

Repulsive Processes: Theory and Applications

By

José Javier Quinlan Binelli

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR IN STATISTICS

AT

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
SANTIAGO, CHILE

MARCH 2017

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
DEPARTMENT OF STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Mathematics for acceptance a thesis entitled “**Repulsive Processes: Theory and Applications**” by **José Javier Quinlan Binelli** in partial fulfillment of the requirements for the degree of **Doctor in Statistics**.

Dated: March, 2017

Research Supervisor: _____

Fernando Quintana
Pontificia Universidad Católica de Chile

External Supervisor: _____

Garritt Page
Brigham Young University

Examining Committee: _____

Luis Gutiérrez
Pontificia Universidad Católica de Chile

Alejandro Jara
Pontificia Universidad Católica de Chile

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Date: **March, 2017**

Author: **José Javier Quinlan Binelli**
Title: **Repulsive Processes: Theory and Applications**
Department: **Statistics**
Degree: **Doctor in Statistics**
Convocation: **March**
Year: **2017**

Permission is herewith granted to Pontificia Universidad Católica de Chile to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

Contents

Acknowledgments	i
List of Figures	iii
List of Tables	v
1 Introduction	1
1.1 The Research Context	1
1.2 Finite Point Processes	2
1.2.1 Poisson Point Process	3
1.2.2 Repulsive Point Processes	3
1.3 A Class of Repulsive Distributions	5
1.4 Outline of this Dissertation	6
2 Density Estimation using Repulsive Distributions	7
2.1 Chapter Overview	7
2.2 Introduction	8
2.3 Probability Repulsive Distributions	12
2.3.1 Background and Preliminaries	13
2.3.2 $\text{Rep}_{k,d}(f_0, C_0, \rho)$ Distribution	15
2.3.3 $\text{Rep}_{k,d}(f_0, C_0, \rho)$ Properties	18

2.4	Gaussian Mixture Models and $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ Distribution	21
2.4.1	Repulsive Gaussian Mixture Models (RGMM)	21
2.4.2	Theoretical Properties	26
2.4.3	Sampling From $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$	29
2.5	Simulation Study	30
2.6	Data Illustrations	35
3	Regression Estimation using Repulsive Distributions	39
3.1	Chapter Overview	39
3.2	Introduction	40
3.3	Covariate Dependent RGMM (RGMMx)	42
3.3.1	Parameter Calibration	46
3.3.2	Computation	47
3.4	Simulation Study	49
3.5	Data Illustration	50
4	Discussion and Future Work	60
4.1	Density Estimation using Repulsive Distributions	60
4.2	Regression Estimation using Repulsive Distributions	62
A	Supplementary Material for Chapter 2	63
A.1	Algorithm RGMM	63
A.2	Proof of Lemma 2.3.1.	65
A.3	Proof of Proposition 2.3.2.	65
A.4	Proof of Lemma 2.4.1.	66
A.5	Proof of Lemma 2.4.2.	69
A.6	Proof of Proposition 2.4.3.	70

A.7 Proof of Lemma 2.4.4.	71
A.8 Proof of Lemma 2.4.5.	72
A.9 Proof of Proposition 2.4.6.	74
B Supplementary Material for Chapter 3	77
B.1 Algorithm RGMMx	77
Bibliography	80

Acknowledgments

Muchas personas han estado presente conmigo a lo largo de este proceso y que, gracias a ellos me encuentro escribiendo estas sinceras palabras de agradecimiento.

A Javiera Orellana, mi compañera de vida, que me ha brindado amor y apoyo incondicionales. Siempre creíste en mí, impulsándome a superar mis temores e inseguridades. Te amo.

A Bianca, nuestra mascota, que me ha servido de terapia para apaciguar la angustia. Es curioso cómo me has enseñado a cultivar la paciencia y crear un cariño tan grande hacia una criatura como tú.

A mis padres (Jorge y Stella) y mis hermanos (Ignacio, Francisca y Antonia), por confiar en mis capacidades y ayudarme constantemente a ser una mejor persona. Los quiero y admiro.

A mis amigos Erik Contreras, Rodrigo Rubio, Bastian Galasso, Luis Muñoz y Sebastián Zúñiga, quienes me entregaron la fuerza necesaria para trabajar día a día. Sus consejos, bromas y experiencias de vida las guardo en mi corazón con mucho cariño. Gracias por aceptarme y quererme tal cual soy.

A mi tutor, Fernando Quintana y mi co-tutor, Garritt Page por guiarme en este camino. Han sido mi fuente de inspiración académica, estando siempre pendientes de mis avances. Fue maravillosa la experiencia de trabajar junto a ustedes. Nunca olvidaré la preocupación que tuvieron por mi bienestar en momentos difíciles de mi vida.

A Ricardo Olea, Reinaldo Arellano, Gregorio Moreno y Duván Henao por nuestras conversaciones entretenidas y sus palabras de aliento cuando el final de esta etapa se veía incierta.

A las personas del Departamento de Estadística de Brigham Young University, quienes me acogieron con alegría durante mi pasantía en Provo. A pesar de que fue corta mi visita,

disfruté y aprendí mucho.

Finalmente, a la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) por todo el apoyo financiero otorgado a través del programa Becas para Estudios de Doctorado en Chile (folio 21120153). Sin ustedes no hubiera sido posible concretar una pasantía y exponer el trabajo de tesis en congresos nacionales e internacionales, experiencias únicas y enriquecedoras para mi formación académica.

Dios, gracias por estar en mi vida. Cuida de mis seres queridos que partieron a tus brazos.

José Javier Quinlan Binelli

Santiago, Chile

March, 2017

List of Figures

2.1	Data simulated from the mixture of 4 bivariate normal densities in (2.2.2). 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 DPM.	10
2.2	The graph and Laplacian matrix for a possible interaction for $k = 4$ coordinates.	20
2.3	Boxplots that resume the behavior of LPML for each of the four models. . .	34
2.4	Boxplots that resume the behavior of MSE for each of the four models. . . .	34
2.5	Boxplots that resume the behavior of L_1 -metric for each of the four models. .	35
2.6	Side-by-side boxplots of the average number of occupied mixture components for each of the procedure.	35
2.7	Side-by-side boxplots that display the average standard deviation associated with the posterior distribution of occupied mixture components for each of the four procedures.	36
2.8	Posterior distribution for the active number of clusters in (a) Galaxy and (b) Air Quality data. Black (gray) bars correspond to RGMM (DPMM).	38
2.9	Posterior predictive densities for (a) Galaxy and (b) Air Quality data. Black solid (gray dashed) curves correspond to RGMM (DPMM).	38

3.1	Boxplots that display LPML, L_1 -metric, the average number of occupied mixture components, and the average standard deviation associated with the distribution of occupied mixture components for each value of τ	51
3.2	Side-by-side box-plots of the posterior distribution for the active number of clusters associated with the Geysler data.	55
3.3	Estimated regression curve (gray solid) for Geysler data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals.	56
3.4	Estimated partitions using Dahl's least squares clustering algorithm for (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2.	57
3.5	Estimated conditional densities (black solid) for Geysler data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals. Here, the selected time eruptions (<i>duration</i>) are 2 and 3 minutes.	58
3.6	Estimated conditional densities (black solid) for Geysler data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals. Here, the selected time eruptions (<i>duration</i>) are 4 and 4.5 minutes.	59

List of Tables

2.1	Summary statistics related to model fit and the number of clusters for Galaxy and Air Quality data based on DPMM and RGMM.	37
3.1	Summary statistics related to model fit and the number of clusters for Geyser data based on WDDP and RGMMx.	54

Chapter 1

Introduction

1.1 The Research Context

The general focus of this thesis is the development of a class of probability distributions that explicitly parametrizes what is often referred to as *repulsion*. Developing *repulsive distributions* arised in the context of density and regression estimation using mixtures of Gaussian distributions in the Bayesian framework. Many of the approaches available in the literature (Escobar and West (1995); Müller et al. (1996); McLachlan and Peel (2000); Frühwirth-Schnatter (2006) for example) assume that a priori the location parameters of each mixture component are i.i.d. generated by an appropriate probability law. Because of the independence assumption, any pair of location parameters can be a priori very close to each other. This has repercussions on model complexity and out of sample prediction, leading to eventual overfitting. Our work uses repulsion as a mechanism to obtain parsimonious models without sacrificing too much goodness of fit.

Since there are strong connections between the class of repulsive distributions we develop and ideas found in the theory of Finite Point Processes (FPPs), we begin by making explicit connections between the two. Doing this will hopefully provide context and also connect the

ideas we develop to well established statistical concepts.

1.2 Finite Point Processes

A finite point process X can be thought of as a finite random configuration of points that lie in a suitable space. From a technical point of view, the elements of the random set X live in a measure space $(S, \mathcal{B}(S), \nu)$, where $\mathcal{B}(S)$ is the Borel σ -algebra of subsets of a locally compact Polish space S (complete and separable metric space) and ν is a finite diffuse measure, i.e. $0 < \nu(S) < \infty$ and $\nu(\{\mathbf{s}\}) = 0$ for all $\mathbf{s} \in S$. This last property implies that two points can not share the same location in S . FPPs have been widely used to describe random patterns in biology, ecology, agronomy and physics, among others (Møller and Waagepetersen 2003; Illian et al. 2008; Diggle 2013). Daley and Vere-Jones (2002) (Chapter 5, Proposition 5.3.II.) give a natural and constructive way to define FPPs X on S under the following (sufficient) assumptions:

- A discrete distribution $\{p_n\}_{n \in \mathbb{N}_0}$ that determines the total number of points.
- A family of probability densities $\{\pi_n\}_{n \in \mathbb{N}}$ with respect to the n -fold product of ν that determines the locations of the points in S , given that their total number is n . π_n must be invariant under permutations of its argument for all $n \geq 2$.

The presence of $\{\pi_n\}_{n \in \mathbb{N}}$ connects FPPs with probability distributions that are invariant to permutations of their arguments. This guarantees that the order in which the points are observed is irrelevant. This is a natural feature of FPPs considering that they are random sets of points.

1.2.1 Poisson Point Process

A process that serves as a basis to construct a wide variety of more complicated processes is the so called (finite) Poisson Point Process (PPP). The PPP can be expressed using

$$p_n = \exp\{-\nu(S)\} \frac{\nu(S)^n}{n!}$$
$$\pi_n(\mathbf{s}_1, \dots, \mathbf{s}_n) = \prod_{i=1}^n \frac{f(\mathbf{s}_i)}{\nu(S)},$$

where f is a non-negative measurable function such that $\nu(A) = \int_A f(\mathbf{s})d\mu(\mathbf{s})$ for all $A \in \mathcal{B}(S)$ and some dominating measure μ on $(S, \mathcal{B}(S))$ that guarantees the mentioned properties of ν . In this setting, ν and f are often referred to as an intensity measure and intensity function, respectively. A space S that is frequently used in modelling is \mathbb{R}^d for some $d \in \mathbb{N}$. In this case $\nu(A) = \int_A f_\theta(\mathbf{s})d\mathbf{s}$, where $f_\theta(\cdot)$ is a non-negative (Lebesgue) integrable function and $\theta \in \Theta$ is a parameter that controls ν . Several techniques are available to estimate θ for an observed set of points. See, for example, Gaetan et al. (2010) and Gelfand et al. (2010).

When additional covariate information $\mathbf{x} \in \mathcal{X}$ is collected at each point it is of interest to learn how \mathbf{x} influences the patterns available from the PPP probability model. This can be done via the intensity measure $\nu(A) = \int_A f_\theta(\mathbf{s}; \mathbf{x})d\mathbf{s}$, where $\{f_\theta(\cdot; \mathbf{x}) : \theta \in \Theta\}$ is a family of non-negative (Lebesgue) integrable functions indexed by \mathbf{x} . Just as when covariates are not available, several techniques to estimate θ for an observed set of points (with covariates) have been developed (Gaetan et al. 2010; Gelfand et al. 2010).

1.2.2 Repulsive Point Processes

A consequence of the PPP definition is that it is fairly restrictive about the types of patterns it permits. In fact, for a given fixed number of points, elements of X are i.i.d. according to the probability measure $\frac{\nu(\cdot)}{\nu(S)}$. This produces point patterns that exhibit “random scatter”

and in most applications this feature is unrealistic (incompatible with the nature of the observed data). It is then necessary to introduce FPPs that can generate patterns where the points tend to be separated and/or grouped. A popular type of point processes that produce regular patterns (more spread point configurations) due to repulsion are Determinantal Point Processes in $S = \mathbb{R}^d$ (DPPs). The intuition behind the construction of DPPs is that their Janossy densities (Daley and Vere-Jones 2002) are defined through the determinant of a matrix \mathbf{M}_C whose entries depend on a continuous complex covariance function $C : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{C}$. If any pairwise distinct points $\mathbf{s}_i, \mathbf{s}_j \in \mathbb{R}^d$ are such that $\|\mathbf{s}_i - \mathbf{s}_j\|_d \approx 0$, then \mathbf{M}_C has almost linear dependent columns which implies $\det(\mathbf{M}_C) \approx 0$. Here, $\|\cdot\|_d$ is the d -dimensional Euclidean norm. For excellent reviews of such stochastic processes and their applications in statistical modelling and inference see Hough et al. (2006), Lavancier et al. (2015) and Xu et al. (2016).

Another useful example of such FPPs on which our work is based are (finite) Gibbs Point Processes (GPPs) on $S = \mathbb{R}^d$. These types of processes arise in Statistical Mechanics to model patterns exhibiting inter-particle interactions. The exact forms of p_n and π_n that correspond to GPPs are

$$p_n = \frac{C(\beta)}{n!} \int_{\mathbb{R}_n^d} \exp\{-\beta U_n(\mathbf{s}_1, \dots, \mathbf{s}_n)\} d\mathbf{s}_1 \cdots d\mathbf{s}_n \quad \text{with} \quad p_0 = C(\beta)$$

$$\pi_n(\mathbf{s}_1, \dots, \mathbf{s}_n) \propto \exp\{-\beta U_n(\mathbf{s}_1, \dots, \mathbf{s}_n)\},$$

where $\mathbb{R}_n^d = \prod_{i=1}^n \mathbb{R}^d$. The parameter $\beta \in (0, \infty)$ is related to the temperature of the particle system and the proportionality constant $C(\beta) \in (0, \infty)$ is known as the partition function. Here, $U_n : \mathbb{R}_n^d \rightarrow [-\infty, \infty]$ is a measurable function that is exchangeable in its arguments for each $n \in \mathbb{N}$. This function, called potential energy, is fairly crucial to our methodology as it models the interaction between n particles located at $\mathbf{s}_1, \dots, \mathbf{s}_n$. In order for GPPs to be well-defined, the following conditions are necessary and sufficient (Daley and Vere-Jones

2002):

$$\int_{\mathbb{R}_n^d} \exp\{-\beta U_n(\mathbf{s}_1, \dots, \mathbf{s}_n)\} d\mathbf{s}_1 \cdots d\mathbf{s}_n < \infty \quad \text{and}$$

$$1 + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{\mathbb{R}_n^d} \exp\{-\beta U_n(\mathbf{s}_1, \dots, \mathbf{s}_n)\} d\mathbf{s}_1 \cdots d\mathbf{s}_n = \frac{1}{C(\beta)}.$$

Suitable choices for U_n induce repulsion, i.e. particles are encouraged to be separated.

1.3 A Class of Repulsive Distributions

In order to construct repulsive distributions, we take advantage of GPPs potential energy U_n (setting $\beta = 1$) using the following particular form:

$$U_n(\mathbf{s}_1, \dots, \mathbf{s}_n) = \sum_{i=1}^n \varphi(\mathbf{s}_i) + \sum_{j < k}^n \phi\{\rho(\mathbf{s}_j, \mathbf{s}_k)\},$$

for suitable measurable functions $\varphi : \mathbb{R}^d \rightarrow (0, \infty)$ and $\phi : [0, \infty) \rightarrow (0, \infty]$, and a metric $\rho : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$ in \mathbb{R}^d . The functions φ and ϕ are called potential functions of first and second order, correspondingly. The remarkable aspect of the previous specification is that

$$\pi_n(\mathbf{s}_1, \dots, \mathbf{s}_n) \propto \exp \left[- \sum_{i=1}^n \varphi(\mathbf{s}_i) - \sum_{j < k}^n \phi\{\rho(\mathbf{s}_j, \mathbf{s}_k)\} \right] \leq \prod_{i=1}^n \exp\{-\varphi(\mathbf{s}_i)\}.$$

Because of the above inequality, $\exp\{-\varphi(\cdot)\}$ can be associated to a continuous density that emulates the i.i.d. scheme while ϕ models repulsion by penalizing small inter-particle distances. This idea allows the construction of a general class of probability distributions called (second order) Gibbs measures (Illian et al. 2008) which shares the same basis of GPPs.

1.4 Outline of this Dissertation

In Chapter 2 we more fully develop these ideas by discussing how (second order) Gibbs measures can be used to define probability measures with repulsive properties. We will briefly describe some global characteristics to get a better understanding of the nature of the repulsion, and then use them in (Bayesian) Gaussian Mixture Models for density estimation. We show a simple way to obtain posterior samples from our model, and prove theoretical results relative to the Kullback-Leibler support of the prior and posterior convergence rate under regularity conditions. This chapter concludes with a simulation study and illustrations of our methodology applied to real data sets.

In this thesis we also consider the influence that $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$ has on the location of points using a particular approach, details of which are found in Chapter 3. This chapter is completely methodological. Using the ideas developed in Chapter 2 we construct a covariate dependent (Bayesian) Gaussian Mixture Model that includes a repulsion component for regression estimation. As in Chapter 2, the repulsion encourages the location parameters of each mixture component to be well separated. Our approach is to model the response and covariates jointly, which requires treating covariates as random quantities. From the joint distribution of the response and covariates we induce a conditional probability law that is a function of the covariates. The key aspect is that the repulsion is directly inherited to the conditional distribution. We show mechanisms to generate posterior samples from the joint distribution and how to use them for regression estimation. This chapter concludes with illustrations of our methodology using simulated and real data sets.

Finally, Chapter 4 contains some overall conclusions and discussion of possible future work (theoretical aspects and generalizations) that we want to study.

Chapter 2

Density Estimation using Repulsive Distributions

2.1 Chapter Overview

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.

2.2 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

$$\mathbf{y}_i \mid \boldsymbol{\theta}_i \stackrel{ind.}{\sim} k(\mathbf{y}_i; \boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n \stackrel{i.i.d.}{\sim} \sum_{k=1}^N \pi_k \delta_{\boldsymbol{\phi}_k}, \quad \boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N \stackrel{i.i.d.}{\sim} G_0, \quad (2.2.1)$$

where $k(\cdot; \boldsymbol{\theta})$ is a suitable kernel density indexed by $\boldsymbol{\theta}$, $1 \leq N \leq \infty$, component weights π_1, \dots, π_N are nonnegative and $\sum_{k=1}^N \pi_k = 1$ with probability 1, and G_0 is a suitable probability distribution. 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 (2.2.1) 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 (DP) (Ferguson 1973) and the Pitman-Yor Process (Pitman and Yor 1997).

A common feature of models like (2.2.1) is the use of i.i.d. atoms $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N$. 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 (2.2.1) 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:

$$\mathbf{y} \sim 0.2\mathcal{N}_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.3\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.3\mathcal{N}_2(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) + 0.2\mathcal{N}_2(\boldsymbol{\mu}_4, \boldsymbol{\Sigma}_4), \quad (2.2.2)$$

with

$$\begin{aligned} \boldsymbol{\mu}_1 &= (0, 0)^\top, & \boldsymbol{\mu}_2 &= (3, 3)^\top, & \boldsymbol{\mu}_3 &= (-3, -3)^\top, & \boldsymbol{\mu}_4 &= (-3, 0)^\top \\ \boldsymbol{\Sigma}_1 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & \boldsymbol{\Sigma}_2 &= \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, & \boldsymbol{\Sigma}_3 &= \begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}, & \boldsymbol{\Sigma}_4 &= \begin{pmatrix} 3 & -2 \\ -2 & 2 \end{pmatrix}. \end{aligned}$$

The left panel in Figure 2.1 shows the original data and clusters, labeled with different numbers and colors. We fit to these data the variation of model (2.2.1) 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 2.1 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 (2.2.1): 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?

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

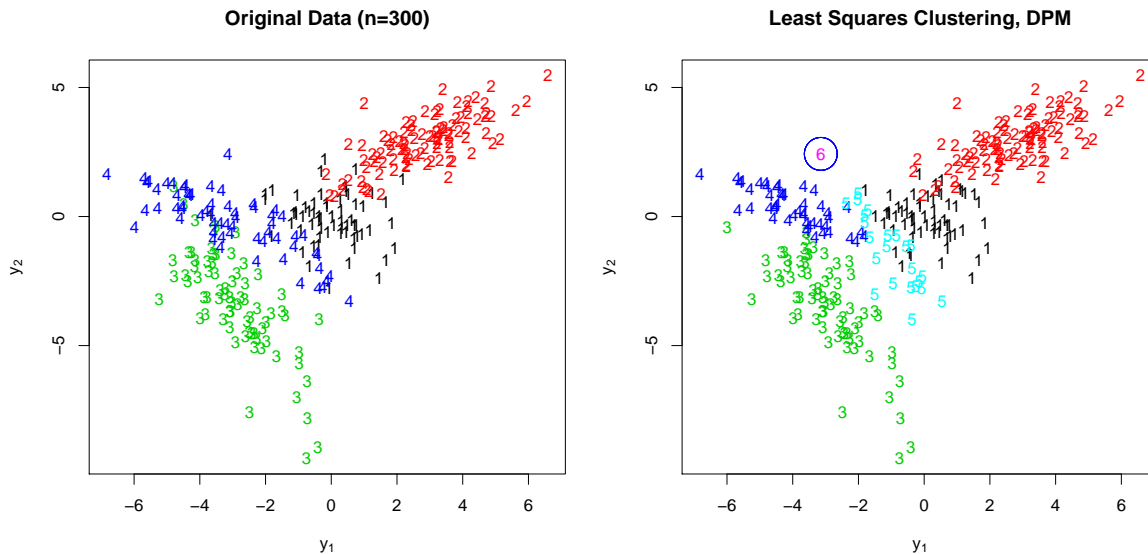


Figure 2.1: Data simulated from the mixture of 4 bivariate normal densities in (2.2.2). The left panel shows the original $n = 300$ data points with colors and numbers indicating the original cluster. The right panel shows the clustering resulting from applying Dahl’s least squares clustering algorithm to a DPM.

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 2.1 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 that inform pattern configurations. Among these, to our knowledge, only Determinantal Point Processes 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 probability distribution that explicitly parameterizes repulsion. Along these lines Fúquene et al. (2016) develop a family of probability densities called Non-Local Priors that incorporates repulsion by penalizing small relative distances between coordinates. Our approach to incorporating repulsion is to model coordinate interactions through potentials (functions that describe the ability to interact) found in so called (second order) Gibbs measures. As will be shown, this allows us to control the strength of repulsion and also consider a large variety of types of repulsion.

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 in a Gibbs measure (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 measures and Determinantal Point Processes 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 distribution 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 chapter will be organized as follows. In Section 2.3 we formally introduce the notion of repulsion in the context of a probability distribution and discuss several resulting properties. In Section 2.4, we detail how the repulsive probability distributions can be employed in hierarchical mixture modeling for density estimation. Section 2.5 contains results from a small simulation study that compares the repulsive mixture model we develop to DPM and finite mixture models. In Section 2.6 we apply the methodology to two well known datasets. Proofs of all technical results and computational strategies are provided in Appendix A.

2.3 Probability Repulsive Distributions

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

2.3.1 Background and Preliminaries

We will use the k -fold product space of \mathbb{R}^d denoted by $\mathbb{R}_k^d = \prod_{i=1}^k \mathbb{R}^d$ and $\mathcal{B}(\mathbb{R}_k^d)$ its associated σ -algebra as the reference space on which the class of distributions we derive will be defined. Here, $k \in \mathbb{N}$ ($k \geq 2$) and $d \in \mathbb{N}$. Let $\mathbf{x}_{k,d} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ with $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^d$. The k -tuple $\mathbf{x}_{k,d}$ can be thought of as k ordered objects of dimension d jointly allocated in \mathbb{R}_k^d . We add to the measurable space $(\mathbb{R}_k^d, \mathcal{B}(\mathbb{R}_k^d))$ a σ -finite measure λ_d^k , that is the k -fold product of the d -dimensional Lebesgue measure λ_d . To represent integrals with respect to λ_d^k , we will use $d\mathbf{x}_{k,d}$ instead of $d\lambda_d^k(\mathbf{x}_{k,d})$. 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 $f : \Omega_1 \rightarrow \Omega_2$. 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. distributions depends heavily on Gibbs measures where dependence (and hence repulsion) between the coordinates of $\mathbf{x}_{k,d}$ is introduced via functions that model interactions between them. More formally, consider $\varphi_1 : \mathbb{R}^d \rightarrow [-\infty, \infty]$ a measurable function and $\varphi_2 : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [-\infty, \infty]$ a measurable and symmetric function. Define

$$\nu_G \left(\prod_{i=1}^k A_i \right) = \int_{\prod_{i=1}^k A_i} \exp \left\{ - \sum_{i=1}^k \varphi_1(\mathbf{x}_i) - \sum_{r < s}^k \varphi_2(\mathbf{x}_r, \mathbf{x}_s) \right\} d\mathbf{x}_{k,d}, \quad (2.3.1)$$

where $\prod_{i=1}^k A_i$ is the cartesian product of Borel sets A_1, \dots, A_k in \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 \mathbf{x}_i while φ_2 controls the interaction between pairs of coordinates \mathbf{x}_r and \mathbf{x}_s . If φ_1 and φ_2 are selected so that $\nu_G(\mathbb{R}_k^d)$ 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 (2.3.1), is called a (second order) Gibbs measure. The normalizing

constant (total mass of \mathbb{R}_k^d under ν_G)

$$\nu_G(\mathbb{R}_k^d) = \int_{\mathbb{R}_k^d} \exp \left\{ - \sum_{i=1}^k \varphi_1(\mathbf{x}_i) - \sum_{r < s}^k \varphi_2(\mathbf{x}_r, \mathbf{x}_s) \right\} d\mathbf{x}_{k,d}$$

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 $\mathbf{x}_{k,d}$. In general, $\nu_G(\mathbb{R}_k^d)$ is (almost) intractable mainly because of the presence of φ_2 .

Note that symmetry of φ_2 (i.e., $\varphi_2(\mathbf{x}_r, \mathbf{x}_s) = \varphi_2(\mathbf{x}_s, \mathbf{x}_r)$) means that ν_G defines a symmetric measure. This implies that the order of coordinates is immaterial. If $\varphi_2 = 0$ 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 $\varphi_2(\mathbf{x}, \mathbf{y})$ only depends on the relative distance between \mathbf{x} and \mathbf{y} (Daley and Vere-Jones 2002). More formally, let $\rho : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$ be a metric on \mathbb{R}^d and $\phi : [0, \infty) \rightarrow [-\infty, \infty]$ 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 $\varphi_2(\mathbf{x}, \mathbf{y}) = \phi\{\rho(\mathbf{x}, \mathbf{y})\}$, interactions will be smooth if, for example, $\phi \in C([0, \infty); [-\infty, \infty])$. Following this general idea, Petralia et al. (2012) use $\phi(r) = \tau(1/r)^\nu : \tau, \nu \in (0, \infty)$ 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 $\phi(r) = \infty \mathbb{I}_{[0,b]}(r) : b \in (0, \infty)$ (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 \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 $\mathbf{x}_{k,d}$ 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 parameters which inform the types of patterns produced. It would therefore be natural to estimate these parameters using observed data. However, $\nu_G(\mathbb{R}_k^d)$ is typically a function of the unknown parameters which makes deriving closed form expressions of $\nu_G(\mathbb{R}_k^d)$ practically impossible and renders Bayesian or frequentist estimation procedures intractable. To avoid this complication, pseudo-maximum likelihood methods have been proposed to approximate $\nu_G(\mathbb{R}_k^d)$ when carrying out estimation (Ogata and Tanemura 1981; Penttinen 1984). We provide details of a Bayesian approach in subsequent sections.

2.3.2 $\text{Rep}_{k,d}(f_0, C_0, \rho)$ Distribution

As mentioned, our principal objective is to construct a family of probability densities for $\mathbf{x}_{k,d}$ 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

$$\phi(r) = -\log\{1 - \exp(-cr^2)\} : c \in (0, \infty) \tag{2.3.2}$$

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 (2.3.2) with ν_G is straightforward: if we take

$$\varphi_2(\mathbf{x}, \mathbf{y}) = -\log[1 - C_0\{\rho(\mathbf{x}, \mathbf{y})\}], \quad C_0(r) = \exp(-cr^2) : c \in (0, \infty)$$

then ν_G will have a “pairwise-interaction term” given by

$$\exp \left\{ - \sum_{r < s}^k \varphi_2(\mathbf{x}_r, \mathbf{x}_s) \right\} = \prod_{r < s}^k [1 - C_0\{\rho(\mathbf{x}_r, \mathbf{x}_s)\}]. \quad (2.3.3)$$

The right-hand side of (2.3.3) induces a particular interaction structure that separates the coordinates of $\mathbf{x}_{k,d}$, 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 $C_0 : [0, \infty) \rightarrow (0, 1]$ that satisfy the following properties:

- A1. $C_0 \in C([0, \infty); (0, 1])$.
- A2. $C_0(0) = 1$.
- A3. $C_0(r) \rightarrow 0$ (right-side limit) when $x \rightarrow \infty$.
- A4. For all $r_1, r_2 \in [0, \infty)$, if $r_1 < r_2$ then $C_0(r_1) > C_0(r_2)$.

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 $\mathbf{x}_{k,d}$.

Lemma 2.3.1. *Given a metric $\rho : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$ such that singletons are not open sets in the topology induced by ρ , the function $R_C : \mathbb{R}_k^d \rightarrow [0, 1)$ defined by*

$$R_C(\mathbf{x}_{k,d}) = \prod_{r < s}^k [1 - C_0\{\rho(\mathbf{x}_r, \mathbf{x}_s)\}] \quad (2.3.4)$$

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

Through out the article we will refer to (2.3.4) as the *repulsive component*. We finish the construction of repulsive probability measures by specifying a distribution supported on \mathbb{R}^d which will be common for all the coordinates of $\mathbf{x}_{k,d}$. Let $f_0 \in C(\mathbb{R}^d; (0, \infty))$ be a probability density function, then under $\varphi_1(\mathbf{x}) = -\log\{f_0(\mathbf{x})\}$, ν_G will have a “base component term” given by

$$\exp \left\{ -\sum_{i=1}^k \varphi_1(\mathbf{x}_i) \right\} = \prod_{i=1}^k f_0(\mathbf{x}_i). \quad (2.3.5)$$

Incorporating (2.3.3) and (2.3.5) into (2.3.1) we get

$$\nu_G \left(\prod_{i=1}^k A_i \right) = \int_{\prod_{i=1}^k A_i} \left\{ \prod_{i=1}^k f_0(\mathbf{x}_i) \right\} R_C(\mathbf{x}_{k,d}) d\mathbf{x}_{k,d}.$$

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

Proposition 2.3.2. *Let $f_0 \in C(\mathbb{R}^d; (0, \infty))$ be a probability density function. The function*

$$g_{k,d}(\mathbf{x}_{k,d}) = \left\{ \prod_{i=1}^k f_0(\mathbf{x}_i) \right\} R_C(\mathbf{x}_{k,d}) \quad (2.3.6)$$

is measurable and integrable for all $d \in \mathbb{N}$ and $k \in \mathbb{N}$ ($k \geq 2$).

With Proposition 2.3.2 it is now straightforward to construct a probability measure with the desired repulsive structure; small relative distances are penalized in a smooth way. Notice

that the support of (2.3.6) 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 (2.3.6) defines a valid joint probability density function which we now provide.

Definition 2.3.1. The probability distribution $\text{Rep}_{k,d}(f_0, C_0, \rho)$ has probability density function

$$\text{Rep}_{k,d}(\mathbf{x}_{k,d}) = \frac{1}{c_{k,d}} \left\{ \prod_{i=1}^k f_0(\mathbf{x}_i) \right\} R_C(\mathbf{x}_{k,d}), \quad (2.3.7)$$

$$c_{k,d} = \int_{\mathbb{R}_k^d} \left\{ \prod_{i=1}^k f_0(\mathbf{x}_i) \right\} R_C(\mathbf{x}_{k,d}) d\mathbf{x}_{k,d}. \quad (2.3.8)$$

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

2.3.3 $\text{Rep}_{k,d}(f_0, C_0, \rho)$ Properties

In this section we will investigate a few general properties of the $\text{Rep}_{k,d}(f_0, C_0, \rho)$ class. The distributional results are provided to further understanding regarding characteristics of (2.3.7) from a qualitative and analytic point of view. As a first observation, because of symmetry, $\text{Rep}_{k,d}(\mathbf{x}_{k,d})$ is an exchangeable distribution in $\mathbf{x}_1, \dots, \mathbf{x}_k$. This facilitates the study of computational techniques motivated by $\text{Rep}_{k,d}(f_0, C_0, \rho)$. However, it is worth noting that $\{\text{Rep}_{k,d}(f_0, C_0, \rho)\}_{k \geq 2}$ does not induce a sample-size consistent sequence of finite-dimensional distributions, meaning that

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

This makes predicting locations of new coordinates problematic. In Section 2.4 we address how this may be accommodated in modeling contexts. To simplify notation, in what follows we will use $[m] = \{1, \dots, m\}$, with $m \in \mathbb{N}$.

Normalizing Constant

Because $R_C(\mathbf{x}_{k,d})$ is invariant under permutations of the coordinates of $\mathbf{x}_{k,d}$, an interaction's direction is immaterial to whether it is present or absent (i.e., \mathbf{x}_r interacts with \mathbf{x}_s if and only if \mathbf{x}_s interacts with \mathbf{x}_r). Therefore it is sufficient to represent the interaction between \mathbf{x}_r and \mathbf{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 $\mathbf{x}_{k,d}$ and $\ell_k = \text{card}(I_k) = \frac{k(k-1)}{2}$, where $\text{card}(E)$ is the cardinality of a set E . Now, expanding (2.3.4) term-by-term results in

$$R_C(\mathbf{x}_{k,d}) = 1 + \sum_{l=1}^{\ell_k} (-1)^l \sum_{\substack{A \subseteq I_k \\ \text{card}(A)=l}} \left[\prod_{(r,s) \in A} C_0\{\rho(\mathbf{x}_r, \mathbf{x}_s)\} \right]. \quad (2.3.9)$$

The right-side of (2.3.9) is connected to graph theory in the following way: $A \subseteq I_k$ can be interpreted as a non-directed graph whose edges are $(r, s) \in A$.

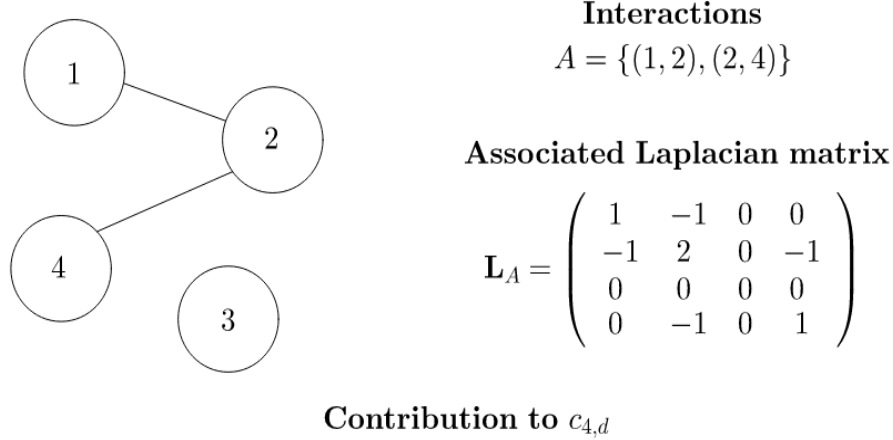
Using (2.3.9), it can be shown that expression (2.3.8) in Definition 2.3.1 has the following form:

$$c_{k,d} = 1 + \sum_{l=1}^{\ell_k} (-1)^l \sum_{\substack{A \subseteq I_k \\ \text{card}(A)=l}} \Psi_{k,d}(A) \quad (2.3.10)$$

$$\Psi_{k,d}(A) = \int_{\mathbb{R}_k^d} \left\{ \prod_{i=1}^k f_0(\mathbf{x}_i) \right\} \left[\prod_{(r,s) \in A} C_0\{\rho(\mathbf{x}_r, \mathbf{x}_s)\} \right] d\mathbf{x}_{k,d}. \quad (2.3.11)$$

Note that representing A as a graph or Laplacian matrix can help develop intuition on how each summand contributes to the expression (2.3.10). Figure 2.2 shows one particular case of how 3 of $k = 4$ coordinates in \mathbb{R}^d might interact by providing the respective Laplacian

matrix together with the contribution that (2.3.11) brings to calculating $c_{4,d}$ according to (2.3.10).



$$\Psi_{4,d}(A) = \int_{\mathbb{R}_4^d} f_0(\mathbf{x}_1)f_0(\mathbf{x}_2)f_0(\mathbf{x}_3)f_0(\mathbf{x}_4)C_0\{\rho(\mathbf{x}_1, \mathbf{x}_2)\}C_0\{\rho(\mathbf{x}_2, \mathbf{x}_4)\}d\mathbf{x}_{4,d}$$

Figure 2.2: The graph and Laplacian matrix for a possible interaction for $k = 4$ coordinates.

Equation (2.3.10) retains connections with the probabilistic version of the Inclusion-Exclusion Principle. This result, which is very useful in Enumerative Combinatorics, says that in any probability space $(\Omega, \mathcal{F}, \mathbb{P})$

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

with A_1, \dots, A_k events on \mathcal{F} and A_i^c denoting the complement of A_i . With this in mind, $c_{k,d}$ is the result of adding/subtracting all the contributions $\Psi_{k,d}(A)$ that emerge for every non-empty set $A \subseteq I_k$. 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 $\text{Rep}_{k,d}(f_0, C_0, \rho)$ distribution is from the (unattainable) extreme case $C_0 = 0$ (i.e., the coordinates $\mathbf{x}_1, \dots, \mathbf{x}_k$ 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 (2.3.11) becomes prohibitive as it requires carrying out (at least) $2^{\ell_k} - 1$ numerical calculations. In Subsection 2.4.1 we highlight a particular choice of f_0 , C_0 and ρ that produces a closed form expression for (2.3.11).

2.4 Gaussian Mixture Models and $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ 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.

2.4.1 Repulsive Gaussian Mixture Models (RGMM)

Consider $n \in \mathbb{N}$ experimental units whose responses $\mathbf{y}_1, \dots, \mathbf{y}_n$ 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 j th cluster ($j \in [k]$) is modeled through a Gaussian density $N_d(\cdot; \boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j)$ 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 $d \times d$. We let $\boldsymbol{\theta}_{k,d} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k) \in \mathbb{R}_k^d$ and $\boldsymbol{\Lambda}_{k,d} = (\boldsymbol{\Lambda}_1, \dots, \boldsymbol{\Lambda}_k) \in \mathbb{S}_k^d$ where \mathbb{S}_k^d is the k -fold product space of \mathbb{S}^d . Next let $\boldsymbol{\pi}_{k,1} = (\pi_1, \dots, \pi_k) \in \Delta_{k-1}$, where Δ_{k-1} is the standard $(k-1)$ -simplex ($\Delta_0 = \{1\}$), denote a set of weights that reflect the probability of allocating $\mathbf{y}_i : i \in [n]$ to a cluster. Then the standard Gaussian Mixture Model

is

$$\mathbf{y}_i \mid \boldsymbol{\pi}_{k,1}, \boldsymbol{\theta}_{k,d}, \boldsymbol{\Lambda}_{k,d} \stackrel{i.i.d.}{\sim} \sum_{j=1}^k \pi_j \text{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j). \quad (2.4.1)$$

It is common to restate (2.4.1) by introducing latent cluster membership indicators $z_1, \dots, z_n \in [k]$ such that \mathbf{y}_i is drawn from the j th mixture component if and only if $z_i = j$:

$$\mathbf{y}_i \mid z_i, \boldsymbol{\theta}_{k,d}, \boldsymbol{\Lambda}_{k,d} \stackrel{ind.}{\sim} \text{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_{z_i}, \boldsymbol{\Lambda}_{z_i}) \quad (2.4.2)$$

$$z_i \mid \boldsymbol{\pi}_{k,1} \stackrel{i.i.d.}{\sim} \mathbb{P}(z_i = j) = \pi_j. \quad (2.4.3)$$

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

Specifying a prior distribution for $k \in \mathbb{N}$ is possible. For example, DPM models by construction induce a prior distribution on the number of clusters k . Alternatively, Reversible Jump MCMC (Green 1995; Richardson and Green 1997) or Birth-Death Chains (Stephens 2000) could be employed after assigning a particular prior for k . These methods do not translate well to the non-i.i.d. case and so we employ a case-specific upper bound $k \geq 2$.

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 (2.4.1) 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

$$\boldsymbol{\theta}_{k,d} \sim \text{Rep}_{k,d}(f_0, C_0, \rho)$$

$$f_0(\mathbf{x}) = N_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{S}^d \quad (2.4.4)$$

$$C_0(r) = \exp(-0.5\tau^{-1}r^2) : \tau \in (0, \infty) \quad (2.4.5)$$

$$\rho(\mathbf{x}, \mathbf{y}) = \{(\mathbf{x} - \mathbf{y})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{y})\}^{1/2}. \quad (2.4.6)$$

The specific forms of f_0 , C_0 and ρ are admissible according to Definition 2.3.1. The repulsive distribution parameterized by (2.4.4)–(2.4.6) 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 $(\mathbf{y}_{n,d}, \mathbf{z}_{n,1}, \boldsymbol{\Lambda}_{k,d})$, with $\mathbf{y}_{n,d} = (\mathbf{y}_1, \dots, \mathbf{y}_n) \in \mathbb{R}_n^d$ and $\mathbf{z}_{n,1} = (z_1, \dots, z_n) \in [k]^n$. The parameter τ in (2.4.5) controls the strength of repulsion associated with coordinates in $\boldsymbol{\theta}_{k,d}$ via (2.4.6): as $\tau \rightarrow 0$ (right-side limit), the repulsion becomes weaker. The selection of (2.4.4) mimics the usual i.i.d. multivariate normal assumption.

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

$$\mathbf{y}_i \mid z_i, \boldsymbol{\theta}_{k,d}, \boldsymbol{\Lambda}_{k,d} \stackrel{i.i.d.}{\sim} N_d(\mathbf{y}_i; \boldsymbol{\theta}_{z_i}, \boldsymbol{\Lambda}_{z_i}) \quad (2.4.7)$$

$$z_i \mid \boldsymbol{\pi}_{k,1} \stackrel{i.i.d.}{\sim} \mathbb{P}(z_i = j) = \pi_j \quad (2.4.8)$$

together with the following mutually independent prior distributions:

$$\boldsymbol{\pi}_{k,1} \sim \text{Dir}(\boldsymbol{\alpha}_{k,1}) : \boldsymbol{\alpha}_{k,1} \in (0, \infty)^k \quad (2.4.9)$$

$$\boldsymbol{\theta}_{k,d} \sim \text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau) : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{S}^d, \tau \in (0, \infty) \quad (2.4.10)$$

$$\boldsymbol{\Lambda}_j \stackrel{i.i.d.}{\sim} \text{IW}_d(\boldsymbol{\Psi}, \nu) : \boldsymbol{\Psi} \in \mathbb{S}^d, \nu \in (0, \infty). \quad (2.4.11)$$

In what follows we will refer to the model in (2.4.7)–(2.4.11) as the (Bayesian) Repulsive Gaussian Mixture Model (abbreviated as RGMM).

Parameter Calibration

We briefly discuss strategies of selecting values for parameters that control the prior distributions in (2.4.9)–(2.4.11). We select values for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and τ of the RGMM instead of treating them as unknown and assigning them hyperprior distributions because of computational cost. First notice that $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ acts as a location/scale parameter: if $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\top$ is the corresponding Cholesky decomposition for $\boldsymbol{\Sigma}$, then $\boldsymbol{\theta}_{k,d} \sim \text{NRep}_{k,d}(\mathbf{0}_d, \mathbf{I}_d, \tau)$ implies that

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

where \mathbf{I}_d is the $d \times d$ identity matrix and $\mathbf{0}_d, \mathbf{1}_d \in \mathbb{R}^d$ are d -dimensional vectors of zeroes and ones, respectively. Although a Gaussian hyperprior for $\boldsymbol{\mu}$ is a reasonable candidate (the full conditional distribution is also Gaussian), it is not straightforward how to select its associated hyperparameters. A slightly more complicated problem occurs with $\boldsymbol{\Sigma}$, since this parameter participates in the repulsive component and no closed form is available for its posterior distribution. Even more problematic, the induced full conditional distribution for τ turns out to be doubly-intractable (Murray et al. 2006) and as a result the standard MCMC algorithms do not apply. To see this, it can be shown using (2.3.10), (2.3.11) and the Gaussian integral that the normalizing constant of $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ 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 is the $k \times k$ identity matrix, \mathbf{L}_A denotes the Laplacian matrix associated to the set of interactions $A \subseteq I_k$ (see Subsection 2.3.3) and \otimes is the matrix Kronecker product, making

it a function of τ .

To facilitate hyperparameter selection we standardize the \mathbf{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 $\boldsymbol{\alpha}_{k,1} = k^{-1}\mathbf{1}_d$ produces a weakly informative prior for $\boldsymbol{\pi}_{k,1}$. 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 $\psi \in (0, \infty)$ 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 $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k$ are a random sample coming from $N_d(\mathbf{0}_d, \mathbf{I}_d)$. To favor separation among these random vectors we can use (2.4.5) and (2.4.6) with $\boldsymbol{\Sigma} = \mathbf{I}_d$ to choose τ such that for all $r \neq s \in [k]$

$$\mathbb{P}[1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\} \leq u] = p,$$

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

$$\mathbb{P}\{G \leq w(u)\tau\} = p, \quad G = \frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s) \sim G(d/2, 1/2). \quad (2.4.12)$$

Creating a grid of points in $(0, \infty)$ it is straightforward to find a τ that fulfills criterion (2.4.12). 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 (2.4.10), thus making computation numerically more stable. In

practice, we apply the procedure outlined above to the vectors coming from the repulsive distribution (2.4.10), treating them as if they were sampled from a multivariate Gaussian distribution. This gives us a simple procedure to approximately achieve the desired goal of prior separation with a pre-specified probability.

2.4.2 Theoretical Properties

In this section we explore properties associated with the support and posterior consistency of (2.4.1) under (2.4.9)–(2.4.11). 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 $\boldsymbol{\xi}_k = \boldsymbol{\pi}_{k,1} \times \boldsymbol{\theta}_{k,1} \times \{\lambda\} = (\pi_1, \dots, \pi_k) \times (\theta_1, \dots, \theta_k) \times \{\lambda\}$, $\boldsymbol{\Theta}_k = \Delta_{k-1} \times \mathbb{R}_k^1 \times (0, \infty)$ and

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

Let $B_p(\mathbf{x}, r)$ with $\mathbf{x} \in \mathbb{R}_k^1$ and $r \in (0, \infty)$ denote an open ball centered on \mathbf{x} , and with radius r , and $D_p(\mathbf{x}, r)$ its closure relative to the Euclidean L_p -metric ($p \geq 1$) on \mathbb{R}_k^1 .

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

- B1. The true data generating density $f_0(\cdot; \boldsymbol{\xi}_{k_0}^0)$ belongs to \mathcal{F}_{k_0} for some fixed $k_0 \geq 2$, where $\boldsymbol{\xi}_{k_0}^0 = \boldsymbol{\pi}_{k_0,1}^0 \times \boldsymbol{\theta}_{k_0,1}^0 \times \{\lambda_0\} = (\pi_1^0, \dots, \pi_{k_0}^0) \times (\theta_1^0, \dots, \theta_{k_0}^0) \times \{\lambda_0\}$.
- B2. The true locations $\theta_1^0, \dots, \theta_{k_0}^0$ satisfy $\min(|\theta_r^0 - \theta_s^0| : r \neq s \in [k_0]) \geq v$ 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 $\boldsymbol{\xi}_k \sim \text{Dir}(k^{-1}\mathbf{1}_k) \times \text{NRep}_{k,1}(\mu, \sigma^2, \tau) \times \text{IG}(a, b)$. In the case that $k = 1$, $\boldsymbol{\xi}_k \sim \delta_1 \times \mathcal{N}(\mu, \sigma^2) \times \text{IG}(a, b)$ with δ_1 a Dirac measure centred on 1. In both scenarios $\mu \in \mathbb{R}$ and $\sigma^2, \tau, a, b \in (0, \infty)$ are fixed values.

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 Θ_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

$$\boldsymbol{\xi} \mid K = k \sim \boldsymbol{\xi}_k, \quad K \sim \kappa. \quad (2.4.13)$$

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, for all $\varepsilon > 0$

$$\Pi \left\{ \left(f \in \mathcal{F} : \int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_*)} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx < \varepsilon \right) \right\} > 0, \quad (2.4.14)$$

where $\boldsymbol{\xi}_* \in \bigcup_{k=1}^{\infty} \Theta_k$. Condition (2.4.14) 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 2.4.1. *Under condition B1, let $\varepsilon > 0$. Then there exists $\delta > 0$ such that*

$$\int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_{k_0})} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx < \varepsilon$$

for all $\boldsymbol{\xi}_{k_0} \in B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta) \times B_1(\boldsymbol{\pi}_{k_0,1}^0, \delta) \times (\lambda_0 - \delta, \lambda_0 + \delta)$.

Lemma 2.4.2. *Assume condition B2 and let $\boldsymbol{\theta}_{k_0,1} \sim \text{NRep}_{k_0,1}(\mu, \sigma^2, \tau)$. Then there exists $\delta_0 > 0$ such that*

$$\mathbb{P}\{\boldsymbol{\theta}_{k_0,1} \in B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta)\} > 0.$$

for all $\delta \in (0, \delta_0]$. This result remains valid even when replacing $B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta)$ with $D_1(\boldsymbol{\theta}_{k_0,1}^0, \delta)$.

Using Lemmas 2.4.1 and 2.4.2 we are able to prove the following Proposition.

Proposition 2.4.3. *Assume that conditions B1–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 2.4.4. *For each $k \geq 2$ the coordinates of $\boldsymbol{\theta}_{k,1} \sim \text{NRep}_{k,1}(\mu, \sigma^2, \tau)$ share the same functional form. Moreover, there exists $\gamma \in (0, \infty)$ such that*

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

for all $t \in [\gamma, \infty)$ and $i \in [k]$. Here, $c_k = c_{k,1}$ is the normalizing constant of $\text{NRep}_{k,1}(\mu, \sigma^2, \tau)$ with $c_1 = 1$.

Lemma 2.4.5. *The sequence $\{c_k : k \in \mathbb{N}\}$ defined in Lemma 2.4.4 satisfies*

$$0 < \frac{c_{k-1}}{c_k} \leq A_1 \exp(A_2 k)$$

for all $k \in \mathbb{N}$ ($k \geq 2$) and some constants $A_1, A_2 \in (0, \infty)$.

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 Lemmas 2.4.4 and 2.4.5, we now state the following

Proposition 2.4.6. *Assume that conditions B1, B2 and B4 hold. Replace condition B3 with:*

B3'. There exists $B_1 \in (0, \infty)$ such that for all $k \in \mathbb{N}$, $0 < \kappa(k) \leq B_1 \exp\{-B_2 k\}$, where $B_2 > A_2$ and $A_2 \in (0, \infty)$ is given by Lemma 2.4.5.

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

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

Here we describe an algorithm that can be used to sample from $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$. Upon introducing component labels, sampling marginally from the joint posterior distribution of $\boldsymbol{\theta}_{k,d}$, $\boldsymbol{\Lambda}_{k,d}$, $\boldsymbol{\pi}_{k,1}$ and $\mathbf{z}_{n,1}$ can be done with a Gibbs sampler. However, 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 from $\text{NRep}_{k,d}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ via a Metropolis–Hastings step inside the Gibbs sampling scheme. In Appendix A we detail the entire MCMC algorithm (Algorithm RGMM), but here we focus on the nonstandard aspects.

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

$$(\boldsymbol{\theta}_{k,d} \mid \cdots) \propto \left\{ \prod_{j=1}^k \text{N}_d(\boldsymbol{\theta}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right\} \prod_{r < s}^k [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\}]$$

where $\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Lambda}_j^{-1}\mathbf{s}_j)$, $\mathbf{s}_j = \sum_{i=1}^n \mathbb{I}_{\{j\}}(z_i)\mathbf{y}_i$, $\boldsymbol{\Sigma}_j = (\boldsymbol{\Sigma}^{-1} + n_j\boldsymbol{\Lambda}_j^{-1})^{-1}$ and $n_j = \text{card}(i \in [n] : z_i = j)$. Now, the complete conditional distributions $(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \cdots)$ for $j \in [k]$

and $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_l : l \neq j) \in \mathbb{R}_{k-1}^d$, have the following form

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

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

1. Let $\boldsymbol{\theta}_{k,d}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_k^{(0)}) \in \mathbb{R}_k^d$ be the actual state for $\boldsymbol{\theta}_{k,d}$.
2. For $j = 1, \dots, k$:
 - (a) Generate a candidate $\boldsymbol{\theta}_j^{(1)}$ from $N_d(\boldsymbol{\theta}_j^{(0)}, \boldsymbol{\Gamma}_j)$ with $\boldsymbol{\Gamma}_j \in \mathbb{S}^d$.
 - (b) Set $\boldsymbol{\theta}_j^{(0)} = \boldsymbol{\theta}_j^{(1)}$ with probability $\min(1, \beta_j)$, where

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

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

$$\boldsymbol{\Gamma}_j = \frac{1}{B} \sum_{t=1}^B \{\boldsymbol{\Sigma}^{-1} + n_j^{(t)} (\boldsymbol{\Lambda}_j^{(t)})^{-1}\}^{-1} : n_j^{(t)} = \text{card}(i \in [n] : z_i^{(t)} = j), \quad (2.4.15)$$

where $t \in [B]$ is the t th iteration of the burn-in period with length $B \in \mathbb{N}$.

2.5 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. Gaussian Mixture Model (GMM) and a Dirichlet Process Gaussian Mixture Model (DPMM). This is done by treating the following as a data generating mechanism:

$$y \sim f_0 = 0.3N(-5, 1.0^2) + 0.05N(0, 0.3^2) + 0.25N(1, 0.3^2) + 0.4N(4, 0.8^2). \quad (2.5.1)$$

Using (2.5.1) we simulate 100 data sets with sample sizes 500, 1000 and 5000. For each of these scenarios, we compare the following 4 models (abbreviated by M1, M2, M3 y M4) to estimate f_0 :

M1. GMM corresponding to (2.4.7)–(2.4.8) with prior distributions given by (2.4.9)–(2.4.11), replacing (2.4.10) by $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k \stackrel{i.i.d.}{\sim} N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In this case:

- $k = 10, d = 1, \boldsymbol{\alpha}_{k,1} = 10^{-1}\mathbf{1}_{10}, \boldsymbol{\mu} = 0, \boldsymbol{\Sigma} = 1, \boldsymbol{\Psi} = 0.06$ 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 2.4.1 and setting $u = 0.5$ and $p = 0.95$. The remaining prior parameters are:

- $k = 10, d = 1, \boldsymbol{\alpha}_{k,1} = 10^{-1}\mathbf{1}_{10}, \boldsymbol{\mu} = 0, \boldsymbol{\Sigma} = 1, \tau = 5.45, \boldsymbol{\Psi} = 0.06$ 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 2.4.1 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:

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

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

M4. DPMM given by:

$$\mathbf{y}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \stackrel{ind.}{\sim} N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (2.5.2)$$

$$(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid H \stackrel{i.i.d.}{\sim} H \quad (2.5.3)$$

$$H \mid \alpha, H_0 \sim DP(\alpha, H_0) \quad (2.5.4)$$

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

$$H_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N_d(\boldsymbol{\mu}; \mathbf{m}_1, k_0^{-1}\boldsymbol{\Sigma}) IW_d(\boldsymbol{\Sigma}; \boldsymbol{\Psi}_1, \nu_1) : \nu_1 \in (0, \infty). \quad (2.5.5)$$

To complete the model specification given by (2.5.2)–(2.5.5), the following independent hyperpriors are assumed:

$$\alpha \mid a_0, b_0 \sim G(a_0, b_0) : a_0, b_0 \in (0, \infty) \quad (2.5.6)$$

$$\mathbf{m}_1 \mid \mathbf{m}_2, \mathbf{S}_2 \sim N_d(\mathbf{m}_2, \mathbf{S}_2) : \mathbf{m}_2 \in \mathbb{R}^d, \mathbf{S}_2 \in \mathbb{S}^d \quad (2.5.7)$$

$$k_0 \mid \tau_1, \tau_2 \sim G(\tau_1/2, \tau_2/2) : \tau_1, \tau_2 \in (0, \infty) \quad (2.5.8)$$

$$\boldsymbol{\Psi}_1 \mid \boldsymbol{\Psi}_2, \nu_2 \sim IW_d(\boldsymbol{\Psi}_2, \nu_2) : \boldsymbol{\Psi}_2 \in \mathbb{S}^d, \nu_2 \in (0, \infty). \quad (2.5.9)$$

In the simulation study we set $d = 1$. The selection of hyperparameters found in (2.5.6)–(2.5.9) was based on similar strategies as outlined in Escobar and West (1995) which produced:

- $a_0 = 2$, $b_0 = 5$, $\nu_1 = 4$, $\nu_2 = 4$, $\mathbf{m}_2 = 0$, $\mathbf{S}_2 = 1$, $\boldsymbol{\Psi}_2 = 1$, $\tau_1 = 2.01$ 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**. 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 occupied mixture components, we recorded the following numeric indicators:

- Average number of occupied mixture components.
- Standard deviation of the average number of occupied mixture components.

Figures 2.3, 2.4 and 2.5 contain side-by-side boxplots of the LPML, MSE and L_1 -metric respectively as the sample size grows. Notice that trends seen here indicate that M1 and M4 tend to fit better, but M2 and M3 are very competitive with the advantage of being more parsimonious. In other words, very little model fit was sacrificed for the sake of parsimony.

Figures 2.6 and 2.7 show that the average number of occupied mixture 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 occupied

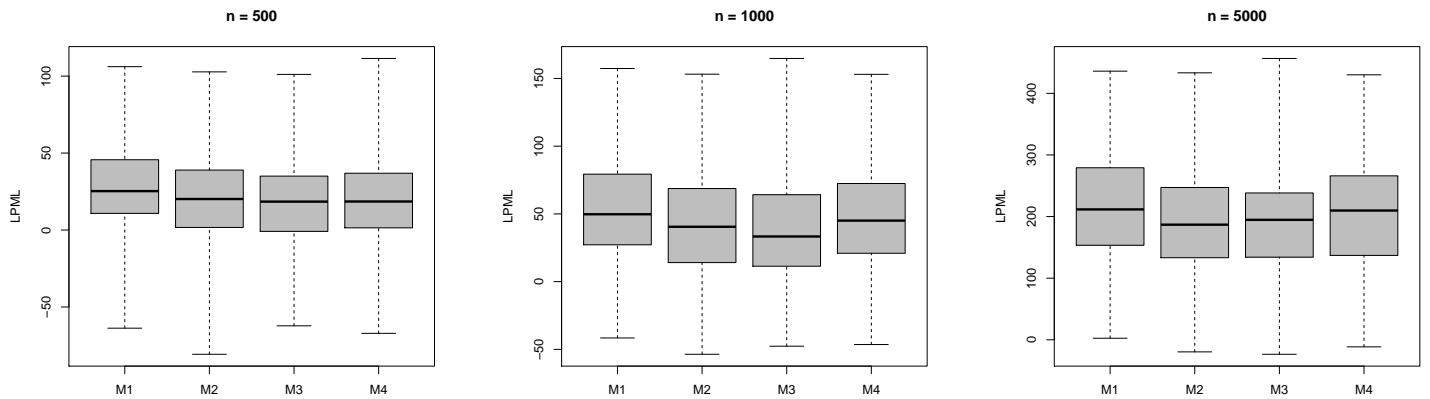


Figure 2.3: Boxplots that resume the behavior of LPML for each of the four models.

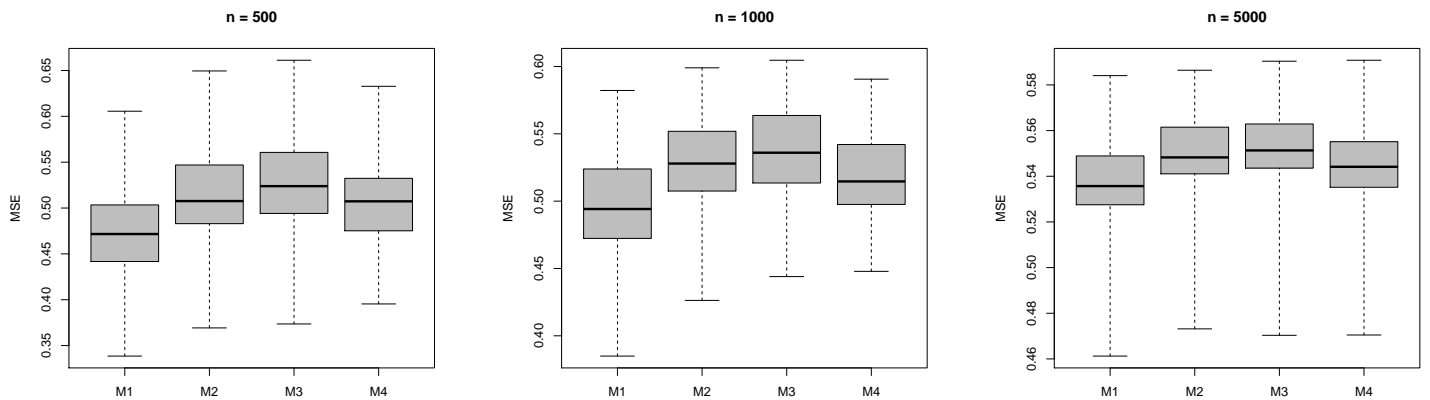


Figure 2.4: Boxplots that resume the behavior of MSE for each of the four models.

mixture components for M2 and M3 are also highly concentrated around 3, 4 and 5 (recall that the data were generated using a mixture of four components). Conversely, M1 and M4 require many more occupied mixture components to achieve the same goodness-of-fit, a trend that persists when the sample size grows.

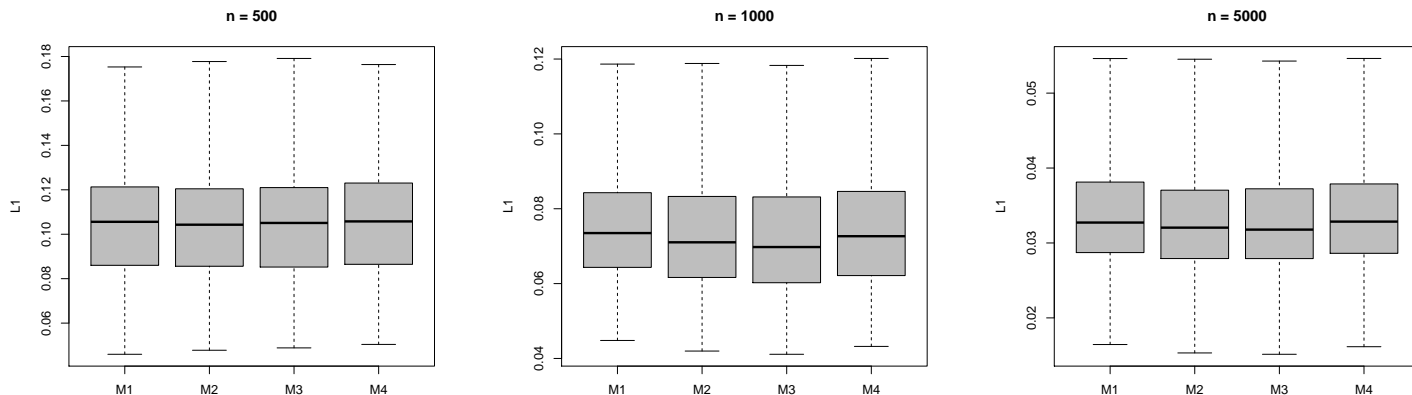


Figure 2.5: Boxplots that resume the behavior of L_1 -metric for each of the four models.

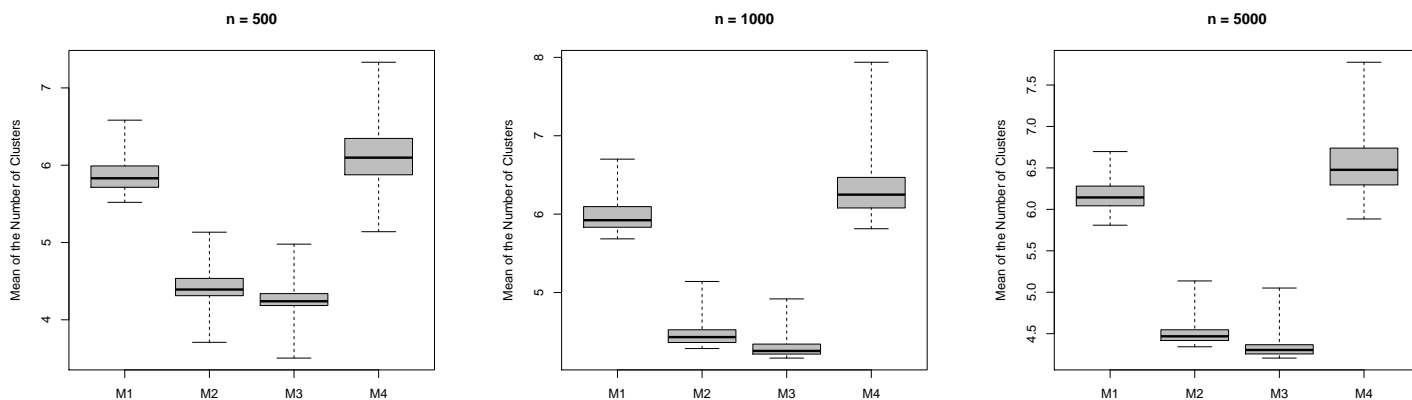


Figure 2.6: Side-by-side boxplots of the average number of occupied mixture components for each of the procedure.

2.6 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 DPMM to those from the RGMM. For each procedure we report the LPML as a measure of goodness-of-fit, a brief summary regarding

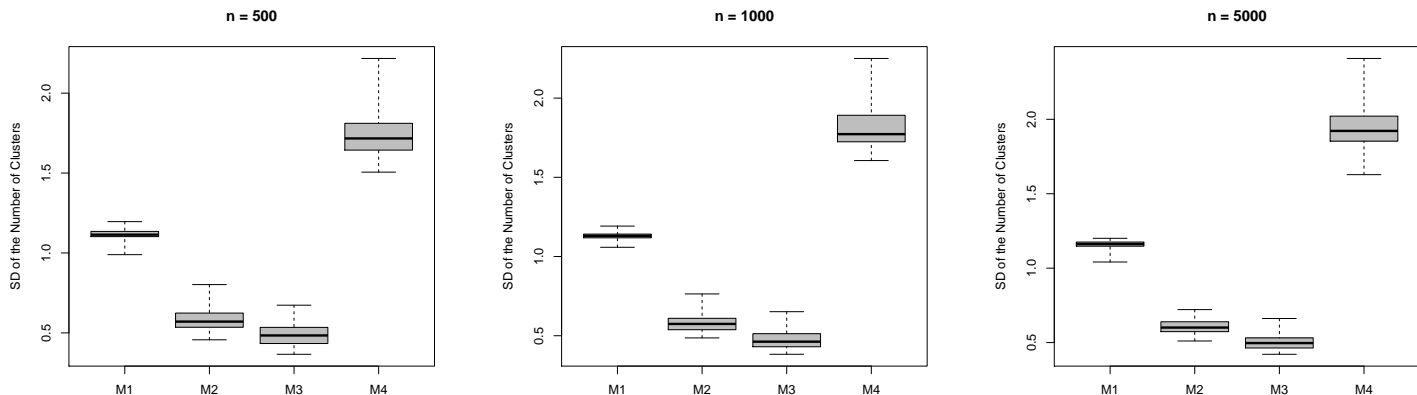


Figure 2.7: Side-by-side boxplots that display the average standard deviation associated with the posterior distribution of occupied mixture components for each of the four procedures.

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. DPMM: 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 (2.5.2)–(2.5.9). 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 model prior parameter values are now provided:

(a) Galaxy: $d = 1$, $a_0 = 2$, $b_0 = 2$, $\nu_1 = 4$, $\nu_2 = 4$, $\mathbf{m}_2 = \mathbf{0}$, $\mathbf{S}_2 = \mathbf{1}$, $\Psi_2 = 0.15$, $\tau_1 = 2.01$ and $\tau_2 = 1.01$.

(b) Air Quality: $d = 2$, $a_0 = 1$, $b_0 = 3$, $\nu_1 = 4$, $\nu_2 = 4$, $\mathbf{m}_2 = \mathbf{0}_2$, $\mathbf{S}_2 = \mathbf{I}_2$, $\Psi_2 = \mathbf{I}_2$, $\tau_1 = 2.01$ 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 2.4.1: $(u, p) = (0.5, 0.95)$ and $(u, p) = (0.05, 0.95)$ for Galaxy and Air Quality data respectively. Parameter selection for model components (2.4.9)–(2.4.11) were carried out according to the methods in Subsection 2.4.1. Specific details now follow:

- (a) Galaxy: $k = 10$, $d = 1$, $\boldsymbol{\alpha}_{k,1} = 10^{-1}\mathbf{1}_{10}$, $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \mathbf{1}$, $\tau = 5.45$, $\boldsymbol{\Psi} = 0.15$ and $\nu = 5$.
- (b) Air Quality: $k = 10$, $d = 2$, $\boldsymbol{\alpha}_{k,1} = 10^{-1}\mathbf{1}_{10}$, $\boldsymbol{\mu} = \mathbf{0}_2$, $\boldsymbol{\Sigma} = \mathbf{I}_2$, $\tau = 116.76$, $\boldsymbol{\Psi} = 3\mathbf{I}_2$ and $\nu = 6$.

Results of the fits are provided in Table 2.1. Notice that the fit associated with RGMM is better relative to the DPMM, which corroborates the argument that RGMM sacrifices no appreciable model fit for the sake of model parsimony. Figure 2.8 further reinforces the idea that RGMM is more parsimonious relative to DPMM. 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 DPMM. Graphs of the estimated densities (provided in Figure 2.9) show that the cost of parsimony is negligible as density estimates are practically the same.

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

Table 2.1: Summary statistics related to model fit and the number of clusters for Galaxy and Air Quality data based on DPMM and RGMM.

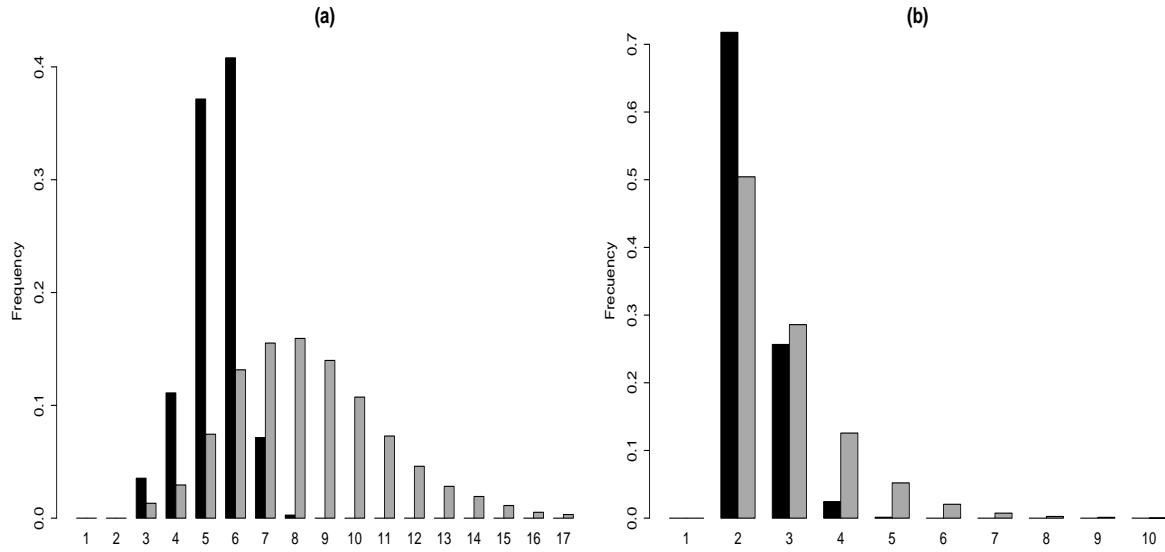


Figure 2.8: Posterior distribution for the active number of clusters in (a) Galaxy and (b) Air Quality data. Black (gray) bars correspond to RGMM (DPMM).

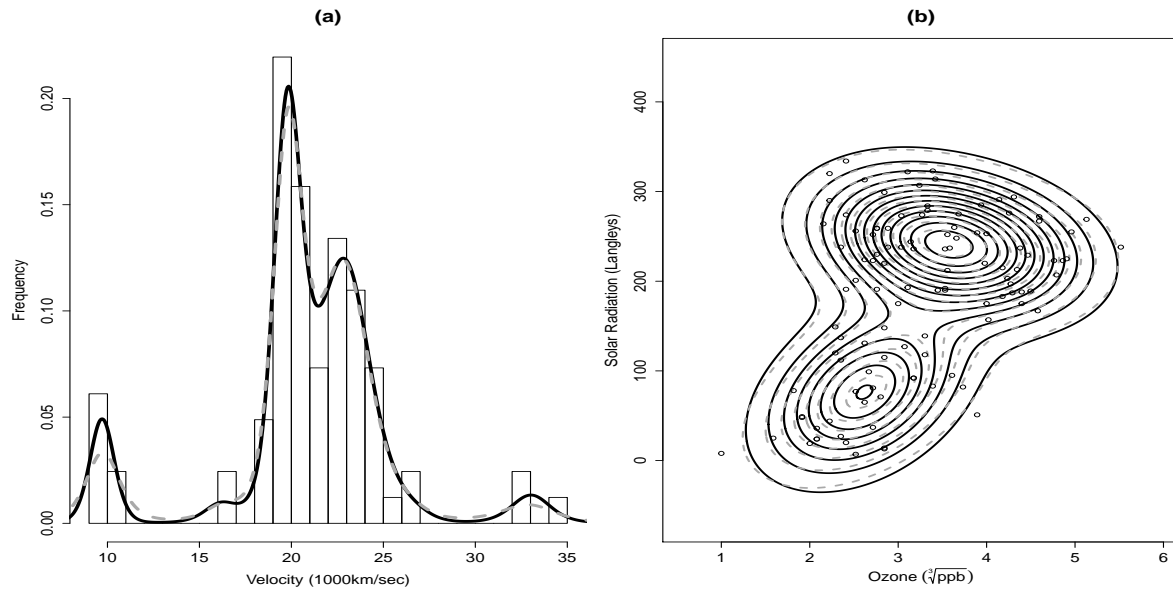


Figure 2.9: Posterior predictive densities for (a) Galaxy and (b) Air Quality data. Black solid (gray dashed) curves correspond to RGMM (DPMM).

Chapter 3

Regression Estimation using Repulsive Distributions

3.1 Chapter Overview

Flexible regression is a traditional motivation for the development of nonparametric Bayesian models. A popular approach for this involves a joint model for responses and covariates, from which the desired result arises by conditioning on the covariates. Many such models involve the convolution of a continuous kernel with a discrete random probability measures defined as an infinite mixture of i.i.d. atoms. Following this strategy, we propose a flexible model that involves the concept of repulsion between atoms. We show that this results in a more parsimonious representation of the regression than the i.i.d. counterpart. The key aspect is that repulsion discourages mixture components that are near each other, thus favoring parsimony. We show that the conditional model retains the repulsive features, thus facilitating interpretation of the resulting flexible regression, and with little or no sacrifice of model fit compared to the infinite mixture case. We show the utility of the methodology by way of small simulation study and a well known application.

3.2 Introduction

In Chapter 2 we defined a class of probability distributions whose coordinates are encouraged to be mutually separated, a property we refer to as *repulsion*. The origins of this class of distributions can be traced back to (finite) Gibbs Point Processes (Illian et al. 2008). The main motivation for developing the class of *repulsive distributions* was the desire to make Bayesian finite mixture models more parsimonious.

In Bayesian parametric and nonparametric hierarchical models, a common assumption is that parameters (atoms) in the latent level of a hierarchy are assumed to be mutually independent and sampled from a common distribution. This is particularly true for finite and infinite mixture models where these typically correspond to component (cluster) location parameters (see, for example, Frühwirth-Schnatter (2006) and Hjort et al. (2010)). A consequence of the independence assumption is the creation of a large number of clusters, often times making the models unnecessarily complex. This could result in overfitting which may negatively impact out-of-sample prediction and model interpretability. To counteract this, in Chapter 2 we showed how to use the class of repulsive distributions to model component location parameters in a Bayesian mixture model. Therefore, cluster locations are not modeled independently, but rather are encouraged to repel each other producing a more parsimonious mixture model. The key component of the probability law that produces repulsion is that small relative distances between the centers of the mixture components are penalized by way of a single parameter that controls the strength of the repulsion.

In many studies, it is common for researchers to collect additional covariate information on each experimental unit or subject. This is the case in the well known application that we consider in Section 3.5. These data consist of duration times of Old Faithful geyser eruptions and the waiting time until the next eruption occurs. Interest lies in being able to learn how time until the next eruption influences eruption duration. There are a number of meth-

ods developed in Bayesian nonparametrics that are available to model such data. Among them are approaches classified as nonparametric residual distributions, nonparametric mean functions, and fully nonparametric regression which is often times referred to as Bayesian density regression (Dunson et al. 2007). For more details and references, we direct the reader to Chapter 4 of Müller et al. (2015). A motivation for considering these methods is the desire to flexibly accommodate arbitrarily shaped mean curves associated with the distribution of the response given a covariate. However, flexibility comes at a cost as it is common that a large number of clusters are created to carry this out, many of which are redundant. The focus of this chapter then is to incorporate the same repulsive modeling ideas as considered in Chapter 2, but now include covariate information. Since repulsion in statistical models has only recently been studied (apart from Chapter 2 of this thesis, see Petralia et al. (2012), Xu et al. (2016), and Fúquene et al. (2016)), this will be the first attempt (as far as we know) of including covariate information in repulsive modeling. Because directly making the class of repulsive distributions covariate dependent renders computation intractable, our approach will be similar to what was done in Chapter 2. Specifically we will incorporate covariates in a Bayesian mixture model and then model component centers with a repulsive distribution.

The remainder of this chapter is organized as follows: in Section 3.3 we provide a concise description of dependent Gaussian Mixture Models for continuous covariates and discuss a novel Bayesian approach in which repulsion is introduced at a latent level in the joint distribution for responses and covariates (location parameters). We also provide guidance (similar to that found in Chapter 2) to calibrate the hyperparameters of our model. Details associated with posterior sampling are also provided in this section. Sections 3.4 and 3.5 illustrate the performance of our proposal applied to synthetic and real data sets. Computational strategies are provided in Appendix B

3.3 Covariate Dependent RGMM (RGMM_x)

Consider $n \in \mathbb{N}$ experimental units where on each the ordered pairs $(\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_n, \mathbf{x}_n)$ are recorded. In this case, $\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{R}^d$ are the responses and $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ are the corresponding subject-specific covariates. The motivation for considering $\mathbf{x}_1, \dots, \mathbf{x}_n$ is that they provide additional information regarding the distribution of $\mathbf{y}_1, \dots, \mathbf{y}_n$. It is common to assume mutual independence among the responses. However, assuming that responses are identically distributed is not tenable because of the dependence on covariates. The challenge is to model how the covariates guide the evolution of the mean response. This task can be solved by expressing it as a nonparametric Bayesian regression problem. In what follows, to simplify the notation, we will use $[m] = \{1, \dots, m\}$ with $m \in \mathbb{N}$.

The fundamental idea in the context of nonparametric Bayesian regression models is to estimate the average behavior of a response variable $\mathbf{y} \in \mathbb{R}^d$ as an (unknown) function of available covariates $\mathbf{x} \in \mathbb{R}^p$ for some $d, p \in \mathbb{N}$. Müller and Quintana (2004) provide a nice overview that details a number of possible Bayesian nonparametric regression approaches. Among those is the approach of Müller et al. (1996) which we adopt. With the flexibility of Gaussian Mixture Models to emulate smooth densities accurately in mind, they propose a statistical model that reduces regression estimation to a density estimation problem. Specifically they treat $\mathbf{u} = (\mathbf{y}, \mathbf{x}) \in \mathbb{R}^d \times \mathbb{R}^p$ as a random vector generated by a Dirichlet Process Gaussian Mixture Model (DPMM). The joint distribution is used to estimate the regression mean curve through the implied conditional distribution $(\mathbf{y} \mid \mathbf{x})$. To make these ideas concrete, consider the hierarchical model

$$\mathbf{u}_i \mid (\boldsymbol{\theta}_i, \boldsymbol{\Lambda}_i) \stackrel{i.i.d.}{\sim} N_{d+p}(\boldsymbol{\theta}_i, \boldsymbol{\Lambda}_i) \quad (3.3.1)$$

$$(\boldsymbol{\theta}_i, \boldsymbol{\Lambda}_i) \mid H \stackrel{i.i.d.}{\sim} H \quad (3.3.2)$$

$$H \mid \alpha, H_0 \sim \text{DP}(\alpha, H_0), \quad (3.3.3)$$

where the $\text{DP}(\alpha, H_0)$ denotes a Dirichlet Process with base measure H_0 (which is often times selected to be the conjugate Normal-Inverse-Wishart) and dispersion parameter $\alpha \in (0, \infty)$. Being that $\text{DP}(\alpha, H_0)$ is a discrete random probability measure (see Sethuraman (1994)), Müller et al. (1996) show that the posterior predictive conditional density for $(\mathbf{y} \mid \mathbf{x})$ takes the form of a locally weighted mixture of linear regressions, also known as WDDP (Weight Dependent Dirichlet Process). For more technical and computational details see Müller et al. (2015) and Jara et al. (2011).

In order to capture flexible mean structures the above proposal tends to produce a large number of covariate dependent clusters making the models unnecessary complex. To make the models more parsimonious, we propose a straightforward method similar to WDDP that introduces repulsion in the location parameters of the joint distribution for $\mathbf{u} = (\mathbf{y}, \mathbf{x})$. Thus, we will consider \mathbf{x} as a random quantity lying in \mathbb{R}^p and therefore the stochastic behavior of \mathbf{u} can be modeled on the product space $\mathbb{R}^d \times \mathbb{R}^p = \mathbb{R}^{d+p}$. Instead of employing a DPMM for \mathbf{u} , we use the following Gaussian Mixture Model (see Chapter 2 for justifications behind this model choice):

$$\mathbf{u} \mid \boldsymbol{\pi}_{k,1}, \boldsymbol{\theta}_{k,d+p}, \boldsymbol{\Lambda}_{k,d+p} \sim \sum_{j=1}^k \pi_j \mathbb{N}_{d+p}(\mathbf{u}; \boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j), \quad (3.3.4)$$

where $\boldsymbol{\pi}_{k,1} = (\pi_1, \dots, \pi_k) \in \Delta_{k-1}$, $\boldsymbol{\theta}_{k,d+p} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k) \in \mathbb{R}_k^{d+p} = \prod_{j=1}^k \mathbb{R}^{d+p}$ and $\boldsymbol{\Lambda}_{k,d+p} = (\boldsymbol{\Lambda}_1, \dots, \boldsymbol{\Lambda}_k) \in \mathbb{S}_k^{d+p} = \prod_{j=1}^k \mathbb{S}^{d+p}$. Here Δ_{k-1} is the standard $(k-1)$ -simplex ($\Delta_0 = \{1\}$) and \mathbb{S}^d is the space of real, symmetric and positive-definite matrices of dimension $d \times d$. Since \mathbf{u} follows a Gaussian Mixture Model, the conditional distribution $(\mathbf{y} \mid \mathbf{x})$ will also be a Gaussian Mixture Model. To see this, let \mathbf{u} be modeled as in (3.3.4) and consider the

following partitioned $\boldsymbol{\theta}_j$ and $\boldsymbol{\Lambda}_j$:

$$\boldsymbol{\theta}_j = \begin{pmatrix} \boldsymbol{\theta}_j^y \\ \boldsymbol{\theta}_j^x \end{pmatrix}, \quad \boldsymbol{\Lambda}_j = \begin{pmatrix} \boldsymbol{\Lambda}_j^{yy} & \boldsymbol{\Lambda}_j^{yx} \\ \boldsymbol{\Lambda}_j^{xy} & \boldsymbol{\Lambda}_j^{xx} \end{pmatrix}.$$

Using standard properties of the Gaussian distribution it can be shown that the conditional distribution implied by (3.3.4) corresponds to the following weighted Gaussian regression mixture

$$\mathbf{y} \mid \mathbf{x}, \boldsymbol{\pi}_{k,1}, \boldsymbol{\theta}_{k,d+p}, \boldsymbol{\Lambda}_{k,d+p} \stackrel{ind.}{\sim} \sum_{j=1}^k \pi_j(\mathbf{x}) \mathcal{N}_d(\mathbf{y}; \boldsymbol{\theta}_j(\mathbf{x}), \boldsymbol{\Lambda}_j(\mathbf{x})), \quad (3.3.5)$$

where $\pi_j(\mathbf{x})$, $\boldsymbol{\theta}_j(\mathbf{x})$ and $\boldsymbol{\Lambda}_j(\mathbf{x})$ have the following forms:

$$\pi_j(\mathbf{x}) \propto \pi_j \mathcal{N}_p(\mathbf{x}; \boldsymbol{\theta}_j^x, \boldsymbol{\Lambda}_j^{xx}) \quad (3.3.6)$$

$$\boldsymbol{\theta}_j(\mathbf{x}) = \boldsymbol{\theta}_j^y + \boldsymbol{\Lambda}_j^{yx} (\boldsymbol{\Lambda}_j^{xx})^{-1} (\mathbf{x} - \boldsymbol{\theta}_j^x) \quad (3.3.7)$$

$$\boldsymbol{\Lambda}_j(\mathbf{x}) = \boldsymbol{\Lambda}_j^{yy} - \boldsymbol{\Lambda}_j^{yx} (\boldsymbol{\Lambda}_j^{xx})^{-1} \boldsymbol{\Lambda}_j^{xy}. \quad (3.3.8)$$

To complete the Bayesian model, we need priors for $\boldsymbol{\pi}_{k,1}$, $\boldsymbol{\theta}_{k,d+p}$ and $\boldsymbol{\Lambda}_{k,d+p}$. It is common to assume mutual independent conjugate priors for these parameters. The location parameters $\boldsymbol{\theta}_{k,d+p}$ are often assumed to come from a common distribution and are mutually independent (e.g., $\boldsymbol{\theta}_j \stackrel{i.i.d.}{\sim} \mathcal{N}_{d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^{d+p}$ and $\boldsymbol{\Sigma} \in \mathbb{S}^{d+p}$). Although this approach generates flexible structures that capture non-linear patterns for the mean response, as mentioned previously, the independent assumption can encourage the creation of an unnecessary number of mixture components. To avoid this, we propose modeling $\boldsymbol{\theta}_{k,d+p}$ by means of the $\text{NRep}_{k,d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ class. The repulsive feature is naturally inherited by $(\mathbf{y} \mid \mathbf{x})$, encouraging more parsimonious models. The definition of $\text{NRep}_{k,d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ is provided in Chapter 2, but for the sake of completeness we provide it here.

Definition 3.3.1. The probability density function of $\text{NRep}_{k,d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ is

$$\text{NRep}_{k,d+p}(\boldsymbol{\theta}_{k,d+p}) \propto \left\{ \prod_{j=1}^k \text{N}_{d+p}(\boldsymbol{\theta}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\} \text{R}_C(\boldsymbol{\theta}_{k,d+p}), \quad (3.3.9)$$

$$\text{R}_C(\boldsymbol{\theta}_{k,d+p}) = \prod_{r < s}^k [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\}]. \quad (3.3.10)$$

The parameters of this distribution are $d, p, k \in \mathbb{N}$ with $k \geq 2$, $\boldsymbol{\mu} \in \mathbb{R}^{d+p}$, $\boldsymbol{\Sigma} \in \mathbb{S}^{d+p}$ and $\tau \in (0, \infty)$.

(3.3.10) introduces dependence between $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k$ by penalizing small relative distances through the expression $0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)$. The parameter τ controls the strength of the repulsion: as $\tau \rightarrow 0$ (right-side limit), (3.3.9) converges functionally to an i.i.d. model for $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k$ with each following a common Gaussian distribution $\text{N}_{d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note further that this probability density equals 0 when $\boldsymbol{\theta}_r = \boldsymbol{\theta}_s$ for some $r \neq s$. Because of repulsion, the implied conditional distribution $(\mathbf{y} \mid \mathbf{x})$ will tend to fit flexible regression curves by using information from a small number of active clusters.

Notice that the joint likelihood derived from (3.3.4) involves an expansion into k^n terms, which is computationally expensive. An approach that simplifies the previous problem is based on introducing mutually independent auxiliary variables $z_1, \dots, z_n \in [k]$ called mixture component indicators, such that $\mathbf{u}_1, \dots, \mathbf{u}_n$ are conditionally independent given z_1, \dots, z_n . The auxiliary variables can be thought of cluster labels: \mathbf{u}_i is generated by the j th cluster if and only if $z_i = j$. Notice that from the following hierarchical stochastic model

$$\mathbf{u}_i \mid z_i, \boldsymbol{\theta}_{k,d+p}, \boldsymbol{\Lambda}_{k,d+p} \stackrel{\text{i.i.d.}}{\sim} \text{N}_{d+p}(\mathbf{u}_i; \boldsymbol{\theta}_{z_i}, \boldsymbol{\Lambda}_{z_i}) \quad (3.3.11)$$

$$z_i \mid \boldsymbol{\pi}_{k,1} \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}(z_i = j) = \pi_j. \quad (3.3.12)$$

the model (3.3.4) is recovered after marginalizing over each z_i in the joint distribution defined

by (3.3.11) and (3.3.12).

Under this framework, the covariate dependent (Bayesian) Repulsive Gaussian Mixture Model is completely specified by (3.3.11) and (3.3.12) with the following prior distributions (mutually independent):

$$\boldsymbol{\pi}_{k,1} \sim \text{Dir}(\boldsymbol{\alpha}_{k,1}) : \boldsymbol{\alpha}_{k,1} \in (0, \infty)^k \quad (3.3.13)$$

$$\boldsymbol{\theta}_{k,d+p} \sim \text{NRep}_{k,d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau) : \boldsymbol{\mu} \in \mathbb{R}^{d+p}, \boldsymbol{\Sigma} \in \mathbb{S}^{d+p}, \tau \in (0, \infty) \quad (3.3.14)$$

$$\boldsymbol{\Lambda}_j \stackrel{i.i.d.}{\sim} \text{IW}_{d+p}(\boldsymbol{\Psi}, \nu) : \boldsymbol{\Psi} \in \mathbb{S}^{d+p}, \nu \in (0, \infty). \quad (3.3.15)$$

The conditional model derived from (3.3.11)–(3.3.15) from now on will be called RGMMx.

3.3.1 Parameter Calibration

One advantage of starting with (3.3.4) and then inducing (3.3.5) is that we can exploit essentially the same recommendations described in Chapter 2 to elicit the hyperparameters in (3.3.13)–(3.3.15). For completeness, we provide details.

To begin with, standardizing the response and covariates makes selecting values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ straightforward. This technique, which is suggested in Gelman et al. (2014), justifies the assignment of $\boldsymbol{\mu} = \mathbf{0}_{d+p}$ and $\boldsymbol{\Sigma} = \mathbf{I}_{d+p}$ where $\mathbf{0}_{d+p} \in \mathbb{R}^{d+p}$ is the vector whose entries are all equal 0 and \mathbf{I}_{d+p} is the identity matrix of dimension $(d+p) \times (d+p)$. The same authors suggest that fixing $\boldsymbol{\alpha}_{k,1} = k^{-1}\mathbf{1}_{d+p}$ produces a weakly informative prior for the weights when the number of mixture components is relatively high. Here, $\mathbf{1}_{d+p} \in \mathbb{R}^{d+p}$ is a vector with entries equal 1. As for ν and $\boldsymbol{\Psi}$, their values are critical since misspecification can result in masking the repulsion effect: large variances can produce an overlap between mixture components, even though their location parameters are well-separated by the presence of repulsion. We suggest fixing $\nu = p + d + 4$ and $\boldsymbol{\Psi} = 3\psi\mathbf{I}_{d+p}$ with $\psi \in (0, \infty)$. This choice guarantees that $\boldsymbol{\Lambda}_j$ is centered at $\psi\mathbf{I}_{d+p}$ and has finite variance that is controlled by ψ .

Finally, specifying a value for τ is of principal interest because it guides the repulsion between the centers of each mixture components. In this case, we propose the following criterion: for fixed values $u, p \in (0, 1)$ choose τ such that

$$\mathbb{P}\{G \leq -2 \log(1 - u)\tau\} = p, \quad G \sim G(d/2 + p/2, 1/2). \quad (3.3.16)$$

After creating a grid of points in $(0, \infty)$ it is straightforward to find τ that satisfies (3.3.16). For more details about the motivation behind (3.3.16) see Chapter 2.

3.3.2 Computation

In this section we describe the sampling mechanism to obtain posterior samples from RGMMx. Although we are not interested in making inference on parameters in (3.3.4), these realizations can be used directly to sample from the induced conditional distribution (3.3.5).

As a starting point, the posterior sampling procedure for $\boldsymbol{\pi}_{k,1}$, $\boldsymbol{\theta}_{k,d+p}$ and $\boldsymbol{\Lambda}_{k,d+p}$ is simply a Gibbs Sampler. Due to conjugacy, the full conditional distributions for $\boldsymbol{\pi}_{k,1}$ and $\boldsymbol{\Lambda}_{k,d+p}$ have known closed forms and are easy to sample from. Unfortunately, this is not the case for $\boldsymbol{\theta}_{k,d+p}$. However, the fact that the coordinates of $\boldsymbol{\theta}_{k,d+p} \sim \text{NRep}_{k,d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ are exchangeable implies that all the associated full conditional distributions share the same functional form. Moreover, evaluating the respective densities has a low computational cost. Because of this, we incorporate a Metropolis–Hastings step inside the Gibbs Sampler. To illustrate we need the full conditional distribution $(\boldsymbol{\theta}_{k,d+p} \mid \dots)$ which given by

$$(\boldsymbol{\theta}_{k,d+p} \mid \dots) \propto \left\{ \prod_{j=1}^k N_{d+p}(\mathbf{u}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right\} \prod_{r < s}^k [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\}],$$

where $\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Lambda}_j^{-1}\mathbf{s}_j)$, $\mathbf{s}_j = \sum_{i=1}^n \mathbb{I}_{\{j\}}(z_i)\mathbf{u}_i$, $\boldsymbol{\Sigma}_j = (\boldsymbol{\Sigma}^{-1} + n_j\boldsymbol{\Lambda}_j^{-1})^{-1}$ and $n_j = \text{card}(i \in [n] : z_i = j)$. With this information, the full conditional distributions $(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \dots)$

for $j \in [k]$ and $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_l : l \neq j) \in \mathbb{R}_{k-1}^{d+p}$ have corresponding densities

$$f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \dots) \propto \text{N}_{d+p}(\boldsymbol{\theta}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \prod_{l \neq j}^k [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_l)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_l)\}].$$

The following pseudo-code could then be used to sample from $(\boldsymbol{\theta}_{k,d+p} \mid \dots)$ using $f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \dots)$ through a random walk Metropolis–Hastings step inside the Gibbs Sampler:

1. Let $\boldsymbol{\theta}_{k,d+p}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_k^{(0)}) \in \mathbb{R}_k^{d+p}$ be the actual state for $\boldsymbol{\theta}_{k,d+p}$.
2. For $j = 1, \dots, k$:
 - (a) Generate a candidate $\boldsymbol{\theta}_j^{(1)}$ coming from $\text{N}_{d+p}(\boldsymbol{\theta}_j^{(0)}, \boldsymbol{\Gamma}_j)$ with $\boldsymbol{\Gamma}_j \in \mathbb{S}^{d+p}$.
 - (b) Set $\boldsymbol{\theta}_j^{(0)} = \boldsymbol{\theta}_j^{(1)}$ with probability $\min(1, \beta_j)$, where

$$\beta_j = \frac{\text{N}_{d+p}(\boldsymbol{\theta}_j^{(1)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\text{N}_{d+p}(\boldsymbol{\theta}_j^{(0)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \prod_{l \neq j}^k \left[\frac{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(1)} - \boldsymbol{\theta}_l^{(0)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(1)} - \boldsymbol{\theta}_l^{(0)})\}}{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(0)} - \boldsymbol{\theta}_l^{(0)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(0)} - \boldsymbol{\theta}_l^{(0)})\}} \right].$$

Since $\boldsymbol{\Gamma}_j$ controls the variability of the generated candidates, care must be taken when selecting it. An approach that seems to work well in practice is to fix

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

where $t \in [B]$ is the t th iteration of the burn-in phase of length $B \in \mathbb{N}$. In Appendix B we provide the complete pseudo-code associated with the RGMMx model to obtain posterior samples for the parameters that appear in (3.3.4) and how they can be used to estimate regression densities and regression curves. We will refer to the algorithm in Appendix B as Algorithm RGMMx.

3.4 Simulation Study

As a means to explore the proposed methodology we take on the simulation situation of Dunson et al. (2007). More specifically we generate data sets from the following density

$$f_0(y | x) = \exp(-2x)\mathcal{N}(y; x, 0.01) + \{1 - \exp(-2x)\}\mathcal{N}(y; x^4, 0.05) : y \in \mathbb{R},$$

where $x \in (0, 1)$ is taken as the covariate. This Gaussian regression mixture has weights that vary smoothly in x , different variances for each mixture component, and a non-linear mean in the second component. The mean regression curve m_0 associated with f_0 is

$$m_0(x) = \exp(-2x)x + \{1 - \exp(-2x)\}x^4 : x \in (0, 1). \quad (3.4.1)$$

Notice that by construction f_0 can take on a diverse number of shapes, ranging from unimodal (symmetric or asymmetric) to bimodal densities.

In this small simulation study we focus on proof of concept associated with the RGMMx by exploring how values of τ influence the repulsiveness of our methodology and ultimately the number of estimated clusters and goodness-of-fit. We do this by fitting the RGMMx to the generated data using different values for τ , namely, 0.01, 0.1, 1 and 10. Goodness-of-fit will be evaluated using the following metrics:

- Log Pseudo Marginal Likelihood (LPML) for the joint model (Christensen et al. (2011)) which is a model fit metric that takes into account model complexity. We calculate the LPML by first estimating all the corresponding conditional predictive ordinates (Gelfand et al. 1992) using the method in Chen et al. (2000).
- L_1 -metric between the estimated mean regression curve and m_0 .

To see how τ influences the posterior distribution associated with the number of clusters, 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.

We will generate 100 data sets, each of size 500, by first generating x_i from $U(0, 1)$ and for each x_i generating a realization y_i using $f_0(y | x_i)$. For each value of τ we fit RGMMx to data by collecting 5000 MCMC iterates after discarding the first 5000 as burn-in and thinning by 25. The rest of the prior parameter values in (3.3.13)–(3.3.15) will be set to the following:

- $k = 10$, $d = 1$, $p = 1$, $\alpha_{k,1} = 10^{-1}\mathbf{1}_{10}$, $\mu = \mathbf{0}_2$, $\Sigma = \mathbf{I}_2$, $\Psi = 3^{-1}\mathbf{I}_2$ and $\nu = 6$.

In Figure 3.1 we provide the results of each metric for each value of τ considered by way of side-by-side box-plots. Interestingly, LPML is not a monotonic function of τ . There is an initial increase and then a decrease of LPML as τ increases (and thus the number of clusters decreases). The L_1 -metric does not seem to be influenced by τ . As expected, increasing τ results in stronger repulsion and therefore less clusters. In addition, stronger repulsion produces less variability in the number of components. This makes sense because as the strength of repulsion increases, there are less locations at which cluster centers can exist.

3.5 Data Illustration

Azzalini and Bowman (1990) analyzed a data set concerning the eruptions from the Old Faithful geyser in Yellowstone National Park, Wyoming. These were continuous measurements from August 1 to August 15, 1985. The recorded data represents time eruptions (*duration*) and waiting times for each eruption (*waiting*), both in minutes. In this illustration, the first (last) variable is considered as the covariate (response). Of the 299 observations we removed 78 observations that were collected at night (coded as 2, 3 or 4 minutes) and originally de-

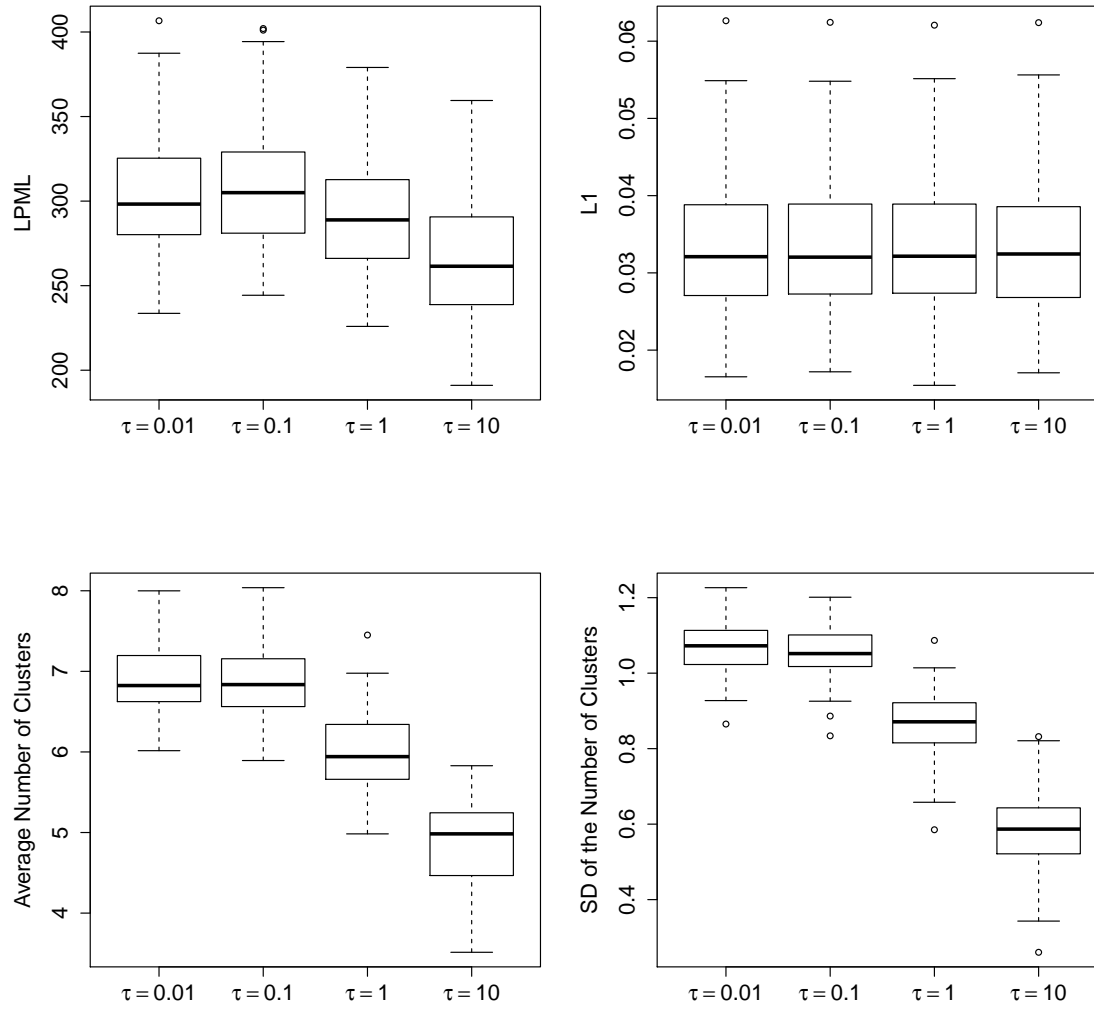


Figure 3.1: Boxplots that display LPML, L_1 -metric, the average number of occupied mixture components, and the average standard deviation associated with the distribution of occupied mixture components for each value of τ .

scribed as “short”, “medium” or “long”. The original data can be found in **R** under the name *geyser* (MASS library).

We will implement two WDDP and two RGMMx to compare the respective regression mean curves for *waiting* in terms of *duration*. 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. We standardized the data before fitting the above models. Our main aim here is to assess the effect that the prior specification on the repulsion parameter has on the reported inference. Specific details now follow:

1. RGMMx: We coded Algorithm RGMMx in `Fortran` to generate posterior draws from this model. For both model specifications (which will be referred to as RGMMx1 and RGMMx2) we collected 5000 MCMC iterates after discarding the first 10000 as burn-in and thinning by 20. We use the same prior parameter values in (3.3.13) to (3.3.15) for both models. Specifically, we use: $k = 10$, $d = 1$, $p = 1$, $\boldsymbol{\alpha}_{k,1} = 10^{-1}\mathbf{1}_{10}$, $\boldsymbol{\mu} = \mathbf{0}_2$, $\boldsymbol{\Sigma} = \mathbf{I}_2$, $\boldsymbol{\Psi} = 3^{-1}\mathbf{I}_2$ and $\nu = 6$. The respective values for τ , selected by the calibration criterion from Subsection 3.3.1, are provided below.

(a) RGMMx1: $\tau = 0.2$ with $(u, p) = (0.999, 0.5)$.

(b) RGMMx2: $\tau = 4.6$ with $(u, p) = (0.5, 0.8)$.

We emphasize the fact that the setting $\tau = 0.2$ in RGMMx1 produces a fairly weak repulsive behavior when modeling the response and covariate jointly. The motivation behind this selection is to avoid underfitting as large values of τ could result in mixture models with a small number of occupied components. This would have serious repercussions regarding the quality of model fit and flexibility in estimating the regression curve. On the other hand, overfitting can be partially avoided by fixing the number of components to a reasonable value k that is not too large. In addition, since (3.3.10) models repulsion *softly* (see Ogata and Tanemura (1981)), the strength of repulsion (i.e., the value of τ) would need to be very small for the active number of components to be large. As for $\tau = 4.6$ in RGMMx2, this value forces the number of occupied components to be smaller than RGMMx1.

2. WDDP: The baseline distribution H_0 in (3.3.3) that we will use is the conjugate Normal-Inverse-Wishart

$$H_0(\boldsymbol{\theta}, \boldsymbol{\Lambda}) = N_{d+p}(\boldsymbol{\theta}; \mathbf{m}_1, k_0^{-1}\boldsymbol{\Lambda}) IW_{d+p}(\boldsymbol{\Lambda}; \boldsymbol{\Psi}_1, \nu_1) : \nu_1 \in (0, \infty). \quad (3.5.1)$$

To complete the model specification given by (3.3.1)–(3.3.3) with (3.5.1), the following independent hyperpriors are assumed:

$$\alpha \mid a_0, b_0 \sim G(a_0, b_0) : a_0, b_0 \in (0, \infty) \quad (3.5.2)$$

$$\mathbf{m}_1 \mid \mathbf{m}_2, \mathbf{S}_2 \sim N_{d+p}(\mathbf{m}_2, \mathbf{S}_2) : \mathbf{m}_2 \in \mathbb{R}^{d+p}, \mathbf{S}_2 \in \mathbb{S}^{d+p} \quad (3.5.3)$$

$$k_0 \mid \tau_1, \tau_2 \sim G(\tau_1/2, \tau_2/2) : \tau_1, \tau_2 \in (0, \infty) \quad (3.5.4)$$

$$\boldsymbol{\Psi}_1 \mid \boldsymbol{\Psi}_2, \nu_2 \sim IW_{d+p}(\boldsymbol{\Psi}_2, \nu_2) : \boldsymbol{\Psi}_2 \in \mathbb{S}^{d+p}, \nu_2 \in (0, \infty). \quad (3.5.5)$$

We employed the R function `DPcdensity` available in `DPpackage` (Jara et al. 2011). Decisions on hyperprior parameter values were guided by Escobar and West (1995). For each of the following model specifications, named WDDP1 and WDDP2, we collected 5000 MCMC iterates after discarding the first 5000 as burn-in and thinning by 3. The respective prior hyperparameter values in (3.5.2) to (3.5.5) are provided below.

(a) WDDP1: $d = 1, p = 1, a_0 = 10, b_0 = 1, \nu_1 = 4, \nu_2 = 4, \mathbf{m}_2 = \mathbf{0}_2, \mathbf{S}_2 = \mathbf{I}_2, \boldsymbol{\Psi}_2 = \mathbf{I}_2, \tau_1 = 6.01$ and $\tau_2 = 2.01$.

(b) WDDP2: $d = 1, p = 1, a_0 = 2, b_0 = 4, \nu_1 = 4, \nu_2 = 4, \mathbf{m}_2 = \mathbf{0}_2, \mathbf{S}_2 = \mathbf{I}_2, \boldsymbol{\Psi}_2 = \mathbf{I}_2, \tau_1 = 2.01$ and $\tau_2 = 1.01$.

These values make WDDP1 and RGMMx1 (WDDP2 and RGMMx2) “similar” in terms of the distribution for the number of occupied components *a priori*: more (less) spread around a high (small) average value.

Model	LPML	Mean (Clusters)	SD (Clusters)
RGMMx1	-195.14	6.15	0.97
RGMMx2	-208.62	4.90	0.74
WDDP1	-185.82	11.51	3.11
WDDP2	-226.73	5.64	1.39

Table 3.1: Summary statistics related to model fit and the number of clusters for Geyser data based on WDDP and RGMMx.

Table 3.1 and Figure 3.2 show that for a small number of clusters RGMMx tends to fit the Geyser data better than WDDP. On the other hand, WDDP is able to fit the data slightly better than RGMMx when the number of clusters increases. This slight increase in the LPML value comes at a substantial model complexity cost as the number of active mixture components is doubled. Figure 3.3 shows that there is no appreciable difference between the mean regression curves, but the estimated 95% point-wise credible bands under RGMMx are slightly wider than those under WDDP. Figure 3.4 displays an estimated partition for each procedure using Dahl’s least squares method (Dahl 2006). Notice that partitions associated with model specifications that produce the highest LPML values (i.e., panels “b” (RGMMx2) and “d” (WDDP2)) agree, inducing four well-formed groups and one isolated point. Panel “a” (RGMMx1) reveals that a small repulsion effect allows grouping data into a moderate number of clusters that are fairly separated. Different is the case in panel “c” (WDDP1), where the number of clusters is quite high and some of these overlap.

Figures 3.5 and 3.6 provide a few estimated conditional densities for four different values of time eruptions (2, 3, 4 and 4.5), and 95% point-wise credible bands. As expected, all plots labeled with “a” (RGMMx1) exhibit similar densities to those labeled with “c” (WDDP1). The same scenario occurs between plots labeled with “b” (RGMMx2) and “d” (WDDP2). However, there are slight differences in the width of the point-wise credible bands with WDDP1 and WDDP2 being slightly narrower.

The take home message from this application is that the RGMMx, being a parametric

model, is a simple, parsimonious alternative to WDDP in being able to capture flexible regression curves.

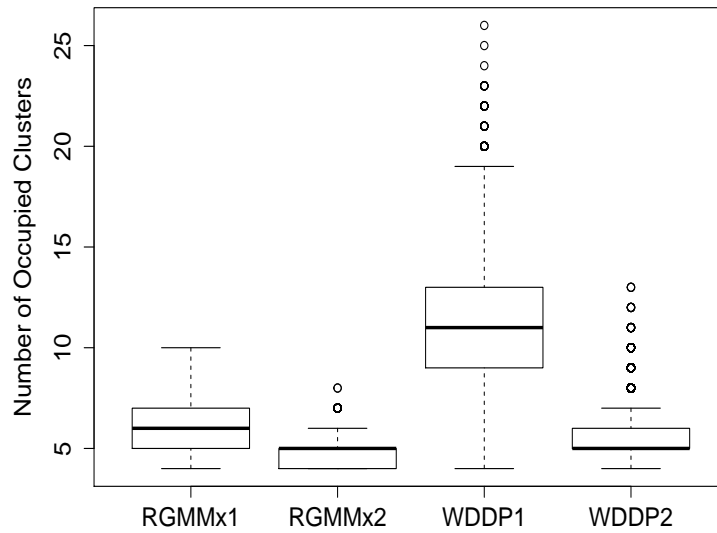


Figure 3.2: Side-by-side box-plots of the posterior distribution for the active number of clusters associated with the Geyser data.

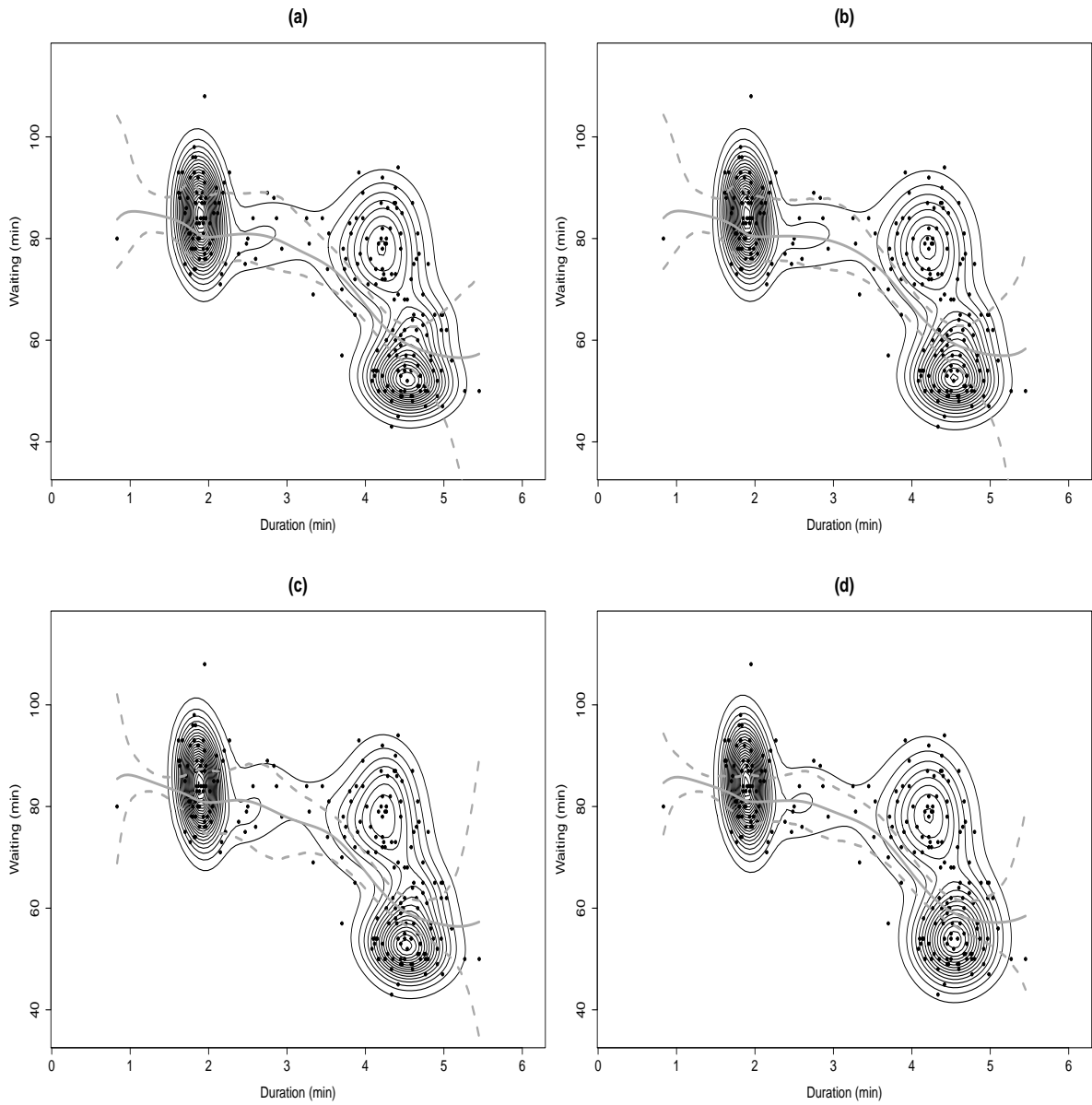


Figure 3.3: Estimated regression curve (gray solid) for Geyser data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals.

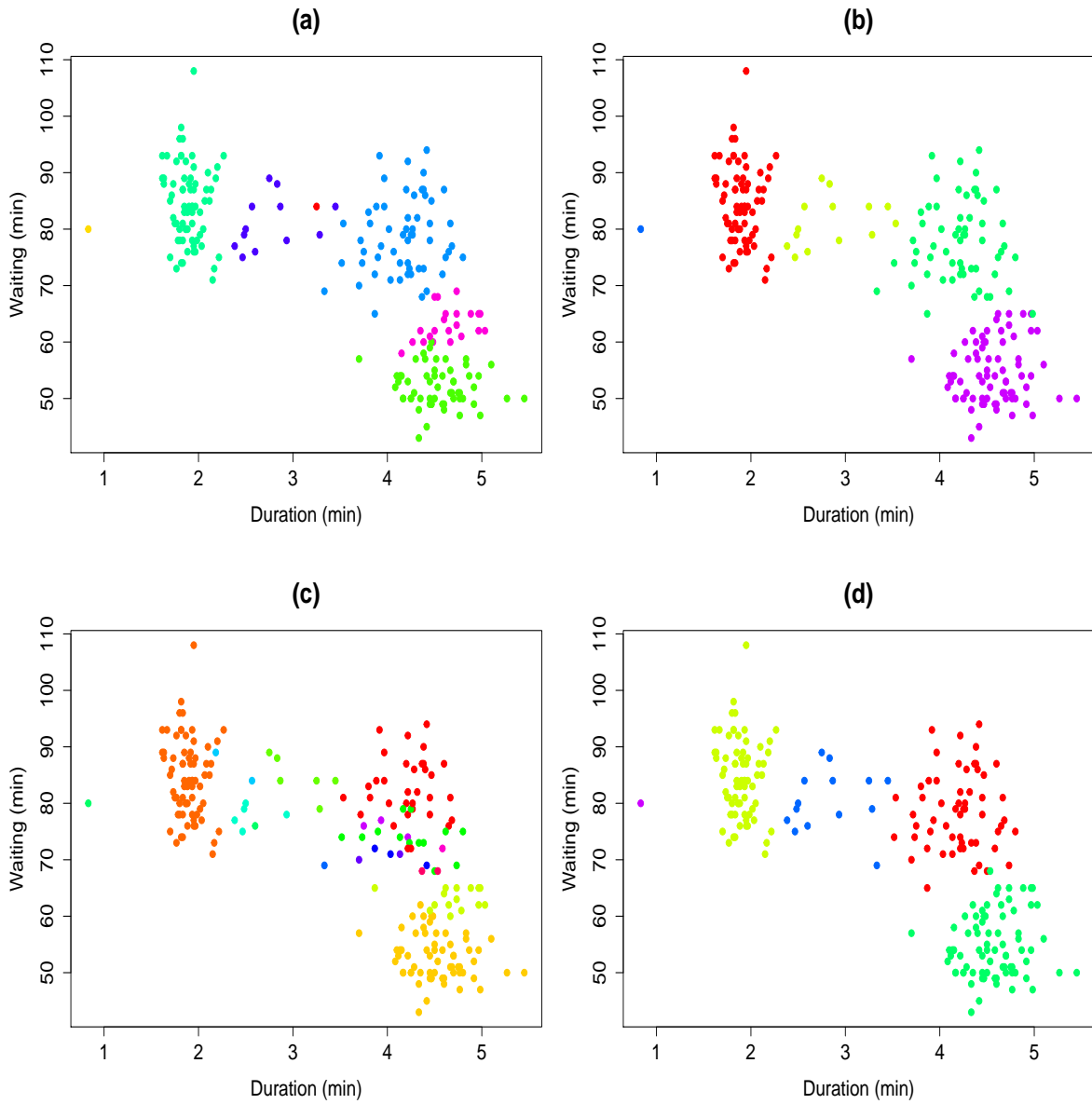


Figure 3.4: Estimated partitions using Dahl's least squares clustering algorithm for (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2.

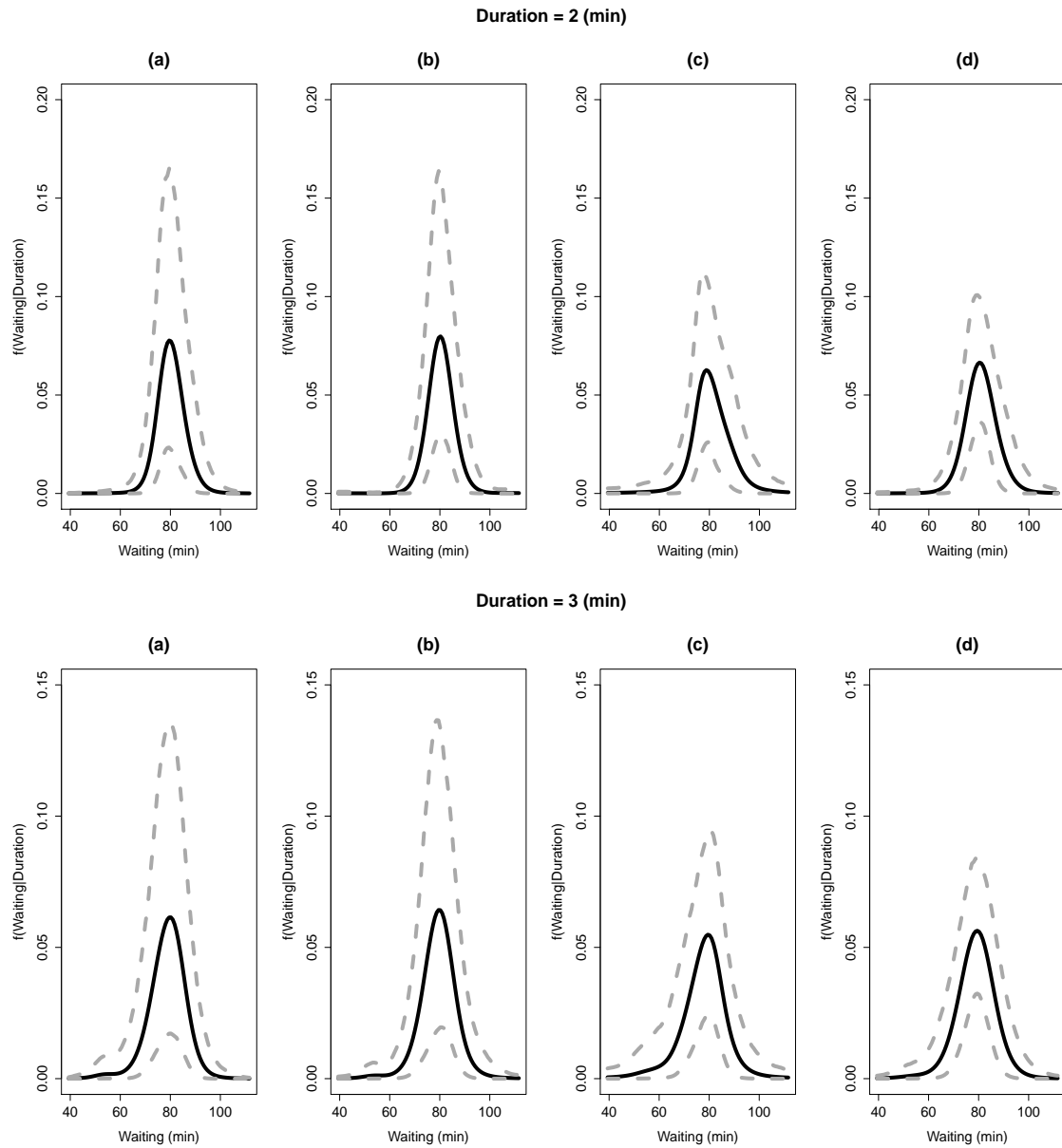


Figure 3.5: Estimated conditional densities (black solid) for Geyser data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals. Here, the selected time eruptions (*duration*) are 2 and 3 minutes.

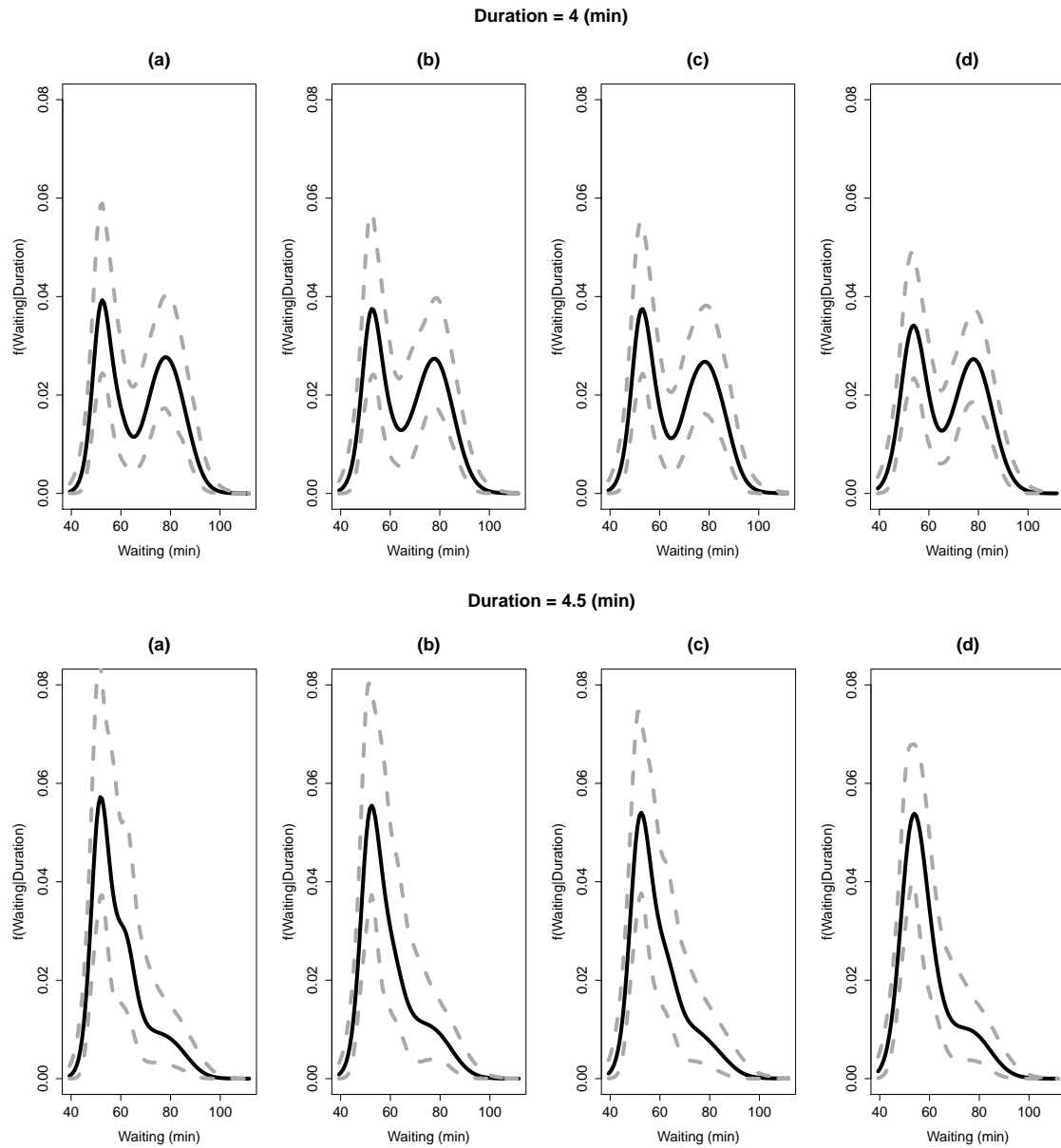


Figure 3.6: Estimated conditional densities (black solid) for Geyser data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals. Here, the selected time eruptions (*duration*) are 4 and 4.5 minutes.

Chapter 4

Discussion and Future Work

4.1 Density Estimation using Repulsive Distributions

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 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 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, Propositions 2.4.3 and 2.4.6 were established for Gaussian mixtures of dimension $d = 1$ with mixture components sharing the same variance. Extending results to the general d dimensional case would be a natural progression. Additionally, we are exploring the possibility of relaxing the assumption of common variance between mixture components and adapting the mentioned theoretical results to a larger class of potential functions. Studying the influence of the metric on the repulsive component in Definition 2.3.1 and allowing the number of mixture components to be random are also topics of future research. 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.

4.2 Regression Estimation using Repulsive Distributions

This chapter contains extensions to repulsive mixture modeling that are completely methodological. Using a similar approach to Müller et al. (1996) we propose a Bayesian mixture of Gaussian distributions to jointly model responses and (continuous) covariates where the associated location parameters are driven by a probability distribution which encourages them to repel each other. The joint model naturally induces a conditional distribution, which is a weighted mixture of Gaussian regressions that inherits the repulsion effect. An important consequence of this is that the conditional distribution allows estimation of regression curves in a flexible and parsimonious way, i.e. using a reduced number of mixture components at almost no cost in terms of goodness-of-fit. It is worth noting that a downside of both methodologies is the curse of dimensionality. That is, when responses and/or covariates lie on high dimensional Euclidean spaces, computation becomes very expensive.

Future research will be dedicated to studying the topological support associated with RGMMx to determine the class of regression curves that can be approximated (a study that is similar to what which was done in Barrientos et al. (2012)). Additionally, since τ seems to influence model fit and predictions it would be natural to assign it a prior and treat it as an unknown. In this way the data could guide τ 's location and its uncertainty taken into account. Doing this however will come at a formidable computational cost because it can be shown that τ 's posterior distribution is doubly intractable. Finally, we would like to develop a method that avoids focusing on the joint distribution of a response and covariate and instead incorporates repulsion directly in the conditional distribution. One possible way of carrying this out is to employ a probit stick-breaking prior (Rodriguez and Dunson 2011) for the mixture weights and model the centers of each component by linear regressions. Repulsion would then influence the vector of regression coefficients producing well separated clusters of linear regressions.

Appendix A

Supplementary Material for Chapter 2

A.1 Algorithm RGMM

In what follows we describe the Gibbs Sampler for the RGMM in its entirety. 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 $z_i^{(0)} : i \in [n]$, $\boldsymbol{\pi}_{k,1}^{(0)}$ 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 $t = 0, \dots, B - 1$:
 1. $(z_i^{(t+1)} \mid \dots) \sim \mathbb{P}(z_i^{(t+1)} = j) = \pi_j^{(t,i)}$ independently for each $i \in [n]$, where

$$\pi_j^{(t,i)} = \left\{ \sum_{l=1}^k \pi_l^{(t)} \mathbb{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_l^{(t)}, \boldsymbol{\Lambda}_l^{(t)}) \right\}^{-1} \pi_j^{(t)} \mathbb{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Lambda}_j^{(t)}) : j \in [k].$$

2. $(\boldsymbol{\pi}_{k,1}^{(t+1)} \mid \dots) \sim \text{Dir}(\boldsymbol{\alpha}_{k,1}^{(t)})$, where

$$\begin{aligned}\boldsymbol{\alpha}_{k,1}^{(t)} &= (\alpha_1 + n_1^{(t+1)}, \dots, \alpha_k + n_k^{(t+1)}) \\ n_j^{(t+1)} &= \text{card}(i \in [n] : z_i^{(t+1)} = j) : j \in [k].\end{aligned}$$

3. For $j = 1, \dots, k$:

3.1. Generate a candidate $\boldsymbol{\theta}_j^{(*)}$ from $N_d(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Omega}_j^{(t)})$, where

$$\boldsymbol{\Omega}_j^{(t)} = \{\boldsymbol{\Sigma}^{-1} + n_j^{(t+1)}(\boldsymbol{\Lambda}_j^{(t)})^{-1}\}^{-1}.$$

3.2. Update $\boldsymbol{\theta}_j^{(t)} \rightarrow \boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(*)}$ with probability $\min(1, \beta_j)$, where

$$\beta_j = \frac{N_d(\boldsymbol{\theta}_j^{(*)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{N_d(\boldsymbol{\theta}_j^{(t)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})} \prod_{l \neq j}^k \left[\frac{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(*)} - \boldsymbol{\theta}_l^{(t)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(*)} - \boldsymbol{\theta}_l^{(t)})\}}{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_l^{(t)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_l^{(t)})\}} \right].$$

In the above expression for β_j

$$\begin{aligned}\boldsymbol{\Sigma}_j^{(t)} &= \{\boldsymbol{\Sigma}^{-1} + n_j^{(t+1)}(\boldsymbol{\Lambda}_j^{(t)})^{-1}\}^{-1} \\ \boldsymbol{\mu}_j^{(t)} &= \boldsymbol{\Sigma}_j^{(t)} \{\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + (\boldsymbol{\Lambda}_j^{(t)})^{-1} \mathbf{s}_j^{(t)}\} : \mathbf{s}_j^{(t)} = \sum_{i=1}^n \mathbb{I}_{\{j\}}(z_i^{(t+1)}) \mathbf{y}_i.\end{aligned}$$

Otherwise, set $\boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(t)}$.

3.3. Update $\boldsymbol{\Gamma}_j \rightarrow \boldsymbol{\Gamma}_j + B^{-1} \boldsymbol{\Omega}_j^{(t)}$.

4. $(\boldsymbol{\Lambda}_j^{(t+1)} \mid \dots) \sim \text{IW}_d(\boldsymbol{\Psi}_j^{(t)}, \nu_j^{(t)})$ independently for each $j \in [k]$, where $\nu_j^{(t)} = \nu + n_j^{(t+1)}$ and

$$\boldsymbol{\Psi}_j^{(t)} = \boldsymbol{\Psi} + \sum_{i=1}^n \mathbb{I}_{\{j\}}(z_i^{(t+1)}) (\mathbf{y}_i - \boldsymbol{\theta}_j^{(t+1)}) (\mathbf{y}_i - \boldsymbol{\theta}_j^{(t+1)})^\top.$$

- (Save samples) For $t = B, \dots, ST + B - 1$: 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

3.1a. Generate a candidate $\boldsymbol{\theta}_j^{(*)}$ from $N_d(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Gamma}_j)$.

Finally, save the generated samples every T th iteration.

- (Posterior predictive estimate) With the T saved samples, compute

$$f(\mathbf{y} \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \approx \frac{1}{T} \sum_{t=1}^T \left\{ \sum_{j=1}^k \pi_j^{(t)} N_d(\mathbf{y}; \boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Lambda}_j^{(t)}) \right\}.$$

A.2 Proof of Lemma 2.3.1.

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

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

A.3 Proof of Proposition 2.3.2.

Notice that $g_{k,d} \in C(\mathbb{R}_k^d; (0, \infty))$ by construction (see Lemma 2.3.1). Because of the continuity, measurability follows. Using conditions A1–A4 of C_0 -properties it follows that for all $x \in [0, \infty)$, $\{1 - C_0(x)\} \in [0, 1)$. By Tonelli's Theorem

$$\int_{\mathbb{R}_k^d} g_{k,d}(\mathbf{x}_{k,d}) d\mathbf{x}_{k,d} \leq \left(\int_{\mathbb{R}^d} f_0(\mathbf{x}) d\mathbf{x} \right)^k = 1.$$

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

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

because for all $\mathbf{x}_{k,d} \in \mathbb{R}_k^d$, $g_{k,d}(\mathbf{x}_{k,d}) \geq 0$ by construction. To prove the above inequality, fix $\mathbf{x}_{k,d}^0 \in \mathbb{R}_k^d$ such that $\mathbf{x}_r^0 \neq \mathbf{x}_s^0$ for $r \neq s \in [k]$. Then $g_{k,d}(\mathbf{x}_{k,d}^0) > 0$. Because $g_{k,d}$ is a continuous function on \mathbb{R}_k^d , there exists $r_0 \in (0, \infty)$ such that for all $\mathbf{x}_{k,d} \in B(\mathbf{x}_{k,d}^0, r_0)$

$$g_{k,d}(\mathbf{x}_{k,d}) > 0,$$

where $B(\mathbf{x}_{k,d}^0, r_0)$ is the cartesian product of $B_2(\mathbf{x}_1^0, r_0), \dots, B_2(\mathbf{x}_k^0, r_0)$. Further, $B(\mathbf{x}_{k,d}^0, r_0) \in \mathcal{B}(\mathbb{R}_k^d)$ and $\lambda_d^k\{B(\mathbf{x}_{k,d}^0, r_0)\} = (\pi^{kd/2} r_0^{kd}) \Gamma(1 + d/2)^{-k} \in (0, \infty)$ by the Volume Formula, where $\Gamma(\cdot)$ is the Gamma function. Thus

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

A.4 Proof of Lemma 2.4.1.

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

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

and

$$u(\lambda, \boldsymbol{\theta}_{k_0,1}; x, \lambda_0) = \frac{1}{(2\pi)^{1/2}} \left[\frac{k_0}{\lambda \lambda_0^{1/2} + \lambda_0 \lambda^{1/2}} + \frac{\lambda_0^{1/2}}{2\lambda \lambda_0^2} \sum_{j=1}^{k_0} (x - \theta_j)^2 \right]$$

for all $(\lambda, \boldsymbol{\theta}_{k_0,1}) \in (0, \infty) \times \mathbb{R}_{k_0}^1$, with $\|\cdot\|_1$ being the Euclidean L_1 -norm in $\mathbb{R}_{k_0}^1$. Because $u(\lambda, \boldsymbol{\theta}_{k_0,1}; x, \lambda_0)$ is continuous at $(\lambda_0, \boldsymbol{\theta}_{k_0,1}^0)$,

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

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

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

point-wise in x when $\boldsymbol{\xi}_{k_0} \rightarrow \boldsymbol{\xi}_{k_0}^0$.

By condition B2, we can assume that $\theta_1^0 < \dots < \theta_{k_0}^0$ (possibly after an appropriate relabeling). Choose $t_1^0, t_2^0 \in \mathbb{R}$ and $l_1^0, l_2^0 \in (0, \infty)$ such that $\lambda_0 \in [l_1^0, l_2^0]$ and, for all $x \in (-\infty, t_1^0) \cup (t_2^0, \infty)$

$$f_0(x; \boldsymbol{\xi}_{k_0}^0) < 1, \quad \theta_j \in (t_1^0, t_2^0) : j \in [k_0].$$

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

$$I_1 = \int_{[t_1^0, t_2^0]} |\log\{f_0(x; \boldsymbol{\xi}_{k_0}^0)\}|f_0(x; \boldsymbol{\xi}_{k_0}^0)dx \in (0, \infty).$$

Fix $\delta_1 \in (0, 1)$, $\delta_2 = 0.5 \min(t_2^0 - \theta_{k_0}^0, \theta_1^0 - t_1^0)$ and define $V_0 = D_1(\boldsymbol{\pi}_{k_0,1}^0, \delta_1) \times D_1(\boldsymbol{\theta}_{k_0,1}^0, \delta_2) \times [l_1^0, l_2^0]$. 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 $M_0 = \max(M(x, \boldsymbol{\xi}_{k_0}) : (x, \boldsymbol{\xi}_{k_0}) \in [t_1^0, t_2^0] \times V_0) \in (0, \infty)$ and

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

By the Triangle Inequality

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

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 Gaussian distribution

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

Again, using the Triangle Inequality

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

The previous arguments show that $|\log\{f_0(x; \boldsymbol{\xi}_{k_0}^0)\} - \log\{f(x; \boldsymbol{\xi}_{k_0})\}| f_0(x; \boldsymbol{\xi}_{k_0}^0)$ for all $(x, \boldsymbol{\xi}_{k_0}) \in \mathbb{R} \times V_0$ is bounded above by a positive and integrable function that depends only in $x \in \mathbb{R}$.

As a consequence of Lebesgue's Dominated Convergence Theorem

$$\int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_{k_0})} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx \rightarrow 0$$

as $\boldsymbol{\xi}_{k_0} \rightarrow \boldsymbol{\xi}_{k_0}^0$. In other words, for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_{k_0})} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx < \varepsilon$$

provided that $\boldsymbol{\xi}_{k_0} \in B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta) \times B_1(\boldsymbol{\pi}_{k_0,1}^0, \delta) \times (\lambda_0 - \delta, \lambda_0 + \delta)$.

A.5 Proof of Lemma 2.4.2.

Set $\delta_{00} = 0.25vk_0$ with $v > 0$ specified by condition B2. Notice that

$$\boldsymbol{\theta}_{k_0,1} \in B_\delta = \prod_{i=1}^{k_0} \left(\theta_i^0 - \frac{\delta}{k_0}, \theta_i^0 + \frac{\delta}{k_0} \right) \subseteq B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta).$$

for all $\delta \in (0, \delta_{00}]$. Using the definition of $\text{NRep}_{k_0,1}(\mu, \sigma^2, \tau)$ and denoting $c_{k_0} = c_{k_0,1}$ the associated normalizing constant, we have that

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

for all $\delta \in (0, \delta_{00}]$. Now

$$\prod_{r < s}^{k_0} \left[1 - \exp \left\{ -\frac{(\theta_r - \theta_s)^2}{2\tau\sigma^2} \right\} \right] \geq \left[1 - \exp \left\{ -\frac{v_0}{2\tau\sigma^2} \right\} \right]^{\ell_{k_0}} = R_0 \in (0, \infty)$$

for all $\boldsymbol{\theta}_{k_0,1} \in B_\delta$, with $v_0 = (v - 2\delta_{00}k_0^{-1})^2$ and $\ell_{k_0} = 0.5k_0(k_0 - 1)$. Using this information and Fubini's Theorem

$$\mathbb{P}\{\boldsymbol{\theta}_{k_0,1} \in B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta)\} \geq \frac{R_0}{c_{k_0}} \prod_{i=1}^{k_0} \left\{ \Phi\left(\frac{\theta_i^0 - \mu}{\sigma} + \frac{\delta}{k_0\sigma}\right) - \Phi\left(\frac{\theta_i^0 - \mu}{\sigma} - \frac{\delta}{k_0\sigma}\right) \right\}$$

for all $\delta \in (0, \delta_{00}]$. Because for each $i \in [k_0]$

$$\frac{1}{\delta} \left\{ \Phi\left(\frac{\theta_i^0 - \mu}{\sigma} + \frac{\delta}{k_0\sigma}\right) - \Phi\left(\frac{\theta_i^0 - \mu}{\sigma} - \frac{\delta}{k_0\sigma}\right) \right\} \rightarrow \frac{2}{k_0\sigma} \mathcal{N}\left(\frac{\theta_i^0 - \mu}{\sigma}; 0, 1\right) = S_i^0 \in (0, \infty)$$

as $\delta \rightarrow 0$ (right-side limit), there exists $\delta_{0i} > 0$ such that

$$\left\{ \Phi\left(\frac{\theta_i^0 - \mu}{\sigma} + \frac{\delta}{k_0\sigma}\right) - \Phi\left(\frac{\theta_i^0 - \mu}{\sigma} - \frac{\delta}{k_0\sigma}\right) \right\} \geq \frac{S_i^0}{2}.$$

for all $\delta \in (0, \delta_{0i}]$. Finally, choose $\delta_0 = \min(\delta_{0j} : j \in \{0\} \cup [k_0])$ to conclude that

$$\mathbb{P}\{\boldsymbol{\theta}_{k_0,1} \in B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta)\} \geq \frac{R_0}{c_{k_0}} \left(\prod_{i=1}^{k_0} \frac{S_i^0}{2} \right) \exp\{-k_0 \log(1/\delta)\} \in (0, \infty).$$

for all $\delta \in (0, \delta_0]$.

Remark: The previous inequality also applies replacing $B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta)$ by $D_1(\boldsymbol{\theta}_{k_0,1}^0, \delta)$.

A.6 Proof of Proposition 2.4.3.

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

$$B_{\text{KL}}(f_0, \varepsilon) = \left\{ f \in \mathcal{F} : \int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_*)} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx < \varepsilon \right\}$$

with $\boldsymbol{\xi}_* \in \bigcup_{k=1}^{\infty} \Theta_k$. Using the stochastic representation (2.4.13),

$$\mathbb{P}\{B_{\text{KL}}(f_0, \varepsilon)\} \geq \kappa(k_0) \mathbb{P}\left(\boldsymbol{\xi}_{k_0} \in \Theta_{k_0} : \int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_{k_0})} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx < \varepsilon\right).$$

By condition B3, $\kappa(k_0) > 0$. In this case, to guarantee (2.4.14) it is sufficient to show that

$$\mathbb{P}\left(\boldsymbol{\xi}_{k_0} \in \Theta_{k_0} : \int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_{k_0})} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx < \varepsilon\right) > 0.$$

Lemma 2.4.1 guaranties the existence of $\delta_1 > 0$ such that for all $\boldsymbol{\xi}_{k_0} \in B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta_1) \times B_1(\boldsymbol{\pi}_{k_0,1}^0, \delta_1) \times (\lambda_0 - \delta_1, \lambda_0 + \delta_1)$

$$\int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_{k_0})} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx < \varepsilon.$$

Choose $\delta = \min(\delta_0, \delta_1)$ where $\delta_0 > 0$ is given by Lemma 2.4.2. Now $p_1 = \mathbb{P}\{\boldsymbol{\theta}_{k_0,1} \in B_1(\boldsymbol{\theta}_{k_0,1}^0, \delta)\} > 0$. The same holds for $p_2 = \mathbb{P}\{\boldsymbol{\pi}_{k_0,1} \in B_1(\boldsymbol{\pi}_{k_0,1}^0, \delta)\}$ and $p_3 = \mathbb{P}\{\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta)\}$. Thus, independence between $\boldsymbol{\pi}_{k_0,1}$, $\boldsymbol{\theta}_{k_0,1}$ and λ implies

$$\mathbb{P}\left(\boldsymbol{\xi}_{k_0} \in \Theta_{k_0} : \int_{\mathbb{R}} \log \left\{ \frac{f_0(x; \boldsymbol{\xi}_{k_0}^0)}{f(x; \boldsymbol{\xi}_{k_0})} \right\} f_0(x; \boldsymbol{\xi}_{k_0}^0) dx < \varepsilon\right) \geq p_1 p_2 p_3 > 0.$$

A.7 Proof of Lemma 2.4.4.

As already mentioned at the beginning of Subsection 2.3.3, $\boldsymbol{\theta}_{k,1} \sim \text{NRep}_{k,1}(\mu, \sigma^2, \tau)$ is an exchangeable distribution in $\theta_1, \dots, \theta_k$ 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 for all $t \in (0, \infty)$

$$\mathbb{P}(|\theta_i| > t) \leq \frac{c_{k-1}}{c_k} \int_{A_t} \text{N}(x; \mu, \sigma^2) dx = \frac{c_{k-1}}{c_k} \int_{B_t} \text{N}(s; 0, 1) ds.$$

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

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

Set $\gamma = \max\{2|\mu| + 1, (2 + \sqrt{2})|\mu|\} \in (0, \infty)$. By Mill's Inequality, for all $t \in [\gamma, \infty)$

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

Using the previous information

$$\mathbb{P}(|\theta_i| > t) \leq \frac{2}{(2\pi)^{1/2}} \sigma (|\mu| + 1)^{-1} \exp\{-(4\sigma^2)^{-1}t^2\}$$

for all $t \in [\gamma, \infty)$ and $i \in [k]$.

A.8 Proof of Lemma 2.4.5.

By the Change of Variables Theorem and Fubini's Theorem, it can be shown that for all $k \geq 2$ ($k \in \mathbb{N}$)

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

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

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

Notice that $F_{k-1} \in C(\mathbb{R}_{k-1}^1; (0, 1))$ (as a consequence of Lebesgue's Dominated Convergence Theorem) and $F_{k-1}(\boldsymbol{\theta}_{-1,1}) \rightarrow 1$ as $\|\boldsymbol{\theta}_{-1,1}\| \rightarrow \infty$. By Jensen's Inequality, for all $\boldsymbol{\theta}_{-1,1} \in \mathbb{R}_{k-1}^1$

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

Now

$$\left| \int_{\mathbb{R}} N(\theta_1; 0, 1) \log \left[1 - \exp \left\{ -\frac{(\theta_1 - \theta_j)^2}{2\tau} \right\} \right] d\theta_1 \right| \leq -2 \frac{\tau^{1/2}}{\pi^{1/2}} \int_0^\infty \log\{1 - \exp(-\theta_1^2)\} d\theta_1.$$

Using the substitution $\theta_1(z) = z^{1/2} : z \in (0, \infty)$ and then integrating by parts

$$\int_0^\infty \log\{1 - \exp(-\theta_1^2)\} d\theta_1 = - \int_0^\infty \frac{z^{3/2-1}}{\exp(z) - 1} dz = -\Gamma(3/2)\zeta(3/2) \in (-\infty, 0)$$

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

$$\left| \int_{\mathbb{R}} N(\theta_1; 0, 1) \log \left[1 - \exp \left\{ -\frac{(\theta_1 - \theta_j)^2}{2\tau} \right\} \right] d\theta_1 \right| \leq 2.6124\tau^{1/2} \in (0, \infty).$$

With this bound, defining $A_2 = 2.6124\tau^{1/2}$ and $A_1^{-1} = \exp(A_2)$ the following holds: for all $\boldsymbol{\theta}_{-1,1} \in \mathbb{R}_{k-1}^1$

$$\log\{F_{k-1}(\boldsymbol{\theta}_{-1,1})\} \geq -(k-1)A_2$$

which implies

$$F_{k-1}(\boldsymbol{\theta}_{-1,1}) \geq A_1^{-1} \exp(-A_2 k).$$

To conclude the proof, notice that

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

Using the previous equation it follows that for all $k \geq 2$ ($k \in \mathbb{N}$)

$$c_k \geq A_1^{-1} \exp(-A_2 k) c_{k-1} > 0,$$

the above being equivalent to

$$0 < \frac{c_{k-1}}{c_k} \leq A_1 \exp(A_2 k).$$

A.9 Proof of Proposition 2.4.6.

Following Theorem 3.1 in Scricciolo (2011) $p = 2$ induce a (finite) Gaussian Mixture Model, $\lambda \sim \text{IG}(a, b) : a, b \in (0, \infty)$ satisfy (i) and $\boldsymbol{\pi}_{k,1} \sim \text{Dir}(k^{-1} \mathbf{1}_k)$ 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 $\boldsymbol{\theta}_{k,1} \sim \text{NRep}_{k,1}(\mu, \sigma^2, \tau)$. 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 $E(K)\Pi([-a_n, a_n]^c)$. This quantity is part of the chain of inequalities

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

under the conditions (ii) and (iv). In our case, $\rho(i) = \kappa(i)$ for $i \in \mathbb{N}$. By way of Lemma 2.4.4

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

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

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

and by Lemma 2.4.5

$$\sum_{i=1}^{k_n} i \rho(i) \frac{c_{i-1}}{c_i} \leq A_1 B_1 \sum_{i=1}^{\infty} i \exp\{-(B_2 - A_2)i\} \in (0, \infty).$$

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):

$$\mathbb{P}\{\boldsymbol{\theta}_{k_0} \in B(\boldsymbol{\theta}_{k_0}^0; \varepsilon)\} = \Pi^{\otimes k_0} \{B(\boldsymbol{\theta}_{k_0}^0; \varepsilon)\} \gtrsim \exp\{-d_1 k_0 \log(1/\varepsilon)\}$$

In our case, $\boldsymbol{\theta}_{k_0} = \boldsymbol{\theta}_{k_0,1}$, $\boldsymbol{\theta}_{k_0}^0 = \boldsymbol{\theta}_{k_0,1}^0$ and $B(\boldsymbol{\theta}_{k_0}^0; \varepsilon) = D_1(\boldsymbol{\theta}_{k_0,1}^0, \varepsilon)$. At the end of the proof of Lemma 2.4.2 it is shown that for every $\delta = \varepsilon \in (0, \delta_0]$

$$\mathbb{P}\{\boldsymbol{\theta}_{k_0,1} \in D_1(\boldsymbol{\theta}_{k_0,1}^0, \varepsilon)\} \geq \frac{R_0}{c_{k_0}} \left(\prod_{i=1}^{k_0} \frac{S_i^0}{2} \right) \exp\{-k_0 \log(1/\varepsilon)\}.$$

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

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

Appendix B

Supplementary Material for Chapter 3

B.1 Algorithm RGMMx

In what follows we describe the Gibbs Sampler for the RGMMx in its entirety. 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 $z_i^{(0)} : i \in [n]$, $\boldsymbol{\pi}_{k,1}^{(0)}$ and $\boldsymbol{\theta}_{k,d+p}^{(0)}, \boldsymbol{\Lambda}_{k,d+p}^{(0)} : j \in [k]$. Set $\boldsymbol{\Gamma}_j = \mathbf{O}_{d+p} : j \in [k]$, where \mathbf{O}_{d+p} is the null matrix of dimension $(d+p) \times (d+p)$.
- (Burn-in phase) For $t = 0, \dots, B - 1$:
 1. $(z_i^{(t+1)} \mid \dots) \sim \mathbb{P}(z_i^{(t+1)} = j) = \pi_j^{(t,i)}$ independently for each $i \in [n]$, where,

$$\pi_j^{(t,i)} = \left\{ \sum_{l=1}^k \pi_l^{(t)} \mathbb{N}_{d+p}(\mathbf{y}_i; \boldsymbol{\theta}_l^{(t)}, \boldsymbol{\Lambda}_l^{(t)}) \right\}^{-1} \pi_j^{(t)} \mathbb{N}_{d+p}(\mathbf{y}_i; \boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Lambda}_j^{(t)}) : j \in [k].$$

2. $(\boldsymbol{\pi}_k^{(t+1)} \mid \dots) \sim \text{Dir}(\boldsymbol{\alpha}_{k,1}^t)$, where

$$\begin{aligned}\boldsymbol{\alpha}_{k,1}^{(t)} &= (\alpha_1 + n_1^{(t+1)}, \dots, \alpha_k + n_k^{(t+1)}) \\ n_j^{(t+1)} &= \text{card}(i \in [n] : z_i^{(t+1)} = j) : j \in [k].\end{aligned}$$

3. For $j = 1, \dots, k$:

3.1. $\boldsymbol{\theta}_j^{(*)}$ from $N_{d+p}(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Omega}_j^{(t)})$, where

$$\boldsymbol{\Omega}_j^{(t)} = \{\boldsymbol{\Sigma}^{-1} + n_j^{(t+1)}(\boldsymbol{\Lambda}_j^{(t)})^{-1}\}^{-1}.$$

3.2. Update $\boldsymbol{\theta}_j^{(t)} \rightarrow \boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(*)}$ with probability $\min(1, \beta_j)$, where

$$\beta_j = \frac{N_{d+p}(\boldsymbol{\theta}_j^{(*)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{N_{d+p}(\boldsymbol{\theta}_j^{(t)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})} \prod_{l \neq j}^k \left[\frac{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(*)} - \boldsymbol{\theta}_l^{(t)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(*)} - \boldsymbol{\theta}_l^{(t)})\}}{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_l^{(t)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_l^{(t)})\}} \right].$$

In the above expression for β_j

$$\begin{aligned}\boldsymbol{\Sigma}_j^{(t)} &= \{\boldsymbol{\Sigma}^{-1} + n_j^{(t+1)}(\boldsymbol{\Lambda}_j^{(t)})^{-1}\}^{-1} \\ \boldsymbol{\mu}_j^{(t)} &= \boldsymbol{\Sigma}_j^{(t)} \{\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + (\boldsymbol{\Lambda}_j^{(t)})^{-1} \boldsymbol{s}_j^{(t)}\} : \boldsymbol{s}_j^{(t)} = \sum_{i=1}^n \mathbb{I}_{\{j\}}(z_i^{(t+1)}) \boldsymbol{y}_i.\end{aligned}$$

Otherwise, set $\boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(t)}$.

3.3. Update $\boldsymbol{\Gamma}_j \rightarrow \boldsymbol{\Gamma}_j + B^{-1} \boldsymbol{\Omega}_j^{(t)}$.

4. $(\boldsymbol{\Lambda}_j^{(t+1)} \mid \dots) \sim \text{IW}_{d+p}(\boldsymbol{\Psi}_j^{(t)}, \nu_j^{(t)})$ independently for each $j \in [k]$, 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 $t = B, \dots, ST + B - 1$: 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 $\boldsymbol{\theta}_j^*$ from $N_{d+p}(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Gamma}_j)$.

Finally, save the generated samples every T th iteration.

- (Posterior conditional predictive estimates) With the T saved samples, compute

$$f(\mathbf{y} \mid \mathbf{x}, \mathbf{u}_1, \dots, \mathbf{u}_n) \approx \frac{1}{T} \sum_{t=1}^T \left\{ \sum_{j=1}^k \pi_j^{(t)}(\mathbf{x}) N_{d+p}(\mathbf{u}; \boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Lambda}_j^{(t)}) \right\}$$

$$\mathbb{E}[\mathbf{y} \mid \mathbf{x}, \mathbf{u}_1, \dots, \mathbf{u}_n] \approx \frac{1}{T} \sum_{t=1}^T \left\{ \sum_{j=1}^k m_j^{(t)}(\mathbf{x}) \pi_j^{(t)}(\mathbf{x}) N_p(\mathbf{x}; (\boldsymbol{\theta}_j^{\mathbf{x}})^{(t)}, (\boldsymbol{\Lambda}_j^{\mathbf{x}\mathbf{x}})^{(t)}) \right\}$$

where $\mathbf{u} = (\mathbf{y}, \mathbf{x})$ and

$$\pi_j^{(t)}(\mathbf{x}) = \left[\frac{1}{T} \sum_{s=1}^T \left\{ \sum_{l=1}^k \pi_l^{(s)} N_p(\mathbf{x}; (\boldsymbol{\theta}_l^{\mathbf{x}})^{(s)}, (\boldsymbol{\Lambda}_l^{\mathbf{x}\mathbf{x}})^{(s)}) \right\} \right]^{-1} \pi_j^{(t)}$$

$$m_j^{(t)}(\mathbf{x}) = (\boldsymbol{\theta}_j^{\mathbf{y}})^{(t)} + (\boldsymbol{\Lambda}_j^{\mathbf{y}\mathbf{x}})^{(t)} \{ (\boldsymbol{\Lambda}_j^{\mathbf{x}\mathbf{x}})^{(t)} \}^{-1} (\mathbf{x} - (\boldsymbol{\theta}_j^{\mathbf{x}})^{(t)}).$$

Bibliography

- Azzalini, A. and Bowman, A. W. (1990), “A Look at Some Data on the Old Faithful Geyser,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39, 357–365.
- Barrientos, A. F., Jara, A., and Quintana, F. A. (2012), “On the Support of MacEachern’s Dependent Dirichlet Processes and Extensions,” *Bayesian Anal.*, 7, 277–310.
- 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 Process 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.
- Diggle, P. J. (2013), *Statistical analysis of spatial and spatio-temporal point patterns*, CRC Press.

- Dunson, D. B., Pillai, N., and Park, J.-H. (2007), “Bayesian density regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 163–183.
- 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,” .
- Gaetan, C., Guyon, X., and Bleakley, K. (2010), *Spatial statistics and modeling*, vol. 81, Springer.
- 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.
- Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010), *Handbook of spatial statistics*, CRC press.
- 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.
- Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010), *Bayesian Nonparametrics*, vol. 28, Cambridge University Press.
- Hough, J. B., Krishnapur, M., Peres, Y., and Virág, B. (2006), “Determinantal Processes and Independence,” *Probability Surveys*, 3, 206–229.
- 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.
- MacEachern, S. N. (2000), “Dependent dirichlet processes,” *Unpublished manuscript, Department of Statistics, The Ohio State University*, 1–40.

- Mateu, J. and Montes, F. (2000), “Approximate maximum likelihood estimation for a spatial point pattern,” *Qüestiió*, 24, 3–25.
- McLachlan, G. and Peel, D. (2005), *Finite Mixture Models*, John Wiley & Sons, Inc.
- McLachlan, G. J. and Peel, D. (2000), *Finite mixture models*, New York: Wiley Series in Probability and Statistics.
- Møller, J. and Waagepetersen, R. P. (2003), *Statistical inference and simulation for spatial point processes*, CRC Press.
- Müller, P., Erkanli, A., and West, M. (1996), “Bayesian curve fitting using multivariate normal mixtures,” *Biometrika*, 83, 67–79.
- Müller, P., Quintana, F., and Rosner, G. L. (2011), “A Product Partition Model With Regression on Covariates,” *Journal of Computational and Graphical Statistics*, 20, 260–277.
- Müller, P. and Quintana, F. A. (2004), “Nonparametric Bayesian data analysis,” *Statistical Science*, 95–110.
- Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, Springer.
- 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.
- Neal, R. M. (2000), “Markov Chain Sampling Methods for Dirichlet Process Mixture Models,” *Journal of Computational and Graphical Statistics*, 9, 249–265.

- 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.
- Robert, C. P. and Casella, G. (2005), *Monte Carlo Statistical Methods (Springer Texts in Statistics)*, Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Roberts, G. O. and Rosenthal, J. S. (2009), “Examples of Adaptive MCMC,” *Journal of Computational and Graphical Statistics*, 18, 349–367.
- Rodriguez, A. and Dunson, D. B. (2011), “Nonparametric Bayesian models through probit stick-breaking processes,” *Bayesian analysis (Online)*, 6.
- 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.
- Sethuraman, J. (1994), “A constructive definition of Dirichlet priors,” *Statistica Sinica*, 639–650.
- 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.