichlet process mixture model (DPMM).
Bayesian nonparametric models are a class of models with a potentially infinite number of parameters. This property allows a higher modeling flexibility than with parametric models, making them particularly attractive for numerous applications where the underlying structure of the data is not known or is assumed to grow as more data are observed. In fact, because of their infinite-dimensional parameterization, these models incorporate the uncertainty about the appropriate model complexity to adopt for the observed data, thereby automatically addressing the problem of model selection when performing posterior inference. Hence, the analysis does not rely on heuristic methods as it is the case when using models with a finite set of parameters. Thanks to the recent developments in inferential procedures, in particular Markov chain Monte Carlo methods, these models have gained widespread popularity and have been increasingly applied to many machine learning problems in a wide range of fields, including business, medicine and genetics. Clustering is one of the problems where the benefits of the Bayesian nonparametric approach are immediately evident, because there is no more need to specify the number of clusters in advance of analyzing the data, being it estimated with posterior inference.
In this paper, we will discuss the theory behind the Dirichlet process mixture model and its application to clustering, starting from the background knowledge necessary for its understanding. Thus, before presenting the main results concerning the DPMM and its clustering property, we will discuss the general Bayesian nonparametric framework and its theoretical foundation, we will analyze in more detail the clustering problem and the downsides of addressing it with finite-dimensional models, and we will review the Dirichlet process and the general MCMC scheme. In the final part, after presenting a procedure for performing posterior inference under the DPMM, we will analyze the sensitivity of the model to hyperparameters in an application to wholesale customer data.

Chapter 1

Background

1.1 Bayesian Nonparametrics

A *statistical problem* arises when we observe data, i.e. a collection of random variables (or random vectors) $x_1, x_2, ..., x_n$, that are usually assumed to be drawn independently from some probability distribution F and there is uncertainty about F. In order to make inference about the *generative process* of the data, we have to assume that the underlying probability distribution F is a member of a family of probability distributions

$$\mathcal{F} = \{F_\theta \ : \ \theta \in \Theta\}$$

indexed by a parameter θ from a set Θ. Such a family of probability distributions is called *statistical model*.

In the Bayesian paradigm, the statistical model is completed with the specification of a probability distribution, known as *prior*, on the parameter space Θ, i.e. the parameter is assumed to be a (latent) random variable, thereby specifying the joint probability distribution of the hidden and observed random variables. The objective is to fit the model to the observed data, i.e. use the observed data in order to get more information about the latent random parameter, computing the conditional distribution of the latent random parameter given the observed data, known as *posterior*. This process is called *posterior inference*.

Statistical models indexed by a finite-dimensional parameter θ are known as *finite dimensional* or *parametric models*. A persistent issue about dealing with parametric models is that in many situations it is not known *a priori* what the "right" parametric form is and how complex the model should be in order to sufficiently capture the important structure of the data without capturing sample-specific unnecessary details. In other words, in many statistical problems there is high risk of underfitting, overfitting and model misspecification. The problems of determining appropriate model families are referred to as *model selection* or *model adaptation*. Examples of such problems are selecting the number

of clusters in clustering problems and selecting the number of hidden states in Hidden Markov Models. These problems are commonly addressed by first fitting several models with different parametric forms and then choosing the best according to model comparison metrics, including a goodness-of-fit component and a penalty factor that penalizes more complex models, i.e. models with a higher number of parametric quantities (Claeskens and Hjort, 2008). However, such comparisons often rely on heuristics, therefore they are not very robust and when there is no clue about the underlying complexity of the data there might still be high risk of mis-specification of statistical model.
In such cases, it might be convenient to follow another approach to this problem, which consists in relaxing the parametric assumptions and defining probability models indexed by an infinite-dimensional parameter, completed with a prior on the parameter, allowing for greater modeling flexibility and robustness to model mis-specification. Such models, known as *Bayesian nonparametric models* (Ghosal and van der Vaart, 2017), determine their complexity appropriately from the data, thereby sidestepping the explicit approximate model selection process, and automatically allow growth in complexity as more data are observed.

1.2 Clustering and the issue of the number of clusters

Unsupervised learning consists in determining the latent structure that generated the observed data. In many applications, such as in marketing (Ližbetinová et al., 2019) and in security (Hodge and Austin, 2004), we might want to partition the observed data into more homogeneous subgroups. Such a process is defined as *clustering*. Formally, given a set of experimental units $[n] = \{1, ..., n\}$ with respective observed data $x = (x_1, ..., x_n)$, we want to find a partition (*cluster arrangement*) $C_n = \{S_1, ..., S_K\}$ of $[n]$ $(K \leq n)$, i.e. subsets (*clusters*) S_i of $[n]$ for any $i \in \{1, ..., K\}$ such that $S_i \cap S_{i'} = \emptyset$ for any $i, i' \in \{1, ..., K\}$ s.t. $i \neq i'$ and $\bigcup_{i=1}^{K} S_i = [n]$, in such a way that there is high similarity between any two elements in S_i for any $i \in \{1, ..., K\}$, and low similarity between any element in S_i and any element in $S_{i'}$ for any $i, i' \in \{1, ..., K\}$ s.t. $i \neq i'$, according to some criterion. Equivalently, we can describe partitions by cluster membership indicators,

by defining $c_i = j$ if the experimental unit $i \in S_j$, for any $i \in [n]$, with the convention that clusters are labeled by appearance of the observations, meaning that $c_1 = 1$, $c_{i_2} = 2$ for the lowest $i_2 > 1$ such that $i_2 \notin S_1$, and so on. There is a one-to-one relation between C_n and $(c_1, ..., c_n)$.

The statistical modeling approach to this problem from a high-level perspective consists in defining a prior on the set of partitions $p(C_n)$, thereby defining a random partition of the experimental units, and a sampling density $p(x \mid C_n)$, and then making inference by finding the posterior distribution of the random partition given the observed data $p(C_n \mid x)$ (Müller et al., 2015). The most popular such approach is *finite model-based clustering* or *finite mixture modeling* (Fraley and Raftery, 2002). In finite mixture modeling, we assume that there is a fixed finite number of clusters H and each cluster h is associated with a parameter θ_h for any $h \in \{1, ..., H\}$. Each observation x_i is assumed to be generated by first generating the cluster membership indicator $c_i \in \{1, ..., H\}$ and then generating the observation from the observation distribution $F(\cdot \mid \theta_{c_i})$. By introducing the latent categorical variables $c_1, ..., c_n$ representing the cluster membership indicators, we implicitly defined a prior on the set of partitions $p(C_n)$. Thus, we have the following hierarchical model

$$\begin{aligned} x_i \mid c_i, (\theta_1, ..., \theta_H) &\overset{\text{ind}}{\sim} f(\cdot \mid \theta_{c_i}) && i = 1, ..., n \\ c_i \mid (w_1, ..., w_H) &\sim \text{Discrete}(w_1, ..., w_H) && i = 1, ..., n \end{aligned}$$

where $\text{Discrete}(w_1, ..., w_H)$ is the multiple-outcome analogue of a Bernoulli random variable, i.e. $\mathbb{P}(c_i = h \mid (w_1, ..., w_H)) = w_h$ for $h = 1, ..., H$, completed by a prior over the mixing proportions $w_1, ..., w_H$ ($w_h \geq 0 \quad \forall h \in \{1, ..., H\}$, $\sum_{h=1}^{H} w_h = 1$), $w = (w_1, ..., w_H) \sim W$, and a prior over the cluster parameters $\theta_1, ..., \theta_H$, $\theta_1, ..., \theta_H \overset{\text{iid}}{\sim} G_0$. It is possible that $K \leq H$ distinct clusters are observed in the data. By defining $S_j = \{i \in [n] : c_i = j\}$ for any $j \in \{1, ..., H\}$, thereby getting the explicit random partition $C_n = \{S_1, ..., S_H\}$, the sampling density is given by

$$p(x \mid C_n) = \int \left\{ \prod_{j=1}^{H} \prod_{i \in S_j} f(x_i \mid \theta_j) \right\} \left\{ \prod_{h=1}^{H} G_0(\theta_h) \right\} d\theta_1 \cdots d\theta_H$$

with the convention that $\prod_{i \in S_j} f(x_i \mid \theta_j) = 1$ if $S_j = \emptyset$.

One limitation of this model is that the implied prior of the random partition $p(C_n)$

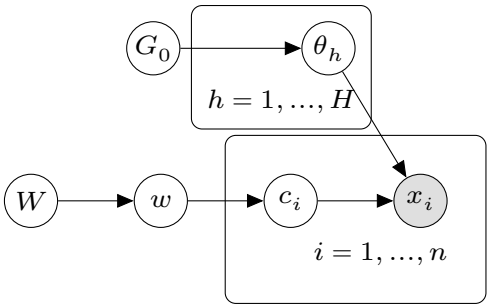

Figure 1.1: Graphical model of finite mixture modeling using plate notation (Buntine, 1994)

has an upper bound on the number of clusters, because it only allows to have up to H clusters. In some situations, however, we may want to allow the number of clusters to grow as more data are observed and therefore it would be more natural to approach the problem with a nonparametric model, in which we assume that the number of clusters in the population is infinite and that only a subset of such clusters is generated in the observed finite sample. Another reason for using nonparametric models is that inference on the number of clusters with these models is much more convenient than working with finite models with unknown number of clusters, by using complex algorithms for inference such as the reversible jump Markov Chain Monte Carlo or heuristic approaches consisting of extensive crossvalidation (Rasmussen, 1999).

The most common nonparametric model for model-based clustering is the *Dirichlet process mixture model (DPMM)*. In the next sections, we introduce some concepts that are necessary for the understanding of the Dirichlet process and the Dirichlet process mixture model.

1.3 Dirichlet distribution

We now introduce the *Dirichlet distribution*, which generalizes the Beta distribution in the multivariate case (Bilodeau and Brenner, 1999). Let $Y_1, ..., Y_k$ be independent random variables with $Y_j \sim \text{Gamma}(\alpha_j, 1)$, $\alpha_j > 0$, i.e. with density

$$f_j(y) = \frac{1}{\Gamma(\alpha_j)} y^{\alpha_j - 1} e^{-y} \mathbb{1}_{\mathbb{R}_+}(y)$$

where Γ indicates the Gamma function[1], for $j = 1, ..., k$. Then, let $W = (W_1, ..., W_k)$ such that

$$W_j = \frac{Y_j}{\sum_{i=1}^k Y_i} \quad \forall j \in \{1, ..., k\}$$

Thus, we have $|W| = 1$, where $|W|$ is the l_1-norm[2] of W, and the vector W has density

$$f_k(w; \alpha) = \frac{\Gamma(|\alpha|)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k w_i^{\alpha_i - 1} \mathbb{1}_{\Delta_{k-1}}(w)$$

where $w = (w_1, ..., w_k)$, $\alpha = (\alpha_1, ..., \alpha_k)$ and Δ_{k-1} is the unit $k-1$ simplex, i.e.

$$\Delta_{k-1} = \left\{ w = (w_1, ..., w_k) : w_i \geq 0 \ \ \forall i \in 1, ..., k \ \text{ and } \ \sum_{i=1}^k w_i = 1 \right\}$$

We say W has a Dirichlet distribution with parameters $\alpha = (\alpha_1, ..., \alpha_k)$ and write $W = (W_1, ..., W_k) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_k)$. The Dirichlet distribution has some important properties.

- *Marginal distributions are Beta distributions*
  Let $W = (W_1, ..., W_k) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_k)$. Then

  $$W_j \sim \text{Beta}(\alpha_j, |\alpha| - \alpha_j) \quad \forall j \in \{1, ..., k\}$$

  where $\alpha = (\alpha_1, ..., \alpha_k)$.

1. The Gamma function (for real numbers) is defined as the function $\Gamma : \mathbb{R}_+ \longrightarrow \mathbb{R}$ such that $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt \ \ \forall z \in \mathbb{R}_+$.
2. For any vector $a \in \mathbb{R}^n$, we define the l_1-norm of a as $\|a\|_1 = |a| = \sum_{i=1}^n a_i$.

- *Aggregation property*

  Let $W = (W_1, ..., W_k) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_k)$. Then, if for any $p \in \{1, ..., k\}$, $r_1, ..., r_p \in \mathbb{N}$ are such that $0 < r_1 < ... < r_p = k$, we have

  $$\left(\sum_{i=1}^{r_1} W_i, \sum_{i=r_1+1}^{r_2} W_i, \ ..., \sum_{i=r_{p-1}+1}^{r_p} W_i \right) \sim \text{Dirichlet} \left(\sum_{i=1}^{r_1} \alpha_i, \sum_{i=r_1+1}^{r_2} \alpha_i, \ ..., \sum_{i=r_{p-1}+1}^{r_p} \alpha_i \right)$$

- *Decimative property*

  The converse of the aggregation property is true as well. For example, if $W = (W_1, ..., W_k) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_k)$ and $V = (V_1, V_2) \sim \text{Dirichlet}(\alpha_1\beta_1, \alpha_1\beta_2)$ with $\beta_1 + \beta_2 = 1$, then we have

  $$(W_1V_1, W_1V_2, W_2, ..., W_k) \sim \text{Dirichlet}(\alpha_1\beta_1, \alpha_1\beta_2, \alpha_2, ..., \alpha_k)$$

The Dirichlet distribution is at the base of the definition of the Dirichlet process provided by Ferguson using Kolmogorov's consistency theorem. Because of this definition, that we will see later, the Dirichlet process is regarded as the infinite-dimensional analogue of the Dirichlet distribution.

1.4 Random probability measures, exchangeability and de Finetti's representation theorem

The infinite-dimensional parameter in the Dirichlet process mixture model is a *random probability measure* whose law acts as a Bayesian nonparametric prior (BNP). In the specific case, such a random probability measure is the Dirichlet process. We now formally define this object.

Definition (Random probability measure)**.** *Let $\mathcal{P}_{\mathbb{R}}$ denote the space of probability measures on $\mathbb{R}$ and $\mathcal{B}(\mathcal{P}_{\mathbb{R}})$ denote the corresponding Borel σ-algebra. Then, any random element P from some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and with values in $(\mathcal{P}_{\mathbb{R}}, \mathcal{B}(\mathcal{P}_{\mathbb{R}}))$ is a random probability measure.*

The random probability measure P can also be described as a stochastic process $\{P(E) : E \in \mathcal{B}(\mathbb{R})\}$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$.

Besides the definition of random probability measure, prior to defining any BNP it is necessary to introduce the notion of *exchangeability*, which, thanks to *de Finetti's representation theorem* (Hewitt and Savage, 1955), constitutes a motivation of Bayesian statistical theory and in particular of the search for Bayesian nonparametric priors.

Definition (Exchangeability)**.** *A sequence of (real-valued) random elements* $(X_n)_{n\geq 1}$ *is (infinitely) exchangeable if for any* $n \geq 1$ *and permutation* σ *of* $\{1, ..., n\}$ *we have*

$$(X_1, ..., X_n) \stackrel{d}{=} (X_{\sigma(1)}, ..., X_{\sigma(n)})$$

i.e. the finite-dimensional distributions of $(X_n)_{n\geq 1}$ *are invariant with respect to permutations of its elements.*

It is easy to see from the above definition that independence and identity in distribution implies exchangeability, but the reverse implication is not necessarily true. Exchangeability is a plausible assumption (much weaker than iid) in many machine learning and statistical applications. Moreover, exchangeability is also a crucial property in many algorithms for posterior simulation in the Dirichlet process mixture model. We can now state de Finetti's representation theorem.

Theorem (De Finetti's representation theorem)**.** *A sequence of (real-valued) random elements* $(X_n)_{n\geq 1}$ *is (infinitely) exchangeable if and only if there exists a probability measure* Q *on* $(\mathcal{P}_{\mathbb{R}}, \mathcal{B}(\mathcal{P}_{\mathbb{R}}))$ *such that for any* $n \geq 1$

$$\mathbb{P}[(X_1, ..., X_n) \in A] = \int_{\mathcal{P}_{\mathbb{R}}} \prod_{i=1}^{n} p(A_i) dQ(p) \quad \forall A = (A_1 \times \cdots \times A_n) \in \mathcal{B}(\mathbb{R}^n)$$

The central role of this theorem in Bayesian statistics is due to the fact that, given a random probability measure P with some distribution Q, an exchangeable sequence of observations is conditionally independent and identically distributed (iid), i.e. this sequence can be viewed as a mixture of iid sequences with some mixing distribution Q, known as *de Finetti's measure*. Thus, exchangeability implies the existence of a hierarchical Bayesian model with latent random probability measure P. This can be written

as

$$X_i \mid P \overset{\text{iid}}{\sim} P \quad i = 1, ..., n$$
$$P \sim Q$$

for any $n \geq 1$. Moreover, the key thing to notice is that Q may not degenerate on a finite-dimensional subspace of $\mathcal{P}_{\mathbb{R}}$, i.e. the associated parameterization of probability measures on $\mathbb{R}$ may be infinite-dimensional. Thus, de Finetti's representation theorem represents the theoretical foundation of the Bayesian nonparametric approach.

1.5 Dirichlet process

One of the most popular BNP priors where the infinite-dimensional parameter is a random probability measure is the *Dirichlet process (DP)* prior (Müller et al., 2015). It is used for density estimation, semi-parametric modelling, sidestepping model selection and averaging for clustering, topic modeling and other applications. It is defined as follows.

Definition (Dirichlet process)**.** *Let $M > 0$ and G_0 be a probability measure defined on $\mathbb{R}^n$ for some integer $n > 0$. A Dirichlet process (DP) with parameters (M, G_0), denoted as DP(M, G_0) or DP(MG_0), is a random probability measure G defined on $\mathbb{R}^n$ such that for any integer $k \geq 1$ and any finite partition $\{B_1, ..., B_k\}$ of $\mathbb{R}^n$, we have*

$$(G(B_1), ..., G(B_k)) \sim \textit{Dirichlet}(MG_0(B_1), ..., MG_0(B_k))$$

It has been proved that such a process exists using Kolmogorov's consistency theorem (Ferguson, 1973). A Dirichlet process has then two parameters with important properties:

- G_0 is called *centering measure* and is the the "mean" of the process; we have

$$\mathbb{E}[G(B)] = G_0(B) \quad \forall B \subseteq \mathbb{R}^n$$

- M is called *precision* or *total mass parameter* and is like an "inverse-variance" of the process; we have

$$Var(G(B)) = \frac{G_0(B)(1 - G_0(B))}{1 + M} \quad \forall B \subseteq \mathbb{R}^n$$

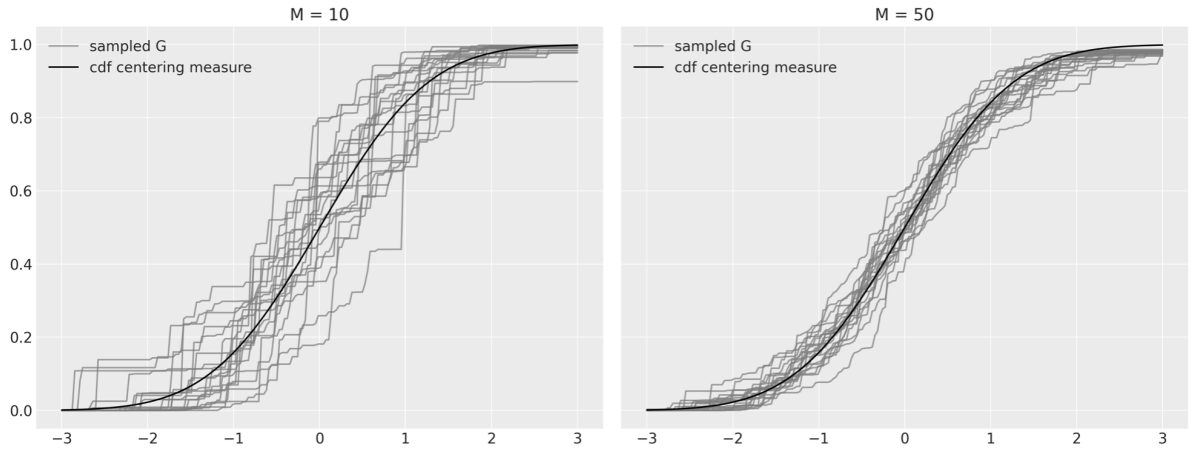

Figure 1.2: Plot of sample cdf's from $\mathrm{DP}(M, G_0)$ with $G_0 = \mathrm{N}(0, 1)$ for $M = 10$ and $M = 50$ (The code for the simulations can be found in the GitHub repository https://github.com/luiginoto/dpmm_clustering)

the larger M, the more G is concentrated about G_0; moreover, as $M \to +\infty$, the process degenerates on G_0.

The product MG_0 is defined as the *base measure* of the Dirichlet process. The crucial property of the Dirichlet process is that the probability law Q of the DP-distributed random probability measure G is concentrated on the space of discrete probability measures on $\mathbb{R}^n$, meaning that G is almost surely discrete. Given its discrete nature, we can then write G as an infinite weighted sum of point masses called *atoms*

$$G(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{m_k}(\cdot)$$

where $w_1, w_2, \ldots$ are the probability weights and $\delta_m(\cdot)$ is the Dirac measure[3] at m, i.e. w_k is the probability assigned to the k-th atom and m_k is the value of that atom. Thus, the proper definition suggests that the Dirichlet process can be considered as the infinite-dimensional analogue of the Dirichlet distribution. However, this definition does not

3. For any integer $n > 0$, the Dirac measure on $\mathbb{R}^n$ is defined as

$$\delta_x(A) = \mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \qquad \forall x \in \mathbb{R}^n \;\; \forall A \subseteq \mathbb{R}^n$$

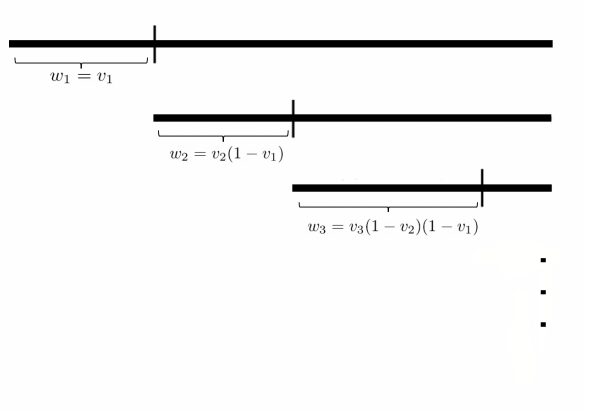

Figure 1.3: Stick-breaking representation of the Dirichlet process

provide a constructive representation of the process. Based on the discrete nature of $G \sim \text{DP}(M, G_0)$, the so-called *stick-breaking representation* provides a useful constructive definition of the process (Sethuraman, 1994). The construction works as follows. The locations $m_1, m_2, ...$ are iid draws from G_0. Then, we consider a stick with unit length and divide it into an infinite number of segments $w_1, w_2, ...$ with the following process.

- Simulate a Beta random variable $v_1 \sim \text{Beta}(1, M)$ and break off a fraction v_1 of the stick, i.e. $w_1 = v_1$.
- For each step $k = 2, 3, ...$, simulate another Beta random variable $v_k \sim \text{Beta}(1, M)$ and break off a fraction v_k of the remainder of the stick, i.e. $w_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$.

Assuming $v_k \overset{\text{iid}}{\sim} \text{Beta}(1, M)$ and $m_k \overset{\text{iid}}{\sim} G_0$ for $k = 1, 2, ...$, where $\{v_k\}_{k\geq 1}$ and $\{m_k\}_{k\geq 1}$ are independent, the resulting random probability measure

$$G(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{m_k}(\cdot)$$

has distribution $\text{DP}(M, G_0)$. The distribution of the infinite-dimensional random vector $(w_1, w_2, ...)$ is the so-called *Griffiths-Engen-McCloskey (GEM) distribution* with parameter M, written $(w_1, w_2, ...) \sim \text{GEM}(M)$.

Another interesting property of the Dirichlet process is that it is conjugate with respect

to iid sampling, i.e. given the Bayesian model

$$x_i \mid G \overset{\text{iid}}{\sim} G \quad i = 1, ..., n$$
$$G \sim \text{DP}(M, G_0)$$

the posterior distribution of G is a Dirichlet process. In particular, we have

$$G \mid x_1, ..., x_n \sim \text{DP}\left(M + n, \frac{M}{M+n} G_0 + \frac{n}{M+n} \hat{G}_n\right)$$

where

$$\hat{G}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

The posterior DP centering measure is a weighted average of the prior DP centering measure G_0 and the empirical distribution $\hat{G}_n$ and the posterior precision parameter increases to $M + n$. Using the above result about the posterior distribution of G, we can easily find the marginal distribution of the data

$$p(x_1, ..., x_n) = \int \prod_{i=1}^{n} G(x_i) dQ(G)$$

where Q is the probability law corresponding to $\text{DP}(M, G_0)$, by writing the above distribution as

$$p(x_1) \prod_{i=2}^{n} p(x_i \mid x_1, ..., x_{i-1})$$

and noting that

$$p(x_1) = \mathbb{E}[p(x_1 \mid G)] = \mathbb{E}[G(x_1)] = G_0(x_1)$$

and

$$\begin{aligned} p(x_n \mid x_1, ..., x_{n-1}) &= \mathbb{E}[p(x_n \mid x_1, ..., x_{n-1}, G) \mid x_1, ..., x_{n-1}] \\ &= \mathbb{E}[G(x_n) \mid x_1, ..., x_{n-1}] \\ &= \frac{M}{M+n-1} G_0(x_n) + \frac{1}{M+n-1} \sum_{k=1}^{n-1} \delta_{x_k}(x_n) \end{aligned}$$

Since the observations $x_1, ..., x_n$ are iid given G, by de Finetti's representation theorem the random vector $(x_1, ..., x_n)$ is exchangeable, i.e. the probabilities are the same for

any permutation of the indices $\{1, ..., n\}$. From this representation we devise a sampling scheme for generating $x_1, ..., x_n$ from an exchangeable sequence whose de Finetti measure is a Dirichlet process with parameters (M, G_0). The procedure is the following

- Generate $v_1, ..., v_n \overset{iid}{\sim} G_0$;
- Set $x_1 = v_1$;
- For $i = 2, ..., n$, set

$$x_i = \begin{cases} v_i & \text{with probability } \frac{M}{M+i-1} \\ x_1 & \text{with probability } \frac{1}{M+i-1} \\ \vdots & \vdots \quad \vdots \\ x_{i-1} & \text{with probability } \frac{1}{M+i-1} \end{cases}$$

This sampling scheme is called *Blackwell-MacQueen Pólya urn scheme* (Blackwell and MacQueen, 1973). It is based on the fact that there is a positive probability of ties among $x_1, ..., x_n$ because of the almost sure discreteness of $G \sim \text{DP}(M, G_0)$. The Pólya urn metaphor is the following: we initially place a ball with a random color in an empty urn; then, in each step, we have two options, either randomly pick a ball from the urn with replacement and placing one more ball of the same color, with probability proportional to the number of balls already in the urn, or place a new ball with a new (random) color in the urn, with probability proportional to M.
The Blackwell-MacQueen Pólya urn representation will play a central role later to find the distribution on clusters of the observed data implied by the DP prior in the Dirichlet process mixture model and to determine the procedure for posterior inference.

1.6 Markov chain Monte Carlo

The greater modeling flexibility of Bayesian nonparametric models obviously comes at a cost in terms of computational complexity and analytical tractability. A common question that might arise is how to perform inference (without truncating the model representation)

in a model with an infinite number of parameters. The answer is that in many such models, and in particular in the Dirichlet Process Mixture Model, it is possible to integrate out all but a finite subset of the parameters, so that inference can be performed with techniques used for Bayesian parametric models, such as *Markov chain Monte Carlo (MCMC)* methods. In this section, we will briefly discuss Markov chain Monte Carlo (MCMC), the Metropolis-Hastings algorithm and Gibbs sampling (Ross, 2014).

We will analyze the discrete case. Some care is needed when moving to the continuous case, but the intuition from the discrete case is useful. It often turns out that we want to sample from a random element X with probability mass function π on some sample space $\mathcal{X}$ that is difficult to simulate. Moreover, π is often known up to a multiplicative constant, i.e. it is given in the form

$$\pi(x) = Cb(x) \quad \forall x \in \mathcal{X}$$

where b is a known function and the normalizing constant

$$C = \frac{1}{\sum_{x \in \mathcal{X}} b(x)}$$

cannot be computed. This is typical when we want to sample from the posterior distribution of some parameter (or vector of parameters) θ given the observed data $x = (x_1, ..., x_n)$ in a Bayesian model. In such cases, it is possible to overcome the problem of simulating $X \sim \pi$ by using the theory of Markov chains.

Let us recall that a probability mass function π on a sample space $\mathcal{X}$ ($\pi(x) \geq 0 \ \forall x \in \mathcal{X}$ and $\sum_{x \in \mathcal{X}} \pi(x) = 1$) is a stationary distribution for a Markov chain with state space $\mathcal{X}$ and transition probabilities $(P_{ij})_{i,j \in \mathcal{X}}$ if

$$\pi(j) = \sum_{i \in \mathcal{X}} \pi(i) P_{ij} \quad \forall j \in \mathcal{X}$$

If we could design and simulate an ergodic Markov chain $(X_t)_{t \geq 0}$ with stationary distribution π, then by the convergence theorem we would have

$$\mathbb{P}(X_t = x) \to \pi(x) \ \text{ as } \ t \to +\infty \ \ \forall x \in \mathcal{X}$$

and

$$\lim_{T \to +\infty} \frac{|\{t \in \{0, ..., T\} : X_t = x\}|}{T} = \pi(x) \ \text{ with probability 1 } \ \forall x \in \mathcal{X}$$

so that for large t we can use the Markov chain states as approximate samples from π. The easiest way to obtain a Markov chain with stationary distribution π is to design (and simulate) a π-reversible Markov chain. We recall that a Markov chain with state space $\mathcal{X}$ and transition probabilities $(P_{ij})_{i,j\in\mathcal{X}}$ is reversible with respect to the distribution π on $\mathcal{X}$ if

$$\pi(i)P_{ij} = \pi(j)P_{ji} \quad \forall i, j \in \mathcal{X}$$

The simulation of a π-reversible Markov chain is precisely what the *Metropolis-Hastings* algorithm achieves. The idea at the base of the algorithm is to find an irreducible Markov chain $(Y_t)_{t\geq 0}$ that we can simulate, with state space $\mathcal{X}$ and transition probabilities $(Q_{ij})_{i,j\in\mathcal{X}}$, known as the *proposal distribution*, and, given the current state of $(X_t)_{t\geq 0}$, propose a transition according to $(Y_t)_{t\geq 0}$, that is accepted with some appropriate probability α and rejected otherwise. Formally, given the current state $X_t = i$, the next state of the move proposal is determined by drawing a random variable $Y \sim p(Y_{t+1} \mid Y_t = i)$, i.e.

$$\mathbb{P}(Y = j \mid X = i) = P(Y_{t+1} = j \mid Y_t = i) = Q_{ij} \quad \forall j \in \mathcal{X}$$

Then, given $Y = j$, the move proposal is accepted with some probability $\alpha(i, j)$, in which case we have $X_{t+1} = j$, otherwise the proposal is rejected and we have $X_{t+1} = i$, meaning that the Markov chain stays in its current state. The resulting transition probabilities of the Markov chain $(X_t)_{t\geq 0}$ are given by

$$P_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i) = \mathbb{P}(Y = j \mid X_t = i)\alpha(i, j) = Q_{ij}\alpha(i, j) \quad \forall i, j \in \mathcal{X}$$

It turns out that the choice of the following acceptance probability

$$\alpha(i, j) = \min\left(1, \frac{\pi(j)Q_{ji}}{\pi(i)Q_{ij}}\right) = \min\left(1, \frac{Cb(j)Q_{ji}}{Cb(i)Q_{ij}}\right) \quad \forall i, j \in \mathcal{X}$$

guarantees that the resulting Markov chain $(X_t)_{t\geq 0}$ is π-reversible, therefore the simulated states can be used as approximate samples from the distribution of interest π. In fact, we have

$$\begin{aligned}\pi(i)P_{ij} = \pi(i)Q_{ij}\min\left(1, \frac{\pi(j)Q_{ji}}{\pi(i)Q_{ij}}\right) &= \min\left(\pi(i)Q_{ij}, \pi(j)Q_{ji}\right) \\ &= \pi(j)Q_{ji}\min\left(1, \frac{\pi(i)Q_{ij}}{\pi(j)Q_{ji}}\right) = \pi(j)P_{ji} \quad \forall i, j \in \mathcal{X}\end{aligned}$$

Moreover, it can be noticed that C is simplified in the formula for $\alpha(i,j)$, meaning that we are able to run the algorithm even when we know the distribution of interest up to a normalizing costant, which is often the case with posterior distributions in Bayesian statistics. Thus, we have the following algorithm.

Algorithm 1: Metropolis-Hastings sampling

Input: transition probabilities $(Q_{ij})_{i,j\in\mathcal{X}}$, $b(x) \;\; \forall x \in \mathcal{X}$

Output: approximate samples from the distribution $\pi(x) = Cb(x) \;\; \forall x \in \mathcal{X}$

1. Initialize $X_0 = i$ for some $i \in \mathcal{X}$
2. **For** $t = 0, 1, ..., T-1$ **do**:
   - Given $X_t = i$, simulate Y such that
   $$\mathbb{P}(Y = j \mid X_t = i) = Q_{ij} \quad \forall j \in \mathcal{X}$$
   - Given $Y = j$, compute $\alpha(i,j) = \min\left(1, \frac{b(j)Q_{ji}}{b(i)Q_{ij}}\right)$
   - Set $X_{t+1} = j$ with probability $\alpha(i,j)$ and $X_{t+1} = i$ otherwise
3. **Return** $X_1, ..., X_T$

A frequently used version of the Metropolis-Hastings algorithm is *Gibbs sampling*. In this case, we assume we want to sample from a random vector $X = (X_1, ..., X_n)$ with pmf π on a sample space $\mathcal{X}$ known up to a multiplicative constant, i.e. $\pi(x) = Cb(x) \;\; \forall x \in \mathcal{X}$ where $b(x)$ is known and C cannot be computed. Moreover, we suppose that for any $i \in \{1, ..., n\}$ and values $x_{-i} = (x_j)_{j\neq i} \in \mathcal{X}^{n-1}$ we can generate a random variable X with pmf $p(X_i \mid X_{-i} = x_{-i})$. The algorithm consists in applying the Metropolis-Hastings algorithm on a Markov chain with transition probabilities defined as follows. Given the current state $X_t = x$, an integer $i \in \{1, ..., n\}$ is (uniformly) chosen at random. Given the choice of i, a random variable X with pmf $p(X_i \mid X_{-i} = x_{-i})$ is generated. Given $X = x$, the candidate state for X_{t+1} is $y = (x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_n)$. The proposal probability is then given by

$$Q_{xy} = \frac{1}{n}\mathbb{P}(X_i = x \mid X_{-i} = x_{-i})$$

It turns out that the candidate state y is accepted with probability 1. The Gibbs sampling procedure will be used later for posterior inference in the Dirichlet Process Mixture Model.

Chapter 2

Clustering with the DP mixture model

2.1 Dirichlet process mixture model

As stated in Section 1.5, dedicated to the Dirichlet process, the random probability measure generated by the DP is almost surely discrete. This clearly poses a strong limitation for this model, since apparently it cannot be used in problems involving continuous distributions, such as density estimation. However, this limitation can be eliminated by using a DP-distributed random probability measure as the mixing measure of a certain parametric form with continuous kernel (Lo, 1984). Let Θ be a finite-dimensional parameter space and

$$\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$$

be a parametric form with continuous kernel, where $f(\cdot \mid \theta)$ indicates the continuous pdf associated with parameter θ. Then, by placing a probability distribution G on the parameter space Θ with Dirichlet process prior, we obtain the mixture pdf of $\mathcal{F}$ with respect to G

$$f_G(x) = \int f(x \mid \theta) dG(\theta)$$

with $G \sim \mathrm{DP}(M, G_0)$. By writing the random probability measure G as an infinite weighted sum of point masses, i.e. $G(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{\theta_k^*}(\cdot)$, then f_G becomes

$$f_G(x) = \sum_{k=1}^{\infty} w_k f(x \mid \theta_k^*)$$

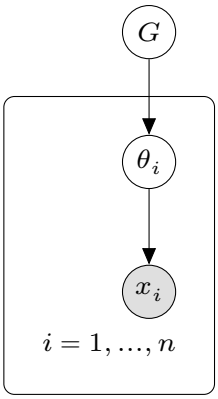

Figure 2.1: Graphical model of the Dirichlet process mixture model (DPMM) using plate notation

By introducing observations $x_i \mid G \overset{\text{iid}}{\sim} f_G$ for $i = 1, ..., n$, the above model is equivalent to the following hierarchical model

$$\begin{aligned} x_i \mid \theta_i &\overset{\text{iid}}{\sim} f(\cdot \mid \theta_i) && i = 1, ..., n \\ \theta_i \mid G &\overset{\text{iid}}{\sim} G && i = 1, ..., n \\ G &\sim \text{DP}(M, G_0) \end{aligned}$$

This model is known as *Dirichlet process mixture model (DPMM)*. The appropriate choice of continuous kernel depends on the application, but in most cases such mixtures define a rich family of distributions. An interesting property of this hierarchical model is that the posterior distribution of the random probability measure G is a mixture of DP models (Antoniak, 1974). With reference to the above model, we have

$$G \mid x_1, ..., x_n \sim \int \text{DP}\left(M + n, \frac{M}{M+n} G_0 + \frac{n}{M+n} \hat{G}_n\right) dp(\theta_1, ..., \theta_n \mid x_1, ..., x_n)$$

where

$$\hat{G}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\theta_i}$$

This means that, conditional on $\theta_1, ..., \theta_n$, the posterior distribution of G is a Dirichlet process with precision parameter $M + n$ and un-normalized centering measure $MG_0 + \sum_{i=1}^{n} \delta_{\theta_i}$.

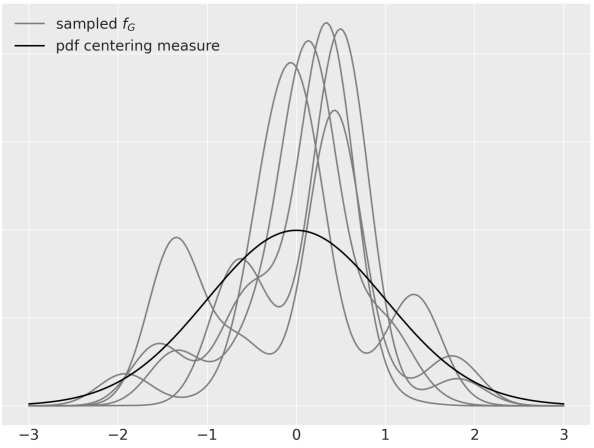

Figure 2.2: Plot of sample pdf's f_G from a DPMM with $x_i \mid \theta_i \overset{\text{iid}}{\sim} N(\theta_i, (0.3)^2)$, $\theta_i \mid G \overset{\text{iid}}{\sim} G$, and $G \sim \text{DP}(2, N(0,1))$

2.2 Clustering property of the DPMM

Consider the Dirichlet process mixture model defined in Section 2.1 and denote by $[n] = \{1, ..., n\}$ the set of the experimental units associated with the observations $x = (x_1, ..., x_n)$, respectively. An important property of the DPMM is that it implicitly defines a probability model for clustering (Müller et al., 2015). The almost sure discreteness of the DP-distributed random probability measure G implies a positive probability of ties among $\theta_1, ..., \theta_n$. Thus, we let $K \leq n$ denote the number of unique values in $\{\theta_1, ..., \theta_n\}$ and

$$\left\{\theta_j^* : j = 1, ..., K\right\}$$

denote the set of such unique values. Then, given the randomness of $\theta_i \mid G \overset{\text{iid}}{\sim} G$ for $i = 1, ..., n$, we have that

$$C_n = \left\{S_1, ..., S_K\right\}$$

where $S_j = \{i : \theta_i = \theta_j^*\}$ for any $j = 1, ..., K$ defines a random partition (or clustering) of the experimental units $[n]$. In other words, through the definition of the DPMM, we implicitly defined a probability prior $p(C_n)$ on the set of all possible partitions of $[n]$ and, together with the sampling density $f(\cdot \mid \theta_i)$, we get the infinite-dimensional analogue of the model-based clustering seen in Section 1.2 (with the DP as prior on the infinite-dimensional discrete distribution), known as *infinite model-based clustering* or *infinite*

mixture modeling, where we assume that there is an infinite number of clusters in the population and, on a finite set of observations, only a finite, but varying, number of clusters is observed. Such a higher flexibility then makes the DPMM very suitable for clustering applications where we do not know the number of clusters in advance or we assume that the number of clusters grows as more data are observed. In order to make the implied model $p(C_n)$ explicit, i.e. find the implied prior probability of each random partition of $[n]$, it is convenient to use the representation of partitions through the cluster membership indicators $c_1, ..., c_n$, labeled by order of appearance, as defined in Section 1.2. In addition we let $n_j = |S_j|$ for any $j \in \{1, ..., K\}$, $\theta^*_{i,j}$ denote the j-th unique value in $\{\theta_1, ..., \theta_i\}$, k_i denote the number of unique values in $\{\theta_1, ..., \theta_i\}$ and $n_{i,j} = |\{\theta_l = \theta^*_{i,j} : l = 1, ..., i\}|$ $(i \leq n)$. By the Blackwell-MacQueen Pólya urn representation of the Dirichlet process, the probability distribution of the conditional $\theta_i \mid \theta_1, ..., \theta_{i-1}$ for any $i \leq n$ is given by

$$p(\theta_i \mid \theta_1, ..., \theta_{i-1}) = \frac{M}{M+i-1} G_0(\theta_i) + \frac{1}{M+i-1} \sum_{j=1}^{k_{i-1}} n_{i-1,j} \delta_{\theta^*_{i-1,j}}(\theta_i)$$

From this expression, it is easy to derive the distribution of the increasing conditionals (predictive distributions) involving the cluster membership indicators, i.e. $c_i \mid c_1, ..., c_{i-1}$ for any $i \leq n$, which is given by

$$\mathbb{P}(c_i = j \mid c_1, ..., c_{i-1}) = \begin{cases} \frac{n_{i-1,j}}{M+i-1} & \text{for } \; j = 1, ..., k_{i-1} \\ \frac{M}{M+i-1} & \text{for } \; j = k_{i-1} + 1 \end{cases}$$

By combining the iid assumption on $\theta_1, ..., \theta_n$ and the above conditional probability, we get the marginal prior probability of each C_n (recall that $c_1 = 1$ by definition)

$$p(C_n) = p(c_1, ..., c_n) = \prod_{i=2}^{n} p(c_i \mid c_1, ..., c_{i-1}) = \frac{M^{K-1} \prod_{j=1}^{K} (n_j - 1)!}{(M+1) \cdots (M+n-1)}$$

Such a representation of the distribution of a random partition of $[n]$ by increasing conditionals of the cluster membership indicators is usually called *Chinese restaurant process (CRP)*, written $C_n \sim \text{CRP}(M)$ or $(c_1, ..., c_n) \sim \text{CRP}(M)$, because of the analogy to the apparently limitless capacity of Chinese restaurants in San Francisco given by Jim Pitman and Lester Dubins (Pitman, 2006). We suppose there is a Chinese restaurant with an

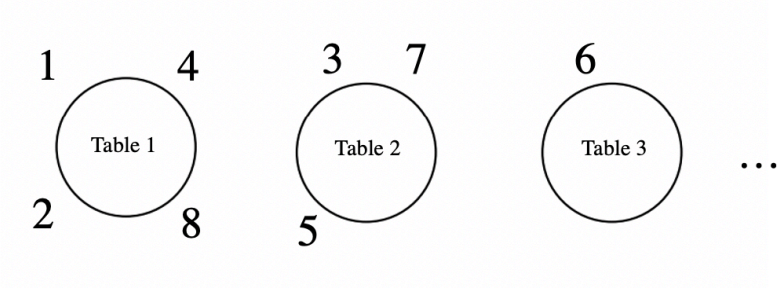

Figure 2.3: Example of sample partition from the *Chinese restaurant process*, where the circles represent the tables and the numbers around the circles represent the customers

infinite number of tables and each table has infinite capacity. We then imagine there is a sequence of customers who enter the restaurant one by one and sit at one of the tables. The first customer enters the restaurant and sits at a table. The second customer enters the restaurant and sits at the table where the first customer is sitting, with probability $\frac{1}{1+M}$, or at a new table, with probability $\frac{M}{1+M}$. At each of the following steps i, the i-th customer enters the restaurant and sits at one of the tables with already at least one customer, with probability proportional to the number customers sitting at that table, or at a new table, with probability proportional to M.

Finally, we remind that the sequence of latent parameters $\theta_1, ..., \theta_n$ in the DPMM is exchangeable, therefore, the probabilities are invariant to permutations of the indices of $\theta_1, ..., \theta_n$. We let $\theta_{-i} = (\theta_j)_{j \neq i} \in \Theta^{n-1}$, $\theta^*_{-i,j}$ denote the j-th unique value in $\{\theta_1, ..., \theta_n\} \backslash \{\theta_i\}$, k_{-i} denote the number of unique values in $\{\theta_1, ..., \theta_n\} \backslash \{\theta_i\}$ and $n_{-i,j} = |\{\theta_l = \theta^*_{-i,j} : l \in \{1, ..., n\} \backslash \{i\}\}|$. Then, we get

$$p(\theta_i \mid \theta_{-i}) = \frac{M}{M+n-1} G_0(\theta_i) + \frac{1}{M+n-1} \sum_{j=1}^{k_{-i}} n_{-i,j} \delta_{\theta^*_{-i,j}}(\theta_i)$$

Alternatively, without highlighting the unique values in the sequence and their multiplicity, we can write $p(\theta_i \mid \theta_{-i})$ as follows

$$p(\theta_i \mid \theta_{-i}) = \frac{M}{M+n-1} G_0(\theta_i) + \frac{1}{M+n-1} \sum_{j \neq i} \delta_{\theta_j}(\theta_i)$$

2.3 Posterior inference for the DPMM

We now analyze one of the procedures used for posterior inference in the Dirichlet process mixture model, considering the model defined in Section 2.1 with experimental units $[n]$ and observations $x = (x_1, ..., x_n)$. Although we may be generally interested in the posterior distribution of all the variables involved in the DPMM, in clustering applications we are primarily interested in the posterior of the latent parameters $\theta = (\theta_1, ..., \theta_n)$, $p(\theta_1, ..., \theta_n \mid x_1, ..., x_n)$, from which we can analyze the uncertainty associated with the cluster membership indicators and the number of observed clusters after observing the data. In fact, our objective is to assign similar observations, i.e. observations with the same latent parameter, to the same cluster, and therefore we want to use this distribution to choose the appropriate number of clusters in the observed data. We point out that, due to the discreteness of the DP-distributed random probability measure G, exact computation of posterior distributions and expectations is infeasible when we have more than a few observations. Thus, we have to resort to approximation methods to sample from the posterior distribution and compute posterior expectations. The most popular methods are based on MCMC (Neal, 2000). The direct approach is to simulate a Markov chain that has the posterior distribution of $\theta = (\theta_1, ..., \theta_n)$ given the data $x_1, ..., x_n$ as its stationary distribution. In this way, for a large number of states in the simulated trajectory of such a Markov chain, the last state of the trajectory is approximately distributed according to the target posterior distribution. Moreover, all the MCMC samples after the burn-in period can be used as approximate samples from this distribution. The easiest approach to simulate such a Markov chain and return the desired samples is to apply Gibbs sampling with the full conditional posterior distribution $p(\theta_i \mid \theta_{-i}, x)$ for any $i = 1, ..., n$, obtained by applying Bayes' theorem

$$p(\theta_i \mid \theta_{-i}, x) \propto p(x \mid \theta) p(\theta_i \mid \theta_{-i})$$

Knowing that

$$p(\theta_i \mid \theta_{-i}) = \frac{M}{M+n-1} G_0(\theta_i) + \frac{1}{M+n-1} \sum_{j \neq i} \delta_{\theta_j}(\theta_i)$$

the desired posterior distribution is given by

$$p(\theta_i \mid \theta_{-i}, x) \propto Mr_0(x_i)H_0(\theta_i) + \sum_{j \neq i} f(x_i \mid \theta_j)\delta_{\theta_j}(\theta_i) \tag{2.1}$$

where $H_0(\theta_i) \propto f(x_i \mid \theta_i)G_0(\theta_i)$ and

$$r_0(x_i) = \int f(x_i \mid \theta)G_0(d\theta)$$

In order for the Gibbs sampling method with such a full conditional distribution to be feasible, G_0 must be a conjugate prior for θ with respect to the kernel $f(\cdot \mid \theta)$, since without conjugacy $r_0(x_i)$ cannot typically be computed analytically. The algorithm, based on a slightly different version of the general Gibbs sampling procedure than the one briefly presented in Section 1.6, is the following.

Algorithm 2: Gibbs sampling for DPMM

Input: full conditional posterior distribution $p(\theta_i \mid \theta_{-i}, x)$

Output: approximate samples from the posterior distribution $p(\theta \mid x)$

1. Initialize $\theta^0 = (\theta_1^0, ..., \theta_n^0)$ for some $(\theta_1^0, ..., \theta_n^0) \in \Theta^n$
2. **For** $t = 1, ..., T$ **do**:
   - **For** $i = 1, ..., n$ **do**: draw θ_i^t from $p(\theta_i \mid \theta_{-i}^{t-1}, x)$ as specified in (2.1)
3. **Return** $\theta^1, ..., \theta^T$

This algorithm successfully returns the desired approximate samples from the posterior distribution of $\theta = (\theta_1, ..., \theta_n)$, from which we can obtain also the approximate samples from the posterior of the clustering membership indicators $c = (c_1, ..., c_n)$ and the posterior of the number of clusters K. However, it has been noticed that this Gibbs sampler suffers from a slowly mixing Markov chain, meaning that convergence to the target posterior distribution is slow. More efficient algorithms have been obtained by introducing the cluster membership indicators in the Gibbs sampling procedure.

2.4 Application: customer segmentation

In this section, we put the previously discussed theory into practice to get a sense of how it performs when applied to some data[1]. This application consists of a cluster analysis to infer the appropriate number of clusters for the data being analyzed. Moreover, we carried out a sensitivity analysis of the model to the prior parameter specification, i.e. we observed how the results change by changing the model hyperparameters. The dataset used is the *Wholesale customers* dataset of the UCI Machine Learning Repository[2]. This dataset is about the customers of a wholesale distributor from the Horeca (Hotel/Restaurant/Café) and the Retail channels, distributed in the regions of two large Portuguese cities, Lisbon and Porto, and a complementary region. There are 6 numerical variables in the dataset, reporting the annual spending in monetary units (m.u.) of each customer on various product categories: fresh products (Fresh), milk products (Milk), grocery (Grocery), frozen products (Frozen), detergents and paper products (Detergents_Paper) and delicatessen (Delicassen). Moreover, there are two categorical variables, indicating each customer's channel (Channel) and region (Region) as previously defined.
We focused on the numerical variables for the cluster analysis. After eliminating some outliers of the 6 numerical variables from the dataset, we got the following summary statistics and histograms.

	Mean	Std. deviation
Fresh	9718	8200
Milk	3989	3261
Grocery	5564	4623
Frozen	1800	1613
Detergents_Paper	1841	2207
Delicassen	984	785

From the histograms, it can be seen that the empirical marginal distributions of all the six variables are quite skewed. However, it was not necessary to apply a logarithmic

1. The code written for this analysis can be found in the GitHub repository https://github.com/luiginoto/dpmm_clustering
2. The dataset is available at https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

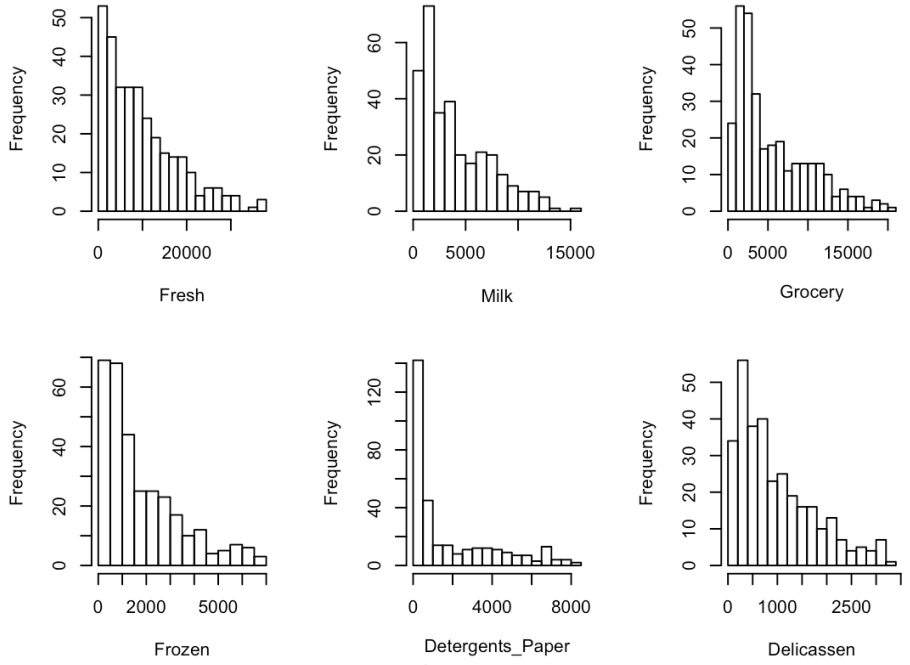

Figure 2.4: Histograms of the 6 numerical variables of the *Wholesale customers* dataset

transformation to the data, since the DPMM is a quite flexible model, providing a rich family of densities when using common kernels for model-based clustering like the normal kernel. Moreover, it is not easy to detect a clustering pattern qualitatively by looking at the histograms, therefore we proceed with the description of the analysis.

In more detail, in this analysis we defined a DPMM on the numerical data and we carried out posterior inference through MCMC. We got the approximate samples from the posterior distribution of the cluster parameters and the cluster membership indicators obtained by running the Gibbs sampling algorithm for a sufficiently large number of steps for convergence to the target posterior distributions. Starting from this samples, at each MCMC step the number of clusters was observed, thereby obtaining approximate samples from the posterior distribution of the number of clusters. This was our distribution of interest, explaining the uncertainty associated with the number of clusters after observing the data. We estimated it with the relative frequencies obtained from the MCMC samples (empirical distribution) and chose the Maximum a Posteriori (MAP) estimate as the point estimation for the number of clusters in the observations. In order to get a sense of the

sensitivity of the model to the prior parameters, i.e. the parameters of the centering measure of the DP prior, we performed the analysis more times using prior parameters which gradually made the prior non-informative (more diffuse) and investigated the variation in the estimates. The analysis has been carried out in R using the `dirichletprocess` package[3], an intuitive package for creating Dirichlet process objects and easily running the Gibbs sampling algorithm.

Univariate case

We first performed the analysis described above in the univariate case, by applying the model to the Grocery variable. We defined a DPMM with univariate normal sampling density and normal-inverse-gamma distribution as the centering measure of the DP prior, since it is the conjugate prior of a normal distribution with unknown mean and variance. Using the same notation as in Section 2.1, the model is then given by $\theta = (\mu, \sigma^2)$ and

$$f(x_i \mid \theta) = \mathrm{N}(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

$$G_0(\theta \mid \gamma) = \mathrm{N}\left(\mu \mid \mu_0, \frac{\sigma^2}{k_0}\right) \text{Inv-Gamma}(\sigma^2 \mid \alpha_0, \beta_0)$$

where $\gamma = (\mu_0, k_0, \alpha_0, \beta_0)$ are the model hyperparameters. The model object has been created with the function `DirichletProcessGaussian`, getting as input the data and the hyperparameters γ, and the Gibbs sampling algorithm was run for 2500 iterations by calling the function `Fit`, which outputs the updated model object with the samples of the cluster parameters and the clustering assignments. Before running the algorithm, we scaled the Grocery variable to have zero mean and unit standard deviation. In order to test the sensitivity of the model to hyperparameters, we performed the analysis with the following lists of hyperparameters, which gradually make the prior more diffuse, non-informative:

- case 1: $\mu_0 = 0$, $k_0 = 1$, $\alpha_0 = 1$, $\beta_0 = 1$; MAP estimate for the number of clusters equal to 6;

3. The package documentation is available at https://CRAN.R-project.org/package=dirichletprocess

- case 2: $\mu_0 = 0$, $k_0 = \frac{1}{5}$, $\alpha_0 = 1$, $\beta_0 = 5$; MAP estimate for the number of clusters equal to 2;
- case 3: $\mu_0 = 0$, $k_0 = \frac{1}{10}$, $\alpha_0 = 1$, $\beta_0 = 10$; MAP estimate for the number of clusters equal to 2;
- case 4: $\mu_0 = 0$, $k_0 = \frac{1}{100}$, $\alpha_0 = 1$, $\beta_0 = 100$; MAP estimate for the number of clusters equal to 1.

The results are represented in Figures 2.5 and 2.6.

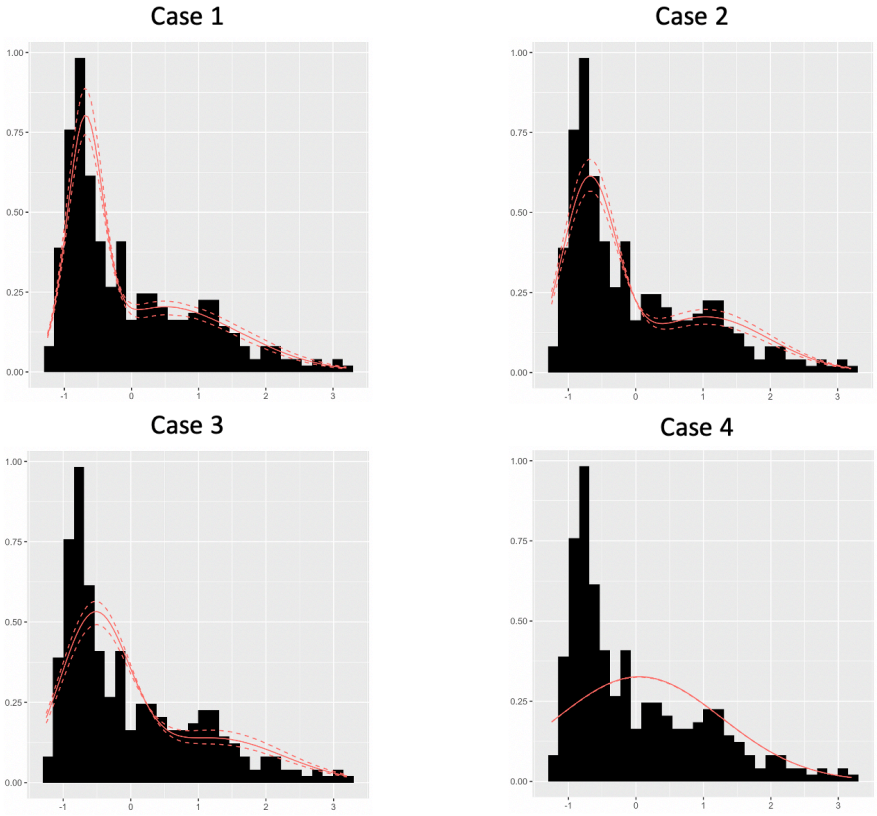

Figure 2.5: Histogram of the Grocery variable and plot of the DPMM posterior mean and credible intervals

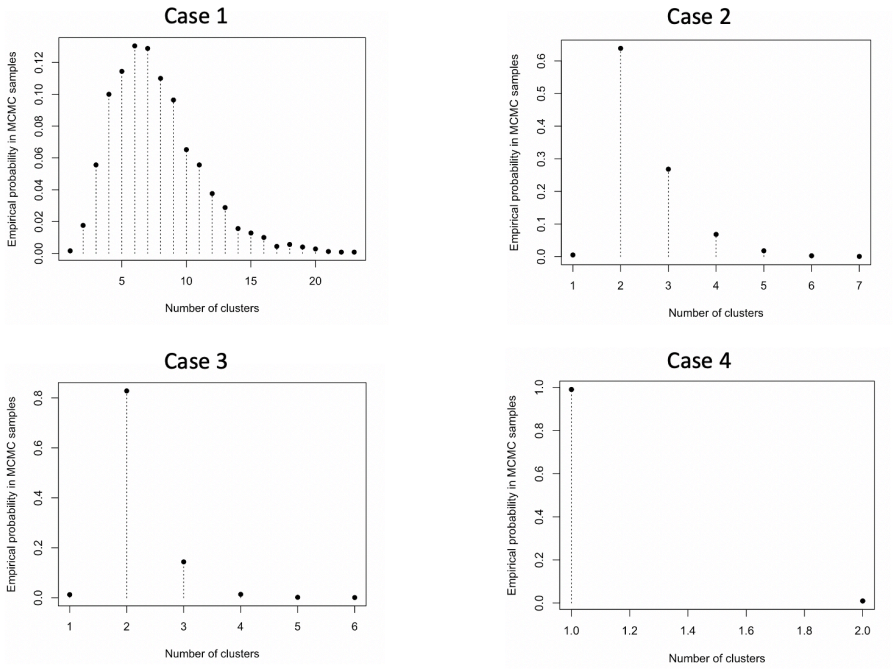

Figure 2.6: Approximation of the posterior distribution of the number of clusters (Univariate case)

Multivariate case

The multivariate case was covered as well. We scaled the data and applied PCA on the numerical variables. The results are collected in the table below.

	PC1	PC2	PC3	PC4	PC5	PC6
Std. deviation	1.64	1.16	0.91	0.83	0.57	0.37
Cumulative prop.	0.45	0.67	0.81	0.92	0.98	1

Based on these results, we decided to use the first three principal components for the analysis, explaining 81% of the total variance in the dataset. A scatterplot of these components is displayed below.

The variables have been modeled through a DPMM with multivariate normal sampling density. The normal-Wishart distribution has been chosen as the centering measure of the DP prior, since it is the conjugate prior of a multivariate normal distribution with unknown mean vector and precision matrix (the inverse of the covariance matrix). This is equivalent to choosing the normal-inverse-Wishart distribution as centering measure with

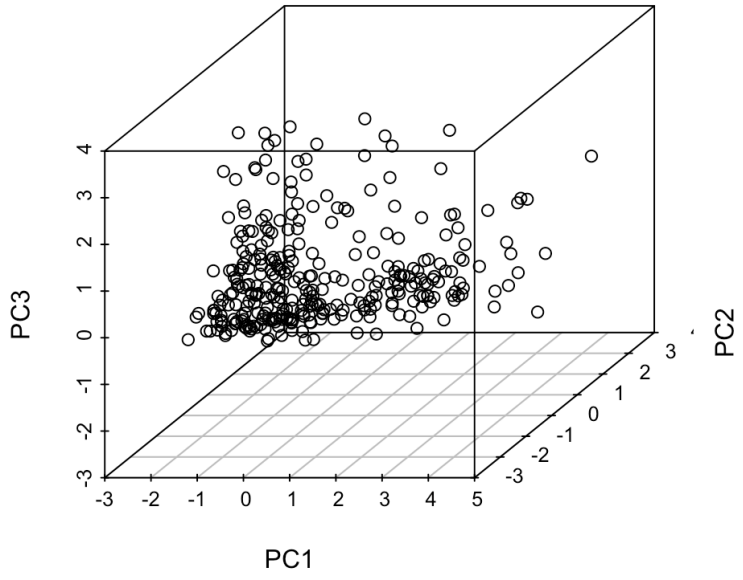

Figure 2.7: Plot of the first three principal components of the dataset

the multivariate normal distribution parametrized by mean vector and covariance matrix. The model is given by $\theta = (\mu, W)$, where μ is the mean vector and W is the precision matrix, and

$$f(x_i \mid \theta) = \frac{|W|^{\frac{1}{2}}}{(2\pi)^{\frac{3}{2}}} \exp\left\{ -\frac{1}{2}(x_i - \mu)^\top W (x_i - \mu) \right\}$$

$$G_0(\mu, W \mid \mu_0, k_0, \nu, \Lambda) = \mathrm{N}(\mu \mid \mu_0, (k_0 W)^{-1}) \mathrm{Wi}(W \mid \Lambda, \nu)$$

where the vector μ_0, the numbers k_0 and ν and the matrix Λ are the model hyperparameters. Similarly to the univariate case, the model object has been created with the function `DirichletProcessMvnormal`, getting as input the data and the hyperparameters, and the Gibbs sampling algorithm was run for 2500 iterations by calling the function `Fit` returning the updated model object. The analysis has been carried out with the following lists of hyperparameters for sensitivity analysis:

- case 1: μ_0 equal to the sample mean of the data, $k_0 = 3$, $\nu = 3$, $\Lambda = I$; MAP estimate for the number of clusters equal to 5;
- case 2: μ_0 equal to the sample mean of the data, $k_0 = \frac{1}{5}$, $\nu = 3$, $\Lambda = 5I$; MAP estimate for the number of clusters equal to 3;

- case 3: μ_0 equal to the sample mean of the data, $k_0 = \frac{1}{10}$, $\nu = 3$, $\Lambda = 10I$; MAP estimate for the number of clusters equal to 3;
- case 4: μ_0 equal to the sample mean of the data, $k_0 = \frac{1}{100}$, $\nu = 3$, $\Lambda = 100I$; MAP estimate for the number of clusters equal to 1.

Figures 2.8 and 2.9 show the results in this case.

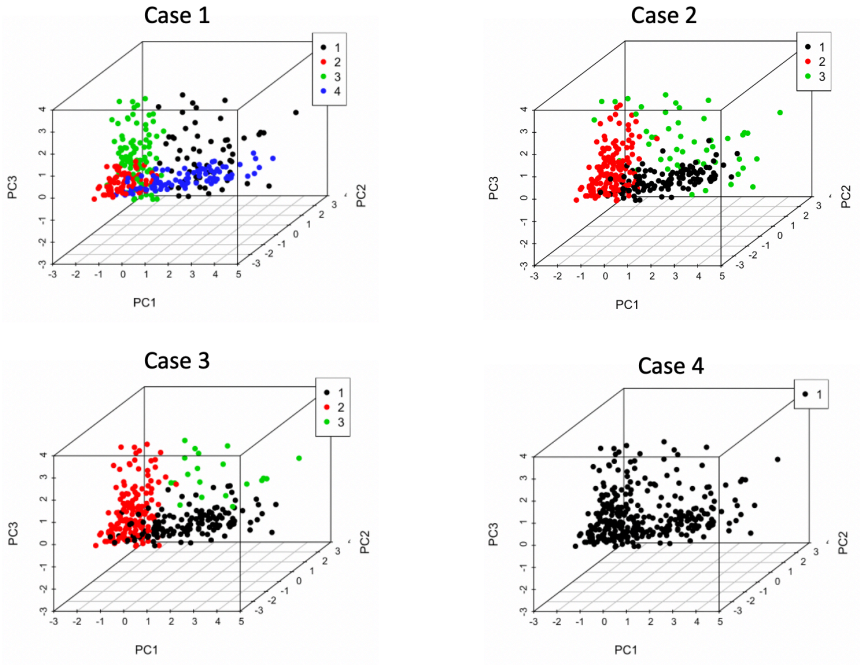

Figure 2.8: Clustering obtained from the last MCMC sample

Comments

In this application, we have seen how the DPMM can be used in practice, highlighting the advantages of using such a sophisticated model but also some potential limitations that it is important to be aware of. The main benefit of using this model is clearly the fact that it naturally "formalizes" the process of inferring the number of clusters in the data, by automatically incorporating the uncertainty about this variable. In this way, we were able to precisely choose the number of clusters through an estimate from its approximate posterior distribution. This is not easy when applying finite mixture modeling, because of the upper bound on the number of clusters in the population. On the other hand, in

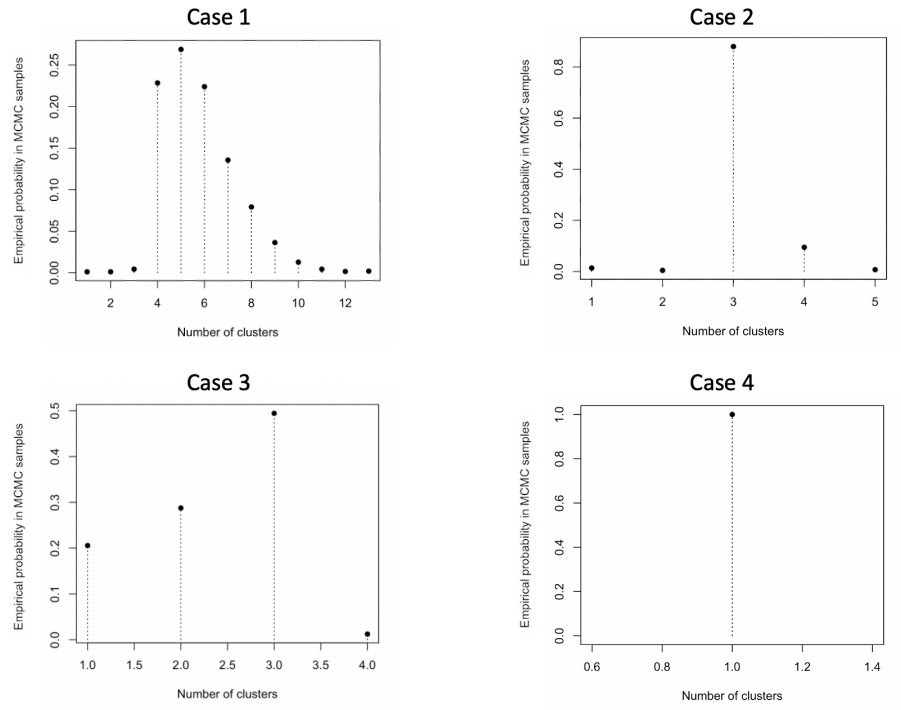

Figure 2.9: Approximation of the posterior distribution of the number of clusters (Multivariate case)

both the univariate and the multivariate cases, we noticed that the results of the analysis changed considerably when using different hyperparameters which gradually made the prior non-informative, showing high sensitivity. Thus, despite the high flexibility of this model and nonparametric models in general, the choice of the prior parameters might substantially affect the inference outcome. It is then important to carefully choose the prior parameters or add an additional layer of uncertainty in the model by introducing hyperpriors. Nevertheless, it is important to point out that we analyzed a small dataset and it might be the case that the model would show less sensitivity when working with datasets of greater size.

Conclusions

To conclude, in this paper we have analyzed the theory behind the Dirichlet process mixture model and its application to clustering. The objective was to provide a complete picture about this topic. According to this approach, the theoretical part was structured in order to cover also the background necessary for a deep understanding of the model. A practical application of the model has also been described, in order to better understand the benefits of the model for cluster analysis with respect to its finite counterpart, but also highlight some potential drawbacks.

Finally, it must be stressed that we had to leave out important results and details of the covered topics that would be worth mentioning in more extended reports. As for the inference part, we would have included a discussion on the MCMC algorithms for dealing with the nonconjugate case in the DPMM and the variational inference algorithms, representing an alternative to MCMC in Bayesian inference. Moreover, it would have been interesting to talk about posterior inference with hyperparameters, as well as the concepts of variation of information and least squares clustering to obtain a point estimation from the posterior $p(C_n \mid x)$, that could have been integrated in the practical application to include a comparison with the output of the k-means algorithm.

Bibliography

Antoniak, C. E. 1974. "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *The Annals of Statistics* 2(6): 1152–1174.

Bilodeau, M. and Brenner, D. 1999. *Theory of Multivariate Statistics.* Chap. 3. New York: Springer-Verlag.

Blackwell, D. and MacQueen, J. B. 1973. "Ferguson Distributions Via Polya Urn Schemes." *The Annals of Statistics* 1(2): 353–355.

Buntine, W. L. 1994. "Operations for Learning with Graphical Models." *Journal of Artificial Intelligence Research* 2(1): 159–225.

Claeskens, G. and Hjort, N. L. 2008. *Model Selection and Model Averaging.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

dirichletprocess package. <https://CRAN.R-project.org/package=dirichletprocess>. [last access: 13/06/21].

Ferguson, T. S. 1973. "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics* 1(2): 209–230.

Fraley, C. and Raftery, A. E. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association* 97(458): 611–631.

Ghosal, S. and van der Vaart, A. 2017. *Fundamentals of Nonparametric Bayesian Inference.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Hewitt, E. and Savage, L. J. 1955. "Symmetric measures on Cartesian products." *Transactions of the American Mathematical Society* 80: 470–501.

Hodge, V. J. and Austin, J. 2004. "A Survey of Outlier Detection Methodologies." *Artificial Intelligence Review* 22: 85–126.

Ližbetinová, L., Štarchoň, P., Lorincová, S., Weberova, D., and Průša, P. 2019. "Application of Cluster Analysis in Marketing Communications in Small and Medium-Sized Enterprises: An Empirical Study in the Slovak Republic." *Sustainability* 11(8): 2302.

Lo, A. Y. 1984. "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates." *The Annals of Statistics* 12(1): 351–357.

Müller, P., Quintana, F. Andrés, Jara, A., and Hanson, T. 2015. *Bayesian Nonparametric Data Analysis.* Chap. 2. Cham: Springer.

Neal, R. M. 2000. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics* 9(2): 249–265.

Pitman, J. 2006. *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002.* Chap. 3. Berlin: Springer-Verlag.

Rasmussen, C. E. 1999. "The Infinite Gaussian Mixture Model." In *Proceedings of the 12th International Conference on Neural Information Processing Systems,* 554–560. Cambridge: MIT Press.

Ross, S. M. 2014. *Introduction to Probability Models.* 11th ed. Chap. 4. San Diego: Academic Press.

Sethuraman, J. 1994. "A constructive definition of Dirichlet priors." *Statistica Sinica* 4(2): 639–650.