FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Pre-Publicación MATUC - 2017 - 6

PARSIMONIOUS HIERARCHICAL MODELING USING REPULSIVE DISTRIBUTIONS

José Quinlan [1], Fernando A. Quintana[1], Garritt L. Page[2]

[1] Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Chile

[2] Department of Statistics, Brigham Young University, USA

Parsimonious Hierarchical Modeling Using Repulsive Distributions

José Quinlan
Departamento de Estadística
Pontificia Universidad Católica de Chile
jjquinla@mat.puc.cl

Fernando A. Quintana
Departamento de Estadística
Pontificia Universidad Católica de Chile
quintana@mat.uc.cl

Garritt L. Page
Department of Statistics
Brigham Young University
page@stat.byu.edu

January 18, 2017

Abstract

Employing nonparametric methods for density estimation has become routine in Bayesian statistical practice. Models based on discrete nonparametric priors such as Dirichlet Process Mixture (DPM) models are very attractive choices due to their flexibility and tractability. However, a common problem in fitting DPMs or other discrete models to data is that they tend to produce a large number of (sometimes) redundant clusters. In this work we propose a method that produces parsimonious mixture models (i.e. mixtures that discourage the creation of redundant clusters), without sacrificing flexibility or model fit. This method is based on the idea of repulsion, that is, that any two mixture components are encouraged to be well separated. We propose a family of d-dimensional probability densities whose coordinates tend to repel each other in a smooth way. The induced probability measure has a close relation with Gibbs measures, Graph theory and Point Processes. We investigate its global properties and explore its use in the context of mixture models for density estimation. Computational techniques are detailed and we illustrate its usefulness with some well-known data sets and a small simulation study.

Key Words: Gibbs measures, Graph Theory, Mixture Models, Repulsive Point Processes.

1 Introduction

Hierarchical mixture models have been very successfully employed in a myriad of applications of Bayesian modeling. A typical formulation for such models adopts the basic form

$$y_i|\theta_i \overset{ind}{\sim} k(y_i;\theta_i), \qquad \theta_1,\ldots,\theta_n \overset{iid}{\sim} \sum_{\ell=1}^{N} w_\ell \delta_{\phi_\ell}, \qquad \phi_1,\ldots,\phi_N \overset{iid}{\sim} G_0, \tag{1.1}$$

where $k(y;\theta)$ is a suitable kernel density, $1 \leq N \leq \infty$, component weights $w_1,\ldots,w_N$ are nonnegative and $\sum_{\ell=1}^{N} w_\ell = 1$ with probability 1. Here N could be regarded as fixed or random and in the latter case a

prior $p(N)$ would need to be specified. Depending on the modeling goals and data particularities, the model could have additional parameters and levels in the hierarchy. The generic model (1.1) includes, as special cases, finite mixture models (Frühwirth-Schnatter 2006) and species sampling mixture models (Pitman 1996; Quintana 2006), in turn including several well-known particular examples such as the Dirichlet process (Ferguson 1973) and the Pitman-Yor process (Pitman and Yor 1997).

A common feature of models like (1.1) is the use of *i.i.d.* parameters $\phi_1, \ldots, \phi_N$. This choice seems to have been largely motivated by the resulting tractability of the models, specially in the nonparametric case ($N = \infty$). There is also a substantial body of literature concerning important properties such as wide support, posterior consistency, and posterior convergence rates, among others. See, for instance, Ghosal and van der Vaart (2007) and Shen et al. (2013).

While the use of *i.i.d.* atoms in the mixture in (1.1) is technically (and practically) convenient, a typical summary of the induced posterior clustering will usually contain a number of very small clusters or even some singletons. As a specific example, we considered a synthetic data set of $n = 300$ independent observations simulated from the following mixture of 4 bivariate normal distributions:

$$\boldsymbol{y} \sim 0.2 \times \mathrm{N}_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.3 \times \mathrm{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.3 \times \mathrm{N}_2(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) + 0.2 \times \mathrm{N}_2(\boldsymbol{\mu}_4, \boldsymbol{\Sigma}_4),$$

with

$$\boldsymbol{\mu}_1 = (0,0)^\top, \quad \boldsymbol{\mu}_2 = (3,3)^\top, \quad \boldsymbol{\mu}_3 = (-3,-3)^\top, \quad \boldsymbol{\mu}_4 = (-3,0)^\top$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}, \quad \boldsymbol{\Sigma}_4 = \begin{pmatrix} 3 & -2 \\ -2 & 2 \end{pmatrix}.$$

The left panel in Figure 1 shows the original data and clusters, labeled with different numbers and colors. We fit to these data the variation of model (1.1) implemented in the function `DPdensity` of `DPpackage` (Jara et al. 2011), which is the bivariate version of the DP-based model discussed in Escobar and West (1995). The right panel of Figure 1 shows the same data but now displays the cluster configuration resulting from the least squares algorithm described in Dahl (2006). The estimated partition can be thought of as a particular yet useful summary of the posterior distribution of partitions for this model. What we observe is a common situation in the application of models like (1.1): we find 6 clusters (the simulation truth involved 4 clusters), one of which is a singleton. Such small clusters are very hard to interpret and a natural question arises, is it possible to limit and ideally, avoid such occurrences?

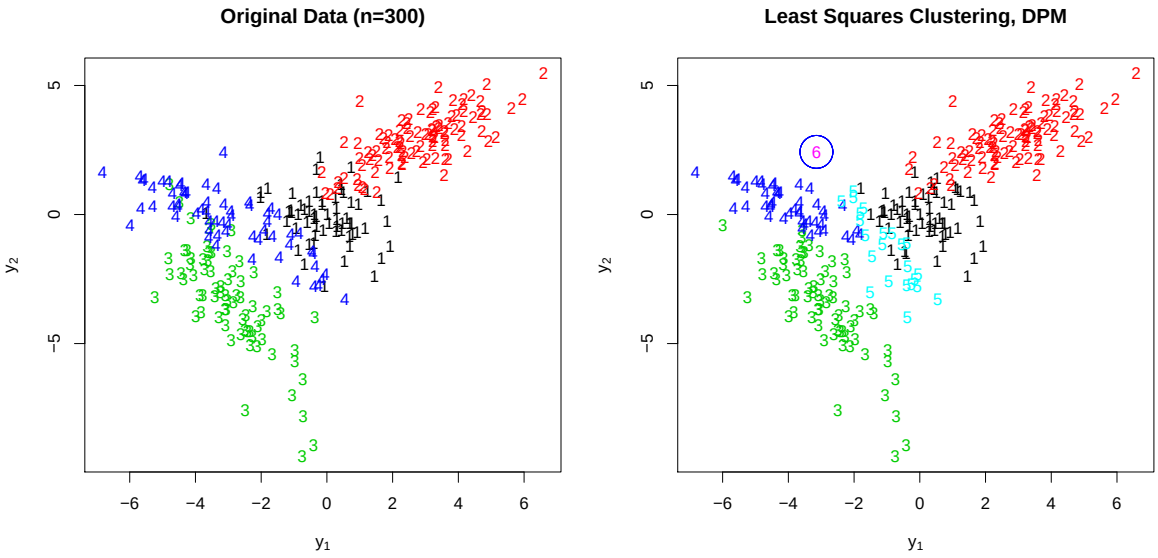

Figure 1: Data simulated from the mixture of 4 bivariate normal densities in (1). The left panel shows the original $n = 300$ data points with colors and numbers indicating the original cluster. The right panel shows the clustering resulting from applying Dahl's least squares clustering algorithm to a DP mixture. See the text for more details.

In an example like what is described above, our main motivation is not pinning down the "true" number of simulated clusters. What we actually want to accomplish is to develop a model that encourages joining such small clusters with other larger ones. This would certainly facilitate interpretation of the resulting clusters. Doing so has another conceptual advantage, which is sparsity. The non-sparse behavior shown in the right panel of Figure 1 is precisely facilitated by the fact that the atoms in the mixture are *i.i.d.* and therefore, can move freely with respect to each other. Thus to achieve our desired goal, we need atoms that mutually *repel* each other.

Colloquially, the concept of *repulsion* among a set of objects implies that the objects tend to separate rather than congregate. This notion of repulsion has been studied in the context of point processes. For example, Determinantal Point Processes (Lavancier et al. 2015), Strauss Point Processes (Mateu and Montes 2000; Ogata and Tanemura 1985) and Matérn-type Point Processes (Rao et al. 2016) are all able to generate point patterns that exhibit more repulsion than that expected from a Poisson Point Process (Daley and Vere-Jones 2002). Given a fixed number of points within a bounded (Borel) set, the Poisson Point Process can generate point configurations such that two points can be very close together simply by chance. The repulsion in Determinantal, Strauss and Matérn-type processes discourages such behavior and is controlled

by a set of parameters (finite/infinite dimensional) that inform pattern configurations. Among these, to our knowledge, only Determinantal Point Processes (DPP) have been employed to introduce the notion of repulsion in statistical modeling (see Xu et al. 2016).

An alternative way to incorporate the notion of repulsion in modeling is to construct a multivariate probability distribution that explicitly parameterizes repulsion. Along these lines Fúquene et al. (2016) develop a family of probability densities called non local priors (NLPs) that incorporates repulsion by penalizing small relative distances between coordinates. Our approach to incorporating repulsion is to model coordinate interactions through the potential functions (functions that describe the ability to interact) found in so called Gibbs measures. As will be shown, this provides more control on the strength of repulsion and more flexibility on the types of repulsion that can be considered.

Gibbs measures have been widely studied and used for describing phenomena from mechanical statistics (Daley and Vere-Jones 2002). Essentially, they are used to model the average macroscopic behavior of particle systems through a set of probability and physical laws that are imposed over the possible microscopic states of the system. Through the action of potentials, Gibbs measures can induce attraction or repulsion between particles. A number of authors have approached repulsive distributions by specifying a particular potential function in a Gibbs measures (though the connections to Gibbs measures was not explicitly stated). For example, Petralia et al. (2012) use a Lennard-Jones type potential (Jones 1924) to introduce repulsion. Interestingly, there is even a connection between Gibbs models and DPPs via versions of Papangelou intensities (Papangelou 1974). See Georgii and Yoo (2005) for more details. It is worth noting that in each of the works just cited, the particles (following the language in mechanical statistics) represent location parameters in mixture models.

Similar to the works just mentioned, we focus on a particular potential specification that introduces repulsion via a joint distribution. There are at least three benefits to employing the class of repulsive distributions we develop for statistical modeling:

(i) The repulsion is explicitly parameterized in the model and produces a flexible and smooth repulsion effect.

(ii) The normalizing constant and induced probability distributions have closed forms, they are (almost) tractable and provide intuition regarding the presence of repulsion.

(iii) The computational aspects related to simulation are fairly simple to implement.

In what follows, we discuss theoretical and applied aspects of the proposed class of repulsive distributions

and in particular we emphasize how the repulsive class of distributions achieves the three properties just listed.

The remainder of this article will be organized as follows. In Section 2 we formally introduce the notion of repulsion in the context of a probability distribution and discuss several resulting properties. In Section 3, we detail how the repulsive probability distributions can be employed in hierarchical mixture modeling. Section 4 contains results from a small simulation study that compares the repulsive mixture models we develop to DPMs and finite mixture models. In Section 5 we apply the methodology to two well known datasets. Section 6 provides a discussion and directions for future work. Proofs of all results are provided in the Appendix.

2 Probability Repulsive Distributions

We start by providing contextual background and introducing notation that will be used throughout.

2.1 Background and Preliminaries

We will use the k-fold product space of $\mathbb{R}^d$ denoted by $\mathbb{R}^d_k = \prod_{i=1}^k \mathbb{R}^d$ and $\mathcal{B}(\mathbb{R}^d_k) = \bigotimes_{i=1}^k \mathcal{B}(\mathbb{R}^d)$ its associated σ-algebra as the reference space on which the class of distributions we derive will be defined. Here, $k \in \mathbb{N}$ ($k \geq 2$), $d \in \mathbb{N}$ and $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-algebra on $\mathbb{R}^d$. Let $\boldsymbol{x}_{k,d} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^d$. $\boldsymbol{x}_{k,d}$ can be thought of as k ordered objects of dimension d jointly allocated in $\mathbb{R}^d_k$. We add to the measurable space $(\mathbb{R}^d_k, \mathcal{B}(\mathbb{R}^d_k))$ a σ-finite measure $\lambda_d^k = \bigotimes_{i=1}^k \lambda_d$, that is the k-fold product of the Lebesgue measure λ_d on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. To represent integrals with respect to λ_d^k, we will use $\mathrm{d}\boldsymbol{x}_{k,d}$ instead of $\mathrm{d}\lambda_d^k(\boldsymbol{x}_{k,d})$. When $d = 1$ we will use the common notation $\mathbb{R}^1 = \mathbb{R}$, $\mathcal{B}(\mathbb{R}^1) = \mathcal{B}(\mathbb{R})$, $\mathbb{R}^1_k = \mathbb{R}^k$, $\mathcal{B}(\mathbb{R}^1_k) = \mathcal{B}(\mathbb{R}^k)$, $\boldsymbol{x}_{k,1} = \boldsymbol{x}_k$, $\lambda_1^k = \lambda^k$ and $\lambda_1 = \lambda$. Also, given two metric spaces (Ω_1, d_1) and (Ω_2, d_2) we denote by $C(\Omega_1; \Omega_2)$ the class of all continuous functions $f : \Omega_1 \to \Omega_2$. In what follows we use the term "repulsive distribution" to reference a distribution that formally incorporates the notion of repulsion.

As mentioned previously, our construction of non-*i.i.d.* multivariate distributions depends heavily on Gibbs measures where dependence (and hence repulsion) between the coordinates of $\boldsymbol{x}_{k,d}$ is introduced via functions that model interactions between them. More formally, let $\overline{\mathbb{R}}$ be the extended real line (i.e., $\mathbb{R} \cup \{-\infty, \infty\}$) and $\mathcal{B}(\overline{\mathbb{R}})$ its associated Borel σ-algebra generated by the order topology. Consider $\varphi_1 : \mathbb{R}^d \to \overline{\mathbb{R}}$

a measurable function and $\varphi_2 : \mathbb{R}^d \times \mathbb{R}^d \to \overline{\mathbb{R}}$ a measurable and symmetric function. Define

$$\nu_{\mathrm{G}}\Big(\prod_{i=1}^{k} A_i\Big) = \int_{\prod_{i=1}^{k} A_i} \exp\Big[-\sum_{i=1}^{k} \varphi_1(\boldsymbol{x}_i) - \sum_{r<s}^{k} \varphi_2(\boldsymbol{x}_r, \boldsymbol{x}_s)\Big] \mathrm{d}\boldsymbol{x}_{k,d}, \tag{2.1}$$

where $\prod_{i=1}^{k} A_i$ is the cartesian product of $A_1, \ldots, A_k \in \mathcal{B}(\mathbb{R}^d)$. Here, φ_1 can be thought of as a physical force that controls the influence that the environment has on each coordinate $\boldsymbol{x}_i$ while φ_2 controls the interaction between pairs of coordinates $\boldsymbol{x}_r$ and $\boldsymbol{x}_s$. If φ_1 and φ_2 are selected so that $\nu_{\mathrm{G}}(\mathbb{R}_k^d)$ is finite, then by Caratheodory's Theorem ν_{G} defines a unique finite measure on $(\mathbb{R}_k^d, \mathcal{B}(\mathbb{R}_k^d))$. The induced probability measure corresponding to the normalized version of (2.1), is called a (second-order) Gibbs measure. The normalizing constant (total mass of $\mathbb{R}_k^d$ under ν_{G})

$$\nu_{\mathrm{G}}(\mathbb{R}_k^d) = \int_{\mathbb{R}_k^d} \exp\Big[-\sum_{i=1}^{k} \varphi_1(\boldsymbol{x}_i) - \sum_{r<s}^{k} \varphi_2(\boldsymbol{x}_r, \boldsymbol{x}_s)\Big] \mathrm{d}\boldsymbol{x}_{k,d}$$

is commonly known as partition function (Pathria and Beale 2011) and encapsulates important qualitative information about the interactions and the degree of disorder present in the coordinates of $\boldsymbol{x}_{k,d}$. In general, $\nu_{\mathrm{G}}(\mathbb{R}_k^d)$ is (almost) intractable mainly because of the presence of φ_2.

Note that symmetry of φ_2 (i.e., $\varphi_2(\boldsymbol{x}_r, \boldsymbol{x}_s) = \varphi_2(\boldsymbol{x}_s, \boldsymbol{x}_r)$) means that ν_{G} defines a symmetric measure. This implies that the order of coordinates is immaterial. If $\varphi_2 \equiv 0$ then ν_{G} reduces to a structure where coordinates do not interact and are only subject to environmental influence through φ_1. When $\varphi_2 \neq 0$, it is common that $\varphi_2(\boldsymbol{x}, \boldsymbol{y})$ only depends on the relative distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ (Daley and Vere-Jones 2002). More formally, let $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+$ be a metric on $\mathbb{R}^d$ and $\phi : \mathbb{R}_0^+ \to \overline{\mathbb{R}}$ a measurable function. To avoid pathological or degenerate cases, we consider metrics that do not treat singletons as open sets in the topology induced by ρ. Then letting $\varphi_2(\boldsymbol{x}, \boldsymbol{y}) = \phi(\rho(\boldsymbol{x}, \boldsymbol{y}))$, interactions will be smooth if, for example, $\phi \in C(\mathbb{R}_0^+; \overline{\mathbb{R}})$. Following this general idea, Petralia et al. (2012) use $\phi(r) = \tau(1/r)^\nu$ with $r \in \mathbb{R}_0^+$ and $\tau, \nu \in \mathbb{R}^+$ to construct repulsive probability densities, which is a particular case of the Lennard-Jones type potential (Jones 1924) that appears in molecular dynamics. Another potential that can be used to define repulsion is the (Gibbs) hard-core potential $\phi(r) = \infty \mathbb{I}_{[0,b]}(r)$ with $b \in \mathbb{R}^+$ (Illian et al. 2008), which is a particular case of the Strauss potential (Strauss 1975). Here, $\mathbb{I}_A(r)$ is the indicator function over a Borel set $A \in \mathcal{B}(\mathbb{R})$. This potential, used in the context of point processes, generates disperse point patterns whose points are all separated by a distance greater than b units. However, the threshold of separation b prevents the repulsion from being smooth (Daley and Vere-Jones 2002). Other examples of repulsive potentials can be found in Ogata and

Tanemura (1981, 1985). The key characteristic that differentiates the behavior of the potentials provided above is the action near 0; the faster the potential function goes to infinity as relative distance between coordinates goes to zero, the stronger the repulsion that the coordinates of $\boldsymbol{x}_{k,d}$ will experiment when they are separated by small distances. Even though Fúquene et al. (2016) do not employ a potential to model repulsion, the repulsion that results from their model is very similar to that found in Petralia et al. (2012) and tends to push coordinates far apart.

It is often the case that φ_1 and φ_2 are indexed by a set of finite-dimensional parameters which inform the types of patterns produced. It would therefore be natural to estimate these parameters using observed data. However, $\nu_{\mathrm{G}}(\mathbb{R}_k^d)$ is typically a function of the unknown parameters which makes deriving closed form expressions of $\nu_{\mathrm{G}}(\mathbb{R}_k^d)$ practically impossible and renders Bayesian or frequentist estimation procedures intractable. To avoid this complication, pseudo-maximum likelihood methods have been proposed to approximate $\nu_{\mathrm{G}}(\mathbb{R}_k^d)$ when carrying out estimation (Ogata and Tanemura 1981; Penttinen 1984). We provide details of a Bayesian approach in subsequent sections.

2.2 $\mathrm{Rep}_{k,d}(f_0, C_0, \rho)$ Distribution

As mentioned, our principal objective is to construct a family of probability densities for $\boldsymbol{x}_{k,d}$ that relaxes the *i.i.d.* assumption associated with its coordinates and we will do this by employing Gibbs measures that include an interaction function that mutually separates the k coordinates. Of all the potentials that might be considered in a Gibbs measure, we seek one that permits modeling repulsion flexibly so that a soft type of repulsion is available which avoids forcing large distances among the coordinates. As noted by Daley and Vere-Jones (2002) and Ogata and Tanemura (1981) the following potential

$$\phi(r) = -\log(1 - \exp\{-cr^2\}), \quad r \in \mathbb{R}_0^+ \text{ with } c \in \mathbb{R}^+ \tag{2.2}$$

produces smoother repulsion compared to other types of potentials in terms of "repelling strength" and for this reason we employ it as an example of interaction function in a Gibbs measure. A question that naturally arises at this point relates to the possibility of specifying a tractable class of repulsive distributions that incorporates the features discussed above. Note first that connecting (2.2) with ν_{G} is straightforward: if we take $\varphi_2(\boldsymbol{x}, \boldsymbol{y}) = -\ln(1 - C_0(\rho(\boldsymbol{x}, \boldsymbol{y})))$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ then ν_{G} will have a pairwise-interaction term given by

$$\exp\Big[-\sum_{r<s}^{k} \varphi_2(\boldsymbol{x}_r, \boldsymbol{x}_s)\Big] = \prod_{r<s}^{k} [1 - C_0(\rho(\boldsymbol{x}_r, \boldsymbol{x}_s))]. \tag{2.3}$$

The right-hand side of (2.3) induces a particular interaction structure that separates the coordinates of $\boldsymbol{x}_{k,d}$, thus introducing a notion of repulsion. The degree of separation is regulated by the speed at which C_0 decays to 0. The answer to the question posed earlier can then be given by focusing on functions $C_0 : \mathbb{R}_0^+ \to (0, 1]$ that satisfy the following properties:

A1. $C_0 \in C(\mathbb{R}_0^+; (0, 1])$.

A2. $C_0(0) = 1$.

A3. $\lim_{x\to+\infty} C_0(x) = 0$.

A4. $\forall\, x, y \in \mathbb{R}_0^+ : x < y \Rightarrow C_0(x) > C_0(y)$.

For future reference we will call A1 to A4 the $C_0 -$ Properties. The following Lemma guarantees that the type of repulsion induced by the $C_0 -$ Properties is smooth in terms of $\boldsymbol{x}_{k,d}$.

Lemma 2.1. *Given a metric $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+$ such that singletons are not open sets in the topology induced by ρ, the function $\mathrm{R_C} : \mathbb{R}_k^d \to [0, 1)$ defined by*

$$\mathrm{R_C}(\boldsymbol{x}_{k,d}) = \prod_{r<s}^{k} [1 - C_0(\rho(\boldsymbol{x}_r, \boldsymbol{x}_s))] \tag{2.4}$$

belongs to $C(\mathbb{R}_k^d; [0, 1))$ for $d \geq 1$ and $k \geq 2$.

Through out the article we will refer to (2.4) as the repulsion component. We finish the construction of repulsive probability measures by specifying a distribution on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ that is common for all the coordinates of $\boldsymbol{x}_{k,d}$. Let $f_0 \in C(\mathbb{R}^d; \mathbb{R}^+)$ be a probability density function, then under $\varphi_1(\boldsymbol{x}) = -\log(f_0(\boldsymbol{x}))$, ν_{G} will have a base component given by

$$\exp\Big[-\sum_{i=1}^{k} \varphi_1(\boldsymbol{x}_i)\Big] = \prod_{i=1}^{k} f_0(\boldsymbol{x}_i). \tag{2.5}$$

Incorporating (2.3) and (2.5) into (2.1) we get

$$\nu_{\mathrm{G}}\Big(\prod_{i=1}^{k} A_i\Big) = \int_{\prod_{i=1}^{k} A_i} \prod_{i=1}^{k} f_0(\boldsymbol{x}_i) \mathrm{R_C}(\boldsymbol{x}_{k,d}) \mathrm{d}\boldsymbol{x}_{k,d}.$$

The following Proposition ensures that the repulsive probability measures just constructed are well defined.

Proposition 2.2. *Let $f_0 \in C(\mathbb{R}^d; \mathbb{R}^+)$ be a probability density function. The function*

$$g_{k,d}(\boldsymbol{x}_{k,d}) = \prod_{i=1}^{k} f_0(\boldsymbol{x}_i) \mathrm{R}_{\mathrm{C}}(\boldsymbol{x}_{k,d}) \tag{2.6}$$

is measurable and integrable for $d \geq 1$ and $k \geq 2$.

With Proposition 2.2 it is now straightforward to construct a multivariate probability measure with the desired repulsive structure; small relative distances are penalized in a smooth way. Notice that the support of (2.6) is determined by the shape of the baseline distribution f_0 and then subsequently distorted (i.e. contracted) by the repulsive component. The normalized version of (2.6) defines a valid joint probability density function which we now provide.

Definition 2.1. The probability distribution $\mathrm{Rep}_{k,d}(f_0, C_0, \rho)$ has density function

$$\mathrm{Rep}_{k,d}(\boldsymbol{x}_{k,d}) = \frac{1}{c_{k,d}} \prod_{i=1}^{k} f_0(\boldsymbol{x}_i) \mathrm{R}_{\mathrm{C}}(\boldsymbol{x}_{k,d}), \tag{2.7}$$

$$\text{with} \quad c_{k,d} = \int_{\mathbb{R}_k^d} \prod_{i=1}^{k} f_0(\boldsymbol{x}_i) \mathrm{R}_{\mathrm{C}}(\boldsymbol{x}_{k,d}) \mathrm{d}\boldsymbol{x}_{k,d}. \tag{2.8}$$

Here $\boldsymbol{x}_{k,d} \in \mathbb{R}_k^d$, $f_0 \in C(\mathbb{R}^d; \mathbb{R}^+)$ is a probability density function, $C_0 : \mathbb{R}_0^+ \to (0, 1]$ is a function that satisfies the $C_0 -$ Properties and $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_0^+$ is a metric such that singletons are not open sets in the topology induced by it.

2.3 $\mathrm{Rep}_{k,d}(f_0, C_0, \rho)$ Properties

In this section we will investigate a few general properties of the $\mathrm{Rep}_{k,d}(f_0, C_0, \rho)$ class. The distributional results are provided to further understanding regarding characteristics of (2.7) from a qualitative and analytic point of view. As a first observation, because of symmetry, $\mathrm{Rep}_{k,d}(\boldsymbol{x}_{k,d})$ is an exchangeable distribution in $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. This facilitates the study of computational techniques motivated by $\mathrm{Rep}_{k,d}(f_0, C_0, \rho)$. To simplify notation, in what follows we will use $[m] = \{1, \ldots, m\}$, with $m \in \mathbb{N}$.

Before proceeding it is worth noting that $\{\mathrm{Rep}_{k,d}(f_0, C_0, \rho)\}_{k \geq 2}$ does not induce a sample-size consistent sequence of finite-dimensional distributions, meaning that

$$\int_{\mathbb{R}^d} \mathrm{Rep}_{k+1,d}(\boldsymbol{x}_{k+1,d}) \mathrm{d}\boldsymbol{x}_{k+1} \neq \mathrm{Rep}_{k,d}(\boldsymbol{x}_{k,d}).$$

This makes predicting locations of new coordinates problematic. In Section 3 we address how this may be accommodated in modeling contexts.

2.3.1 Normalizing Constant

Because $\mathrm{R_C}(\boldsymbol{x}_{k,d})$ is invariant under permutations of the coordinates of $\boldsymbol{x}_{k,d}$, an interaction's direction is immaterial to whether it is present or absent (i.e., $\boldsymbol{x}_r$ interacts with $\boldsymbol{x}_s$ if and only if $\boldsymbol{x}_s$ interacts with $\boldsymbol{x}_r$). Therefore it is sufficient to represent the interaction between $\boldsymbol{x}_r$ and $\boldsymbol{x}_s$ as $(r,s) \in I_k$ where $I_k = \{(r,s) : 1 \leq r < s \leq k\}$. In this setting, I_k reflects the set of all pairwise interactions between the k coordinates of $\boldsymbol{x}_{k,d}$ and $\ell_k \overset{\text{def}}{=} \mathrm{card}(I_k) = \frac{k(k-1)}{2}$. Now, expanding (2.4) term-by-term results in

$$\mathrm{R_C}(\boldsymbol{x}_{k,d}) = 1 + \sum_{l=1}^{\ell_k} (-1)^l \sum_{\substack{A \subseteq I_k \\ \mathrm{card}(A)=l}} \prod_{(r,s)\in A} C_0(\rho(\boldsymbol{x}_r, \boldsymbol{x}_s)). \tag{2.9}$$

The right-side of (2.9) is connected to graph theory in the following way: $A \subseteq I_k$ can be interpreted as a non-directed graph whose edges are $(r,s) \in A$.

Using (2.9), it can be shown that expressions (2.7) and (2.8) in Definition 2.1 have the following forms:

$$\mathrm{Rep}_{k,d}(\boldsymbol{x}_{k,d}) = \prod_{i=1}^{k} f_0(\boldsymbol{x}_i) + \sum_{l=1}^{\ell_k} (-1)^l \sum_{\substack{A \subseteq I_k \\ \mathrm{card}(A)=l}} \Big[\prod_{i=1}^{k} f_0(\boldsymbol{x}_i) \prod_{(r,s)\in A} C_0(\rho(\boldsymbol{x}_r, \boldsymbol{x}_s)) \Big] \tag{2.10}$$

$$c_{k,d} = 1 + \sum_{l=1}^{\ell_k} (-1)^l \sum_{\substack{A \subseteq I_k \\ \mathrm{card}(A)=l}} \Psi_{k,d}(A) \tag{2.11}$$

$$\Psi_{k,d}(A) = \int_{\mathbb{R}_k^d} \prod_{i=1}^{k} f_0(\boldsymbol{x}_i) \prod_{(r,s)\in A} C_0(\rho(\boldsymbol{x}_r, \boldsymbol{x}_s)) \mathrm{d}\boldsymbol{x}_{k,d}. \tag{2.12}$$

Note that representing A as a graph or Laplacian matrix can help develop intuition on how each summand contributes to the expression (2.11). Figure 2 shows one particular case of how three of four coordinates in $\mathbb{R}^d$ might interact by providing the respective Laplacian matrix together with the contribution that (2.12) brings to calculating $c_{4,d}$ according to (2.11).

Equation (2.11) retains connections with the probabilistic version of the Inclusion-Exclusion Principle.

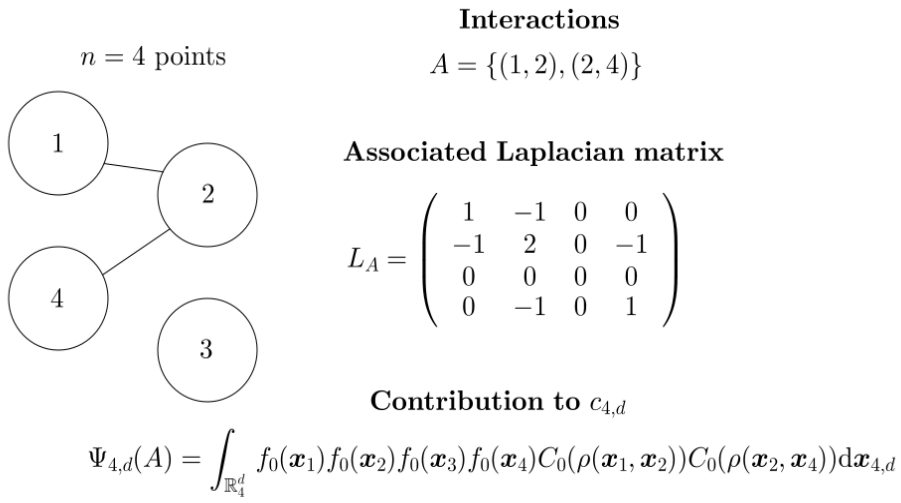

Figure 2: The graph and Laplacian matrix for a possible interaction for $n = 4$ coordinates.

This result, which is very useful in enumerative combinatorics, says that in any probability space $(\Omega, \mathcal{F}, \mathbb{P})$

$$\mathbb{P}\Big(\bigcap_{i=1}^{k} A_i^c\Big) = 1 + \sum_{l=1}^{k} (-1)^l \sum_{\substack{I \subseteq [k] \\ \mathrm{card}(I)=l}} \mathbb{P}\Big(\bigcap_{i \in I} A_i\Big),$$

with $A_1, \ldots, A_k \in \mathcal{F}$ and A_i^c denoting the complement of A_i. With this in mind, $c_{k,d}$ is the result of adding/substracting all the contributions $\Psi_{k,d}(A)$ that emerge for every $A \subseteq I_k$ $(A \neq \emptyset)$. If we think of $c_{k,d}$ as an indicator of the strength of repulsion, $\Psi_{k,d}(A)$ provides the specific contribution from the interactions $(r, s) \in A$. Moreover, it quantifies how distant a $\mathrm{Rep}_{k,d}(f_0, C_0, \rho)$ distribution is from the extreme case of $C_0 \equiv 0$ (i.e., the coordinates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are mutually independent and share a common probability law f_0).

The tractability of $c_{k,d}$ depends heavily on the number of coordinates k since the cost of evaluating (2.12) becomes prohibitive as it requires carrying out (at least) $2^{\ell_k} - 1$ numerical calculations. In Subsection 3.1 we highlight a particular choice of f_0, C_0 and ρ that produces a closed form expression for (2.12).

3 Gaussian Mixture Models and $\mathrm{Rep}_{k,d}(f_0, C_0, \rho)$ Distribution

In this section we will briefly introduce Gaussian Mixture Models, which are very popular in the context of density estimation (Escobar and West 1995) because of their flexibility and computational tractability. Then

we show that repulsion can be incorporated by modeling location parameters with the repulsion distribution described previously.

3.1 Repulsive Gaussian Mixture Models

Consider $n \in \mathbb{N}$ experimental units whose responses $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are d-dimensional and assumed to be exchangeable. Gaussian mixtures can be thought of as a way of grouping the n units into several clusters, say $k \in \mathbb{N}$, each having its own specific characteristics. In this context, the jth cluster ($j \in [k]$) is modeled through a Gaussian density $\mathrm{N}_d(\,\cdot\,; \boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j)$ with location $\boldsymbol{\theta}_j \in \mathbb{R}^d$ and scale $\boldsymbol{\Lambda}_j \in \mathbb{S}^d$. Here, $\mathbb{S}^d$ is the space of real, symmetric and positive-definite matrices of dimension $d \times d$. For notational purpose, let $\boldsymbol{\theta}_{k,d} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \in \mathbb{R}^d_k$ and $\boldsymbol{\Lambda}_{k,d} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_k) \in \mathbb{S}^d_k$ with $\mathbb{S}^d_k = \prod_{i=1}^k \mathbb{S}^d$. Next let $\boldsymbol{\pi}_k = (\pi_1, \ldots, \pi_k)^\top \in \Delta_{k-1}$, where Δ_{k-1} is the $(k-1)$-dimensional simplex ($\Delta_0 = \{1\}$) denote a set of weights that reflect the probability of allocating $\boldsymbol{y}_i : i \in [n]$ to a cluster. Then the standard Gaussian mixture model is

$$\boldsymbol{y}_i | \boldsymbol{\pi}_k, \boldsymbol{\theta}_{k,d}, \boldsymbol{\Lambda}_{k,d} \overset{iid}{\sim} \sum_{j=1}^{k} \pi_j \mathrm{N}_d(\boldsymbol{y}_i; \boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j). \tag{3.1}$$

It is common to restate (3.1) by introducing latent cluster membership indicators $z_1, \ldots, z_n \in [k]$ such that $\boldsymbol{y}_i$ is drawn from the jth mixture componnent if and only if $z_i = j$:

$$\boldsymbol{y}_i | z_i, \boldsymbol{\theta}_{k,d}, \boldsymbol{\Lambda}_{k,d} \overset{ind}{\sim} \mathrm{N}_d(\boldsymbol{y}_i; \boldsymbol{\theta}_{z_i}, \boldsymbol{\Lambda}_{z_i}) \tag{3.2}$$

$$z_i | \boldsymbol{\pi}_k \overset{iid}{\sim} \mathrm{Mult}(1, \boldsymbol{\pi}_k). \tag{3.3}$$

after marginalizing over the z_i indicators. The model is typically completed with conjugate-style priors for all parameters.

It is also possible to consider a prior distribution on $k \in \mathbb{N}$. For example, Dirichlet Process Mixture (DPM) models by construction induce a prior distribution on the number of clusters k. Alternatively, setting a prior specifically on k leads to transdimensional MCMC approaches, such as Reversible Jumps (Green 1995; Richardson and Green 1997) or Birth-Death chains (Stephens 2000). These methodologies allow the number of clusters to change by way of specific movements between parametric spaces indexed by k. In this paper we will use a case-specific upper bound for k.

In the above mixture model, the location parameters associated with each mixture component are typically assumed to be independent a priori. This is precisely the assumption that facilitates the presence of

redundant mixture components. In contrast, our work focuses on employing $\text{Rep}_{k,d}(f_0, C_0, \rho)$ as a model for location parameters in (3.1) which promotes reducing redundant mixture components without sacrificing goodness-of-fit, i.e, more parsimony relative to alternatives with independent locations. Moreover, the responses will be allocated to a few well-separated clusters. This desired behavior can be easily incorporated in the mixture model by assuming

$$\begin{aligned}
&\boldsymbol{\theta}_{k,d} \sim \text{Rep}_{k,d}(f_0, C_0, \rho) \\
&f_0(\boldsymbol{x}) = \text{N}_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{S}^d && (3.4) \\
&C_0(r) = \exp(-0.5\tau^{-1} r^2) : \tau \in \mathbb{R}^+ && (3.5) \\
&\rho(\boldsymbol{x}, \boldsymbol{y}) = \{(\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{y})\}^{1/2}. && (3.6)
\end{aligned}$$

The specific forms of f_0, C_0 and ρ are admissible according to Definition 2.1. The repulsive distribution parameterized by (3.4) - (3.6) will be denoted by $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$. Because $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ introduces dependence a priori (in particular, repulsion) between the coordinates of $\boldsymbol{\theta}_{k,d}$, they are no longer conditionally independent given $(\boldsymbol{y}_{n,d}, \boldsymbol{z}_n, \boldsymbol{\Lambda}_{k,d})$, with $\boldsymbol{y}_{n,d} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \in \mathbb{R}_n^d$ and $\boldsymbol{z}_n = (z_1, \ldots, z_n)^\top \in [k]^n$. The parameter τ in (3.5) controls the strength of repulsion associated with coordinates in $\boldsymbol{\theta}_{k,d}$ via (3.6): as τ approaches 0, the repulsion becomes weaker. The selection of (3.4) reduces to the usual *i.i.d.* multivariate normal assumption when $\tau = 0$.

To facilitate later reference we state the repulsive mixture model in its entirety:

$$\begin{aligned}
&\boldsymbol{y}_i | z_i, \boldsymbol{\theta}_{k,d}, \boldsymbol{\Lambda}_{k,d} \overset{ind}{\sim} \text{N}_d(\boldsymbol{y}_i; \boldsymbol{\theta}_{z_i}, \boldsymbol{\Lambda}_{z_i}) && (3.7) \\
&z_i | \boldsymbol{\pi}_k \overset{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi}_k) && (3.8)
\end{aligned}$$

together with the following mutually independent prior distributions:

$$\begin{aligned}
&\boldsymbol{\pi}_k \sim \text{Dir}(\boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathbb{R}_k^+ && (3.9) \\
&\boldsymbol{\theta}_{k,d} \sim \text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau) : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{S}^d, \tau \in \mathbb{R}^+ && (3.10) \\
&\boldsymbol{\Lambda}_j \overset{iid}{\sim} \text{IW}_d(\boldsymbol{\Psi}, \nu) : \boldsymbol{\Psi} \in \mathbb{S}^d, \nu \in \mathbb{R}^+. && (3.11)
\end{aligned}$$

In what follows we will refer to the model in (3.7) - (3.11) as the (Bayesian) Repulsive Gaussian Mixture Model (RGMM).

3.1.1 Parameter Calibration

Treating $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ of the RGMM as unknowns and assigning them hyperprior distributions is very costly computationally. Even more problematic, it can be shown using (2.11), (2.12) and the Gaussian integral that the normalizing constant of $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ is

$$c_{k,d} = 1 + \sum_{l=1}^{\ell_k} (-1)^l \sum_{\substack{A \subseteq I_k \\ \text{card}(A)=l}} \det(\mathbf{I}_k \otimes \mathbf{I}_d + \mathbf{L}_A \otimes \tau^{-1}\mathbf{I}_d)^{-1/2},$$

where $\mathbf{I}_k, \mathbf{I}_d$ are the $k \times k$ and $d \times d$ identity matrices respectively, $\mathbf{L}_A$ denotes the Laplacian matrix associated to the set of interactions $A \subseteq I_k$ (see Subsection 2.3.1) and $\otimes$ is the matrix Kronecker product making it a function of τ. Because of this, the induced full conditional distribution for τ turns out to be doubly-intractable (Murray et al. 2006) and standard MCMC algorithms do not apply. Finally, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ acts as a location/scale parameter: if $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\top$ is the corresponding Cholesky decomposition for $\boldsymbol{\Sigma}$, then

$$\boldsymbol{\theta}_{k,d} \sim \text{NRep}_{k,d}(\mathbf{0}_d, \mathbf{I}_d, \tau) \Rightarrow \mathbf{1}_k \otimes \boldsymbol{\mu} + (\mathbf{I}_k \otimes \mathbf{C})\boldsymbol{\theta}_{k,d} \sim \text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau),$$

where $\mathbf{0}_d, \mathbf{1}_d \in \mathbb{R}^d$ are d-dimensional vectors of zeroes and ones, respectively. Although a Normal hyperprior for $\boldsymbol{\mu}$ is a reasonable candidate (the full conditional distribution is also Normal), it is not straightforward how to select hyperparameters. To avoid these problems, we briefly outline a strategy that leads to a reasonable selection of hyperparameter values in the RGMM.

To facilitate hyperparameter selection we standardize the $\boldsymbol{y}_i$'s (a common practice in mixture models see, e.g. Gelman et al. 2014). Upon standardizing the response, it is reasonable to assume that $\boldsymbol{\mu} = \mathbf{0}_d$ and $\boldsymbol{\Sigma} = \mathbf{I}_d$. Further Gelman et al. (2014) argue that setting $\boldsymbol{\alpha} = (1/k)\mathbf{1}_d$ produces a weakly informative prior for $\boldsymbol{\pi}_k$. Selecting ν and $\boldsymbol{\Psi}$ is particularly important as they can dominate the repulsion effect. Setting $\nu = d + 4$ and $\boldsymbol{\Psi} = 3\psi\mathbf{I}_d$ with $\psi \in \mathbb{R}^+$ guarantees that each scale matrix $\boldsymbol{\Lambda}_j$ is centered on $\psi\mathbf{I}_d$ and that their entries possess finite variances. The value of ψ can be set to a value that accommodates the desired variability.

To calibrate τ, we follow the strategy outlined in Fúquene et al. (2016). Their approach consists of first specifying the probability that the coordinates of $\boldsymbol{\theta}_{k,d}$ are separated by a certain distance u and then set τ to the value that achieves the desired probability. To formalize this idea, suppose first that $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \overset{iid}{\sim} \text{N}_d(\mathbf{0}_d, \mathbf{I}_d)$. To favor separation among these random vectors we can use (3.5) and (3.6) with $\boldsymbol{\Sigma} = \mathbf{I}_d$ to

choose τ such that

$$\forall\, r \neq s \in [k], \quad \mathbb{P}\left(1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\} \leq u\right) = p,$$

for fixed values $u, p \in (0,1)$. Letting $w(u) = -\log(1-u)$ for $u \in (0,1)$, standard properties of the Normal distribution guarantee that the previous relation is equivalent to

$$\mathbb{P}(G \leq w(u)\tau) = p, \quad \text{with} \quad G = \frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s) \sim \mathrm{G}(d/2, 1/2). \tag{3.12}$$

Creating a grid of points in $\mathbb{R}^+$ it is straightforward to find a τ that fulfills criterion (3.12). This criterion allows the repulsion to be small (according to u), while at the same time preventing it with probability p from being too strong. This has the added effect of avoiding degeneracy of (3.10), thus making computation numerically more stable. In practice, we apply the procedure outlined above to the vectors coming from the repulsive distribution (3.10), treating them as if they were sampled from a multivariate normal distribution. This gives us a simple procedure to approximately achieve the desired goal of prior separation with a pre-specified probability.

3.2 Theoretical Properties

In this section we explore properties associated with the support and posterior consistency of (3.1) under (3.9), (3.10) and (3.11). These results are based on derivations found in Petralia et al. (2012). However, we highlight extensions and generalizations that we develop here. Consider for $k \in \mathbb{N}$ the family of probability densities $\mathcal{F}_k = \{f(\,\cdot\,;\boldsymbol{\xi}_k) : \boldsymbol{\xi}_k \in \boldsymbol{\Theta}_k\}$, where $\boldsymbol{\xi}_k = \boldsymbol{\pi}_k \times \boldsymbol{\theta}_k \times \{\lambda\}$, $\boldsymbol{\Theta}_k = \Delta_{k-1} \times \mathbb{R}^k \times \mathbb{R}^+$ and

$$f(\,\cdot\,;\boldsymbol{\xi}_k) = \sum_{j=1}^{k} \pi_j \mathrm{N}(\,\cdot\,;\theta_j, \lambda).$$

Let $B_p(\boldsymbol{x}, r)$ with $\boldsymbol{x} \in \mathbb{R}^k$ and $r \in \mathbb{R}^+$ denote an open ball centered on $\boldsymbol{x}$, and with radius r, and $\overline{B}_p(\boldsymbol{x}, r)$ its closure relative to the Euclidean L_p-metric ($p \geq 1$) on $\mathbb{R}^k$.

The following four conditions will be assumed to prove the results stated afterwards.

B1. The true data generating density $f_0(\,\cdot\,;\boldsymbol{\xi}^0_{k_0})$ belongs to $\mathcal{F}_{k_0}$ for some fixed $k_0 \geq 2$, where $\boldsymbol{\xi}^0_{k_0} = \boldsymbol{\pi}^0_{k_0} \times \boldsymbol{\theta}^0_{k_0} \times \{\lambda_0\} = (\pi^0_1, \ldots, \pi^0_{k_0})^\top \times (\theta^0_1, \ldots, \theta^0_{k_0})^\top \times \{\lambda_0\}$.

B2. The true locations $\theta^0_1, \ldots, \theta^0_{k_0}$ satisfy $\min\{|\theta^0_r - \theta^0_s| : r \neq s \in [k_0]\} \geq v$ for some $v > 0$.

B3. The number of components $k \in \mathbb{N}$ follows a discrete distribution κ on the measurable space $(\mathbb{N}, 2^{\mathbb{N}})$ such that $\kappa(\{k_0\}) > 0$.

B4. For $k \geq 2$ we have $\boldsymbol{\xi}_k \sim \text{Dir}((1/k)\mathbf{1}_k) \times \text{NRep}_{k,1}(\mu, \sigma^2, \tau) \times \text{IG}(a, b)$. In the case that $k = 1$, $\boldsymbol{\xi}_k \sim \delta_1 \times \text{N}(\mu, \sigma^2) \times \text{IG}(a, b)$ with δ_1 a Dirac measure centred on 1.

Condition B2 requires that the true locations are separated by a minimum (euclidian) distance v, which favors disperse mixture component centroids within the range of the response. For condition B4, the sequence $\{\boldsymbol{\xi}_k : k \in \mathbb{N}\}$ can be constructed (via the Kolmogorov's Extension Theorem) in a way that the elements are mutually independent upon adding to each $\boldsymbol{\Theta}_k$ an appropriate σ-algebra. This guarantees the existence of a prior distribution Π defined on $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$ which correspondingly connects the elements of $\mathcal{F}$ with $\boldsymbol{\xi} = \prod_{k=1}^{\infty} \boldsymbol{\xi}_k$. To calculate probabilities with respect to Π, the following stochastic representation will be useful

$$\boldsymbol{\xi} | K = k \sim \boldsymbol{\xi}_k, \qquad K \sim \kappa. \tag{3.13}$$

Our study of the support of Π employs the Kullback-Leibler (KL) divergence to measure the similarity between probability distributions. We will say that $f_0 \in \mathcal{F}_{k_0}$ belongs to the KL support with respect to Π if

$$\forall \varepsilon > 0 : \quad \Pi\left(\left\{ f \in \mathcal{F} : \int_{\mathbb{R}} \log\left[\frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_\star)}\right] f_0(x; \boldsymbol{\xi}_{k_0}^0) \mathrm{d}x < \varepsilon \right\}\right) > 0, \tag{3.14}$$

where $\boldsymbol{\xi}_\star \in \bigcup_{k=1}^{\infty} \boldsymbol{\Theta}_k$. Condition (3.14) can be understood as Π's ability to assign positive mass to arbitrarily small neighborhoods around the true density f_0. A fundamental step to proving that f_0 lies in the KL support of Π is based on the following Lemmas.

Lemma 3.1. *Under condition* B1*, let* $\varepsilon > 0$*. Then there exists* $\delta > 0$ *such that*

$$\forall\, \boldsymbol{\xi}_{k_0} \in B_1(\boldsymbol{\theta}_{k_0}^0, \delta) \times B_1(\boldsymbol{\pi}_{k_0}^0, \delta) \times (\lambda_0 - \delta, \lambda_0 + \delta) : \qquad \int_{\mathbb{R}} \log\Big[\frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_{k_0})}\Big] f_0(x; \boldsymbol{\xi}_{k_0}^0) \mathrm{d}x < \varepsilon.$$

Lemma 3.2. *Assume condition* B2 *and let* $\boldsymbol{\theta}_{k_0} \sim \text{NRep}_{k_0,1}(\mu, \sigma^2, \tau)$*. Then there exists* $\delta_0 > 0$ *such that*

$$\forall\, \delta \in (0, \delta_0] : \qquad \mathbb{P}\left(\boldsymbol{\theta}_{k_0} \in B_1(\boldsymbol{\theta}_{k_0}^0, \delta)\right) > 0.$$

This result remains valid even when replacing $B_1(\boldsymbol{\theta}_{k_0}^0, \delta)$ *with* $\overline{B}_1(\boldsymbol{\theta}_{k_0}^0, \delta)$*.*

Using these two Lemmas we are able to prove the following Proposition.

Proposition 3.3. *Assume that conditions* B1 *through* B4 *hold. Then* f_0 *belongs to the KL support of* Π.

We next study the rate of convergence of the posterior distribution corresponding to a particular prior distribution (under suitable regularity conditions). To do this, we will use arguments that are similar to those employed in Theorem 3.1 of Scricciolo (2011), to show that the posterior rates derived there are the same here when considering univariate Gaussian mixture models and cluster-location parameters that follow condition B4. First, we need the following two Lemmas.

Lemma 3.4. *For each* $k \geq 2$ *the coordinates of* $\boldsymbol{\theta}_k \sim \text{NRep}_{k,1}(\mu, \sigma^2, \tau)$ *share the same functional form. Moreover, there exists* $\gamma \in \mathbb{R}^+$ *such that*

$$\forall\, t \geq \gamma: \quad \mathbb{P}(|\theta_i| > t) \leq \frac{2}{(2\pi)^{1/2}} \frac{c_{k-1}}{c_k} \left(\frac{\sigma}{|\mu| + 1} \right) \exp\left\{ -\frac{t^2}{4\sigma^2} \right\},$$

with $c_k = c_{k,1}$ *the normalizing constant of* $\text{NRep}_{k,1}(\mu, \sigma^2, \tau)$ *and* $c_1 = 1$.

Lemma 3.5. *The sequence* $\{c_k : k \in \mathbb{N}\}$ *defined in Lemma 3.4 satisfies*

$$\forall\, k \geq 2: \quad 0 < \frac{c_{k-1}}{c_k} \leq A \exp\{ak\},$$

for some constants $a, A \in \mathbb{R}^+$.

Essentially, these results permit us to adapt certain arguments found in Scricciolo (2011) that are applicable when the location parameters of each mixture component are independent and follow a common distribution that is absolutely continuous with respect to the Lebesgue measure, whose support is $\mathbb{R}$ and with tails that decay exponentially. Using the previous Lemmas, we now state the following:

Proposition 3.6. *Assume that conditions* B1, B2 *and* B4 *hold. Replace condition* B3 *with:*

B3′. *There exists* $B \in \mathbb{R}^+$ *such that* $\forall\, k \in \mathbb{N}$, $0 < \kappa(\{k\}) \leq B \exp\{-bk\}$, *where* $b > a$ *and* $a \in \mathbb{R}^+$ *is given by Lemma 3.5.*

Then, the posterior rate of convergence relative to the Hellinger metric is $\varepsilon_n = n^{-1/2} \log(n)$.

3.3 Sampling From $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$

Upon introducing component labels, sampling marginally from the posterior distribution of $(\boldsymbol{\theta}_{k,d}, \boldsymbol{\Lambda}_{k,d}, \boldsymbol{\pi}_k, \boldsymbol{z})$ can be done with a very straightforward Gibbs sampler save for $\boldsymbol{\theta}_{k,d}$. In the Appendix we detail the entire

MCMC algorithm, but here we focus on the nonstandard aspects of the algorithm and describe how to sample from $\mathrm{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$. Notice that the full conditionals of each coordinate of $\boldsymbol{\theta}_{k,d}$ are not conjugate but they are all functionally similar. Because of this, evaluating these densities is computationally cheap making it straightforward to carry out sampling via a Metropolis-Hastings step inside the Gibbs sampling scheme.

To begin, the distribution $(\boldsymbol{\theta}_{k,d}|\cdots)$ is given by

$$\begin{aligned}(\boldsymbol{\theta}_{k,d}|\cdots) &\propto \prod_{i=1}^{n} \mathrm{N}_d(\boldsymbol{y}_i; \boldsymbol{\theta}_{z_i}, \boldsymbol{\Lambda}_{z_i}) \times \prod_{j=1}^{k} \mathrm{N}_d(\boldsymbol{\theta}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{r<s}^{k} [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\}] \\ &= \prod_{j=1}^{k} \mathrm{N}_d(\boldsymbol{\theta}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \prod_{r<s}^{k} [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\}]\end{aligned}$$

where $\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Lambda}_j^{-1}\boldsymbol{s}_j)$, $\boldsymbol{s}_j = \sum_{i=1}^{n} \mathbb{I}_{\{j\}}(z_i)\boldsymbol{y}_i$, $\boldsymbol{\Sigma}_j = (\boldsymbol{\Sigma}^{-1} + n_j\boldsymbol{\Lambda}_j^{-1})^{-1}$ and $n_j = \mathrm{card}(\{i \in [n] : z_i = j\})$. Now, the complete conditional distributions $(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j}, \cdots)$ for $j \in [k]$ and $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_l : l \neq j) \in \mathbb{R}^d_{k-1}$, have the following form

$$f(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j}, \cdots) \propto \mathrm{N}_d(\boldsymbol{\theta}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \prod_{l \neq j}^{k} [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_l)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_l)\}].$$

The following pseudo-code describes how to sample from $f(\boldsymbol{\theta}_{k,d}|\cdots)$ by way of $f(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j}, \cdots)$ via a random walk Metropolis Hastings step within a Gibbs sampler:

1. Let $\boldsymbol{\theta}_{k,d}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_k^{(0)}) \in \mathbb{R}^d_k$ be the actual state for $\boldsymbol{\theta}_{k,d}$.

2. For $j = 1, \ldots, k$:

   (a) Generate a candidate $\boldsymbol{\theta}_j^{(1)}$ from $\mathrm{N}_d(\boldsymbol{\theta}_j^{(0)}, \boldsymbol{\Gamma}_j) : \boldsymbol{\Gamma}_j \in \mathbb{S}^d$.

   (b) Set $\boldsymbol{\theta}_j^{(0)} = \boldsymbol{\theta}_j^{(1)}$ with probability

$$\min\left\{1, \frac{\mathrm{N}_d(\boldsymbol{\theta}_j^{(1)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\mathrm{N}_d(\boldsymbol{\theta}_j^{(0)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \prod_{l \neq j}^{k} \left[\frac{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(1)} - \boldsymbol{\theta}_l^{(0)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(1)} - \boldsymbol{\theta}_l^{(0)})\}}{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(0)} - \boldsymbol{\theta}_l^{(0)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(0)} - \boldsymbol{\theta}_l^{(0)})\}}\right]\right\}.$$

The selection of $\boldsymbol{\Gamma}_j$ can be carried out using adaptive MCMC methods (Roberts and Rosenthal 2009) so that the acceptance rate of the Metropolis-Hastings algorithm is approximately 50% within the burn-in

period for each $j \in [k]$. One approach that works well for the RGMM is to take

$$\boldsymbol{\Gamma}_j = \frac{1}{B}\sum_{t=1}^{B}(\boldsymbol{\Sigma}^{-1} + n_j^{(t)}(\boldsymbol{\Lambda}_j^{(t)})^{-1})^{-1} : n_j^{(t)} = \text{card}(\{i \in [n] : z_i^{(t)} = j\}), \tag{3.15}$$

where $t \in [B]$ is the tth iteration of the burn-in period with length $B \in \mathbb{N}$. As mentioned, in the Appendix we provide pseudo-code that details how the algorithm just described can be incorporated in a Gibbs sampler to generate realizations from joint posterior associated with the parameters found in RGMM. We will refer to the algorithm in the Appendix as Algorithm RGMM.

4 Simulation Study

To provide context regarding the proposed method's performance in density estimation, we conduct a small simulation study. In the simulation we compare density estimates from the RGMM to what is obtained using an *i.i.d.* mixture model and a DPM. This is done by treating the following as a data generating mechanism:

$$y \sim f_0(x) = 0.3\times \text{N}(x; -5, 1.0^2) + 0.05\times \text{N}(x; 0, 0.3^2) + 0.25\times \text{N}(x; 1, 0.3^2) + 0.4\times \text{N}(x; 4, 0.8^2), \quad x \in \mathbb{R}. \tag{4.1}$$

Using (4.1) we simulate 100 data sets with sample sizes 500, 1000 y 5000. For each of these scenarios, we compare the following 4 models (abbreviated by M1, M2, M3 y M4) to estimate f_0:

M1. Mixture of normals with prior distributions given by the regular mixture model that follows by assuming $\tau = 0$ in the RGMM, with

- $k = 10$, $d = 1$, $\boldsymbol{\alpha} = (1/10)\mathbf{1}_{10}$, $\boldsymbol{\mu} = 0$, $\boldsymbol{\Sigma} = 1$, $\boldsymbol{\Psi} = 0.06$ and $\nu = 5$.

We collected 10000 MCMC iterates after discarding the first 1000 as burn-in and thinning by 10.

M2. RGMM with $\tau = 5.45$. This value came from employing the calibration criterion from Section 3.1.1 and setting $u = 0.5$ and $p = 0.95$. The remaining prior parameters are:

- $k = 10$, $d = 1$, $\boldsymbol{\alpha} = (1/10)\mathbf{1}_{10}$, $\boldsymbol{\mu} = 0$, $\boldsymbol{\Sigma} = 1$, $\tau = 5.45$, $\boldsymbol{\Psi} = 0.06$ and $\nu = 5$.

We collected 10000 MCMC iterates after discarding the first 5000 as burn-in and thinning by 20.

M3. RGMM with $\tau = 17.17$. This value came from employing the calibration criterion from Section 3.1.1 and setting $u = 0.2$ and $p = 0.95$. Since τ is bigger here than in M2, M3 has more repulsion than M2. The remaining prior parameters are the same as in M2:

- $k = 10$, $d = 1$, $\boldsymbol{\alpha} = (1/10)\mathbf{1}_{10}$, $\boldsymbol{\mu} = 0$, $\boldsymbol{\Sigma} = 1$, $\tau = 17.17$, $\boldsymbol{\Psi} = 0.06$ and $\nu = 5$.

We collected 10000 MCMC iterates after discarding the first 5000 as burn-in and thinning by 20.

M4. DPM given by:

$$\boldsymbol{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \overset{ind}{\sim} \mathrm{N}_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{4.2}$$

$$(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)|H \overset{iid}{\sim} H \tag{4.3}$$

$$H|\alpha, H_0 \sim \mathrm{DP}(\alpha H_0) \tag{4.4}$$

where the baseline distribution H_0 is the conjugate Normal-Inverse Wishart

$$H_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{N}_d(\boldsymbol{\mu}; \boldsymbol{m}_1, (1/k_0)\boldsymbol{\Sigma}) \times \mathrm{IW}_d(\boldsymbol{\Sigma}; \nu_1, \boldsymbol{\Psi}_1) : \nu_1 \in \mathbb{R}^+. \tag{4.5}$$

To complete the model specification given by (4.2) through (4.5), the following independent hyperpriors are assumed:

$$\alpha|a_0, b_0 \sim \mathrm{G}(a_0, b_0) \tag{4.6}$$

$$\boldsymbol{m}_1|\boldsymbol{m}_2, \mathbf{S}_2 \sim \mathrm{N}_d(\boldsymbol{m}_2, \mathbf{S}_2) \tag{4.7}$$

$$k_0|\tau_1, \tau_2 \sim \mathrm{G}(\tau_1/2, \tau_2/2) \tag{4.8}$$

$$\boldsymbol{\Psi}_1|\nu_2, \boldsymbol{\Psi}_2 \sim \mathrm{IW}_d(\nu_2, \boldsymbol{\Psi}_2) \tag{4.9}$$

In the simulation study we set $d = 1$. The selection of hyperparameters found in (4.6) - (4.9) was based on similar strategies as outlined in Escobar and West (1995) which produced:

- $a_0 = 2$, $b_0 = 5$, $\nu_1 = 4$, $\nu_2 = 4$, $\boldsymbol{m}_2 = 0$, $\mathbf{S}_2 = 1$, $\boldsymbol{\Psi}_2 = 1$, $\tau_1 = 2.01$ and $\tau_2 = 1.01$.

We collected 10000 MCMC iterates after discarding the first 1000 as burn-in and thinning by 10.

Models M2 and M3 were fit using the Algorithm RGMM which was implemented in `Fortran` (see the Appendix). For model M4, density estimates were obtained using the function `DPdensity` which is available in the `DPpackage` of `R` (Jara et al. 2011).

To compare density estimation associated with the four procedures just detailed we employ the following metrics.

- Log Pseudo Marginal Likelihood (LPML) (Christensen et al. 2011) which is a model fit metric that takes into account model complexity. This was computed by first estimating all the corresponding conditional predictive ordinates (Gelfand et al. 1992) using the method in Chen et al. (2000).
- Mean Square Error (MSE).
- L_1-metric between the estimated posterior predictive density and f_0.

Additionally, to explore how the repulsion influences model parsimony in terms of the number of "active" mixture components, we include the following numeric indicators:

- Posterior average number of occupied components, i.e. $n_j = \text{card}(\{i \in [n] : z_i = j\}) > 0$.
- Standard deviation of the posterior average number of occupied components.

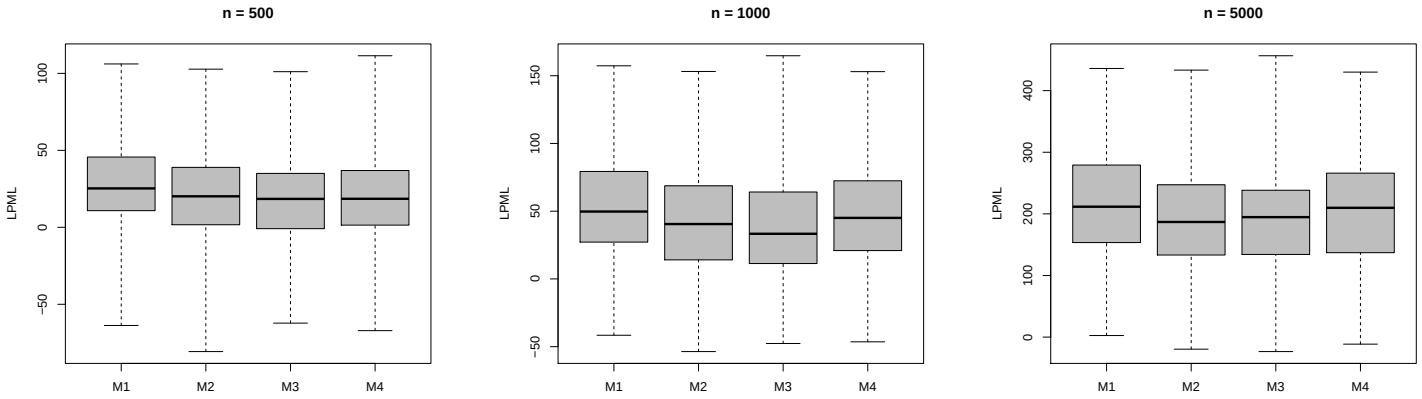

Figure 3: Boxplots that resume the behavior of LPML for each model.

Figures 3, 4 and 5 contain side-by-side boxplots of the LPML, MSE and L_1-metric respectively. Notice that for LPML and L_1-metric the two repulsion models (M2,M3) fit practically as well as the *i.i.d.* fit (M1) and the DPM fit (M4). In other words, very little model fit was sacrificed for the sake of parsimony. Regarding the MSE, M1 did the best (which is to be expected since it is the data generating mechanism) and M4 was next. That said, differences between the procedures tend to get small as the sample size grows. The overall take home message is that M2 and M3 are very competitive to M1 and M4 in terms of model fit.

Figures 6 and 7 show side-by-side boxplots of average number of occupied components and standard deviations associated with the number of occupied mixture components. Notice that the average number of

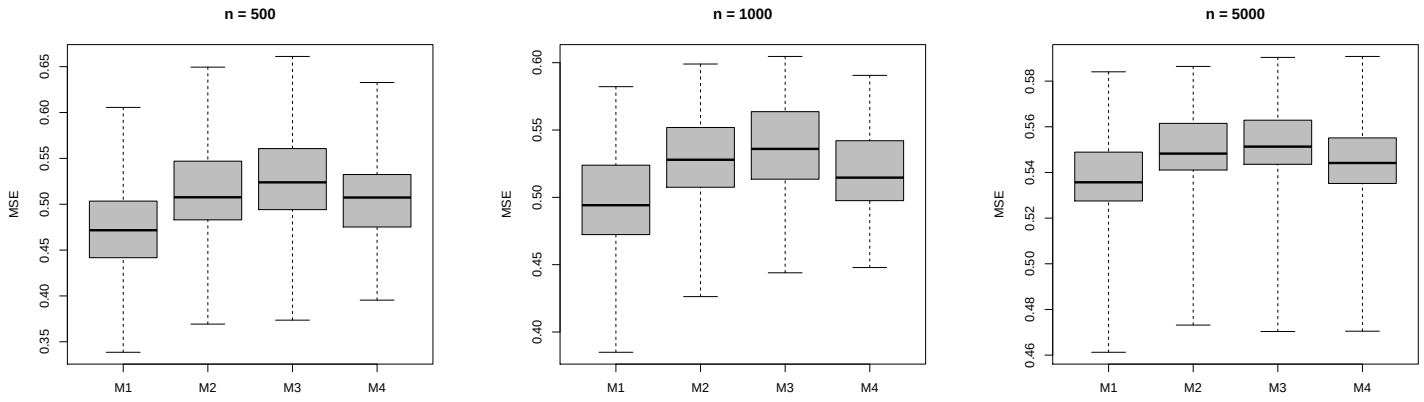

Figure 4: Boxplots that resume the behavior of MSE for each model.

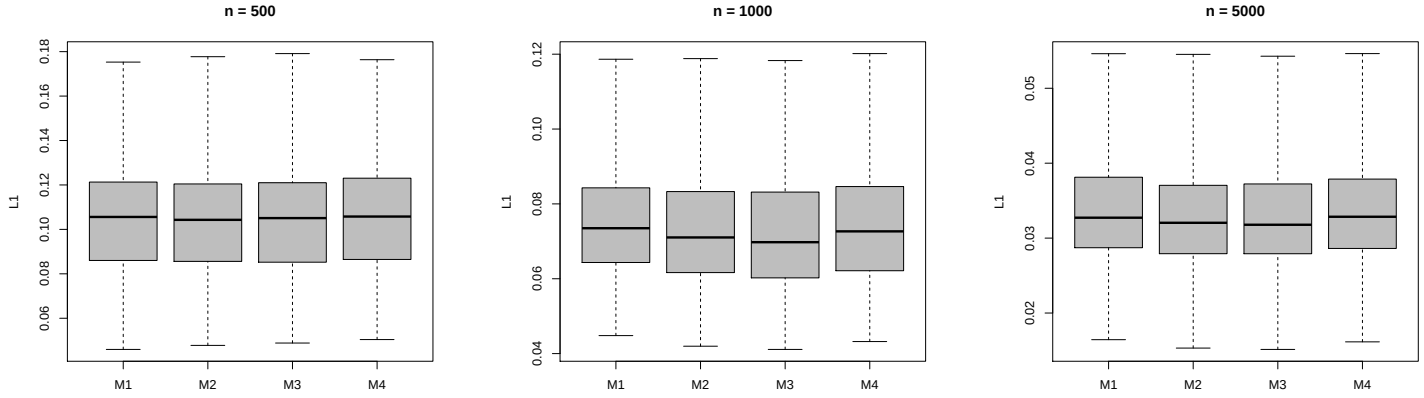

Figure 5: Boxplots that resume the behavior of L_1-metric for each model.

active components is much smaller for M2 and M3 relative to M1 and M4. This pattern persists (possibly becomes more obvious) as the number of observations grows. The number of active components for M2 and M3 are also highly concentrated around $\{3, 4, 5\}$ (recall that the data were generated using a mixture of four components). Conversely, M1 and M4 require many more active components to achieve the same goodness-of-fit, a trend that persists when the sample size grows.

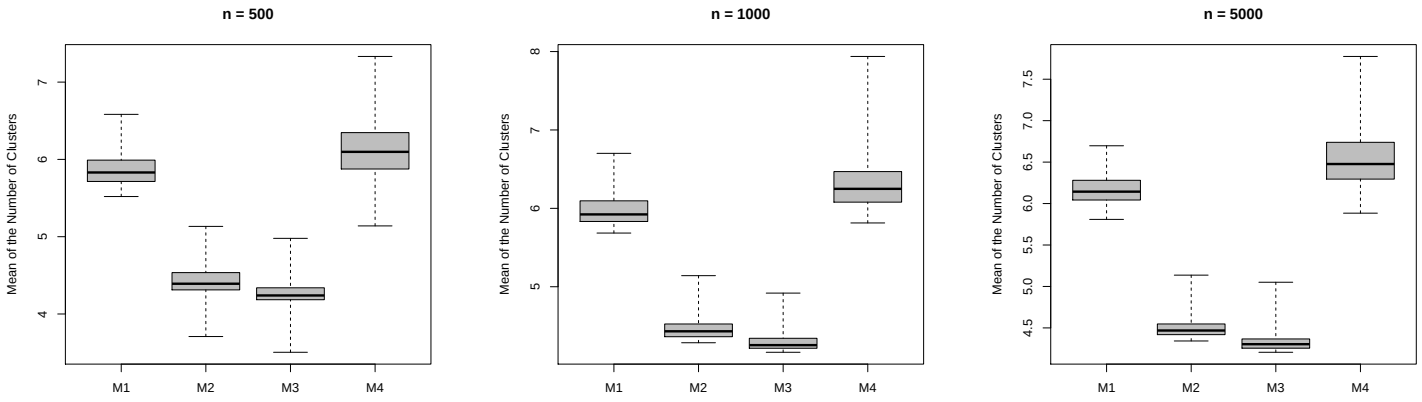

Figure 6: Side-by-side boxplots of the average number of occupied mixture components for each procedure.

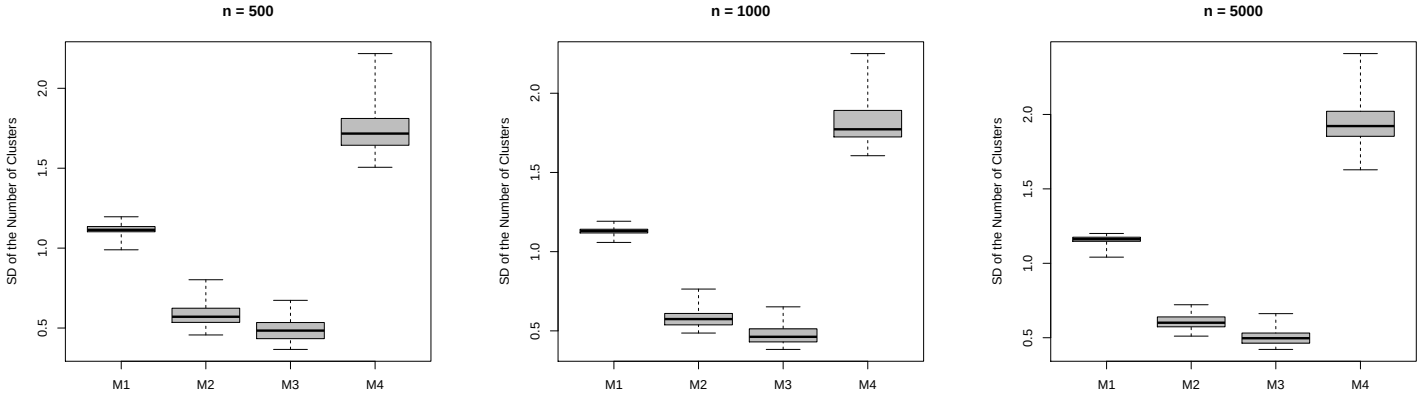

Figure 7: Side-by-side boxplots that display the average standard deviation associated with the distribution of occupied mixture components for each of the four procedures.

5 Data Illustrations

We now turn our attention to two well known data sets. The first is the *Galaxy* data set (Roeder 1990), and the second is bivariate *Air Quality* (Chambers 1983). Both are publicly available in `R`. For the second data set we removed 42 observations that were incomplete. We compare density estimates available from the DPM of Normal distributions (DPM-N) to those from the RGMM. For each procedure we report the LPML as a measure of goodness-of-fit, a brief summary regarding the average number of occupied components, and posterior distribution associated with the number of clusters. It is worth noting that both data sets were standardized prior to model fit. We now provide more details on the two model specifications.

1. DPM-N: We employed the `R` function `DPdensity` available in `DPpackage` (Jara et al. 2011). Decisions on hyperprior parameter values for both data sets were again guided by Escobar and West (1995). In both cases the model is specified by (4.2) - (4.9). We collected 10000 MCMC iterates after discarding the first 1000 (5000) as burn-in for Galaxy (Air Quality) data and thinning by 10. Specific details associated with prior parameter values are now provided:

   (a) Galaxy: $d = 1$, $a_0 = 2$, $b_0 = 2$, $\nu_1 = 4$, $\nu_2 = 4$, $\boldsymbol{m}_2 = 0$, $\mathbf{S}_2 = 1$, $\boldsymbol{\Psi}_2 = 0.15$, $\tau_1 = 2.01$ and $\tau_2 = 1.01$.

   (b) Air Quality: $d = 2$, $a_0 = 1$, $b_0 = 3$, $\nu_1 = 4$, $\nu_2 = 4$, $\boldsymbol{m}_2 = \mathbf{0}_2$, $\mathbf{S}_2 = \mathbf{I}_2$, $\boldsymbol{\Psi}_2 = \mathbf{I}_2$, $\tau_1 = 2.01$ and $\tau_2 = 1.01$.

2. RGMM: We coded Algorithm RGMM in `Fortran` to generate posterior draws for this model. For both data sets, we collected 10000 MCMC iterates after discarding the first 5000 as burn-in and thinning by 50. The values of τ were selected using the procedure outlined in Subsection 3.1.1: $(u, p) = (0.5, 0.95)$ and $(u, p) = (0.05, 0.95)$ for Galaxy and Air Quality data respectively. Parameter selection for model components (3.9), (3.10) and (3.11) were carried out according to the methods in Subsection 3.1.1. Specific details now follow:

   (a) Galaxy: $k = 10$, $d = 1$, $\boldsymbol{\alpha} = (1/10)\mathbf{1}_{10}$, $\boldsymbol{\mu} = 0$, $\boldsymbol{\Sigma} = 1$, $\tau = 5.45$, $\boldsymbol{\Psi} = 0.15$ and $\nu = 5$.

   (b) Air Quality: $k = 10$, $d = 2$, $\boldsymbol{\alpha} = (1/10)\mathbf{1}_{10}$, $\boldsymbol{\mu} = \mathbf{0}_2$, $\boldsymbol{\Sigma} = \mathbf{I}_2$, $\tau = 116.76$, $\boldsymbol{\Psi} = 3\mathbf{I}_2$ and $\nu = 6$.

Results of the fits are provided in Table 1. Notice that the fit associated with RGMM is better relative to the DPM-N, which corroborates the argument that RGMM sacrifices no appreciable model fit for the sake of model parsimony. Figure 8 further reinforces the idea that RGMM is more parsimonious relative to DPM-N. This can be seen as the posterior distribution of the number of clusters (or non-empty components) for RGMM concentrates on values that are smaller relative to the DPM. Graphs of the estimated densities (provided in Figure 9) show that the cost of parsimony is negligible as density estimates are practically the same.

6 Discussion and Future Work

We have created a class of probability models that explicitly parametrizes repulsion in a smooth way. In addition to providing pertinent theoretical properties, we demonstrated how this class of repulsive distributions can be employed to make hierarchical mixture models more parsimonious. A compelling result is

Data	LPML	Mean (Clusters)	SD (Clusters)
Galaxy (DPM-N)	-48.16	8.38	2.64
Galaxy (RGMM)	-36.68	5.37	0.91
Air Quality (DPM-N)	-274.82	2.83	1.11
Air Quality (RGMM)	-274.58	2.30	0.51

Table 1: Summary statistics related to model fit and the number of clusters for *Galaxy* and *Air Quality* data based on DPM-N/RGMM.

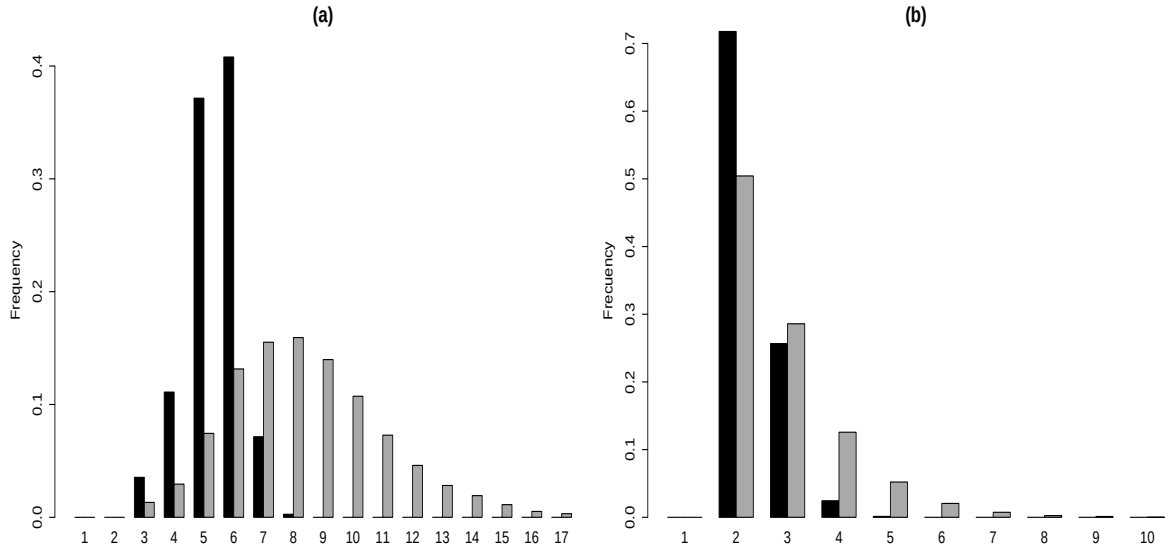

Figure 8: Posterior distribution for the active number of clusters in (a) *Galaxy* and (b) *Air Quality*. Black (gray) bars correspond to RGMM (DPM-N).

that this added parsimony comes at essentially no goodness-of-fit cost. We studied properties of the models, adapting the theory developed in Petralia et al. (2012) to accommodate the potential function we considered. Moreover, we generalized the results to include not only Gaussian mixtures of location but of also of scale (though the scale is constrained to be equal in each mixture component).

Our approach shares the same modeling spirit (presence of repulsion) as in Petralia et al. (2012), Xu et al. (2016) and Fúquene et al. (2016). However, the specific mechanism we propose to model repulsion differs from these works. Petralia et al. (2012) employ a potential (based on Lennard-Jones type potential) that introduces a stronger repulsion than our case, in the sense that in their model, locations are encouraged to be further apart. Xu et al. (2016) is based on Determinantal Point Processes, which introduces repulsion through the determinant of a matrix driven by a Gaussian covariance kernel. By nature of the point process, this approach allows a random number of mixture components (similar to DPM models) something

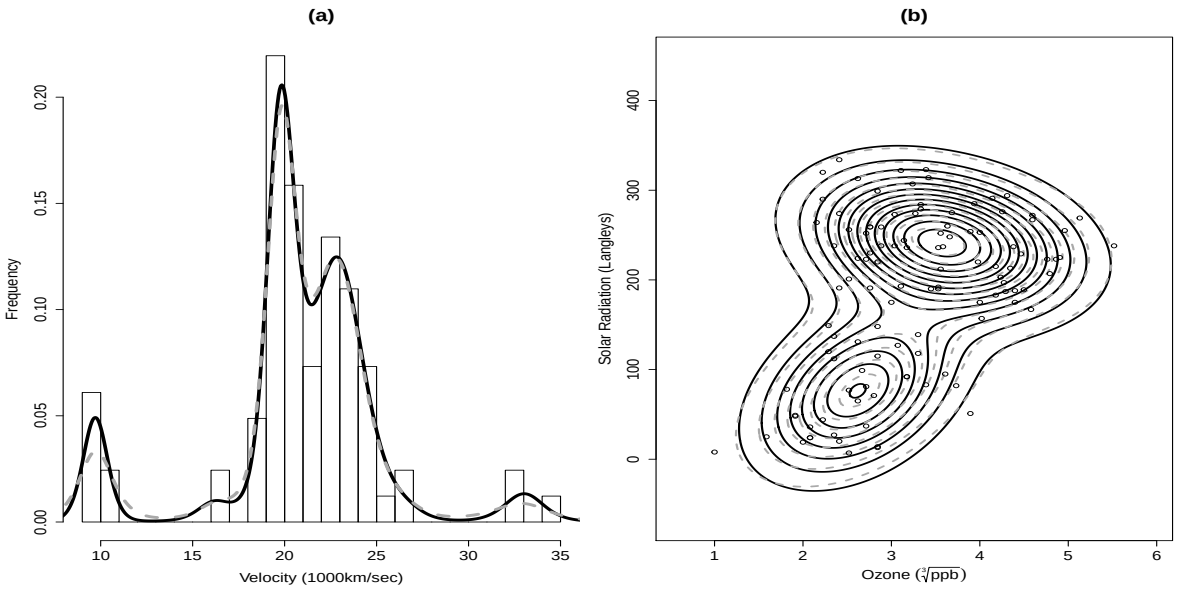

Figure 9: Posterior predictive densities for (a) *Galaxy* and (b) *Air Quality* data. Solid black (dotted gray) curves correspond to RGMM (DPM-N).

that our approach lacks. However, our approach allows a direct modeling of the repulsion that is easier to conceptualize. Finally, the work by Fúquene et al. (2016) defines a family of probability densities that promotes well-separated location parameters through a penalization function, that cannot be re-expressed as a (pure) repulsive potential. However, for small relative distances, the penalization function can be identified as an interaction potential that produces repulsion similar to that found in Petralia et al. (2012).

Presently we are pursuing a few directions of continued research. First, Proposition 3.3 was established for dimension $d = 1$. Extending results to the general d dimensional case would be a natural progression. In terms of modeling, we are exploring the possibility of relaxing the assumption that all mixture components in the Gaussian mixture models have the same variance. In terms of theory, we are exploring the possibility of relaxing the assumption of constant variance and adapting theoretical results to a larger class of potential functions. Rousseau and Mengersen (2011) developed some very interesting results that explore statistical properties associated with mixtures when k is chosen to be conservatively large (overfitted mixtures) with decaying weights associated with these extra mixture components. They did so using a framework that is an alternative to what we developed here. Under some restrictions on the prior and regularity conditions for the mixture component densities, the asymptotic behavior of the posterior distribution on the weights tends to empty the extra mixture components. We are currently exploring connections between these two

approaches.

Acknowledgments: We would like to thank Gregorio Moreno and Duvan Henao for helpful conversations and comments. José Quinlan gratefully recognizes the financial support provided by CONICYT through Doctoral Scholarship Grant 21120153 and Fondecyt Grant 11121131. Fernando A. Quintana was supported by Fondecyt Grant 1141057 and Garritt L. Page was partially supported by Fondecyt Grant 11121131.

References

Chambers, J. (1983), *Graphical methods for data analysis*, Chapman & Hall statistics series, Wadsworth International Group.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000), *Monte Carlo Methods in Bayesian Computation*, Springer New York.

Christensen, R., Johnson, W., Branscum, A. J., and Hanson, T. (2011), *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, CRC Press.

Dahl, D. B. (2006), "Model-Based Clustering for Expression Data via a Dirichlet Pro cess Mixture Model," in *Bayesian Inference for Gene Expression and Proteomics*, eds. Vannucci, M., Do, K. A., and Müller, P., Cambridge University Press, pp. 201–218.

Daley, D. and Vere-Jones, D. (2002), *An Introduction to the Theory of Point Processes*, vol. I: Elementary Theory and Methods, New York: Springer-Verlag, 2nd ed.

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

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Frühwirth-Schnatter, S. (2006), *Finite mixture and Markov switching models*, Springer Series in Statistics, Springer, New York.

Fúquene, J., Steel, M., and Rossell, D. (2016), "On choosing mixture components via non-local priors," .

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), "Model determination using predictive distributions with implementation via sampling-based methods," Tech. rep., DTIC Document.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014), *Bayesian Data Analysis*, London: Chapman and Hall/CRC, 3rd ed.

Georgii, H.-O. and Yoo, H. J. (2005), "Conditional intensity and Gibbsianness of Determinantal Point Processes," *Journal of Statistical Physics*, 118, 55–84.

Ghosal, S. and van der Vaart, A. (2007), "Posterior convergence rates of Dirichlet mixtures at smooth densities," *The Annals of Statistics*, 35, 697–723.

Ghosal, S. and van der Vaart, A. W. (2001), "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities," *Ann. Statist.*, 29, 1233–1263.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008), *Statistical analysis and modelling of spatial point patterns*, Statistics in Practice, John Wiley & Sons, Ltd., Chichester.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011), "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *Journal of Statistical Software*, 40, 1–30.

Jones, J. E. (1924), "On the Determination of Molecular Fields. II. From the Equation of State of a Gas," *Proceedings of the Royal Society of London Series A*, 106, 463–477.

Lavancier, F., Møller, J., and Rubak, E. (2015), "Determinantal point processes models and statistical inference," *Journal of the Royal Statistical Society: Series B*, 77, 853–877.

Mateu, J. and Montes, F. (2000), "Approximate maximum likelihood estimation for a spatial point pattern," *Qüestiió*, 24, 3–25.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006), "MCMC for doubly-intractable distributions," in *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, AUAI Press, pp. 359–366.

Ogata, Y. and Tanemura, M. (1981), "Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure," *Annals of the Institute of Statistical Mathematics*, 33, 315–338.

— (1985), "Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method," *Biometrics*, 41, 421–433.

Papangelou, F. (1974), "The conditional intensity of general point processes and an application to line processes," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 28, 207–226.

Pathria, R. and Beale, P. D. (2011), "10 - Statistical Mechanics of Interacting Systems: The Method of Cluster Expansions," in *Statistical Mechanics (Third Edition)*, eds. Pathria, R. and Beale, P. D., Boston: Academic Press, third edition ed., pp. 299 – 343.

Penttinen, A. (1984), *Modelling interactions in spatial point patterns: parameter estimation by the maximum likelihood method*, vol. 7, Jyväskylän yliopisto.

Petralia, F., Rao, V., and Dunson, D. B. (2012), "Repulsive Mixtures," in *Advances in Neural Information Processing Systems 25*, eds. Pereira, F., Burges, C., Bottou, L., and Weinberger, K., Curran Associates, Inc., pp. 1889–1897.

Pitman, J. (1996), "Some developments of the Blackwell-MacQueen urn scheme," in *Statistics, probability and game theory*, Inst. Math. Statist., Hayward, CA, vol. 30 of *IMS Lecture Notes Monogr. Ser.*, pp. 245–267.

Pitman, J. and Yor, M. (1997), "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *The Annals of Probability*, 25, 855–900.

Quintana, F. A. (2006), "A predictive view of Bayesian clustering," *Journal of Statistical Planning and Inference*, 136, 2407–2429.

Rao, V., Adams, R. P., and Dunson, D. D. (2016), "Bayesian inference for Matérn repulsive processes," *Journal of the Royal Statistical Society: Series B*, n/a–n/a.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society: Series B*, 859, 731–792.

Roberts, G. O. and Rosenthal, J. S. (2009), "Examples of Adaptive MCMC," *Journal of Computational and Graphical Statistics*, 18, 349–367.

Roeder, K. (1990), "Density Estimation with Confidence Sets Exemplified by Superclusters and Voids in the Galaxies," *Journal of the American Statistical Association*, 85, 617–624.

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

Scricciolo, C. (2011), "Posterior rates of convergence for Dirichlet mixtures of exponential power densities," *Electronic Journal of Statistics*, 5, 270–308.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013), "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures," *Biometrika*, 100, 623–640.

Stephens, M. (2000), "Bayesian Analysis of Mixture Models with an Unknown Number of Components An Alternative to Reversible Jump Methods," *The Annals of Statistics*, 28, 40–74.

Strauss, D. J. (1975), "A model for clustering," *Biometrika*, 62, 467–475.

Xu, Y., Müller, P., and Telesca, D. (2016), "Bayesian Inference for Latent Biological Structure with Determinantal Point Processes (DPP)," *Biometrics*, 72, 955–964.

A Algorithm RGMM

In what follows we describe the Gibbs sampler for the RGMM model. Let $B, S, T \in \mathbb{N}$ be the total number of iterations during the burn-in, the number of collected iterates, and the thinning, respectively.

- (Start) Choose initial values $z_i^{(0)} : i \in [n]$, $\boldsymbol{\pi}_k^{(0)}$ and $\boldsymbol{\theta}_j^{(0)}, \boldsymbol{\Lambda}_j^{(0)} : j \in [k]$. Set $\boldsymbol{\Gamma}_j = \mathbf{O}_d : j \in [k]$, where $\mathbf{O}_d$ is the null matrix of dimension $d \times d$.
- (Burn-in phase) For $t = 0, \ldots, B-1$:
  1. $(z_i^{(t+1)} | \cdots) \overset{ind}{\sim} \text{Mult}(1, \boldsymbol{\pi}_{i,k}^{(t)}) : i \in [n]$, where $\boldsymbol{\pi}_{i,k}^{(t)} = (\pi_{i,1}^{(t)}, \ldots, \pi_{i,k}^{(t)})^\top \in \Delta_{k-1}$ is given by
  $$\pi_{i,j}^{(t)} = \Big[\sum_{l=1}^{k} \pi_l^{(t)} \mathrm{N}_d(\boldsymbol{y}_i; \boldsymbol{\theta}_l^{(t)}, \boldsymbol{\Lambda}_l^{(t)})\Big]^{-1} \pi_j^{(t)} \mathrm{N}_d(\boldsymbol{y}_i; \boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Lambda}_j^{(t)}) : j \in [k].$$
  2. $(\boldsymbol{\pi}_k^{(t+1)} | \cdots) \sim \text{Dir}(\boldsymbol{\alpha}^{(t)})$, where
  $$\boldsymbol{\alpha}^{(t)} = (\alpha_1 + n_1^{(t+1)}, \ldots, \alpha_k + n_k^{(t+1)})^\top \in \mathbb{R}_k^+ : n_j^{(t+1)} = \text{card}(\{i \in [n] : z_i^{(t+1)} = j\}).$$
  3. For $j = 1, \ldots, k$:
     3.1. Generate a candidate $\boldsymbol{\theta}_j^{(\star)}$ from $\mathrm{N}_d(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Omega}_j^{(t)})$, where
     $$\boldsymbol{\Omega}_j^{(t)} = (\boldsymbol{\Sigma}^{-1} + n_j^{(t+1)} (\boldsymbol{\Lambda}_j^{(t)})^{-1})^{-1}.$$
     3.2. Update $\boldsymbol{\theta}_j^{(t)} \to \boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(\star)}$ with probability
     $$\min\left\{1, \frac{\mathrm{N}_d(\boldsymbol{\theta}_j^{(\star)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\mathrm{N}_d(\boldsymbol{\theta}_j^{(t)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})} \prod_{l \neq j}^{k} \left[\frac{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(\star)} - \boldsymbol{\theta}_l^{(t)})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_j^{(\star)} - \boldsymbol{\theta}_l^{(t)})\}}{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_l^{(t)})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_l^{(t)})\}}\right]\right\}$$
     where $\boldsymbol{\Sigma}_j^{(t)} = (\boldsymbol{\Sigma}^{-1} + n_j^{(t+1)} (\boldsymbol{\Lambda}_j^{(t)})^{-1})^{-1}$ and
     $$\boldsymbol{\mu}_j^{(t)} = \boldsymbol{\Sigma}_j^{(t)} (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + (\boldsymbol{\Lambda}_j^{(t)})^{-1} \boldsymbol{s}_j^{(t)}) : \boldsymbol{s}_j^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{\{j\}}(z_i^{(t+1)}) \boldsymbol{y}_i$$
     Otherwise, set $\boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(t)}$.
     3.3. Update $\boldsymbol{\Gamma}_j \to \boldsymbol{\Gamma}_j + \frac{1}{B} \boldsymbol{\Omega}_j^{(t)}$.
  4. $(\boldsymbol{\Lambda}_j^{(t+1)} | \cdots) \overset{ind}{\sim} \mathrm{IW}_d(\boldsymbol{\Psi}_j^{(t)}, \nu_j^{(t)}) : j \in [k]$, where $\nu_j^{(t)} = \nu + n_j^{(t+1)}$ and
  $$\boldsymbol{\Psi}_j^{(t)} = \boldsymbol{\Psi} + \sum_{i=1}^{n} \mathbb{I}_{\{j\}}(z_i^{(t+1)}) (\boldsymbol{y}_i - \boldsymbol{\theta}_j^{(t+1)}) (\boldsymbol{y}_i - \boldsymbol{\theta}_j^{(t+1)})^\top.$$
- (Save samples) For $t = B, \ldots, ST + B - 1$: Repeat steps 1, 2 and 4 of the burn-in phase. As for step 3 ignore 3.3, maintain 3.2 and replace 3.1 with

  3.1. Generate a candidate $\boldsymbol{\theta}_j^{(\star)}$ from $\mathrm{N}_d(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Gamma}_j)$.

  Finally, save the generated samples every Tth iteration.

B Proof of Lemma 2.1.

Assign to $\mathbb{R}_k^d$ and $[0,1)$ the metrics $d_1(\boldsymbol{x}_{k,d}, \boldsymbol{y}_{k,d}) = \max\{\rho(\boldsymbol{x}_i, \boldsymbol{y}_i) : i \in [k]\}$ and $d_2(x,y) = |x-y|$, respectively. Continuity of $\mathrm{R_C} : \mathbb{R}_k^d \to [0,1)$ follows from condition A1 of C_0-Properties and the following inequality:

$$|\rho(\boldsymbol{x}_r, \boldsymbol{x}_s) - \rho(\boldsymbol{y}_r, \boldsymbol{y}_s)| < 2d_1(\boldsymbol{x}_{k,d}, \boldsymbol{y}_{k,d}).$$

C Proof of Proposition 2.2.

Notice that $g_{k,d}$ is a $C(\mathbb{R}_k^d; \mathbb{R}^+)$-class function by construction (see Lemma 2.1). Because of the continuity, measurability follows. Using conditions A1 to A4 of C_0-Properties it follows that $\forall\, x \in \mathbb{R}_0^+ : 1 - C_0(x) \in [0,1)$. By Tonelli's Theorem

$$\int_{\mathbb{R}_k^d} g_{k,d}(\boldsymbol{x}_{k,d})\mathrm{d}\boldsymbol{x}_{k,d} \leq \Big(\int_{\mathbb{R}^d} f_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big)^k = 1.$$

The upper bound only proves that $g_{k,d}$ is integrable. However, this does not guarantee that $g_{k,d}$ is well defined, i.e. $\lambda_d^k(\{g_{k,d} > 0\}) = 0$. For this, it is sufficient to show that

$$\int_{\mathbb{R}_k^d} g_{k,d}(\boldsymbol{x}_{k,d})\mathrm{d}\boldsymbol{x}_{k,d} > 0$$

because $\forall\, \boldsymbol{x}_{k,d} \in \mathbb{R}_k^d : g_{k,d}(\boldsymbol{x}_{k,d}) \geq 0$ by construction. To prove the above inequality, fix $\boldsymbol{x}_{k,d}^0 \in \mathbb{R}_k^d$ such that $\boldsymbol{x}_r^0 \neq \boldsymbol{x}_s^0$ for $r \neq s \in [k]$. Then $g_{k,d}(\boldsymbol{x}_{k,d}^0) > 0$. Because $g_{k,d}$ is a continuous function on $\mathbb{R}_k^d$, there exists $r_0 \in \mathbb{R}^+$ such that $\forall\, \boldsymbol{x}_{k,d} \in B(\boldsymbol{x}_{k,d}^0, r_0) : g_{k,d}(\boldsymbol{x}_{k,d}) > 0$, where $B(\boldsymbol{x}_{k,d}^0, r_0) = \prod_{i=1}^k B_2(\boldsymbol{x}_i^0, r_0)$. Further, $B(\boldsymbol{x}_{k,d}^0, r_0) \in \mathcal{B}(\mathbb{R}_k^d)$ and $\lambda_d^k(B(\boldsymbol{x}_{k,d}^0, r_0)) = (\pi^{kd/2} r_0^{kd})\Gamma(1+d/2)^{-k} \in \mathbb{R}^+$ by the volume formula. Thus

$$\int_{\mathbb{R}_k^d} g_{k,d}(\boldsymbol{x}_{k,d})\mathrm{d}\boldsymbol{x}_{k,d} \geq \int_{B(\boldsymbol{x}_{k,d}^0, r_0)} g_{k,d}(\boldsymbol{x}_{k,d})\mathrm{d}\boldsymbol{x}_{k,d} > 0.$$

D Proof of Lemma 3.1.

For any $x \in \mathbb{R}$ we have that

$$|f_0(x; \boldsymbol{\xi}_{k_0}^0) - f(x; \boldsymbol{\xi}_{k_0})| \leq \frac{||\boldsymbol{\pi}_{k_0}^0 - \boldsymbol{\pi}_{k_0}||_1}{\sqrt{2\pi\lambda_0}} + \frac{||\boldsymbol{\theta}_{k_0}^0 - \boldsymbol{\theta}_{k_0}||_1}{\sqrt{2\pi\exp(1)\lambda_0}} + u(\lambda, \boldsymbol{\theta}_{k_0}; x, \lambda_0)|\lambda - \lambda_0|$$

$$u(\lambda, \boldsymbol{\theta}_{k_0}; x, \lambda_0) = \frac{1}{\sqrt{2\pi}}\Big[\frac{k_0}{\lambda\sqrt{\lambda_0} + \lambda_0\sqrt{\lambda}} + \frac{\sqrt{\lambda_0}}{2\lambda\lambda_0^2}\sum_{j=1}^{k_0}(x-\theta_j)^2\Big] : (\lambda, \boldsymbol{\theta}_{k_0}) \in \mathbb{R}^+ \times \mathbb{R}^{k_0},$$

with $||\cdot||_1$ being the Euclidean L_1-norm in $\mathbb{R}^{k_0}$. Because $u(\lambda, \boldsymbol{\theta}_{k_0}; x, \lambda_0)$ is continuous at $(\lambda_0, \boldsymbol{\theta}_{k_0}^0)$,

$$f(x; \boldsymbol{\xi}_{k_0}) \to f_0(x; \boldsymbol{\xi}_{k_0}^0)$$

pointwise in x when $\boldsymbol{\xi}_{k_0} \to \boldsymbol{\xi}_{k_0}^0$. The last statement is equivalent to the condition that

$$|\log[f(x; \boldsymbol{\xi}_{k_0})] - \log[f_0(x; \boldsymbol{\xi}_{k_0}^0)]|f_0(x; \boldsymbol{\xi}_{k_0}^0) \to 0$$

pointwise in x when $\boldsymbol{\xi}_{k_0} \to \boldsymbol{\xi}_{k_0}^0$.

By condition B2, we can assume that $\theta_1^0 < \cdots < \theta_{k_0}^0$ (possibly after an appropriate relabeling). Choose $t_1^0, t_2^0 \in \mathbb{R}$ and $l_1^0, l_2^0 \in \mathbb{R}^+$ such that $\forall\, j \in [k_0] : \theta_j \in (t_1^0, t_2^0)$, $f_0(x; \boldsymbol{\xi}_{k_0}^0) < 1 : x \in (-\infty, t_1^0) \cup (t_2^0, +\infty)$ and

$\lambda_0 \in [l_1^0, l_2^0]$. Since $|\log[f_0(x;\boldsymbol{\xi}_{k_0}^0)]|f_0(x;\boldsymbol{\xi}_{k_0}^0)$ is uniformly continuous for $x \in [t_1^0, t_2^0]$,

$$I_1 = \int_{[t_1^0, t_2^0]} |\log[f_0(x;\boldsymbol{\xi}_{k_0}^0)]|f_0(x;\boldsymbol{\xi}_{k_0}^0)\mathrm{d}x \in \mathbb{R}^+.$$

Fix $\delta_1 \in (0,1)$, $\delta_2 = \frac{1}{2}\min\{t_2^0 - \theta_{k_0}^0, \theta_1^0 - t_1^0\}$ and define $V_0 = \overline{B}_1(\boldsymbol{\pi}_{k_0}^0, \delta_1) \times \overline{B}_1(\boldsymbol{\theta}_{k_0}^0, \delta_2) \times [l_1^0, l_2^0]$. Notice that $M(x, \boldsymbol{\xi}_{k_0}) = |\log[f(x;\boldsymbol{\xi}_{k_0})]|$ is uniformly continuous for $(x, \boldsymbol{\xi}_{k_0}) \in [t_1^0, t_2^0] \times V_0$. Then $M_0 = \max\{M(x, \boldsymbol{\xi}_{k_0}) : (x, \boldsymbol{\xi}_{k_0}) \in [t_1^0, t_2^0] \times V_0\} \in \mathbb{R}^+$ and

$$\int_{[t_1^0, t_2^0]} |\log[f(x;\boldsymbol{\xi}_{k_0})]|f_0(x;\boldsymbol{\xi}_{k_0}^0)\mathrm{d}x \leq I_2 = \int_{[t_1^0, t_2^0]} M_0 f_0(x;\boldsymbol{\xi}_{k_0}^0)\mathrm{d}x \in \mathbb{R}^+.$$

By the triangle inequality

$$\int_{[t_1^0, t_2^0]} |\log[f_0(x;\boldsymbol{\xi}_{k_0}^0)] - \log[f(x;\boldsymbol{\xi}_{k_0})]|f_0(x;\boldsymbol{\xi}_{k_0}^0)\mathrm{d}x \leq I_1 + I_2 \in \mathbb{R}^+.$$

On the other hand, define the following continuous functions:

$$\begin{aligned}
h_1(x) &= 0.5|\log(2\pi\lambda_0)| + 0.5\lambda_0^{-1}(x - \theta_{k_0}^0)^2 : x \in (-\infty, t_1^0) \\
h_2(x) &= 0.5|\log(2\pi l_1^0)| + (2l_1^0)^{-1}(x - \delta_2 - \theta_{k_0}^0)^2 : x \in (-\infty, t_1^0) \\
h_3(x) &= 0.5|\log(2\pi\lambda_0)| + 0.5\lambda_0^{-1}(x - \theta_1^0)^2 : x \in (t_2^0, +\infty) \\
h_4(x) &= 0.5|\log(2\pi l_1^0)| + (2l_1^0)^{-1}(x + \delta_2 - \theta_1^0)^2 : x \in (t_2^0, +\infty).
\end{aligned}$$

Using the initial assumptions

$$\begin{aligned}
|\log[f(x;\boldsymbol{\xi}_{k_0}^0)]| &\leq h_1(x) : x \in (-\infty, t_1^0) \\
|\log[f(x;\boldsymbol{\xi}_{k_0})]| &\leq h_2(x) : (x, \boldsymbol{\xi}_{k_0}) \in (-\infty, t_1^0) \times V_0 \\
|\log[f(x;\boldsymbol{\xi}_{k_0}^0)]| &\leq h_3(x) : x \in (t_2^0, +\infty) \\
|\log[f(x;\boldsymbol{\xi}_{k_0})]| &\leq h_4(x) : (x, \boldsymbol{\xi}_{k_0}) \in (t_2^0, +\infty) \times V_0.
\end{aligned}$$

Taking into account the existence of second order moments of a Normal distribution

$$\begin{aligned}
I_3 &= \int_{(-\infty, t_1^0)} [h_1(x) + h_2(x)]f_0(x;\boldsymbol{\xi}_{k_0}^0)\mathrm{d}x \in \mathbb{R}^+ \\
I_4 &= \int_{(t_2^0, +\infty)} [h_3(x) + h_4(x)]f_0(x;\boldsymbol{\xi}_{k_0}^0)\mathrm{d}x \in \mathbb{R}^+.
\end{aligned}$$

Again, using the triangle inequality

$$\int_{(-\infty, t_1^0)\cup(t_2^0, +\infty)} |\log[f_0(x;\boldsymbol{\xi}_{k_0}^0)] - \log[f(x;\boldsymbol{\xi}_{k_0})]|f_0(x;\boldsymbol{\xi}_{k_0}^0)\mathrm{d}x \leq I_3 + I_4 \in \mathbb{R}^+.$$

The previous arguments show that $|\log[f_0(x;\boldsymbol{\xi}_{k_0}^0)] - \log[f(x;\boldsymbol{\xi}_{k_0})]|f_0(x;\boldsymbol{\xi}_{k_0}^0) : (x, \boldsymbol{\xi}_{k_0}) \in \mathbb{R} \times V_0$ is bounded above by a positive and integrable function that depends only in $x \in \mathbb{R}$. As a consequence of Lebegue's Dominated Convergence Theorem

$$\lim_{\boldsymbol{\xi}_{k_0} \to \boldsymbol{\xi}_{k_0}^0} \int_{\mathbb{R}} \log\Big[\frac{f_0(x;\boldsymbol{\xi}_{k_0}^0)}{f(x;\boldsymbol{\xi}_{k_0})}\Big] f_0(x;\boldsymbol{\xi}_{k_0}^0)\mathrm{d}x = 0.$$

In other words, for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\forall\, \boldsymbol{\xi}_{k_0} \in B_1(\boldsymbol{\theta}^0_{k_0}, \delta) \times B_1(\boldsymbol{\pi}^0_{k_0}, \delta) \times (\lambda_0 - \delta, \lambda_0 + \delta) : \int_{\mathbb{R}} \log\Big[\frac{f_0(x; \boldsymbol{\xi}^0_{k_0})}{f(x; \boldsymbol{\xi}_{k_0})}\Big] f_0(x; \boldsymbol{\xi}^0_{k_0}) \mathrm{d}x < \varepsilon.$$

E Proof of Lemma 3.2.

Set $\delta_{00} = \frac{v k_0}{4}$ with $v > 0$ specified by condition B2. Notice that

$$\forall\, \delta \in (0, \delta_{00}] : \boldsymbol{\theta}_{k_0} \in B_\delta = \prod_{i=1}^{k_0} \Big(\theta^0_i - \frac{\delta}{k_0}, \theta^0_i + \frac{\delta}{k_0}\Big) \subseteq B_1(\boldsymbol{\theta}^0_{k_0}, \delta).$$

Using the definition of $\mathrm{NRep}_{k_0,1}(\mu, \sigma^2, \tau)$ and denoting $c_{k_0} = c_{k_0,1}$ the associated normalizing constant, we have that

$$\forall\, \delta \in (0, \delta_{00}] : \mathbb{P}(\boldsymbol{\theta}_{k_0} \in B_1(\boldsymbol{\theta}^0_{k_0}, \delta)) \geq \frac{1}{c_{k_0}} \int_{B_\delta} \prod_{i=1}^{k_0} \mathrm{N}(\theta_i; \mu, \sigma^2) \prod_{r<s}^{k_0} \Big[1 - \exp\Big\{ - \frac{(\theta_r - \theta_s)^2}{2\tau\sigma^2}\Big\}\Big] \mathrm{d}\boldsymbol{\theta}_{k_0}.$$

Now

$$\forall\, \boldsymbol{\theta}_{k_0} \in B_\delta : \prod_{r<s}^{k_0} \Big[1 - \exp\Big\{ - \frac{(\theta_r - \theta_s)^2}{2\tau\sigma^2}\Big\}\Big] \geq \Big[1 - \exp\Big\{ - \frac{v_0}{2\tau\sigma^2}\Big\}\Big]^{\ell_{k_0}} = R_0 > 0$$

with $v_0 = (v - \frac{2\delta_{00}}{k_0})^2$ and $\ell_{k_0} = \frac{k_0(k_0-1)}{2}$. Using this information and Fubini's Theorem

$$\forall\, \delta \in (0, \delta_{00}] : \mathbb{P}(\boldsymbol{\theta}_{k_0} \in B_1(\boldsymbol{\theta}^0_{k_0}, \delta)) \geq \frac{R_0}{c_{k_0}} \prod_{i=1}^{k_0} \Big[\Phi\Big(\frac{\theta^0_i - \mu}{\sigma} + \frac{\delta}{k_0\sigma}\Big) - \Phi\Big(\frac{\theta^0_i - \mu}{\sigma} - \frac{\delta}{k_0\sigma}\Big)\Big].$$

Because

$$\lim_{\delta \to 0^+} \frac{1}{\delta}\Big[\Phi\Big(\frac{\theta^0_i - \mu}{\sigma} + \frac{\delta}{k_0\sigma}\Big) - \Phi\Big(\frac{\theta^0_i - \mu}{\sigma} - \frac{\delta}{k_0\sigma}\Big)\Big] = \frac{2}{k_0\sigma}\mathrm{N}\Big(\frac{\theta^0_i - \mu}{\sigma}; 0, 1\Big) = S^0_i \in \mathbb{R}^+$$

for each $i \in [k_0]$, there exists $\delta_{0i} > 0$ such that

$$\forall\, \delta \in (0, \delta_{0i}] : \Big[\Phi\Big(\frac{\theta^0_i - \mu}{\sigma} + \frac{\delta}{k_0\sigma}\Big) - \Phi\Big(\frac{\theta^0_i - \mu}{\sigma} - \frac{\delta}{k_0\sigma}\Big)\Big] \geq \frac{S^0_i}{2}.$$

Finally, choose $\delta_0 = \min\{\delta_{0j} : j \in \{0\} \cup [k_0]\}$ to conclude that

$$\forall\, \delta \in (0, \delta_0] : \mathbb{P}(\boldsymbol{\theta}_{k_0} \in B_1(\boldsymbol{\theta}^0_{k_0}, \delta)) \geq \frac{R_0}{c_{k_0}}\Big(\prod_{i=1}^{k_0} \frac{S^0_i}{2}\Big) \exp\Big\{ - k_0 \log\Big(\frac{1}{\delta}\Big)\Big\} \in \mathbb{R}^+.$$

Remark: The previous inequality also applies replacing $B_1(\boldsymbol{\theta}^0_{k_0}, \delta)$ by $\overline{B}_1(\boldsymbol{\theta}^0_{k_0}, \delta)$.

F Proof of Proposition 3.3.

We will follow the proof of Lemma 1 in Petralia et al. (2012) with a few variations. For this, let $\varepsilon > 0$ and define

$$B_{\mathrm{KL}}(f_0, \varepsilon) = \Big\{ f \in \mathcal{F} : \int_{\mathbb{R}} \log\Big[\frac{f_0(x; \boldsymbol{\xi}^0_{k_0})}{f(x; \boldsymbol{\xi}_\star)}\Big] f_0(x; \boldsymbol{\xi}^0_{k_0}) \mathrm{d}x < \varepsilon \Big\}$$

with $\boldsymbol{\xi}_\star \in \bigcup_{k=1}^{\infty} \boldsymbol{\Theta}_k$. Using the stochastic representation (3.13),

$$\Pi(B_{\mathrm{KL}}(f_0,\varepsilon)) \geq \kappa(\{k_0\})\mathbb{P}\Big(\Big\{\boldsymbol{\xi}_{k_0} \in \boldsymbol{\Theta}_{k_0} : \int_{\mathbb{R}} \log\Big[\frac{f_0(x;\boldsymbol{\xi}^0_{k_0})}{f(x;\boldsymbol{\xi}_{k_0})}\Big] f_0(x;\boldsymbol{\xi}^0_{k_0})\mathrm{d}x < \varepsilon\Big\}\Big).$$

By condition B3, $\kappa(\{k_0\}) > 0$. In this case, to guarantee (3.14) it is sufficient to show that

$$\mathbb{P}\Big(\Big\{\boldsymbol{\xi}_{k_0} \in \boldsymbol{\Theta}_{k_0} : \int_{\mathbb{R}} \log\Big[\frac{f_0(x;\boldsymbol{\xi}^0_{k_0})}{f(x;\boldsymbol{\xi}_{k_0})}\Big] f_0(x;\boldsymbol{\xi}^0_{k_0})\mathrm{d}x < \varepsilon\Big\}\Big) > 0.$$

Lemma 3.1 guaranties the existence of $\delta_1 > 0$ such that

$$\forall\, \boldsymbol{\xi}_{k_0} \in B_1(\boldsymbol{\theta}^0_{k_0},\delta_1) \times B_1(\boldsymbol{\pi}^0_{k_0},\delta_1) \times (\lambda_0-\delta_1,\lambda_0+\delta_1) : \int_{\mathbb{R}} \log\Big[\frac{f_0(x;\boldsymbol{\xi}^0_{k_0})}{f(x;\boldsymbol{\xi}_{k_0})}\Big] f_0(x;\boldsymbol{\xi}^0_{k_0})\mathrm{d}x < \varepsilon.$$

Choose $\delta = \min\{\delta_0,\delta_1\}$ where $\delta_0 > 0$ is given by Lemma 3.2. Now $p_1 = \mathbb{P}(\boldsymbol{\theta}_{k_0} \in B_1(\boldsymbol{\theta}^0_{k_0},\delta)) > 0$. The same holds for $p_2 = \mathbb{P}(\boldsymbol{\pi}_{k_0} \in B_1(\boldsymbol{\pi}^0_{k_0},\delta))$ and $p_3 = \mathbb{P}(\lambda \in (\lambda_0-\delta,\lambda_0+\delta))$. Thus, independence between $\boldsymbol{\pi}_{k_0}$, $\boldsymbol{\theta}_{k_0}$ and λ implies

$$\mathbb{P}\Big(\Big\{\boldsymbol{\xi}_{k_0} \in \boldsymbol{\Theta}_{k_0} : \int_{\mathbb{R}} \log\Big[\frac{f_0(x;\boldsymbol{\xi}^0_{k_0})}{f(x;\boldsymbol{\xi}_{k_0})}\Big] f_0(x;\boldsymbol{\xi}^0_{k_0})\mathrm{d}x < \varepsilon\Big\}\Big) \geq p_1p_2p_3 > 0.$$

G Proof of Lemma 3.4.

As already mentioned at the beginning of Subsection 2.3, $\boldsymbol{\theta}_k \sim \mathrm{NRep}_{k,1}(\mu,\sigma^2,\tau)$ is an exchangeable distribution in $\theta_1,\ldots,\theta_k$ for $k \geq 2$. This implies that the probability laws of each $\theta_i : i \in [k]$ are the same. To prove the desired inequality, observe that

$$\forall\, t \in \mathbb{R}^+ : \mathbb{P}(|\theta_i| > t) \leq \frac{c_{k-1}}{c_k}\int_{A_t} \mathrm{N}(x;\mu,\sigma^2)\mathrm{d}x = \frac{c_{k-1}}{c_k}\int_{B_t} \mathrm{N}(s;0,1)\mathrm{d}s.$$

where $A_t = \{x \in \mathbb{R} : |x| > t\}$ and $B_t = \{s \in \mathbb{R} : |\mu+\sigma s| > t\}$. Now

$$B_t \subseteq \{s \in \mathbb{R} : |\mu| + \sigma|s| > t\} = \Big\{s \in \mathbb{R} : |s| > \frac{t-|\mu|}{\sigma}\Big\} = C_t.$$

Set $\gamma = \max\{2|\mu|+1,(2+\sqrt{2})|\mu|\} \in \mathbb{R}^+$. By Mill's Inequality

$$\begin{aligned}\forall\, t \geq \gamma : \int_{C_t} \mathrm{N}(s;0,1)\mathrm{d}s &\leq \frac{2}{(2\pi)^{1/2}}\Big(\frac{\sigma}{t-|\mu|}\Big)\exp\Big\{-\frac{(t-|\mu|)^2}{2\sigma^2}\Big\} \\ &\leq \frac{2}{(2\pi)^{1/2}}\Big(\frac{\sigma}{|\mu|+1}\Big)\exp\Big\{-\frac{t^2}{4\sigma^2}\Big\}.\end{aligned}$$

Using the previous information

$$\forall\, t \geq \gamma : \mathbb{P}(|\theta_i| > t) \leq \frac{2}{(2\pi)^{1/2}}\frac{c_{k-1}}{c_k}\Big(\frac{\sigma}{|\mu|+1}\Big)\exp\Big\{-\frac{t^2}{4\sigma^2}\Big\}.$$

H Proof of Lemma 3.5.

By the Change of Variables Theorem and Fubini's Theorem, it can be shown that

$$\forall\, k \geq 2 : c_k = \int_{\mathbb{R}^{k-1}} F_{k-1}(\boldsymbol{\theta}_{-1}) \prod_{i=2}^{k} \mathrm{N}(\theta_i; 0, 1) \prod_{2 \leq r < s \leq k} \Big[1 - \exp\Big\{ - \frac{(\theta_r - \theta_s)^2}{2\tau} \Big\}\Big] \mathrm{d}\boldsymbol{\theta}_{-1}$$

where $\boldsymbol{\theta}_{-1} = (\theta_i : i \neq 1)^\top \in \mathbb{R}^{k-1}$ and $F_{k-1} : \mathbb{R}^{k-1} \to (0, 1)$ is given by

$$F_{k-1}(\boldsymbol{\theta}_{-1}) = \int_{\mathbb{R}} \mathrm{N}(\theta_1; 0, 1) \prod_{j=2}^{k} \Big[1 - \exp\Big\{ - \frac{(\theta_1 - \theta_j)^2}{2\tau} \Big\}\Big] \mathrm{d}\theta_1.$$

Notice that F_{k-1} is a $C^\infty(\mathbb{R}^{k-1}; (0, 1))$-class positive function (as a consequence of Lebesgue's Dominated Convergence Theorem) and $F_{k-1}(\boldsymbol{\theta}_{-1}) \to 1$ as $||\boldsymbol{\theta}_{-1}|| \to +\infty$. By Jensen's Inequality

$$\forall\, \boldsymbol{\theta}_{-1} \in \mathbb{R}^{k-1} : \log[F_{k-1}(\boldsymbol{\theta}_{-1})] \geq \sum_{j=2}^{k} \int_{\mathbb{R}} \mathrm{N}(\theta_1; 0, 1) \log \Big[1 - \exp\Big\{ - \frac{(\theta_1 - \theta_j)^2}{2\tau} \Big\}\Big] \mathrm{d}\theta_1.$$

Now

$$\Big| \int_{\mathbb{R}} \mathrm{N}(\theta_1; 0, 1) \log \Big[1 - \exp\Big\{ - \frac{(\theta_1 - \theta_j)^2}{2\tau} \Big\}\Big] \mathrm{d}\theta_1 \Big| \leq -2\sqrt{\frac{\tau}{\pi}} \int_{\mathbb{R}^+} \log[1 - \exp\{-\theta_1^2\}] \mathrm{d}\theta_1.$$

Using the substitution $\theta_1(z) = \sqrt{z} : z \in \mathbb{R}^+$ and then integrating by parts

$$\begin{aligned} \int_{\mathbb{R}^+} \log[1 - \exp\{-\theta_1^2\}] \mathrm{d}\theta_1 &= \frac{1}{2} \int_{\mathbb{R}^+} \log[1 - \exp\{-z\}] z_1^{-1/2} \mathrm{d}z \\ &= - \int_{\mathbb{R}^+} \frac{z^{3/2-1}}{\exp\{z\} - 1} \mathrm{d}z \\ &= -\Gamma\Big(\frac{3}{2}\Big) \zeta\Big(\frac{3}{2}\Big) \in \mathbb{R}^- \end{aligned}$$

where $\Gamma(\,\cdot\,)$ and $\zeta(\,\cdot\,)$ are the Gamma and Riemann Zeta functions, respectively. The previous information implies that

$$\Big| \int_{\mathbb{R}} \mathrm{N}(\theta_1; 0, 1) \log \Big[1 - \exp\Big\{ - \frac{(\theta_1 - \theta_j)^2}{2\tau} \Big\}\Big] \mathrm{d}\theta_1 \Big| \leq 2\sqrt{\frac{\tau}{\pi}} \Gamma\Big(\frac{3}{2}\Big) \zeta\Big(\frac{3}{2}\Big) \leq 2.6124\sqrt{\tau} \in \mathbb{R}^+.$$

With this bound, defining $a = 2.6124\sqrt{\tau}$ and $A^{-1} = \exp\{a\}$ the following holds:

$$\forall\, \boldsymbol{\theta}_{-1} \in \mathbb{R}^{k-1} : \log[F_{k-1}(\boldsymbol{\theta}_{-1})] \geq -(k-1)a \Rightarrow F_{k-1}(\boldsymbol{\theta}_{-1}) \geq A^{-1} \exp\{-ak\}.$$

To conclude the proof

$$c_{k-1} = \int_{\mathbb{R}^{k-1}} \prod_{i=2}^{k} \mathrm{N}(\theta_i; 0, 1) \prod_{2 \leq r < s \leq k} \Big[1 - \exp\Big\{ - \frac{(\theta_r - \theta_s)^2}{2\tau} \Big\}\Big] \mathrm{d}\boldsymbol{\theta}_{-1} \Rightarrow c_k \geq A^{-1} \exp\{-ak\} c_{k-1}.$$

Finally

$$\forall\, k \geq 2 : 0 < \frac{c_{k-1}}{c_k} \leq A \exp\{ak\}.$$

I Proof of Proposition 3.6.

Following Theorem 3.1 in Scricciolo (2011) $p = 2$ induce a (finite) Gaussian Mixture Model, $\lambda \sim \mathrm{IG}(a, b) : a, b \in \mathbb{R}^+$ satisfy (i) and $\boldsymbol{\pi}_k \sim \mathrm{Dir}((1/k)\mathbf{1}_k)$ satisfy (iii). Condition B3′ is equivalent to (ii). However, (iv) does not apply because the cluster-location parameters are not *i.i.d.* in our framework.

Along the proof of Theorem 3.1 we identified those steps that can be adapted by the assumption $\boldsymbol{\theta}_k \sim \mathrm{NRep}_{k,1}(\mu, \sigma^2, \tau)$. It is important to mention that Theorem 3.1 appeals to conditions (A.1), (A.2) and (A.3) in Theorem A.1 (Appendix of Scricciolo's paper) which is a powerful result given by Ghosal and van der Vaart (2001). We will check that (A.1) to (A.3) are satisfied:

(A.1) The proof is the same as the arguments presented at page 277 and the first paragraph in page 278. The reason for this is that it only depends on the structure of the mixture, leaving aside the prior distributions for all the involved parameters.

(A.2) What needs to be modified on the first inequality found on page 278 is the term $\mathrm{E}[K]\Pi([-a_n, a_n]^c)$. This quantity is part of the chain of inequalities

$$\sum_{i=1}^{k_n} \rho(i) \sum_{j=1}^{i} \mathbb{P}(|\theta_j| > a_n) = \sum_{i=1}^{k_n} i\rho(i)\Pi([-a_n, a_n]^c) \leq \mathrm{E}[K]\Pi([-a_n, a_n]^c) \lesssim \exp\{-c a_n^{\vartheta}\}$$

under the conditions (ii) and (iv). In our case, $\rho(i) = \kappa(\{i\}) : i \in \mathbb{N}$. By way of Lemma 3.4

$$\sum_{j=1}^{i} \mathbb{P}(|\theta_j| > a_n) \leq \frac{2i}{(2\pi)^{1/2}} \frac{c_{i-1}}{c_i} \Big(\frac{\sigma}{|\mu| + 1} \Big) \exp\Big\{ - \frac{a_n^2}{4\sigma^2} \Big\}$$

under the convention that $c_0 = 1$ and $n \in \mathbb{N}$ is big enough. Thus,

$$\sum_{i=1}^{k_n} \rho(i) \sum_{j=1}^{i} \mathbb{P}(|\theta_j| > a_n) \leq \frac{2}{(2\pi)^{1/2}} \Big(\frac{\sigma}{|\mu| + 1} \Big) \exp\Big\{ - \frac{a_n^2}{4\sigma^2} \Big\} \sum_{i=1}^{k_n} i\rho(i) \frac{c_{i-1}}{c_i}$$

and by Lemma 3.5

$$\sum_{i=1}^{k_n} i\rho(i) \frac{c_{i-1}}{c_i} \leq AB \sum_{i=1}^{k_n} i \exp\{-(b-a)i\} \leq AB \sum_{i=1}^{\infty} i \exp\{-(b-a)i\} \in \mathbb{R}^+.$$

Finally, we obtain the following upper bound (in order), which is analogous to that obtain in Scricciolo (2011):

$$\sum_{i=1}^{k_n} \rho(i) \sum_{j=1}^{i} \mathbb{P}(|\theta_j| > a_n) \lesssim \exp\{-(4\sigma^2)^{-1} a_n^2\}.$$

(A.3) We only need to adapt the following inequality found on page 279, whose validity is deduced from (iv):

$$\mathbb{P}(\boldsymbol{\theta}_{k_0} \in B(\boldsymbol{\theta}_{k_0}^0; \varepsilon)) = \Pi^{\otimes k_0}(B(\boldsymbol{\theta}_{k_0}^0; \varepsilon)) \geq \prod_{i=1}^{k_0} \Pi\Big(\Big[\theta_j^0 - \frac{\varepsilon}{k_0}, \theta_j^0 + \frac{\varepsilon}{k_0} \Big] \Big) \gtrsim \exp\Big\{ - d_1 k_0 \log\Big(\frac{1}{\varepsilon} \Big) \Big\}$$

In our case, $B(\boldsymbol{\theta}_{k_0}^0; \varepsilon) = \overline{B}_1(\boldsymbol{\theta}_{k_0}^0, \varepsilon)$. At the end of the proof of Lemma 3.2 it is shown that for every $\varepsilon \in (0, \delta_0]$

$$\mathbb{P}(\boldsymbol{\theta}_{k_0} \in \overline{B}_1(\boldsymbol{\theta}_{k_0}^0, \varepsilon)) \geq \frac{R_0}{c_{k_0}} \Big(\prod_{i=1}^{k_0} \frac{S_i^0}{2} \Big) \exp\Big\{ - k_0 \log\Big(\frac{1}{\varepsilon} \Big) \Big\}.$$

With this information, we obtain a lower bound (in order) analogous to that obtained in Scricciolo (2011):

$$\mathbb{P}(\boldsymbol{\theta}_{k_0} \in B(\boldsymbol{\theta}^0_{k_0}; \varepsilon)) \gtrsim \exp\Big\{ - k_0 \log\Big(\frac{1}{\varepsilon}\Big)\Big\}.$$