University of Milan – Bicocca

Università della Svizzera italiana

Doctoral Thesis

Bayesian Mixtures for Large Scale Inference

Author:
Francesco Denti

Supervisor:
Prof. Antonietta Mira

Internal supervisor:
Prof. Fulvia Mecatti

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

Department of Statistics and Quantitative Methods at UNIMIB
Department of Economics at USI

February 7, 2020

A chi mette la testa fuori.
A chi ti spinge a farlo.
E a voi tutti, così comprensivi e pazienti.

"We all wanna be big, big stars
Yeah, but we got different reasons for that
Believe in me because I don't believe in anything
and I want to be someone to believe."
Mr. Jones - Counting Crows
(Sung somewhere in Irive, playing Rockband and drinking I.P.A.)

"C'è chi vuole che io taccia adesso
Bene, meglio fare un pezzo strumentale
Che un trattato che non faccia testo
L'importante è che si faccia presto!
Il testo che avrei voluto scrivere
Non è di certo questo."
Il testo che avrei voluto scrivere - Michele Salvemini

"Bicocca, mein herz in flammen
Will dich lieben und verdammen
Bicocca, dein atem kalt
So jung, und doch so alt."
~~*Dutschland*~~ *Bicocca - Rammstein*

Abstract

Francesco Denti

Bayesian Mixtures for Large Scale Inference

Bayesian mixture models are ubiquitous in statistics due to their simplicity and flexibility and can be easily employed in a wide variety of contexts. In this dissertation, we aim at providing a few contributions to current Bayesian data analysis methods, often motivated by research questions from biological applications. In particular, we focus on the development of novel Bayesian mixture models, typically in a nonparametric setting, to improve and extend active research areas that involve large-scale data: the modeling of nested data, multiple hypothesis testing, and dimensionality reduction. Therefore, our goal is two-fold: to develop robust statistical methods motivated by a solid theoretical background, and to propose efficient, scalable and tractable algorithms for their applications.
The thesis is organized as follows. In Chapter 1 we briefly review the methodological background and discuss the necessary concepts that belong to the different areas that we will contribute to with this dissertation.
In Chapter 2 we propose a Common Atoms model (CAM) for nested datasets, which overcomes the limitations of the nested Dirichlet Process, as discussed in Camerlenghi et al. (2019b). We derive its theoretical properties and develop a slice sampler for nested data to obtain an efficient algorithm for posterior simulation. We then embed the model in a Rounded Mixture of Gaussian kernels framework to apply our method to an abundance table from a microbiome study. In Chapter 3 we develop a BNP version of the two-group model (Efron, 2004), modeling both the null density f_0 and the alternative density f_1 with Pitman-Yor process mixture models. We propose to fix the two discount parameters σ_0 and σ_1 so that $\sigma_0 > \sigma_1$, according to the rationale that the null PY should be closer to its base measure (appropriately chosen to be a standard Gaussian base measure), while the alternative PY should have fewer constraints. To induce separation, we employ a non-local prior (Johnson and Rossell, 2010) on the location parameter of the base measure of the PY placed on f_1. We show how the model performs in different scenarios and apply this methodology to data from microbiome and prostate cancer experiments. Chapter 4 presents a second proposal for the two-group model. Here, we make use of non-local distributions to model the alternative density directly in the likelihood. We propose both a parametric and a nonparametric formulation of the model. We provide a theoretical justification for the adoption of this approach and, after comparing the performance of our model with several competitors, we present three applications on real, publicly available genomic datasets. In Chapter 5 we focus on improving the model for intrinsic dimensions (IDs) estimation discussed in Allegra et al. (2019). In particular, the authors estimate the IDs modeling the ratio of the distances from a point to its first and second nearest neighbors (NNs). First, we propose to include more suitable priors in their parametric, finite mixture model. Then, we extend the existing theoretical methodology by deriving closed-form distributions for the ratios of distances from a point to two NNs of generic order. We propose a simple Dirichlet process mixture model, where we exploit the novel theoretical results to extract more information from the data. The chapter is then concluded with simulation studies and the application to real data. Finally, Chapter 6 presents the future directions and conclusions.

Contents

List of Figures ix

List of Tables xiii

1 A review of paradigms and tools used in the Dissertation 1
1.1 Introduction . . . 1
1.2 A review of the Bayesian paradigm and some relevant priors . . . 2
1.2.1 Two examples of parametric priors . . . 3
Non-Local Distributions . . . 3
Repulsive Distributions . . . 4
1.2.2 Bayesian Nonparametrics (BNP) . . . 5
Dirichlet Process . . . 5
Pitman-Yor Process . . . 8
Dirichlet Process Mixture Model . . . 9
Posterior Computations . . . 11
1.2.3 Models for Partially Exchangeable Data . . . 12
Exchangeable Partition Probability Functions and the limitations of the nDP construction . . . 14
A Mixed nested Dirichlet Process approach . . . 16
1.3 Multiple Hypothesis testing . . . 17
1.4 The Estimation of the Intrinsic Dimension of a dataset . . . 20
1.5 Outline and main contributions . . . 22

Appendix 25
1.A Gibbs sampler for the Mixed nested DP . . . 25

2 Bayesian Nonparametric Analysis of Nested Data via Common Atom priors 33
2.1 Introduction . . . 34
2.2 Common Atoms Model for Continuous Measurements . . . 36
2.2.1 Properties of the Common Atoms model . . . 37
2.2.2 Common Atoms Mixture Model . . . 39
2.3 Common Atoms Model for Count Data . . . 39
2.4 Posterior Inference . . . 40
2.5 Analysis of microbial distributions of infants in low-income countries . . . 42
2.6 Simulation study . . . 48
2.7 Discussion . . . 49

Appendix 51
2.A Proofs . . . 51
2.B Truncated Blocked Gibbs Sampler for CAM . . . 55
2.B.1 Error bounds estimation . . . 58
2.C Summary of Differences between nDP, HDP and CAM . . . 61
2.D Densities of the three scenarios considered in the simulation study . . . 62

3 Two-group Poisson-Dirichlet mixtures for multiple testing **63**
3.1 Introduction . . . 64
3.2 A review of the 2PPD process . . . 65
3.3 Methods . . . 66
3.3.1 A two-group 2PPD model . . . 66
3.3.2 Bayesian hierarchical two-group mixture model . . . 67
3.3.3 Posterior inference . . . 69
3.4 Applications . . . 70
3.4.1 Simulation study . . . 70
3.4.2 Case study: Microbiome data . . . 73
3.4.3 Case study: Prostate Cancer Dataset . . . 75
3.5 Discussion and Conclusion . . . 75

Appendix **79**
3.A Posterior inference . . . 79
3.B Proof of Proposition 1. . . . 83
3.C Split-Merge move for the two-group 2PPD model . . . 83
3.D Microbiome data case study: list of differentially abundant taxa . . . 86
3.E Plots of the five scenarios considered in the simulation study . . . 88
3.F Computational Burden . . . 88
3.G Additional simulation study . . . 89

4 Bayesian Two-Group Model: a Non-Local Likelihood Approach **91**
4.1 Introduction . . . 92
4.2 Non-local Likelihood . . . 93
4.2.1 Weighted densities and Non-local Distributions . . . 93
4.2.2 Non-local two-group Model . . . 94
A Bayesian Nonparametric Alternative . . . 96
4.2.3 On the choice of the weight function . . . 97
4.3 Theoretical justifications . . . 99
4.4 Posterior Computation . . . 101
4.4.1 Inference on f_0 and f_1 . . . 101
4.4.2 Gibbs Sampler . . . 102
Parametric model specification . . . 102
Nonparametric model specification . . . 103
4.5 Applications . . . 104
4.5.1 Simulated Data . . . 104
4.5.2 Gene Expression Case Studies . . . 107
Alon Microarray Data . . . 107
Microbiome Abundance table: Kostic Dataset . . . 108
Grouped Proteomics Data: Ubiquitin-protein interactors . . . 109
4.6 Discussion . . . 110

Appendix **113**
4.A Sampling from Weighted Distributions . . . 113
4.B FDR, FNR, Type II error as function the Acceptance Region . . . 114

5 Bayesian Mixture Models for Intrinsic Dimension Estimation **117**
5.1 Introduction . . . 118
5.2 Background: the TWO-NN estimator and Hidalgo . . . 119
5.2.1 Alternative parametric prior specifications for d . . . 122

5.3 Extending Hidalgo to Consecutive Ratios (CRIME) . . . 124
5.3.1 Extension to the Nonparametric Case . . . 127
5.4 Posterior Inference . . . 129
5.4.1 MCMC algorithm . . . 129
5.4.2 Post-processing the MCMC output . . . 130
5.5 Applications . . . 131
5.5.1 Simulation Study . . . 131
5.5.2 Application to real data . . . 132
Leukemia Dataset . . . 132
Schimtz dataset . . . 133
5.6 Future Directions and Conclusions . . . 136
5.6.1 A Dependent Dirichlet Process Approach . . . 136
5.6.2 The representative individual. . . . 137
5.6.3 Discussion and Conclusions . . . 140

Appendix **141**
5.A Proofs of theoretical results . . . 141
5.A.1 Proof of Lemma 1 . . . 141
5.A.2 Proof of Lemma 2 . . . 141
5.A.3 Proof of Lemma 3 . . . 142
5.A.4 Distributions of the distances r_l . . . 145
5.B A general formula for sampling Interval-truncated random variables . . . 146
5.C Additional simulation with truncated priors . . . 149
5.D Five Gaussians Dataset . . . 150
5.E Schmitz dataset: list of 19 interesting genes . . . 150

6 Conclusions and Future Directions **151**

7 Acknowledgements **155**

Bibliography **159**

List of Figures

2.1 Histograms of the microbiome populations of two Mali children. As we can see, the distributions appear very similar and extremely skewed. 36
2.1 Pairwise Posterior Probability matrix of coclustering among the 212 subjects. A partition of the subjects' distributions into five clusters is obtained after minimization of the posterior expected Variation of information loss function. 44
2.2 Violin and boxplot combinations for three distributional summaries stratified by the partition clusters estimated using the Variation of Information criterion: a) α-diversity as measured by the Shannon entropy index; b) the mean of the non-zero entries in the abundance table for each subject; c) the percentage of OTU sequences with zero counts. 45
2.3 The Entropy index $\bar{H}^{i,g}$ described in Section 2.5: for each row, representing a specific OTU, we observe the Entropy indexes stratified by the optimal distributional allocation of subjects, which has been estimated using the Variation of Information criterion. 47
2.D.1 The densities distributions of each unit in three scenarios considered. 62

3.1 Histogram and two-group Posterior 74
3.2 Histogram and two-group Posterior 76
3.E.1 The five Scenarios considered in the simulation study in Section 4.1. We plot the histograms of the simulated data and superimpose the null (blue) and alternative (red) density functions. 88
3.G.1 Posterior probability of inclusion for observations sampled in the first dataset of the five different scenarios, estimated adopting $\sigma_0 = 0.1$ and $\sigma_1 = 0.75$ 90

4.1 Visualization of the dependence between the latent variables. The rectangular leaves of the tree report the variables of the model. The bottom line of rectangular leaves contains the r.v.s representing probabilities: connected to them we find circular leaves, containing the corresponding densities for each scenario. 96
4.2 The panel in the first row report the different behaviours of the weight function w_1 when the *value* of ξ (I) and k (II) changes, keeping the other parameter fixed equal to 2. The second row shows the same for the weight function w_2. 98
4.1 Alon Dataset. Left panel: Density estimation a posteriori of the global density f (black), the null density f_0 (blue) and the alternative density f_1 (red). Right panel: Histogram of the data with the Posterior Probability of Inclusion function $(1 - locfdr(z))$ superimposed, for both Efron's `locfdr` (blue) and Nollik (red). The dashed line represents the threshold controlling for a BFDR of 5%. 108
4.2 Kostic Dataset. Left panel: Density estimation a posteriori of the global density f (black), the null density f_0 (blue) and the alternative density f_1 (red). Right panel: Histogram of the data with the Posterior Probability of Inclusion function $(1 - locfdr(z))$ superimposed, for both Efron's `locfdr` (blue) and Nollik (red). The dashed line represents the threshold controlling for a BFDR of 5%. 109

4.3 Ubiquitin-protein interactors dataset. Histograms of the three groups present in the data with the Posterior Probability of Inclusion function $(1 - locfdr(z))$ superimposed, one for each experiment. The three dashed lines represent the thresholds for each group, controlling for a BFDR of 5%. 111
4.A.1 Histograms referring to the two distributions adopted in (S1) and (S2) 114

5.1 A pictorial example in $\mathbb{R}^3$ of the quantities involved. The points represent the data. The selected observation, $\boldsymbol{x}_i$ is connected by dashed lines representing the distances $r_{i,j}$, $j = 1, 2, 3$ to its first three NNs. The different spherical shells, characterized by different colors, have area $v_{i,j}$, $j = 1, 2, 3$. 119
5.2 The graphs reports four Pareto densities characterized by the same scale parameter equal to 1 and different shape parameters, ranging from 0.1 to 2. Even for very different shapes, the densities overlap to a great extent. 121
5.3 `iris` dataset. The top-left panel reports the posterior means and the 90% credible sets for each observation, after applying Hidalgo with K=3. The panels on the right show the same output when a truncated Gamma (top) or a Uniform (bottom) prior on d is adopted. The bottom-left panel shows the results when a Truncated Gamma is mixed with a point mass in D, with prior mixture proportion equal to $\hat{\rho} = 0.9$ 123
5.4 `growth` dataset. The left panel shows the smoothed growth curves for both male (1, in blue) and female (2, in red). The right panel reports the posterior median of the IDs obtained with Hidalgo, where $K = 2$, comparing the output when a classical independent prior (circles) is adopted against when d is modeled with a repulsive prior (triangles). 124
5.1 Each panel shows the MLE of the ID for an increasing number of consecutive ratios $\mu_{l,i}$ considered in the estimation. The five lines correspond to the average $\hat{d}_{MLE}$, while the shaded area highlights the interval between the 10th and 90th quantile. 126
5.1 `Golub` dataset. The graph shows the boxplots of the median IDs estimated for the 72 patients, stratified by the recovered clusters. The different colors are associated with prognosis. The different horizontal lines highlight the median ID per cluster. 133
5.2 `Schmitz` dataset. Survivial curves based on Kaplan-Meier estimates (top panels) and table of number at risk (bottom panel) in the 3 estimated clusters. 134
5.3 `Schmitz` dataset. Survivial curves based on Kaplan-Meier estimates (top panels) and table of numbers at risk (bottom panel) in the 3 retrieved clusters. 135
5.1 `Iris` dataset - Median Intrinsic Dimension a posteriori obtained fitting a Dependent Dirichlet Process approached applied to Crime model, K=10 137
5.2 Top panel: `Spiral` Dataset - 1000 observations (gray dots) of intrinsic dimension $d = 2$, embedded in a $D = 3$-dimensional space. Bottom panel: `Iris` Dataset - Observations (gray dots) of dimension $d_1 = 4$, $d_2 = 3$, and $d_3 = 3$ embedded in a $D = 4$ dimensional space. The points are projected onto the second and the third dimensions. The undirected edges indicates whether or not two points are connected according to the matrix $\hat{\mathcal{N}}^q$. Representative individuals obtained with different methods are highlighted. In detail: `CW-Average`, `CW-Median` are the coordinate-wise average and median, `MV-Median` represents the multivariate median, `PAM` denotes the overall medoid, `Floyd` and `Spher` label the RIs found applying graphical-based methods 139
5.A.1 The various panels report the shapes of the density in (5.28) for different values of d, n_1 and n_2 144
5.A.2 The various panels report the shapes of the density in (5.29) for different values of d, n_1 and n_2 144

5.B.1 Four different applications of the the inverse c.d.f. method based on interval-truncated c.d.f.s. The plots report the histograms (top panels) sampled from specific distributions (bottom panels). 148
5.C.1 `Uniform clouds` dataset. The top-left panel reports the posterior means and the 90% credible sets for each observation, after applying Hidalgo with K=4. The panels on the right show the same output when a truncated Gamma (top) or a Uniform (bottom) prior on d is adopted. The bottom-left panel shows the results when a Truncated Gamma is mixed with a point mass in D, with prior mixture proportion equal to $\hat{\rho} = 0.9$. The dashed horizontal lines denote the true values of the different IDs. 149
5.D.1 The first three coordinates (out of 10) of the `5 Gaussians` dataset used in Allegra et al. (2019). Evidently, the five distributions are overlapping. The 1-dimensional Gaussian is hidden in the main cloud of points. 150

List of Tables

2.1 Posterior median, 95% credible intervals and posterior probability $\mathbb{P}(\beta > 0|\text{data})$ for each coefficient from a Bayesian regression models, to study the association between diarrhea onset (left) and Age (right) with the estimated cluster allocation. 45
2.2 Summary statistics stratified by observational cluster, considering the $> 30k$ OTU counts as units. 46
2.3 Top three OTU species in Observational Clusters 2 and 3 in terms of presence across individuals. 47
2.1 True number of clusters, detected number of clusters, Adjusted Rand Index and Classification Error computed in the three simulation scenarios to quantitatively compare the clustering performance 50

3.1 Sensitivity to σ_i 71
3.2 Bakeoff table 72
3.D.1 Microbiome data case study: differentially abundant taxa with negative z-scores indicating less abundance in the children with moderate to severe diarrhea. Most are well known commensal bacteria, e.g. *Prevotella* spp. and *Clostridium* spp. Posterior probability that the z-score belongs to the non-null group is given for each taxa. The dotted line highlights the difference between the genes flagged as relevant by our method and the ones found with the *locfdr* model. 86
3.D.2 Microbiome data case study: differentially abundant taxa with positive z-scores indicating greater abundance in the children with moderate to severe diarrhea. Most are well known pathogenic bacteria, e.g. *Shigella* spp. and *E. coli*. Posterior probability that the z-score belongs to the non-null group is given for each taxa. The dotted line highlights the difference between the genes flagged as relevant by our method and the ones found with the *locfdr* model. 87
3.F.1 Computational time in seconds for 100 iterations with the Polya Urn Scheme (PUS) sampler, compared with the Split-Merge (SM) scheme. We investigate how the time varies as the sample size n and the discount parameter of the null process change. 89
3.G.1 Simulation study: sensitivity results across the five simulation scenarios considered in Section 4.1 ($\rho = 0.05$), modeled with PYs processes characterized by $\theta_0 = \theta_1 = 1$ and $\sigma_0 = 0.1 < \sigma_1 = 0.75$. The values in the table represent the average MCC and F_1 scores, the average precision (PRE), specificity (SPEC), accuracy (ACC) and the area under the curve (AUC) of the corresponding receiver operating characteristic curve, over 30 replicates with corresponding standard deviations between brackets. 90

4.1 Elapsed time in seconds of the Nollik parametric and nonparametric models using w_1 as weight function to obtain 1,000 iterations for different values of the sample size. 104

4.1 Simulation study. Posterior Probability of Inclusion thresholded for BFDR=0.05. The *locfdr* and *MixFDR* provide estimates for the $lfdr$, with threshold at 0.2 as suggested in Efron (2007) and Muralidharan (2012). BH adjusted p-values thresholded at 5%. The table shows the performance of 7 different models: Nollik with quadratic (MOM), w_1 and w_2 - as in (4.13) -as weight functions, the BNP Nollik with w_1, the MixFDR of Muralidharan (2012), the LocFDR of Efron (2007) and the classical BH procedure. The performances have been measured in terms of Accuracy (ACC), Specificity (SPE), Precision (PRE), and AUC. Moreover, we compare the Matthew's Correlation Coefficient (MCC) and the F_1 score. The highest MCC and F_1 scores among the *Nollik* models and among the competitors are highlighted. 106

5.1 Median values of the estimated ID and corresponding width of the credible set between the 10th and 90th quantile across 50 different samples, varying the sample size from $n = 50$ to $n = 10000$. In particular, $n_1 = 50$, $n_2 = 500$, $n_3 = 2500$, $n_4 = 5000$, and $n_5 = 10000$. As expected, the underestimation typical of higher dimensionality is less evident as more data are used. Using more than one ratio helps lowering the variance of the estimates. 128

5.1 Simulation Study. The table shows the estimated ID for each of the estimated clusters, recovered minimizing the Variation of Information as in Wade and Ghahramani (2015). GT indicates the ground truth, ARI the Adjusted Rand Index. In these cases, we fix $q = 3$ and $\zeta = 0.75$ 131

5.2 Simulation Study. The table shows the estimated ID for each of the estimated clusters, recovered minimizing the Variation of Information as in Wade and Ghahramani (2015). The number of observation is fixed, equal to n=500. 132

5.3 Simulation Study. `5 Gaussians` dataset. The table shows the estimated ID for each of the estimated clusters. 132

5.4 `Golub` Dataset. Average, standard deviation of the median ID and AML proportion found in each cluster. 133

5.5 `Schmitz` Dataset. Different characteristics of the three estimated clusters. 134

Chapter 1

A review of paradigms and tools used in the Dissertation

"Tu eri per me la consapevolezza
che con l'aiuto del tempo anche un Magikarp è in grado
di diventare Gyarados."
Tetris – Pinguini Tattici Nucleari

"I'm Dr. Ross Geller."
"Ross, please, this is a hospital, okay? That actually means something here."
Friends, s10e13

1.1 Introduction

Bayesian statistics has experienced spectacular growth over the last few decades. There are multiple reasons for its popularity, ranging from its conceptually intuitive paradigm, the advances in computational techniques, to the ease of the interpretation of its results. It is often noted that the way a Bayesian learns from data naturally resembles the way knowledge evolves: an initial (prior) belief, after observing data from real phenomena (likelihood), is updated into a new, more complete idea (posterior) (Bain, 2016).
Despite its theoretical and methodological appeal, Bayesian statistics had fundamental breakthroughs only after the development of computational methods. Before the advent of MCMC techniques for posterior inference, many Bayesian models for complex applications were often analytically intractable.
The wide use of Metropolis Hasting and the Gibbs sampler algorithms (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; Gelfand, 1990) allows the possibility to apply the Bayesian paradigm to investigate ever new and complex datasets, from different fields. It is no exaggeration that the Metropolis–Hastings algorithm and its extensions transformed Bayesian statistics from a theoretical curiosity to its modern place as the inferential paradigm of choice (Dunson and Johndrow, 2019).
In this thesis, we will often focus on applications in biology: the advent of new generation genome sequencing techniques and the objective of personalized medicine have revolutionized the way scientific research is carried out, providing an abundance of data, which more and more accurately describe the various facets of complex natural phenomena. This, in turn, raises many new and interesting research questions. In Efron's words, *progress in statistics is usually at the mercy of our scientific colleagues, whose data is the "nature" from which we work* (Efron, 2012).

Indeed, new research questions pose challenges and opportunities for statisticians, who are called to devise new methods for capturing the hidden patterns in the data, oftentimes with algorithms that can supply results in a timely manner. Therefore, we aim at providing a few contributions to current Bayesian data analysis methods, often motivated by questions from biological applications. More specifically, our interest is in comparing and describing the heterogeneity of samples from observed populations.
This chapter succinctly introduces some of the tools that we will use in later chapters. In detail, in the next section we briefly summarize the Bayesian paradigm, and we consider several classes of parametric priors that will turn out to be useful for our modeling purposes. In particular, we will introduce the notion of non-local priors and repulsive distribution. We will then discuss Bayesian nonparametric (BNP) priors, with a focus on models for partially exchangeable data. In Section 3, we will present some introductory ideas about multiple hypothesis testing. More specifically, we will introduce the two-group model, first discussed in Efron (2004), which will be the focus of two of the projects in this dissertation, where we will use it for Bayesian Hypothesis testing applied to large scale inference. In Section 4, we will examine the concept of the Intrinsic Dimension (ID) of a dataset to pave the way for the last project, where we study how Bayesian nonparametric mixture models can contribute to the analysis of heterogeneous IDs. A more exhaustive outline of the dissertation projects concludes this chapter.

1.2 A review of the Bayesian paradigm and some relevant priors

Amidst all possible reasons that might persuade a statistician to undertake the Bayesian path, the celebrated de Finetti's representation theorem is one of the most convincing, giving a mathematical justification for this learning paradigm. Consider an ideally infinite sequence of observations $\{X_n\}_{n\geq 1}$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $(\mathbb{X}, \mathcal{X})$, where $\mathbb{X}$ is a Polish space and $\mathcal{X}$ its associated Borel σ-field. We need to postulate some degree of dependence among the data: we can perform inference and prediction only if the data show a pattern that we can learn. A fairly reasonable assumption is the *exchangeability* of the data: for every $n \geq 1$ the distribution of the random vector $(X_1, \ldots, X_n)$ is invariant under permutation of its components. In other words, for any permutation σ of the indexes $\{1, \ldots, n\}$, we have

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

In this way, we make a probability statement about the homogeneity of the observations, making possible the extraction of some "common signal" the data may conceal. Now, let $\mathbf{P}_{\mathbb{X}}$ be the space of all the probability measures on $\mathbb{X}$ and denote as $\mathcal{P}_{\mathbb{X}}$ its corresponding σ-field. The so-called de Finetti representation theorem states that the sequence $\{X_n\}_{n\geq 1}$ is exchangeable if and only if, for any $n \geq 1$ and $A_i \in \mathcal{X}$,

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \int_{\mathbf{P}_{\mathbb{X}}} \prod_{i=1}^{n} p\left(A_i\right) Q(dp). \tag{1.1}$$

The measure Q is a probability measure defined on $(\mathbf{P}_{\mathbb{X}}, \mathcal{P}_{\mathbb{X}})$, and it is usually referred to as de Finetti's measure. Q plays a pivotal role since it provides a formal justification of the existence of prior distributions and, more generally, of the entire Bayesian paradigm. This is even more evident if we rewrite the exchangeability assumption in the following hierarchical form:

$$X_i | \tilde{p} \stackrel{i.i.d.}{\sim} \tilde{p}, \qquad \tilde{p} \sim Q, \tag{1.2}$$

where $\tilde{p}$ is a random probability measure defined on $(\Omega, \mathcal{F}, \mathbb{P})$[1]. The representation theorem shows how statistical models emerge in a Bayesian context: in fact, if (and only if) the exchangeability assumption on $\{X_i\}_{i=1}^n$ holds, then *there exists* a random quantity (parameter) $\tilde{p}$ that has law Q. Conditionally on G, $\{X_i\}_{i=1}^n$ is regarded as an i.i.d. sequence. The nature of the model is then determined by the nature of the prior probability Q. If the distribution of Q has finite-dimensional support, then we will call the model (1.2) *parametric*. Otherwise, the model is called *nonparametric*.
The vast majority of Bayesian literature deals with parametric models (Ghosal and Vaart, 2017; Lindley and Schervish, 2006; Robert et al., 2015). A common way to state a parametric model is the following:

$$X_i|\theta \sim \tilde{p}_\theta, \quad \theta \sim \pi,$$

for $\theta \in \Theta \subset \mathbb{R}^d$, where π denotes a finite dimensional distribution. This parametric specification can be re-expressed according to (1.2), once we assume that

$$Q\left(\{p_\theta : \theta \in \Theta\}\right) = 1. \tag{1.3}$$

1.2.1 Two examples of parametric priors

A fundamental step in a Bayesian analysis is the choice of the prior distributions. Often conjugate priors are selected, mostly out of convenience. However, the particular nature of the analysis at hand can raise some specific needs, either from the theoretical and computational points of view. In those situations, it becomes necessary eliciting a clever, but still tractable and admissible, prior distribution. In the remainder of the dissertation, we will make use, in addition to conjugate and nonparametric priors, of two parametric distributions with useful properties, that we now present.

Non-Local Distributions

Non-local distributions were introduced by Johnson and Rossell (2010). In the paper, the authors provide the definition of non-local distribution as:

Definition 1. *Consider a random variable defined on a support Θ and let $\Theta_0 \subset \Theta$ describe a subset of the support. If for every $\varepsilon > 0$ there is $\zeta > 0$ such that*

$$\pi_{NL}(\theta) < \varepsilon \quad \text{for all } \theta \in \Theta : \inf_{\theta_0 \in \Theta_0} |\theta - \theta_0| < \zeta$$

then we define π_{NL} to be a non-local prior density.

Let us call a distribution that does not satisfy the property stated in Definition 1 a *local* distribution. In words, a non-local distribution is characterized by a density that vanishes on a (pre-determined) subset of the support. This property is appealing in the context of Bayesian Hypothesis Testing, formulated as $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs $H_1 : \boldsymbol{\theta} \in \Theta_1$ or, equivalently, $H_0 : \boldsymbol{\theta} \sim \pi(\theta_0)$ vs $H_1 : \boldsymbol{\theta} \sim \pi(\theta_1)$ where the non-local distributions are used as priors for the alternative scenario. The separation between the null distribution π_0 and the alternative distribution π_1 is necessary. Quoting Johnson and Rossell (2010), we require this property because *on a philosophical level [...], an alternative hypothesis, by definition, should reflect a theory that is fundamentally different from the null hypothesis. Local alternative hypotheses do not.* In their paper, the authors argue about the superiority of this prior specification by noting that local alternatives induce an unappealing large sample behavior of the Bayes Factor: as the sample size n increases, the evidence accumulates much more rapidly in favor of true alternative models

[1] With a small abuse of notation, unless otherwise indicated, we denote a distribution via its name or its density indistinctively in the model specifications that will follow.

than in favor of true null models. More precisely, for a true null hypothesis, the Bayes Factor in favor of the alternative hypothesis decreases only at rate $O_p\left(n^{-1/2}\right)$ while for a true alternative hypothesis, the Bayes Factor in favor of the null hypothesis decreases exponentially fast. The adoption of a non-local density for $\pi(\theta_1)$ alleviates this problem.
Examples of non-local priors are the so-called Moment (MOM) prior and the Inverse-Moment ($IMOM$) prior, characterized by the following forms:

$$\pi_{\text{MOM}}(\theta) = \frac{(\theta-\theta_0)^{2k}}{\mathcal{K}}\pi_{\text{L}}(\theta),$$
$$\pi_{\text{IMOM}}(\theta) = \frac{k\,\tau^{\nu/2}}{\Gamma(\nu/2k)}\left\{(\theta-\theta_0)^2\right\}^{-(\nu+1)/2}\exp\left[-\left\{\frac{(\theta-\theta_0)^2}{\tau}\right\}^{-k}\right],$$

where k is a positive integer, $\tau, \nu > 0$, $\mathcal{K}$ is the normalizing constant of π_{MOM} and π_L denotes a local density. We can see how $\pi_{MOM} \to 0$ as $\theta \to \theta_0$. The same happens with π_{IMOM}.
Since the first seminal paper by Johnson and Rossell (2010), many authors have contributed to this research area studying the properties of the non-local distributions in different settings. For example, Johnson and Rossell (2012), Rossell et al. (2013), Rossell and Telesca (2017), and Shi et al. (2019) extend the topic to model selection devising priors for the coefficients of Bayesian regressions in numerous different contexts, Consonni et al. (2012) use non-local priors for Directed Acyclic Graphs (DAGs), and Shi et al. (2019) provide results on model consistency in non-local settings.

Repulsive Distributions

Consider the sample $(X_1, \ldots, X_n)$ and suppose that the classical parametric distributions are not flexible enough to describe the data-generating process. Bayesian mixture models constitute a simple and flexible extension. Let $f(\cdot|\theta)$ be a parametric kernel. We can express a K-components mixture model as

$$X_i|\boldsymbol{\pi}, \boldsymbol{\theta} \overset{i.i.d.}{\sim} \sum_{k=1}^{K} \pi_k\, f(\cdot|\theta_k) \tag{1.4}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the mixture weights and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is the vector of the component-specific parameters. To adopt the Bayesian paradigm, we need to specify prior distributions for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$. Usually the component-specific parameters $\{\theta_k\}$ are assumed to be drawn independently from a common prior P_0. However, it may happen that the estimation process introduces redundant components. In other words, the Bayesian model might over-fit the data making use of similar parameters for different components. These excessive parameters are poorly informative, and have an adverse impact on density estimation, since this leads to an unnecessarily complex model. A solution to address this issue is to introduce some sort of repulsion among the parameters via suitable priors.
One of the approach used in the literature employs the *determinantal point process*, a stochastic point process that favors configurations of well separated points (Affandi et al., 2013; Affandi et al., 2014; Lavancier et al., 2015; Kulesza and Taskar, 2012; Xu et al., 2016).
Another approach consists in the definition of *repulsive densities*, and we will adopt this idea in the following. Petralia et al. (2012) propose to jointly model the entire vector $\boldsymbol{\theta}$ with a prior that is chosen to assign low density to sets of θ_k's located close together. Xie and Xu (2017) extend this approach to the nonparametric case. The deviation from independence is specified a priori by a pair of repulsion parameters. Formally, a repulsive density is defined as follows:

Definition 2. *A density $\pi(\boldsymbol{\theta}) = \pi(\theta_1, \dots, \theta_K)$ is repulsive if for any $\delta > 0$ there is a corresponding $\epsilon > 0$ such that $\pi(\boldsymbol{\theta}) < \delta$ for all $\boldsymbol{\theta} \in \Theta \backslash G_\epsilon$, where*

$$G_\epsilon = \{\boldsymbol{\theta} : \Delta(\theta_s, \theta_j) > \epsilon; s = 1, \dots, K; j < s\}$$

and Δ is a distance.

Once we define a suitable distance $\Delta(\cdot, \cdot)$, which measures the closeness between two mixture component parameters θ_s and θ_j, a convenient repulsive prior that smoothly pushes components apart can be obtained as

$$\begin{aligned} \pi(\boldsymbol{\theta}) &= c_1 \left(\prod_{k=1}^{K} g_0(\theta_k) \right) h(\boldsymbol{\theta}), \qquad h(\boldsymbol{\theta}) = \min_{\{(s,j) \in A\}} g(\Delta(\theta_s, \theta_j)) \\ g\{\Delta(\theta_s, \theta_j)\} &= \exp\left[-\tau (\Delta(\theta_s, \theta_j))^{-\nu} \right], \text{ with } \tau, \nu > 0, \end{aligned} \tag{1.5}$$

where $A = \{(s, j) : s = 1, \dots, K; j < s\}$. The repulsiveness is induced via the function $h(\cdot)$. The multiplicative structure makes it easy to apply a slice sampler (Damien et al., 1999; Walker, 2007) to obtain samples from the prior and the posterior.

1.2.2 Bayesian Nonparametrics (BNP)

Based on (1.3), one can claim that – from a broader point of view – a parametric model assumption corresponds to a strong prior opinion. Indeed, parametric modeling is equivalent to insist on a prior that assigns probability one to a specific subset of the set of all densities. Parametric models make restrictive assumptions about the data-generating mechanism, which may cause serious bias in inference (Ghosal and Vaart, 2017).
The need to avoid such misspecifications and the increasing availability of high-dimensional data, together with advancements in computational resources, have spurred recent interest in the characterization of priors over infinite-dimensional spaces. BNP models can be thought of as "infinite-dimensional priors" (BNP priors), i.e. distributions that cannot be described by a finite number of parameters. Indeed, the use of the "nonparametric" terminology is somehow misleading: to be precise, we should speak of "infinite-parametric" Bayesian statistics. Thus, the objects of interest for the nonparametric Bayesian are the estimates (and the relative uncertainty quantification) of functions of probability densities/distributions. Such a prior specification consequently allows inferring the quantities of interest taking into account the uncertainty over the distribution of the prior. This philosophically reflects the lack of prior information.
The entire concept of nonparametric prior could be easier to understand if we notice that a BNP prior is essentially a stochastic process, reminding that a stochastic process can be seen as a probability distribution over its paths. We refer to Hjort et al. (2010), Müller et al. (2015), Mueller and Rodríguez (2013), and Ghosal and Vaart (2017) for the foundational concepts and extended reviews of the basic BNP methods, along with the discussion of main research directions and open questions. Here, we discuss a few most common and well-know BNP priors, i.e. the Dirichlet and Pitman-Yor processes, which constitute the basic building blocks of BNP density estimation and which we will use in later chapters.

Dirichlet Process

The Dirichlet Process (DP) has been broadly and deeply studied over the last few decades. The DP is the most basic and popular model for random probability measures, and a large number of refinements and extensions have been proposed in the literature. Multiple definitions and characterizations of the DP can be provided, but here we will discuss only the most relevant for

our discussion.

Consider two probability spaces: $(\mathbb{X}, \mathcal{B}, P)$ and $(\mathbf{P}, \mathcal{C}, Q)$. The first one is called *base space* where usually $\mathbb{X} \subseteq \mathbb{R}^d$ and $\mathcal{B}$ is the corresponding Borel σ- algebra. The second space is referred to as *distributional space*, where $\mathbf{P}$ is the set of probability measures over $(\mathbb{X}, \mathcal{B})$. The DP arises as a measure on the latter space where P can be considered as one of its realizations. In other words, the DP is actually a *distribution over distributions.*
Formally, the DP was originally defined in the seminal paper by Ferguson (Ferguson, 1973), where the author construct the process via finite dimensional distributions satisfying the Kolmogorov's consistency conditions:

Definition 3 (Dirichlet Process via finite dimensional distributions)**.** *Given a measurable set S, a base probability distribution H and a positive real number α, the Dirichlet process $DP(\alpha, H)$ is a stochastic process whose sample path is a probability distribution over S, such that, for any measurable finite partition of S, denoted $B_1, B_2, \ldots, B_n$, then*

$$(X(B_1), \ldots, X(B_n)) \sim Dir(\alpha H(B_1), \ldots, \alpha H(B_n)).$$

We indicate a Dirichlet Process with $DP(\alpha, H)$, to highlight that the DP is parameterized by a *precision parameter* $\alpha \in \mathbb{R}^+$, and by a probability measure H, called a *base measure*, around which the DP is centered. It can easily be shown that, given $A \subseteq \mathbb{X}$ measurable set member of a partition of $\mathbb{X}$, if $G \sim DP(\alpha, H)$, then $\mathbb{E}[G(A)] = H(A)$ and $\mathrm{Var}[G(A)] = \frac{G(A)(1-G(A))}{\alpha+1}$.

A second characterization of the DP comes from its predictive rule, also called the Blackwell-MacQueen formula (Blackwell and MacQueen, 1973).

Definition 4 (Dirichlet Process via predictive rule)**.** *Let $X_i|G \sim G$. The following two are equivalent:*

$$G \sim DP(\alpha, H) \iff X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha H + \sum_{i=1}^{n} \delta_{X_i}}{\alpha + n} \tag{1.6}$$

The discrete-time stochastic process induced by this formula is often referred to as the *Chinese Restaurant Process* (expressed in a Pólya Urn Scheme), a distribution over partitions that embodies the assumed prior distribution over the cluster structures (Pitman, 2002). The form of the predictive rule in (1.6) shows that the presence of ties is allowed among the elements of the random vector $\boldsymbol{X}_{n+1} = (X_1, \ldots, X_{n+1})$. As a consequence, a **partition** of the components of $\boldsymbol{X}_{n+1}$ is induced.
This formulation is the basis of numerous models in machine learning (Blei and Frazier, 2011; Teh et al., 2006). The formula from Definition 4 describes what is typically called the *rich gets richer* mechanism in the machine learning literature, and can be explained with a clever metaphor. Imagine a Chinese restaurant with infinite tables. The first customer enters and sits at the first table. The second customer can either sit at the first table or at a new one. Any further customer can sit either at any of the previously occupied tables, or at a new one. However, the probability of sitting at an already occupied table depends positively on the number of customers already sat at that table. Hence, larger tables tend to bring more customers.

The clustering behavior of the DP (1.6) can also be investigated studying the so-called *exchangeable partition probability function* (EPPF), whose general definition is due to Pitman (Pitman, 1995). More formally, let us denote with $K_n = k$ the number of distinct values in $\boldsymbol{X}_n$ and with $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$ their relative frequencies. The EPPF is the probability

of observing a specific sample $(X_1, \ldots, X_n)$ characterized by k distinct values and frequencies $(n_1, \ldots, n_k)$. It is defined as

$$\begin{aligned} \Pi_k^{(n)}(n_1, \cdots, n_k) &:= \mathbb{P}\left((K_n = k) \cap (N_{j,n} = n_j, j = 1, \ldots, k)\right) \\ &:= \int_{\mathbb{X}^k} \mathbb{E}\left[G^{n_1}(\mathrm{d}x_1) \cdots G^{n_k}(\mathrm{d}x_k)\right]. \end{aligned} \tag{1.7}$$

Notice that this probability depends only on the cluster frequencies and not on the actual values of the partition: this motivates the term *exchangeable*. For the DP, the previous formula simplifies down to

$$\Pi_k^{(n)}(n_1, \cdots, n_k) = \frac{c^k}{(c)_n} \prod_{j=1}^{k} (n_j - 1)!\ ,$$

where $(c)_n = \Gamma(c+n)/\Gamma(c)$ is the Pochhammer symbol (or rising factorial). The EPPF is an extremely important object, and will be discussed more later, after introducing the concept of partial exchangeability. Next, let us go back to other DP characterizations.

The last DP characterization of interest here is Sethuraman's stick-breaking representation (Sethuraman, 1994). It is a constructive definition, which is often the most intuitive and useful from a practical point of view. One can build a $DP(\alpha, H)$ according to the following hierarchical model:

Definition 5 (Dirichlet Process via stick-breaking construction)**.** *Consider the following stick-breaking prior on the space* $\mathbf{P}$*:*

$$\begin{aligned} G(\cdot) &= \sum_{i=1}^{K} \omega_k \delta_{x_k}(\cdot) & \omega_k &= u_k \prod_{l=1}^{i-1} (1 - u_l) \\ x_k &\sim H & u_k &\sim \begin{cases} Beta(1, \alpha), & \text{if} \quad k < K, \\ \delta_1, & \text{if} \quad k = K \end{cases} \end{aligned}$$

A Dirichlet Process (DP), characterized by base measure H *and* precision parameter α*, denoted as* $DP(\alpha, H)$ *can be defined by adopting the former construction and letting* $K \uparrow +\infty$*.*

This characterization helps us to understand some crucial properties of the DP. First of all, it is evident that realizations G of a DP consist in a discrete probability distribution with probability 1, where the support is composed by points called atoms, i.e. $G = \sum_{i=1}^{+\infty} \omega_k \delta_{x_k}(\cdot)$ almost surely. The discrete nature of G remarks that ties are allowed among a random sample taken from G, naturally inducing a partition as a byproduct. The distribution of the weights, on the other hand, is influenced only by the concentration parameter α. We will summarize the stick-breaking representation with $\boldsymbol{\omega} \sim SB(\alpha)$ or with $\boldsymbol{\omega} \sim GEM(\alpha)$ (named after Griffiths, Engen, and McCloskey's early work in this field).

The stick-breaking construction can be extended to show that the DP is part of the broader set of random probability measures known as the species sampling models. The definition follows:

Definition 6 (Species Sampling model)**.** *Let* $\{\omega_j\}_{j \geq 1}$ *be a sequence of non-negative random weights with* $\sum_{j \geq 1} \omega_j \leq 1$*, and assume that* $\{\theta_j\}_{j \geq 1}$ *is a sequence of i.i.d. variables independent of the* ω_j*'s and distributed according to a non-atomic probability measure* P_0*. Then the random probability measure*

$$P^* = \sum_{j \geq 1} \omega_j \delta_{\theta_j} + \left(1 - \sum_{j \geq 1} \omega_j\right) P_0$$

is called a species sampling model (SSM). A sequence of random variables $\{X_n\}_{n\geq 1}$ *such that* $X_n|P^* \overset{i.i.d.}{\sim} P^*$ *is termed a species sampling sequence.*

In particular, the DP is a *proper* SSM, since by construction $\sum_{j\geq 1}\omega_j = 1$.

Another property of the DP is its conjugacy. Consider the following model, where the DP is used to model the data directly:

$$\begin{aligned}(X_1,\dots,X_n)\,|G &\sim G\\ G &\sim Q\\ Q &= DP(\alpha,H).\end{aligned} \tag{1.8}$$

Then, the posterior has the following distribution:

$$G|X_1,\dots,X_n \sim \text{DP}\left(\alpha+n, \frac{\alpha}{\alpha+n}G_0 + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{X_i}\right) \tag{1.9}$$

The resulting posterior distribution has a concentration parameter updated by the number of observations in the sample. The new base measure is a convex combination between the prior base measure and the empirical c.d.f. given by the sample.
It is important to note that the DP has many other properties (e.g. self-similarity). We refer the interested reader to Ghosal and Vaart (2017), Müller et al. (2015), and Hjort et al. (2010) for more details.

It can be proven that the expected number of (random) clusters $\mathbb{E}[K_n]$ in which the data in (1.8) are partitioned grows proportionally to $\alpha\log n$. Such logarithmic growth rate might be undesirable in some applications. Furthermore, the random sizes of these K_n clusters show an exponential tail behavior that might be undesirable as well. For example, in a context like language processing, image segmentation, etc., the cluster size distributions exhibit a power-law tail decay (Sudderth and Jordan, 2009). There is a simple extension of the DP that offers a more flexible clustering behavior, the *2 parameters Poisson-Dirichlet* (2PPD) process, or Pitman-Yor (PY) process. We will briefly discuss this DP extension before focusing on Dirichlet Process Mixture models (DPMM).

Pitman-Yor Process

Despite being previously studied outside of statistics (Pitman, 1995; Pitman and Yor, 1997), the PY process began to gain attention in the applied statistical community after Ishwaran and James (2001) introduced flexible Gibbs sampling methods for posterior sampling in models with a PY prior. We will let $PY(\theta,\sigma,H)$ denote a Pitman-Yor process characterized by base measure H, *mass* parameter θ and *discount* parameter σ, where $\theta > -\sigma$ and $\sigma\in[0,1)$. Here, we will report just two characterizations of the PY process that are of use in this dissertation: the Generalized Pólya Urn characterization and the stick-breaking characterization. The first follows from the PY predictive rule.

Definition 7 (Pitman-Yor via predictive rule)**.** *Let* $X_i|G \sim G$. *The following two statements are equivalent:*

$$G \sim PY(\theta, \sigma, H) \iff X_{n+1}|X_1, \dots, X_n \sim \frac{\theta + \sigma K_n}{\theta + (n+1) - 1} H(\cdot) + \sum_{j=1}^{K_n} \frac{n_j^* - \sigma}{\theta + (n+1) - 1} \delta_{X_i^*}(\cdot), \tag{1.10}$$

where $\left(X_1^*, \dots, X_{K_n}^*\right)$ *denotes the* K_n *unique values in* $(X_1, \dots, X_n)$*, each occurring with frequency* n_j^* *for* $j = 1, \dots, K_n$*.*

From (1.10) one can see how the discount parameter affects the clustering behaviour directly. In particular, it can be shown that $\mathbb{E}[K_n] \asymp \frac{\Gamma(\theta+1)}{\sigma\Gamma(\sigma+\theta)} n^\sigma$. Essentially, the number of clusters under a PY process prior grows much more rapidly than the $\log n$ rate offered by a DP. Moreover, the clusters' size distribution also shows a power law under the PY process.
The stick-breaking characterization of the PY process is very similar to (5):

Definition 8. *Let* $X_i|G \sim G$*. The following conditions are equivalent:* $G(\cdot) \sim PY(\theta, \sigma, H) \iff$

$$G(\cdot) = \sum_{i=1}^{\infty} \omega_k \delta_{x_k}(\cdot), \qquad \omega_k = u_k \prod_{l=1}^{i-1}(1 - u_l), \qquad u_k \sim Beta(1 - \sigma, \theta + \sigma k), \qquad x_k \sim H. \tag{1.11}$$

Thus, the PY process is a straightforward and still tractable extension of the DP.

Dirichlet Process Mixture Model

Given its discrete nature, the DP is rarely used to model data directly. Instead, it is commonly employed as **mixing measure** in mixture models with parametric kernel. This clever idea led to the development of the Dirichlet Process Mixture Model (DPMM) (Ferguson, 1983; Lo, 1984; Antoniak, 1974; Escobar and West, 1995)

Definition 9 (DPMM)**.** *Consider the exchangeable vector of data* $(X_1, \dots, X_n)$*. Suppose that the distribution of the data can be modeled via a parametric density which depends on some parameter* θ*. A Dirichlet Process Mixture Model is defined as follows:*

$$X_i|\theta_i \overset{ind.}{\sim} F(\theta_i), \qquad \theta_i|G \overset{i.i.d.}{\sim} G, \qquad G = \sum_{k=1}^{+\infty} \pi_k \delta_{\theta_k} \sim DP(\alpha, H).$$

As we can see, the DP is placed at the level of the parameters θ_i that characterize what will be the kernel function of our mixture, $F(\theta_i)$. If we marginalize out θ_i, we get

$$X_i|G \overset{ind.}{\sim} \int F(\theta) G(d\theta), \qquad G = \sum_{k=1}^{+\infty} \pi_k \delta_{\theta_k} \sim DP(\alpha, H). \tag{1.12}$$

Given the peculiar form of G, using the stick-breaking representation, (1.12) can be translated in

$$X_i|G \overset{i.i.d.}{\sim} \sum_{k=0}^{\infty} \pi_k F(\theta_k) \qquad \boldsymbol{\pi} \sim SB(\alpha), \qquad \theta_k \sim H, \tag{1.13}$$

where we appreciate the fact that such formulation reveals the essence of a DPMM: an infinite mixture model. Relying on an infinite mixture model grants important advantages in terms of robustness and flexibility. Given a sufficient number of components, a mixture of simpler densities can approximate every other distribution, under certain assumptions. Results that discuss this idea can be found in Roberts (2017) and Bochkina and Rousseau (2017) and in the references therein. An important advantage of infinite over finite mixture models is that we are not required to specify a priori the number of mixture components K. In the finite case, either

one fixes multiple values for K beforehand and then compares the goodness of fit of the different models with suitable indexes (e.g. DIC, BIC, AICm, BICm, WAIC (Spiegelhalter et al., 2002; Raftery et al., 2007; Hjort et al., 2010; Watanabe, 2013)) or one places a prior directly on K. This, in turn, implies the usage of some non-trivial sampling scheme to estimate the model, such as Reversible-Jump MCMC (Green, 1995). The former solution is easy to use, but it fails to completely take into account the uncertainty regarding the unknown number of components. The latter instead is coherent from a full-Bayesian modeling framework, but can experience issues from the practical point of view due to its complexity (Bhattacharya, 2008), with chains that exhibit poor mixing.

From (1.13) we can see how the model is still able to produce, as a byproduct, a clustering among the data. This time the partition is induced by the ties that occur among the parameters. Thus, the observations that are estimated to come from the same parametric kernel are grouped together. This is possible since, trivially, $\mathbb{P}(\theta_i = \theta_{i'}|G) > 0$ for two indexes $i \neq i'$ both in $\{1, \ldots, n\}$, due to the discreteness of G.
In recent years, the DP and the DPMM have been widely employed in many applications, in the Bayesian nonparametric literature. Moreover, the conceptual simplicity of this model makes it a very convenient tool that can be combined in more involved structures to address complex data analysis tasks. For heterogeneous and interesting applications see, for example, Canale and Prünster (2017), Lennox et al. (2010), Guindani et al. (2014), Banerjee et al. (2013), Bandyopadhyay and Canale (2016), Hong and Martin (2016), and Shahbaba and Johnson (2013). As an example of how the DPMM can be easily embedded in more complex models, we now discuss the Rounded Mixture of Gaussian kernel model (Canale and Dunson, 2011), which we will apply in Chapter 2.

A nontrivial example of DPMM application: the RGM-approach to count data modeling. Consider a discrete random variable X taking values over a discrete set such as $\mathbb{N}$, and its probability mass function $p_X(x) = \mathbb{P}(X = x)$. Canale and Dunson (2011) propose to define a nonparametric prior distribution Π for the probability mass function $p_X(x)$. Instead of specifying Π directly, they induce the prior distribution on $p_X(x)$ defining a prior Π^* on the space of the possible densities f of a continuous latent random variable X^* and then rounding X^* to obtain X.
Formally, they first choose a sequence of fixed threshold $a_0 < a_1 < a_2 < \ldots < a_\infty$. For example, if the support of X is the entire set $\mathbb{N}$, one can simply choose the set $\{a_j\}_{j=0}^{+\infty}$ as $\{-\infty, 0, 1, 2, \ldots, +\infty\}$. The probability mass function of X is defined via the rounding function $g : \tilde{C} \to \tilde{D}$ where $\tilde{C}$ denotes the space of the densities w.r.t. the Lebesgue measure of the continuous random variable X^* and $\tilde{D}$ represents the space of probability mass functions of a discrete random variable of interest X. We can define

$$p_X[j] = g(f)[j] = \int_{a_j}^{a_{j+1}} f(x^*)\, dx^* \quad j \in \mathbb{N} \tag{1.14}$$

where $a_0 = \min\{x^* : x^* \in \mathcal{X} \subseteq \mathbb{R}\}$ and $a_{+\infty} = \max\{x^* : x^* \in \mathcal{X} \subseteq \mathbb{R}\}$, so that $\int_{a_0}^{a_{+\infty}} f(x^*)\, dx^* = 1$. In order to ensure flexibility, the density f is modeled as a nonparametric mixture of kernels. In most applications, the Normal distribution represents a convenient but still useful choice for the kernel function. Finally, the mixing measure denoted as P, is chosen to be a $DP(\alpha, H)$, where H corresponds to a Normal-Inverse Gamma distribution to exploit conjugacy.
To summarize, the model can be written as:

$$X_i \overset{i.i.d.}{\sim} p_X \qquad\qquad p_X = g(f)$$

$$f = \int N\left(x^*; \mu, \sigma^2\right) dP\left(\mu, \sigma^2\right) \qquad P \sim \tilde{\Pi} = DP\left(\alpha, H\right)$$

We will concisely indicate the previous model as $X_i \sim RGM(\tilde{\Pi})$, where $\tilde{\Pi} = DP\left(\alpha, H\right)$ and $H \sim NIG(\mu_0, \sigma_0^2, a_0, b_0)$.
In the paper the authors show how this model, which suits the flexibility of a DPMM to the discrete case via a simple deterministic rounding function, can outperform established competitors among the models for discrete data, such as nonparametric mixture of Poisson kernels or Negative Binomial kernels. Simulation studies also confirm that this model is useful to address zero-inflated cases.

Posterior Computations

The flexibility of an infinite dimensional prior such as the DP comes at a price. In fact, it is not obvious how to deal with an infinite dimensional object when it comes to real data applications. Many MCMC sampling schemes have been proposed to obtain posterior inference from a DPMM and its variants. We briefly discuss the main ones. The details of the various algorithms can be found in the following chapters, where they are employed. The different methods can be divided in two main categories: **marginal** samplers and **conditional** samplers.

Marginal samplers focus on models where the infinite dimensional random probability measure is integrated out. In other words, they essentially exploit the Pólya-urn scheme provided by the predictive rules of the processes (4 and 1.10). Once combined with the exchangeability assumption on the data, they provide the full conditional distributions for the update of the parameters of interest. This class of algorithms was introduced by Escobar (1988) and Escobar and West (1995), where the authors study the case of Mixture of Gaussians kernels. Many extensions and refinements have been proposed in the literature (Müller et al., 2015; MacEachern, 1994; Maceachern and Müller, 1998; Müller et al., 1996; Favaro and Teh, 2013). Notably, Neal (2000b) summarizes many strategies to deal with both conjugate and non-conjugate models. Marginal methods devise MCMC schemes that explore the space of partitions of the data directly. Oftentimes the chains can get stuck in some local modes, and the mixing can be negatively affected. To obviate this issue, some authors propose steps that may increase the efficiency of the sampler. For example, Jain and Neal (2004), Jain and Neal (2005), Dahl (2005), and Dahl (2003) propose variants of the Split and Merge move. First, two observations are randomly sampled and the current clusters they belong to are considered. Then, if the observations are allocated in the same group, a splitting move is proposed. Alternatively, if the two clusters are different, they are merged together. The new MCMC partition is then accepted or rejected according to the result of a Metropolis-Hastings Step.

Conditional methods, on the other hand, rely on summaries of stochastic realizations of G, the infinite dimensional prior. In this case, most of the samplers are devised starting from the stick-breaking representation of G. Since dealing with an infinite dimensional object directly is infeasible, many authors propose to rely on tractable truncation of the stick-breaking representation. In this spirit, Muliere and Tardella (1998) introduce the ϵ-Dirichlet distribution, a stochastic truncation (i.e. following a random stopping rule) of the DP useful for approximating classes of DP functionals. This truncation ensures that both the Total Variation and the Prohorov metrics between the original DP and its approximation are controlled by an arbitrary small $\epsilon > 0$. Arbel et al. (2019) recently develop a similar construction for the PY process. Ishwaran and James (2001) and Ishwaran and Zarepour (2002) propose to adopt a deterministic truncation of the stick-breaking representation of a large class of random probability models. They carefully studied the truncation error, providing upper bounds and rates of decay. Algorithms that provide an exact truncation, exploiting a stochastic number of components at each

iteration, are also available. Walker (2007) extend the Slice sampler of Damien et al. (1999) and Neal (2003) to Bayesian nonparametric mixtures, while soon after Papaspiliopoulos and Roberts (2008) introduce a similar idea, called Retrospective sampler. In a note (Papaspiliopoulos, 2008) Papaspiliopoulos suggested a way to exploit both samplers' strengths to devise a new, more efficient, algorithm. Some years later, Kalli et al. (2011) improve Walker's slice sampler by developing an Independent Slice-Efficient Sampler that is easy to implement and computationally tractable.

Both marginal and classical methods have their pros and cons. Marginal methods explore the partition space directly, and the number of parameters to be updated at each iteration is deterministic and limited. However, performing posterior inference is not a trivial task. Conditional methods are slightly more difficult to implement, but their MCMC output is considerably richer, allowing for inference of many quantities of interest. Recently, Canale et al. (2019) propose the Importance Conditional Sampler (ICS), a scheme that combines appealing features of both conditional and marginal methods. Like marginal methods, ICS has a simple and interpretable sampling scheme, reminiscent of the Blackwell-McQueen Pòlya urn, while, similarly to conditional methods the algorithm allows for parallelizable parameters update and accounts for straightforward posterior quantification.

1.2.3 Models for Partially Exchangeable Data

As was mentioned in Section 1, many recent datasets present complex structures for which the assumption of "full" exchangeability of the observations is unrealistic and too limiting. Specifically, some datasets present a nested structure where various **individuals** are organized into different **groups** or **units**. Possible examples include patients organized into different hospitals, or microbes recorded from different subjects, or students belonging to different classes. In the BNP literature, many authors put considerable effort into building extensions of the DP to model data that do not satisfy a simple exchangeability hypothesis. The dependent Dirichlet Process (DDP) introduced in MacEachern (2000) and MacEachern (1999) may be considered as the starting point for many developments in this direction. Loosely speaking, we desire to accommodate the dependence among the same-group individuals while at the same time allowing for the borrowing of information across different sub-populations.
A large number of authors contributed to this field: see for example De Iorio et al. (2004), Dunson and Park (2008), Griffin and Steel (2006), Griffin et al. (2013), Lijoi et al. (2014), and Duan et al. (2007). Formally, consider the grouped observations X_{ij}, where $j = 1, \dots, J$ denotes the unit and $i = 1, \dots, n_j$ denotes the subject within unit j. Each term in the sequence $\{n_j\}_{j=1}^{J}$ represents the cardinality of the corresponding unit and $\sum_{j=1}^{J} n_j = N$ is the total number of observations. Furthermore, let $\boldsymbol{X}_j = \left(X_{1j}, \dots, X_{n_j j}\right)$ represent the vector of random variables belonging to unit j. It is reasonable to assume that the random variables within a group are exchangeable and that they are characterized by a distribution $X_{ij} \sim G_j$ for $j = 1, \dots, J$. At the same time, observations across different units are assumed to be conditionally independent. We refer to this setting as *partial exchangeability*. A nice review of mathematical properties of partial exchangeability in the context of hierarchical and nested random probability measures can be found in Camerlenghi (2017).
As in the fully exchangeable case, there exists a representation theorem for partially exchangeable data (Finetti, 1938), that we now recall.
Consider J sequences $\left\{(X_{ij})_{i\geq 1} : j = 1, \dots, J\right\}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $(\mathbb{X}, \mathcal{X})$, for $j = 1, \dots, J$. By de Finetti's representation theorem the collection of

sequences $\left\{(X_{ij})_{i\geq 1} : j = 1, \ldots, J\right\}$ is partially exchangeable if and only if

$$\mathbb{P}\left[\bigcap_{j=1}^{J}\left\{\boldsymbol{X}^{(n_j)} \in A_j\right\}\right] = \int_{\boldsymbol{P}_{\mathbb{X}}^{J}} \prod_{j=1}^{J} p_j^{(n_j)}(A_j)\, Q_J(\mathrm{d}p_1, \ldots, \mathrm{d}p_J)$$

for any integer $n_j \geq 1$ and $A_j \in \mathcal{X}^{n_j}$, where $\boldsymbol{X}^{(n_j)} = \left(X_{1,j}, \ldots, X_{n_j,j}\right)$ and $p^{(J)} = p \times \cdots \times p$ is the J-fold product measure on $\mathbb{X}^J$, for any $J \geq 1$. The de Finetti mixing measure Q_J is a probability measure on $\left(\boldsymbol{P}_{\mathbb{X}}^{J}, \mathcal{P}_{\mathbb{X}}^{J}\right)$. In order to adopt a hierarchical representation, let $(G_1, \ldots, G_J)$ denote a vector of random probability measures defined on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $\left(\boldsymbol{P}_{\mathbb{X}}^{J}, \mathcal{P}_{\mathbb{X}}^{J}\right)$ with distribution Q_J. Then, the notion of partial exchangeability can be re-expressed as

$$\begin{aligned}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J) \,|\, (G_1, \ldots, G_J) &\overset{i.i.d.}{\sim} G_1 \times \cdots \times G_J \\ (G_1, \ldots, G_J) &\overset{i.i.d.}{\sim} Q_J\end{aligned} \tag{1.15}$$

All the models for partially exchangeable data we will present and use in the following are merely different specifications of (1.15). Simply posed, partial exchangeability means modeling data that are heterogeneous because they are generated by different but related distributions or experiments.

First, a class of models that naturally considers partially exchangeable data is based on the superposition of random measures. For example, Müller et al. (2004) – and, more recently, Griffin and Kolossiatis (2010) and Kolossiatis et al. (2013)– define probability mass distributions characterized by a shared part across different units, in order to create dependent measures. In their proposal, each of the random probability measures G_j is defined as a combination of a common F_0 and a study-specific F_j:

$$\begin{aligned}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J) \,|\, (G_1, \ldots, G_J) &\sim G_1 \times \cdots \times G_J \\ G_j | \epsilon, F_j, F_0 &= \epsilon F_0 + (1-\epsilon) F_j, \\ F_0 \sim DP(\alpha, H), &\qquad F_j \sim \mathrm{DP}(\alpha, H), \\ \pi(\epsilon) &= \pi_0 \delta_0 + \pi_1 \delta_1 + (1 - \pi_0 - \pi_1)\,\mathrm{Beta}(a, b).\end{aligned} \tag{1.16}$$

The prior on ϵ includes point masses on 0 and 1, allowing for the two extreme cases of common and conditionally independent G_j across groups.

The idea of modeling dependent distributions has received a lot of attention in the machine-learning community as well, especially for the development of language processing methods. A fundamental model for clustering grouped data is the Hierarchical Dirichlet Process (HDP, Teh et al., 2006). A shared grouping structure is induced among sub-populations using a discrete base measure G_0 in a Dirichlet Process $DP(\alpha, G_0)$. However, choosing the base measure G_0 to be a member of a discrete parametric family would be too restrictive. In order to force G_0 to be discrete and yet have broad support, they refer to a nonparametric hierarchical model in which G_0 is itself a draw from another Dirichlet process $DP(\gamma, H)$. The HDP can be expressed as follows:

$$\begin{aligned}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J) \,|\, (G_1, \ldots, G_J) &\sim G_1 \times \cdots \times G_J \\ G_j | G_0 \sim DP(\alpha, G_0), &\qquad G_0 \sim DP(\gamma, H).\end{aligned} \tag{1.17}$$

Alternatively, we can write $G_j \sim Q$ and $Q \sim HDP(\gamma, \alpha, G_0)$. The HDP allows the different distributions to share the same set of atoms, sampled from G_0, while letting every single set of masses (sticks) be unconstrained. This implies that $\mathbb{P}(G_i = G_j) = 0$ for $i \neq j$. However, ties

between the parameters shared among sub-populations are induced, leading to a clustering of individuals that takes into account the grouped structure of the data and shares information across the units.

In some applications, it might also be interesting to detect sub-populations whose distributional characteristics are similar. For that purpose, the nested Dirichlet Process (nDP) (Rodríguez et al., 2008) allows us to describe two clustering processes: one at the level of the observed measurements (observational clustering), the other on the level of the different units (distributional clustering). We are going to discuss some details of this model, as it is related to one of our applications.
When the nDP is directly applied to the data, we have

$$\begin{aligned}(\boldsymbol{X}_1,\dots,\boldsymbol{X}_J)\,|(G_1,\dots,G_J) &\sim G_1\times\cdots\times G_J\\ G_1,\dots,G_J|Q &\overset{i.i.d.}{\sim} Q\\ Q &= \sum_{k\geq 1}\pi_k\delta_{G_k^*}\\ G_k^* &= \sum_{l\geq 1}\omega_{lk}\delta_{\theta_{lk}}\quad \forall k\geq 1,\end{aligned} \tag{1.18}$$

where the θ_{lk}'s are independently distributed as the base measure H, and where $\boldsymbol{\pi}$ and $\boldsymbol{\omega}_k$ are distributed accordingly to $SB(\alpha)$ and $SB(\beta)$, respectively. As in the DP case, rather than using the nDP to model directly the data, one could employ a (nested) mixture representation:

Definition 10 (nested Dirichlet Processes Mixture Model, nDPMM)**.** *Consider $\boldsymbol{X}_j\sim F_j\ \forall j$. A collection of distribution $\{F_1,\dots,F_J\}$ is said to follow a nDPMM if*

$$\begin{aligned}F_j(\cdot|\phi) &= \int_\Theta p(\cdot|\theta,\phi)\,G_j(d\theta) & G_j(\cdot)\sim Q &\equiv \sum_{k=1}^{\infty}\pi_k\delta_{G_k^*(\cdot)}\\ G_k^*(\cdot) &= \sum_{l=1}^{\infty}\omega_{lk}\delta_{\theta_{lk}}(\cdot) & \theta_{lk}&\sim H,\end{aligned}$$

where $\omega_{lk}=u_{lk}\prod_{s=1}^{l-1}(1-u_{sk})$, $u_{lk}\sim Beta(1,\beta)$ and $\pi_k=v_k\prod_{s=1}^{k-1}(1-v_s)$, $v_k\sim Beta(1,\alpha)$. We can also write $G_j\overset{i.i.d.}{\sim}Q$ and $Q\sim DP(\alpha,DP(\beta,H))$.

Notice how different and independent stick-breaking prior formulations are considered for the weights $\boldsymbol{\pi}$ and $\boldsymbol{\omega}_k\ \forall k$. For any $k\geq 1$, each G_k^* is defined as a DP, while the distributions G_j are **sampled** from a discrete distribution where the atoms constituting the support are the aforementioned distributions G_k^*. This allows $\mathbb{P}(G_i=G_j)>0$ for $i\neq j$, i.e. a partition among the distributions is possible.
The nDP construction implies that either the distributions of two sub-populations share the same atoms and the same probability weights simultaneously, or they do not share either of the two. We further discuss the properties of the two-layered clustering structure of the nDP in the next section.

Exchangeable Partition Probability Functions and the limitations of the nDP construction

We have seen how, once a nonparametric prior with almost surely discrete realizations is used, a partition in the data is induced and it is possible to derive the corresponding EPPF. A similar result holds in the partially exchangeable setting, and we refer to it as a *partially exchangeable partition probability function*(pEPPF). Let us limit ourselves to the nDP case and consider

(1.18). In a recent work, Camerlenghi et al. (2019b) extensively study the properties of the two-layered partition induced by the nDP. The authors are able to derive the pEPPF for the nDP with J units. They also provide an example for $J = 2$ that we now report, adapting the notation.
Consider two vectors of grouped data, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and define:

- $K_{n_1} = k_1$ as the number of distinct values specific of $\boldsymbol{X}_1$ not shared with $\boldsymbol{X}_2$. Let $\boldsymbol{n_1}$ be the vector of associated frequencies, with generic element $n_{l,1}$, $l = 1, \dots, k_1$.
- $K_{n_2} = k_2$ as the number of distinct values specific of $\boldsymbol{X}_2$ not shared with $\boldsymbol{X}_1$. Let $\boldsymbol{n_2}$ be the vector of associated frequencies, with generic element $n_{l,2}$, $l = 1, \dots, k_2$.
- $K_0 = k_0$ as the number of distinct values shared among $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. In particular, let $\boldsymbol{q}_j$ be the vector of frequencies for sample $\boldsymbol{X}_j$, $j = 1, 2$.

By denoting $\pi_1 := \mathbb{P}(G_1 = G_2)$, the pEPPF for the nDP can be rewritten as:

$$\begin{aligned}\Pi_k^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1, \boldsymbol{q}_2) =& \pi_1 \Phi_k^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1 + \boldsymbol{q}_2) + \\ & (1-\pi_1)\, \Phi_{k_0+k_1}^{(n_1+|q_1|)}(\boldsymbol{n}_1, \boldsymbol{q}_1)\, \Phi_{k_0+k_2}^{(n_2+|q_2|)}(\boldsymbol{n}_2, \boldsymbol{q}_2)\, \mathbb{I}_{\{0\}}(k_0)\end{aligned} \tag{1.19}$$

where $\Phi_k^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1 + \boldsymbol{q}_2)$ is the partition probability function relative to the fully exchangeable case, while $\Phi_{k_0+k_j}^{(n_j+|q_j|)}(\boldsymbol{n}_j, \boldsymbol{q}_j)$ with $j = 1, 2$ are the two marginal EPPFs for each sub-population $\boldsymbol{X}_j$. This means that the pEPPF for the nested case when $J = 2$ can be rewritten as a convex combination between independence and exchangeability. Interested readers can find more details in Camerlenghi (2017).
The derivation of this result highlights a subtle but crucial drawback of the nDP: the second part of eq. (1.19) vanishes as soon as $k_0 \geq 1$. This means that if any tie is present between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, the model a posteriori collapses to the full-exchangeable case. The same problem is present when $J > 2$ and unfortunately it also holds for the nDPMM, when the nonparametric nested prior is placed over latent variables. In this latter case, it is not clear how the model behaves a posteriori.
Let us consider the following simple example: suppose that the data are organized into two groups of equal length n, namely $\boldsymbol{X}_1 = (X_{1,1}, \dots, X_{n,1})$ and $\boldsymbol{X}_2 = (X_{1,2}, \dots, X_{n,2})$, generated from two simple mixtures:

$$X_{i,1} \sim \frac{1}{2} N\left(\mu_0, \sigma_0^2\right) + \frac{1}{2} N\left(\mu_1, \sigma_1^2\right) \qquad \text{and} \qquad X_{i,2} \sim \frac{1}{2} N\left(\mu_0, \sigma_0^2\right) + \frac{1}{2} N\left(\mu_2, \sigma_2^2\right),$$

with $i = 1, \dots, n$, μ_0, μ_1, μ_2 real values and $\sigma_0^2, \sigma_1^2, \sigma_2^2$ in $\mathbb{R}^+$. The two mixtures share one atom, namely (μ_0, σ_0^2). When the nDP is applied for the nested density estimation task, two possible scenarios can occur:

- The nDP identifies two different sub-populations correctly, but it will provide at the same time two different estimations for the atom (μ_0, σ_0^2).
- The nDP estimates the common atom (μ_0, σ_0^2)correctly, but collapses the two sub-populations into the same distributional cluster.

This behavior does not affect the marginal posterior density estimates drastically, but it may have an effect on the parsimony of representation and the clustering. Thus, inference on the partitions of the groups and the data will become unreliable. From another perspective, the drawback of the nDP is due to the fact that there is no part of the model (1.18) that addresses shared commonalities between the latent measures.

A Mixed nested Dirichlet Process approach

Our first proposal to fix the mentioned problems with the usual nDP formulation is now discussed. Specifically, we combine the idea of the nDP and the superposition of random measure, by proposing a Mixed nDP model (MnDP). We will discuss an alternative solution in a later chapter of the thesis.
To define the MnDP, we combine the various atoms at the distributional level, namely the DP's G_k^* $\forall k \geq 1$, with a distribution G_0^* common to all the possible units. Since the common component acts at the individual/observational level, we modify the model in (1.18), replacing the distributional atoms G_k^* with G_{0k}^{**}, defined as the following mixture:

$$G_{0k}^{**} = \rho_k G_k^* + (1 - \rho_k)\, G_0^*,$$

so the model becomes

$$\begin{aligned} (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J)\,|(G_1, \ldots, G_J) &\sim G_1 \times \cdots \times G_J \\ G_1, \ldots, G_J|Q &\overset{i.i.d.}{\sim} Q \\ Q &= \sum_{k\geq 1} \pi_k \delta_{\{\rho_k G_k^* + (1-\rho_k) G_0^*\}} \end{aligned} \tag{1.20}$$

where the π_k's and the ρ_k's are random weights and the distributions G_k^* $k \geq 0$ that appear in the mixture distribution $\rho_k G_k^* + (1 - \rho_k)\, G_0^*$ are Dirichlet Processes defined as:

$$G_k^* = \sum_{l\geq 1} \omega_{lk} \delta_{\theta_{lk}} \quad \forall k \geq 1, \qquad G_0^* = \sum_{l\geq 1} \omega_{l0} \delta_{\theta_{l0}}.$$

This model is similar to the one discussed in Camerlenghi (2017), with the difference that we allow the proportion parameter ρ_k to vary across the different units. Evidently, Q is still a DP where the realizations of its base measure are the mixtures described above. Choosing a mixing parameter ρ_k that depends on the *distributional index*, ensures that

$$\mathbb{P}\left(G_i = G_j|Q\right) > 0.$$

Hence, the distributional clustering property is maintained. We can prove that

$$\begin{aligned} \mathbb{P}\left(G_i = G_j|Q\right) &= \sum_{k\geq 1} \mathbb{P}\left(G_i = G_j = G_{0k}^{**}|Q\right) = \sum_{k\geq 1} \mathbb{P}\left(G_i = G_{0k}^{**}, G_j = G_{0k}^{**}|Q\right) \\ &= \sum_{k\geq 1} \mathbb{P}\left(G_i = G_{0k}^{**}|Q\right) \mathbb{P}\left(G_j = G_{0k}^{**}|Q\right) = \sum_{k\geq 1} \pi_k^2 > 0. \end{aligned}$$

Introducing latent variables that control the assignments to the distributional clusters ($S_j \in \{1, 2, \ldots, K, \ldots\}$) and observational clusters ($M_{ij} \in \{1, 2, \ldots, L, \ldots\}$), we can rewrite the model in (1.20) by employing the stick-breaking formulation. In this way, we provide an easy-to-implement algorithm for posterior inference, which permits us to perform reasonably fast MCMC estimation even when the number of groups is large. For the sake of brevity, in what follows we directly report a deterministically truncated version of the model, in the spirit of Ishwaran and James (2001). As in the original nDP paper, we cut the sequence of distributional labels at a value $K < \infty$, meanwhile, we let every G_k^*, with $k = 0, \ldots, K$ be composed by $L < \infty$ atoms.

$$
\begin{aligned}
X_{ij}|\boldsymbol{S},\boldsymbol{M},\boldsymbol{\mu},\boldsymbol{\sigma^2} &\sim N\left(\mu_{S_j,M_{ij}},\sigma^2_{S_j,M_{ij}}\right),\\
M_{ij}|\boldsymbol{S},\boldsymbol{\omega},\boldsymbol{\xi_0} &\sim (1-\xi_0^{ij})\left(\sum_{l=1}^{L}\omega_{l,S_j}\delta_l(\cdot)\right)+\xi_0^{ij}\left(\sum_{l=L+1}^{2L}\omega_{l0}\delta_l(\cdot)\right),\\
\boldsymbol{\omega}_{k,0}|\boldsymbol{S} &\sim \boldsymbol{\omega}_{k,0}=(\boldsymbol{\omega}_k,\boldsymbol{\omega}_0)\sim GEM(\alpha)\times GEM(\alpha_0),\\
S_j|\boldsymbol{\pi} &\sim \sum_{k=1}^{K}\pi_k\delta_k(\cdot), \qquad \boldsymbol{\pi}\sim GEM(\beta), \qquad \xi_0^{ij}|\boldsymbol{\rho},\boldsymbol{S}\sim Bern\left(\rho_{Sj}\right),\\
\rho_{Sj} &\sim Beta\left(A,B\right), \qquad \left(\mu_{lk},\sigma^2_{lk}\right)\sim NIG(m,\kappa,a,b).
\end{aligned}
\tag{1.21}
$$

When the stick-breaking representation is used, the differences with the nDP are even more evident. Practically, in order to introduce a common probability distribution, a latent variable ξ_0^{ij} is introduced for each observation. If $\xi_0^{ij}=0$ the observation X_{ij} is drawn from the common mixing distribution G_0^*, otherwise the observation is generated from the idiosyncratic distribution indicated by the value assumed by the latent variable S_j. We can easily recover the plain nDP setting $\rho_k=0\ \forall k$. The full conditionals of a collapsed Gibbs sampler, extended to take into account covariates in the likelihood and to handle discrete data with (1.14), can be found in the Appendix of this chapter.
Although the proposed Mixed nested Dirichlet Process approach aptly provides a solution to the clustering behavior exhibited by the nDP, it may be unfeasible for complex high-dimensional sharing structures, involving a larger number of groups, due to computational limitations. Indeed, we observed in various experiments that the model struggles to recover the true common mixture components when the number of groups is considerably high. In the next chapter of this dissertation, we will introduce the Common Atom Model (CAM), which is proposed as an efficient alternative to nDP as a model for distributional clustering. To conclude this section we underline that there exist extensions of the different models we have presented so far. Admixture models, for example, combine the idea of HDP and nDP, accounting for three layers (or more) in the clustering structure (Paisley et al., 2015; Tekumalla et al., 2015).

1.3 Multiple Hypothesis testing

Chapter 3 and Chapter 4 will discuss methods for addressing the multiple comparison problem with large-scale data. Advancements in technology have led us to an era where big data are more easily available. Oftentimes, the principal goal of a study may be to discover a small number of "interesting" units among numerous candidates, e.g. genes whose expression levels differ between cancer-affected and healthy patients. Such units, once identified, might be further investigated for causal links or to answer other research questions. The literature in this area is vast, and we refer to Efron (2012) and Efron and Hastie (2016) for general introductions and overviews. The classical frequentist hypothesis testing framework was not originally intended for addressing large-scale inference problems. Controlling for an overall type I error α would lead to a situation where the number of false positive is exceptionally large. Bonferroni (1936) proposes to correct the type I error level by dividing α by the number N of hypotheses at hand, ensuring that the overall type I error would not exceed α. A test of level α for a single null hypothesis H_0 satisfies, by definition, $\alpha=\mathbb{P}[\text{ Reject } H_0 \text{ when true}]$. However, for a collection of null hypotheses H_{0i} , $i=1,\dots,N$, one can only control for the *family-wise error rate (FWER)*, the probability of making even one false rejection, i.e. $FWER=\mathbb{P}\left[\text{reject any true } H_{0i}\right]$. Bonferroni's procedure controls the FWER at level α : let I_0 be the indices of the true H_{0i}, having say N_0 members.

Then

$$\text{FWER} = \mathbb{P}\left[\bigcup_{I_0}\left(p_i \leq \frac{\alpha}{N}\right)\right] \leq \sum_{I_0}\mathbb{P}\left[p_i \leq \frac{\alpha}{N}\right] = N_0\frac{\alpha}{N} \leq \alpha,$$

where p_i is the p-value associated with the i-th hypothesis. The immediate consequence of this correction is that the new threshold α/N is too conservative, resulting in a number of statistical findings that are too small.

Another widely used approach for addressing the multicomparison problem is based on the control of the False Discovery Rate (FDR, Benajmini and Hochberg, 1995). Consider a decision rule $\mathcal{D}$, which classifies a hypothesis as null (0) or non-null (1), and suppose the ground truth is known. We are then able to quantify the number of hypotheses declared significant by $\mathcal{D}$, although they are actually null (False Positives). The False Positive Rate (FPR) is defined as the number of false positives divided by the total number of tests. The FDR is defined as the expected value of the FPR seen as a function of $\mathcal{D}$: $FDR(\mathcal{D}) = \mathbb{E}\left[FDP(\mathcal{D})\right]$.

Following Benajmini and Hochberg (1995), we say that a decision rule $\mathcal{D}$ controls the FDR at level $q \in (0, 1)$ if

$$\text{FDR}(\mathcal{D}) \leq q.$$

Benajmini and Hochberg (1995) show that if the the p-values corresponding to valid null hypotheses are independent of each other, then

$$\text{FDR}\left(\mathcal{D}_q\right) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = N_0/N,$$

where $\mathcal{D}_q$ is the decision that rejects H_{0i} for $i \leq i^*$, being i^* the largest index for which

$$p_{(i)} \leq \frac{i}{N}q.$$

A different procedure for multiple hypotheses testing is represented by the *two-group model* which Efron has developed in multilpe papers, along with the notion of *local false discovery rate*, seen as the "Benjamini and Hochberg equivalent in the (Empirical) Bayesian setting", and focusing on densities rather than tail areas (i.e. p-values) (Efron et al., 2001; Efron and Tibshirani, 2002; Efron, 2004; Efron, 2007). Indeed, the idea of employing mixture models in the multiple hypothesis testing framework has been widely exploited in the literature, even before Efron's seminal papers, both from a Bayesian and a classical point of view (see, for some examples, Pounds and Morris, 2003; Pan et al., 2003; Allison et al., 2002; Liao et al., 2004). In the two-group model framework, each of the N cases, from now on represented by a test statistic z, is modeled with a mixture model:

$$z \sim f(x) = \pi_0\, f_0(z) + (1 - \pi_0)\, f_1(z)$$

where f_0 is the density of the null distribution, from which the non-relevant cases are generated, f_1 is the alternative, non-null distribution, and π_0 represents the proportion of null cases, which is typically expected to be relatively large.

Let F_j be the c.d.f. of f_j, with $j = 0, 1$. The goal of the model is to assign to each z score a probability of being generated from f_1. If the statistic is properly standardized, we expect f_0 to be a theoretical null, i.e. a standard Gaussian. However, Efron (2004) argues that some deviations from the theoretical null should be allowed, since what we are able to observe is only the *empirical null*. In fact, the observed null distribution can depart from the theoretical normality because of measurement error, hidden correlation between the observations/scores, unobserved covariates, large proportion of genuine but uninterestingly small effects. The difference between the theoretical and the observed null distributions may affect the inference, so Efron (2004) proposes to estimate an empirical null distribution as well, allowing for small differences from

a standard Gaussian. In general, fewer constraints are imposed on f_1, which also needs to be estimated. The only requirement is that non-null realizations should occur further away from zero than null realizations, meaning that f_1 should have longer tails than f_0. Of course, the more f_0 and f_1 are separated (the less they overlap), the better the model estimates will be. In the following, we will carefully address this point, employing the aforementioned PY mixture models and non-local distributions to induce separation between f_0 and f_1.

Working in a Bayesian framework, we do not need to rely on tail-area statistics and can shift our attention to the probability of the null hypothesis given the observed z_i scores. To this purpose, we define the *local (Bayesian) false-discovery rate*:

$$lfdr\left(z_0\right) = \mathbb{P}\left(\text{case } i \text{ is null } |z_i = z_0\right),$$

as opposed to the tail-area false-discovery rate (the FDR equivalent in the Bayesian setting)

$$Fdr\left(z_0\right) \equiv \mathbb{P}\left(\text{case } i \text{ is null } z_i \geq z_0\right) = \pi_0 \frac{1 - F_0\left(z_0\right)}{1 - F\left(z_0\right)}.$$

The previous expression of the $lfdr$ can also be written as

$$lfdr\left(z\right) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + (1 - \pi_0) f_1(z)}. \tag{1.22}$$

Then, a decision rule to allocate an instance in the null or in the alternative group is needed. One possibility is to apply the following criterion: a observation z is deemed as non-null if $lfdr(z) < 0.2$. The condition, in turn, implies that

$$\frac{f_1(z)}{f_0(z)} \geq 4\frac{\pi_0}{\pi_1}$$

so we can see that controlling the $lfdr$ has a direct impact on the Bayes Factor (Kass and Raftery, 1995).
Again, the crucial point here is how to estimate the two distributions and consequently, the proportion π_0. Efron proposes an empirical Bayes estimator for this model (Efron, 2007), which estimates the mixture distribution $f(z)$ assuming the following parametric form: $f(z) = \exp\left\{\sum_{j=0}^{J} \beta_j z^j\right\}$. A default choice of $J = 7$ is often adopted. The problem of maximizing this likelihood can be cast in a Poisson regression framework, using Lindsey's method (Efron and Tibshirani, 1996; Lindsey, 1974b; Lindsey, 1974a). The null distribution $f_0(z)$ is either selected to be a standard Gaussian or, when an empirical null is required, the procedure selects a set $\mathcal{A}_0$ near $z = 0$ where it is assumed that *all* the z_i in $\mathcal{A}_0$ are null; in terms of the two-group model, the assumption can be stated as $f_1(z) = 0 \ \forall z \in \mathcal{A}_0$.

A different empirical Bayes approach is the one proposed by Muralidharan (2012). The author models the test statistics as $z_i \sim f_{\delta_i}$, so it is possible to rewrite the $lfdr$ as $\mathbb{P}\left(\delta_i = 0 | z_i\right)$. To do so, Muralidharan (2012) propose the following hierarchical model

$$z|\delta \sim f_\delta(z) \qquad \delta \sim g(\delta) = \sum_{j=0}^{J-1} \pi_j g_j(\delta), \tag{1.23}$$

where g_j's are taken from some parametric family of priors for δ. In this context, g_0 is assumed to be a point mass at zero (to reflect the theoretical null) or it can be estimated, along with putting a constraint on the weights π, forcing $\pi_0 > \pi_j \ \forall j \geq 0$. The quantities of interest

are estimated by $p_j(z) = \frac{\pi_j f^{(j)}(z)}{f(z)}$, where $f^{(j)} = \int f_\delta g_j(\delta) d\delta$ is the j-th group's marginal. In particular, the $lfdr$ can be recovered as $p_0(z)$.
Similarly, Martin and Tokdar (2012) develop a likelihood-based analysis of the two-group model, where they consider a regularized estimation for the parameters of the null distribution (μ, σ) and the null proportion π_0, and a semiparametric specification of the non-null density f_1. In particular, they estimate the empirical null and employ a mixture representation of f_1 that results in heavier tails than f_0 to reflect the belief that z-scores from the non-null cases are likely to be larger in magnitude than those from the null cases:

$$f_1(z) = \int_{\mathcal{U}} \mathrm{N}\left(z|\mu + \tau\sigma u, \sigma^2\right) \psi(u) \mathrm{d}u$$

with ψ a density with respect to the Lebesgue measure on $\mathcal{U} = [-1, 1]$ and $\tau \geqslant 1$, a scaling factor.

Many other authors have built on this idea and improved the estimating procedure, choosing a fully Bayesian approach. An important example is provided by Do et al. (2005), where the authors propose to model both the densities f_0 and f_1 as infinite mixtures of Gaussians, using two DPMMs:

$$\begin{aligned} f_j(z) &= \int N\left(z; \mu, \sigma^2\right) \mathrm{d}G_j(\mu) \qquad G_j \sim DP\left(M, G_j^*\right), \quad j = 0, 1 \\ G_0^* &= N\left(b, \sigma_0^2\right) \qquad G_1^* = 0.5N\left(-b_1, \sigma_1^2\right) + 0.5N\left(b_1, \sigma_1^2\right), \end{aligned} \tag{1.24}$$

and they complete the model with hyperpriors for the paramters $b, b_0, b_1, \sigma, \sigma_0, \sigma_1$. A mixture in the base measure in the multiple hypothesis testing context is also employed in Guindani et al. (2009). After developing a coherent decision framework for multiple comparisons, they discuss a semi-parametric model for which they show that the Bayes rule can be approximated by the Optimal Discovery Procedure (ODP) introduced in Storey (2007). In Chapters 3 and 4 we will contribute to the multiple hypothesis testing literature by introducing two extensions of the two-group model framework described here.

1.4 The Estimation of the Intrinsic Dimension of a dataset

The final topic we investigate is the estimation of the Intrinsic Dimension (ID) of a dataset, namely the dimension d of the latent manifold in which the statistical units, observed in a D-dimensional space, lie. We expect some degree of dependency among the variables of a dataset, thus usually $d < D$. More formally, we refer to the definition of ID provided by Bishop (1995): *a set in D dimensions has an intrinsic dimension equal to d if the data lies within a d -dimensional subspace of $\mathbb{R}^D$ entirely, without information loss.* Another interpretation is provided in pattern recognition literature, where a point set is viewed as a *sample set uniformly drawn from an unknown smooth (or locally smooth) manifold structure, eventually embedded in a higher dimensional space through a non-linear smooth mapping; in this case, the ID to be estimated is the manifold's topological dimension* (Campadelli et al., 2015).

The literature regarding this topic is extremely vast and many different approaches for ID estimation have been proposed. We refer to Facco and Laio (2017) and Campadelli et al. (2015) for comprehensive reviews. In general, ID estimation methods can be divided into four main sub-categories:

- *Projective Methods*, such as Multidimensional Scaling (MDS) (Jolliffe, 2002) and Principal Component analysis (PCA) (Cox and Cox, 2000). MDS tends to preserve the pairwise distances among the data as much as possible, and is usually employed to provide a visual

representation of high-dimensional data. PCA instead aims to reduce the data dimensionality deriving the best projection subspace based on the computation and the thresholding of the N eigenvalues of the covariance matrix of the sample: the ID corresponds to the number of most relevant eigenvectors (called the Principal Components). In both cases, the goal is to find the best projection (w.r.t. some pre-specified loss function) of the data onto a lower dimensional space.

- *Fractal Methods.* The basis of all fractal methods is that the volume of a d-dimensional ball of radius r scales as r^d (Falconer, 2003; Camastra and Vinciarelli, 2002). Thus, the fractal dimension estimators are based on the idea of counting the number of observations in a neighborhood of radius r to estimate its rate of growth $\hat{r}$. Since the estimated growth is assumed to resemble the theoretical growth rate r^d, these methods exploit the connection between the empirical $\hat{r}$ and r^d to estimate the parameter d, regarded as the fractal dimension of the data.

- *Graph based methods.* Theory and algorithms for graphs can be exploited in different manners to estimate the ID of datasets. In particular, graph theory is useful when dealing with non-linear subspaces. Building a graph linking points close to each other can provide valuable insights regarding the geometry of the latent manifold where the data lie. Let $\mathcal{G}(\mathbf{X}_N) = \left(\{\boldsymbol{x}_i\}_{i=1,\dots,N}, \{e_{i,j}\}_{i,j=1,\dots,N}\right)$ denote a graph characterized by points $\{\boldsymbol{x}_i\}_{i=1}^N$ and edges $\{e_{ij}\}_{i,j=1}^N$. Working with distances suited for graphs allows us to uncover a data structure impossible to distinquish with the usual Euclidean methods. An important quantity in this context is the *geodesic distance*, defined as follows. Consider the Riemann manifold M, i.e. a smooth manifold equipped with differentiable inner product g, called Riemann metric, and let $\gamma(s) : [0, S] \to M$ be a geodesic arc parametrized by $s \in [0, S]$. Then, the geodesic distance d_M is defined as

$$d_M(x, y) = \inf\{L(\gamma) | \gamma(0) = x, \gamma(S) = y\} \tag{1.25}$$

  where $L(\gamma) := \int_0^S g\{\gamma'(t), \gamma'(t)\}^{1/2} dt$ is the length of curve γ (Li and Dunson, 2019a). This can be applied by using the geodesic distance to recover the geodesic minimum spanning tree and the ID d of the dataset via a linking equation (Costa and Hero, 2004).

- *Nearest-Neighbors (NNs) methods.* Consider a collection of data points $\{\boldsymbol{x}_i\}$, for $i = 1, \dots, n$. This class of methods comes from the idea that close points are uniformly drawn from d -dimensional balls (hyperspheres) $\mathcal{B}_d(\boldsymbol{x}, r)$ characterized by a small radius $r \to 0 \in \mathbb{R}^+$ centered in point $\boldsymbol{x}$. If $\rho(\boldsymbol{x})$ is a density distribution defined on $\mathbb{R}^d$, the following approximation holds: $\frac{k}{n} \approx \rho(\boldsymbol{x})\, \omega_d\, r^d$, where k is the number of NNs of $\boldsymbol{x}$ within the hypersphere $\mathcal{B}_d(\boldsymbol{x}, r)$, while ω_d is the volume of the d-dimensional unit sphere in $\mathbb{R}^d$ (Pettis et al., 1979). Intuitively this tells that the proportion of points of a given sample which fall into the ball $\mathcal{B}(\boldsymbol{x}, r)$ is approximately $\rho(\boldsymbol{x})$ times the volume of the ball. If the density is constant, one can estimate the intrinsic dimension using only the average distances from a point's k NNs.

Additionally, some model-based geometrical approaches to explore the topology of datasets have recently been proposed. Mukhopadhyay et al. (2019) use Fisher-Gaussian kernels to estimate densities of data embedded in non-linear subspaces. Li et al. (2017) propose to learn the structure of latent manifolds by approximating them with spherelets instead of locally linear approximation, developing a spherical version of PCA. In the same spirit, Li and Dunson (2019a) applies this idea to the classification of data lying on complex, non-linear, overlapping and intersecting supports. Finally, Li and Dunson (2019b) propose to use the spherical PCA to estimate a geodesic distance matrix between the data, which takes into account the structure of the latent

embedding manifolds and create a spherical version of the k-medoids algorithm (Kaufman and Rousseeuw, 1987).
In Chapter 5 we contribute to the ID estimation literature extending the methodology developed in Facco and Laio (2017), Facco et al. (2017), and Allegra et al. (2019). Their methods is based on a property of homogeneous Poisson processes (HPP). Let $\rho(\boldsymbol{x})$ be the density of the points in a dataset and call $r_{i,j}$ the distance between a generic point $\boldsymbol{x}_i$ and its NN of order j. The authors have proved that if the density of the data can be deemed as constant in a neighborhood $\mathcal{B}_d(\boldsymbol{x}_i, r)$ each point i (i.e. $\rho(\boldsymbol{x}) = \rho \quad \forall \boldsymbol{x} \in \mathcal{B}_d(\boldsymbol{x}_i, r)$), all the volumes of the stochastic spherical shells – defined as

$$v_l = \omega_d \left(r_{i,l}^d - r_{i,l-1}^d \right) \quad \forall l = 2, \dots, n,$$

where d is the dimensionality of the space in which the points are embedded and ω_d is the volume of the d-sphere with unitary radius – are exponentially distributed with rate equal to the density ρ. From this observation, which describes the inter-arrival times of HPP in a multivariate setting, they develop a parametric model where the parameter of interest is exactly the ID. Then, the ratio $\frac{r_{i,2}}{r_{i,1}}$ is Pareto distributed, with a shape parameter that is exactly d. Moreover, their method was extended to a Bayesian setting mixture model to take into account heterogeneous IDs in the same dataset. It is worth noticing that, to derive these last results, the hypothesis of homogeneity of the density $\rho(\boldsymbol{x})$ is required only on the scale of the second NN of each point.

1.5 Outline and main contributions

Bayesian mixture models are ubiquitous in statistics, and they can be employed in many different ways in a wide variety of contexts. In this first Chapter we briefly reviewed the methodological concepts and discussed the necessary tools that belong to the different areas that we will contribute to with this dissertation.

In Chapter 2 we propose a Common Atoms model (CAM) for nested datasets, which overcomes the limitations of the nDP that we have outlined in Section 1.2.3. We derive its theoretical properties and develop a slice sampler for nested data to obtain an efficient algorithm for posterior simulation. We then embed the model in a Rounded Mixture of Gaussian kernels framework to recover a meaningful clustering structure among subjects of a microbiome study.

In Chapter 3 we develop a BNP version of the two-group model, modeling both f_0 and f_1 with Pitman-Yor process mixture models. We propose to fix the two discount parameters σ_0 and σ_1 so that $\sigma_0 > \sigma_1$, according to the rationale that the null PY should be closer to its base measure (appropriately chosen to be a standard Gaussian base measure), while the alternative PY should have fewer constraints. We propose a marginal sampler and, to improve the computational efficiency and speed, we also introduce a split and merge move. An important role is played by the separation between the null and the alternative distribution. If the two distributions overlap, the inferential process can be jeopardized. To induce separation, we employ a non-local prior on the location parameter of the base measure of the PY placed on f_1. We show how the model performs in different scenarios and apply this methodology to microbiome and prostate cancer data.

Chapter 4 presents an alternative proposal for the two-group model. Here, we use a non-local distribution to model the alternative density directly in the likelihood formulation. By multiplying a weight function and a local density, we are able to induce separation between the models. The local density is modeled both through a parametric model and a nonparametric model. We are able to provide a theoretical justification for the adoption of the proposed likelihood approach. After comparing the performance of our model with several competitors, we present

three applications on real, publicly available datasets.

Finally, in Chapter 5 we examine different ways to extend the model for ID estimation discussed in Allegra et al. (2019). First, we propose to include more suitable priors in their parametric model, such as truncated and repulsive versions of the original priors to facilitate estimation of the ID parameters in posterior inference. Then, we extend their theoretical methodology by deriving distributions for a generic number of consecutive ratios of distances $\frac{r_{i,j}}{r_{i,j-1}}$ and we model them to include more information in the estimation process. To overcome the choice of a fixed number of K mixture components, we propose a simple Dirichlet process mixture model. The chapter is then concluded with simulation studies and the application to real data.

Chapter 6 concludes the thesis with a few remarks and directions for future research.

Appendix

1.A Gibbs sampler for the Mixed nested DP

Consider the model in (1.21), extended to the discrete data modeling case following (1.14). Let $\boldsymbol{X} = \{X_{ij}\}$ represent the discrete data and $\boldsymbol{X}^*$ represent the corresponding continuous latent variables. Moreover, we take into account the presence of group-specific covariates $\boldsymbol{Z^j}$, modeling the continuous latent r.v.s as $X^*_{ij} \sim N\left(\mu_{S_j,M_{ij}} + \boldsymbol{Z^j\beta}, \sigma^2_{S_j,M_{ij}}\right)$. We can rewrite the posterior as proportional to the joint distribution, which in turn can be decomposed following the structure of previous model.

$$
\begin{aligned}
p\left(\boldsymbol{X^*}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \beta, \boldsymbol{\xi_0}, \boldsymbol{\rho} | \boldsymbol{X}\right) \propto & p\left(\boldsymbol{X^*}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \beta, \boldsymbol{\xi_0}, \boldsymbol{\rho}, \boldsymbol{X}\right) \\
\propto & p\left(\boldsymbol{X} | \boldsymbol{X^*}\right) \times p\left(\boldsymbol{X^*} | \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\beta}\right) \times \\
& p\left(\boldsymbol{M} | \boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\xi_0}\right) \times p\left(\boldsymbol{\omega}\right) \times p\left(\boldsymbol{S} | \boldsymbol{\pi}\right) \times \\
& p\left(\boldsymbol{\pi}\right) \times p\left(\boldsymbol{\mu} | \boldsymbol{\sigma^2}\right) \times p\left(\boldsymbol{\sigma^2}\right) \times p\left(\boldsymbol{\beta}\right) \\
& p\left(\boldsymbol{\xi_0} | \boldsymbol{\rho}\right) \times p\left(\boldsymbol{\rho}\right).
\end{aligned}
$$

To derive the full conditionals, we will collapse some of the random variables whenever possible.

1. Full conditional for the latent random variable $\boldsymbol{X^*}$:

$$
\begin{aligned}
p\left(\boldsymbol{X^*} | \cdots\right) \propto & \, p\left(\boldsymbol{X} | \boldsymbol{X^*}\right) p\left(\boldsymbol{X^*} | \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\beta}\right) \\
\propto & \prod_{i,j} \sum_{j=0}^{+\infty} \delta_j(\cdot) \mathbf{1}_{[a_j, a_{j+1})}\left(X^*_{ij}\right) \prod_{i,j} N\left(\mu_{S_j,M_{ij}} + \boldsymbol{Z^j\beta}, \sigma^2_{S_j,M_{ij}}\right) \iff \\
p\left(X^*_{ij} | \cdots\right) \propto & \sum_{j=0}^{+\infty} \delta_j(\cdot) \mathbf{1}_{[a_j, a_{j+1})}\left(X^*_{ij}\right) N\left(\mu_{S_j,M_{ij}} + \boldsymbol{Z^j\beta}, \sigma^2_{S_j,M_{ij}}\right) \iff \\
p\left(X^*_{ij} | X_{ij} = x_{ij} \cdots\right) \propto & \mathbf{1}_{\left[a_{x_{ij}}, a_{x_{ij}+1}\right)}\left(X^*_{ij}\right) N\left(\mu_{S_j,M_{ij}} + \boldsymbol{Z^j\beta}, \sigma^2_{S_j,M_{ij}}\right) \\
p\left(X^*_{ij} | X_{ij} = x_{ij} \cdots\right) \sim & \, TN(\mu_{S_j,M_{ij}} + \boldsymbol{Z^j\beta}, \sigma^2_{S_j,M_{ij}}; a_{x_{ij}}, a_{x_{ij}+1})
\end{aligned}
$$

2. Joint full conditional for $\boldsymbol{M}$ and $\boldsymbol{\xi_0}$.

$$
\begin{aligned}
p\left(\boldsymbol{M}, \boldsymbol{\xi_0} | \cdots - \boldsymbol{X^*}\right) \propto & \int p\left(\boldsymbol{M}, \boldsymbol{\xi_0}, \boldsymbol{X^*} | \cdots\right) d\boldsymbol{X^*} \\
\propto & \int p\left(\boldsymbol{X} | \boldsymbol{X^*}\right) p\left(\boldsymbol{X^*} | \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\beta}\right) p\left(\boldsymbol{M} | \boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\xi_0}\right) p\left(\boldsymbol{\xi_0} | \boldsymbol{\rho}\right) d\boldsymbol{X^*} \\
\propto & \int \prod_{i,j} \left(\sum_{j=0}^{+\infty} \delta_j(\cdot) \mathbf{1}_{[a_j, a_{j+1})}\left(X^*_{ij}\right) \right) \prod_{i,j} N\left(\mu_{S_j,M_{ij}} + \boldsymbol{Z^j\beta}, \sigma^2_{S_j,M_{ij}}\right) \\
& \times \prod_{i,j} \left((1 - \xi_0^{ij}) \left(\sum_{l=1}^{L} \omega_{l,S_j} \delta_l(\cdot) \right) + \xi_0^{ij} \left(\sum_{l=L+1}^{2L} \omega_{l0} \delta_l(\cdot) \right) \right)
\end{aligned}
$$

$$\times \prod_{ij} \left(\rho_{S_j}^{\xi_0^{ij}} \left(1-\rho_{S_j}\right)^{1-\xi_0^{ij}} \right) d\boldsymbol{X}^* \iff$$

$$p\left(M_{ij}, \xi_0^{ij} | \cdots - X_{ij}^*\right) \propto \int \left(\sum_{j=0}^{+\infty} \delta_j(\cdot) \mathbf{1}_{[a_j, a_{j+1})} \left(X_{ij}^*\right) \right) N\left(\mu_{S_j, M_{ij}} + \boldsymbol{Z^j}\boldsymbol{\beta}, \sigma^2_{S_j, M_{ij}}\right)$$
$$\times \left((1-\xi_0^{ij}) \left(\sum_{l=1}^{L} \omega_{l,S_j} \delta_l(\cdot) \right) + \xi_0^{ij} \left(\sum_{l=L+1}^{2L} \omega_{l0} \delta_l(\cdot) \right) \right)$$
$$\times \left(\rho_{S_j}^{\xi_0^{ij}} \left(1-\rho_{S_j}\right)^{1-\xi_0^{ij}} \right) dX_{ij}^*$$

Let $\Delta\Phi\left(x_{ij}; S_j, M_{ij}\right) = \Phi\left(x_{ij}+1; \mu_{S_j,M_{ij}} + \boldsymbol{Z^j}\boldsymbol{\beta}, \sigma^2_{S_j,M_{ij}}\right) - \Phi\left(x_{ij}; \mu_{S_j,M_{ij}} + \boldsymbol{Z^j}\boldsymbol{\beta}, \sigma^2_{S_j,M_{ij}}\right)$, where $\Phi(x; m, s^2)$ is the c.d.f. of a Normal random variable characterized by mean m and variance s^2. Accordingly with the values assumed by M_{ij} and ξ_0^{ij}, four scenarios are possible:

$$p\left(M_{ij} = l \geq L+1, \xi_0^{ij} = 1 | X_{ij} = x_{ij}, \cdots - \boldsymbol{X_{ij}^*}\right) \propto \rho_{S_j} \omega_{l0} \Delta\Phi\left(x_{ij}; 0, l\right),$$
$$p\left(M_{ij} = l \leq L, \xi_0^{ij} = 1 | X_{ij} = x_{ij}, \cdots - \boldsymbol{X_{ij}^*}\right) \propto 0,$$
$$p\left(M_{ij} = l \geq L+1, \xi_0^{ij} = 0 | X_{ij} = x_{ij}, \cdots - \boldsymbol{X_{ij}^*}\right) \propto 0,$$
$$p\left(M_{ij} = l \leq L, \xi_0^{ij} = 0 | X_{ij} = x_{ij}, \cdots - \boldsymbol{X_{ij}^*}\right) \propto \left(1-\rho_{S_j}\right) \omega_{l,S_j} \Delta\Phi\left(x_{ij}; S_j, l\right).$$

$\boldsymbol{M}$ and $\boldsymbol{\xi}$ must be updated jointly, in order to avoid to be trapped into a deterministic update step.

3. Full conditional for $\boldsymbol{S}$:

$$p\left(\boldsymbol{S} | \cdots - \boldsymbol{X}^*, -\boldsymbol{M}, -\boldsymbol{\xi_0}\right) \propto \int\int\int p\left(\boldsymbol{S}, \boldsymbol{M}, \boldsymbol{X}^*, \boldsymbol{\xi_0} | \cdots\right) d\boldsymbol{X}^* d\boldsymbol{M} d\boldsymbol{\xi_0}$$
$$\propto \int\int\int p\left(\boldsymbol{X} | \boldsymbol{X}^*\right) p\left(\boldsymbol{X}^* | \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \beta\right) p\left(\boldsymbol{M} | \boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\xi_0}\right)$$
$$\times p\left(\boldsymbol{\xi_0} | \boldsymbol{\rho}\right) d\boldsymbol{X}^* d\boldsymbol{M} d\boldsymbol{\xi_0}$$
$$\propto \int\int\int \prod_{i,j} \left(\sum_{j=0}^{+\infty} \delta_j(\cdot) \mathbf{1}_{[a_j, a_{j+1})} \left(X_{ij}^*\right) \right)$$
$$\times \prod_{i,j} N\left(\mu_{S_j,M_{ij}} + \boldsymbol{Z^j}\boldsymbol{\beta}, \sigma^2_{S_j,M_{ij}}\right)$$
$$\times \prod_{i,j} \left((1-\xi_0^{ij}) \left(\sum_{l=1}^{L} \omega_{l,S_j} \delta_l(\cdot) \right) + \xi_0^{ij} \left(\sum_{l=L+1}^{2L} \omega_{l0} \delta_l(\cdot) \right) \right)$$
$$\times \prod_{j} \left(\sum_{k=1}^{K} \pi_k \delta_k(\cdot) \right) \prod_{ij} \left(\rho_{S_j}^{\xi_0^{ij}} \left(1-\rho_{S_j}\right)^{1-\xi_0^{ij}} \right) d\boldsymbol{X}^* d\boldsymbol{M} d\boldsymbol{\xi_0} \iff$$

$$p\left(S_j | \cdots - \boldsymbol{X}_j^*, -\boldsymbol{M}_j, -\boldsymbol{\xi_0^j}\right) \propto \left(\sum_{k=1}^{K} \pi_k \delta_k(\cdot) \right) \prod_i \int\int\int \left(\sum_{j=0}^{+\infty} \delta_j(\cdot) \mathbf{1}_{[a_j, a_{j+1})} \left(X_{ij}^*\right) \right)$$
$$\times N\left(\mu_{S_j,M_{ij}} + \boldsymbol{Z^j}\boldsymbol{\beta}, \sigma^2_{S_j,M_{ij}}\right) dX_{ij}^*$$

$$\times \left((1-\xi_0^{ij}) \left(\sum_{l=1}^{L} \omega_{l,S_j} \delta_l(\cdot) \right) + \xi_0^{ij} \left(\sum_{l=L+1}^{2L} \omega_{l0} \delta_l(\cdot) \right) \right) dM_{ij}$$
$$\times \left(\rho_{S_j}^{\xi_0^{ij}} \left(1-\rho_{S_j}\right)^{1-\xi_0^{ij}} \right) d\xi_0^{ij} \iff$$

Specializing the formula with the values assumed by the other variables, we obtain:

$$\begin{aligned}
p\left(S_j = k | \cdots - \boldsymbol{X}_j^*, -\boldsymbol{M}_j, -\boldsymbol{\xi_0^j}\right) &\propto \pi_k \prod_i \sum_{z_i=0}^{1} \left(\rho_k^{z_i} (1-\rho_k)^{1-z_i} \right) \sum_{m=1}^{2L} \Delta\Phi\left(x_{ij}; k\mathbf{1}_{\{m\le L\}}, m\right) \\
&\times \left((1-z_i) \left(\sum_{l=1}^{L} \omega_{lk} \delta_l(m) \right) + z_i \left(\sum_{l=L+1}^{2L} \omega_{l0} \delta_l(m) \right) \right) \\
&\propto \pi_k \prod_i \left((1-\rho_k) \sum_{m=1}^{2L} \Delta\Phi\left(x_{ij}; k\mathbf{1}_{\{m\le L\}}, m\right) \left(\sum_{l=1}^{L} \omega_{lk} \delta_l(m) \right) \right. \\
&\left. + \rho_k \sum_{m=1}^{2L} \Delta\Phi\left(x_{ij}; k\mathbf{1}_{\{m\le L\}}, m\right) \left(\sum_{l=L+1}^{2L} \omega_{l0} \delta_l(m) \right) \right) \\
&\propto \pi_k \prod_i \left((1-\rho_k) \sum_{m=1}^{L} \omega_{m,k} \Delta\Phi\left(x_{ij}; k, m\right) \right. \\
&\left. + \rho_k \sum_{m=L+1}^{2L} \omega_{m,0} \Delta\Phi\left(x_{ij}; 0, m\right) \right)
\end{aligned}$$

4. Full conditional for $\boldsymbol{\rho}$:

$$p\left(\boldsymbol{\rho} | \cdots\right) \propto p\left(\boldsymbol{\xi_0} | \boldsymbol{\rho}\right) p\left(\boldsymbol{\rho}\right) \propto \prod_{i,j} \rho_{S_j}^{\xi_0^{ij}} \left(1-\rho_{S_j}\right)^{1-\xi_0^{ij}} \prod_k \rho_k^{A-1} (1-\rho_k)^{B-1} \iff$$

$$p\left(\rho_k | \cdots\right) \propto \prod_{\substack{i \\ j:S_j=k}} \rho_k^{\xi_0^{ij}} (1-\rho_k)^{1-\xi_0^{ij}} \rho_k^{A-1} (1-\rho_k)^{B-1} = \rho_k^{\sum_i \xi_0^{ij} + A - 1} (1-\rho_k)^{\sum_i \xi_0^{ij} + B - 1}$$

Define as n^k the number of observations in groups indexed with j such that $S_j = k$. Moreover, let $n_0^k = \sum_{j:S_j=k} \xi_0^{ij}$ denote the number of observations in groups indexed with j such that $S_j = k$ (in other words, the group of observation belonging to the same distributional cluster) *assigned to the shared distribution* G_0^*. Exploiting conjugacy, the full conditionals for the ρ_j's will be

$$\rho_j \sim Beta\left(n_0^k + A, n^k - n_0^k + B\right)$$

Let us focus on the stick-breaking priors and their corresponding full conditionals. The proposed modification does not affect the full conditional for the weights $\boldsymbol{\pi}$. We then recover the same formulas proposed by Rodríguez et al. (2008), using the stick-breaking representation with auxiliary variables V_k, U_l^0, $U_{lk}^{\emptyset}$, a priori distributed respectively as $Beta\left(1, \beta\right)$, $Beta\left(1, \alpha_0\right)$ and

$Beta\,(1,\alpha)$. Notice that – in this case – $k = 1,\dots,K$, $l = 1,\dots,L$ and $l' = L+1,\dots,2L$. We need to define m_k^* as the number of groups assigned to the same cluster k, where $\sum_{k=1}^{K} m_k^*$ equals the total number of observed groups J.

5. Full conditional for $\boldsymbol{\pi}$:

$$p\,(\boldsymbol{\pi}|\cdots) \propto p\,(\boldsymbol{S}|\boldsymbol{\pi})\,p\,(\boldsymbol{\pi}) \propto p\,(\boldsymbol{\pi})\,\pi_1^{m_1^*}\cdots\pi_K^{m_K^*}$$

which can be recovered with the stick-breaking construction, defining the full conditional of V_k, $\forall k = 1,\dots,K$:

$$V_k \sim Beta\left(1 + m_k^*, \beta + \sum_{s=k+1}^{K} m_k^*\right).$$

6. The derivation of the full conditional for $\boldsymbol{\omega}$ requires more care. Let us write down the full conditional resembling the form stated in the previous full conditional.

$$\begin{aligned}(7)\quad p\,(\boldsymbol{\omega}|\cdots) &\propto p\,(\boldsymbol{M}|\boldsymbol{S},\boldsymbol{\omega},\boldsymbol{\xi_0})\,p\,(\boldsymbol{\omega}) \\ &\propto \prod_{k=1}^{K} p\,(\boldsymbol{\omega_k}) \prod_{i,j}\left((1-\xi_0^{ij})\left(\sum_{l=1}^{L}\omega_{l,S_j}\delta_l(\cdot)\right) + \xi_0^{ij}\delta_0(\cdot)\right)\end{aligned}$$

The previous formula can be decomposed into the product of K elements. We can then focus for the case $S_j = k$. Let us define $n_{\emptyset}^k$ as the total number of observations assigned to the distributional cluster k that do not come from the shared distribution G_0^*. $n_{\emptyset}^k$ can be in turn decomposed as the sum $n_{\emptyset}^k = \sum_{l=1}^{L} n_{\emptyset}^{lk}$, for $i = 1,\dots,L$. Define n^0 and n^{l0} analogously.

$$\begin{aligned}p\,(\boldsymbol{\omega_{k,0}}|\cdots) &\propto p\,(\boldsymbol{\omega_{k,0}}) \prod_{i}\left((1-\xi_0^{ik})\left(\sum_{l=1}^{L}\omega_{lk}\delta_l(\cdot)\right) + \xi_0^{ik}\left(\sum_{l=L+1}^{2L}\omega_{l0}\delta_l(\cdot)\right)\right) \\ &\propto p\,(\boldsymbol{\omega_k}) \times p\,(\boldsymbol{\omega_0}) \times \omega_{1,k}^{n_{\emptyset}^{1,k}}\cdots\omega_{L,k}^{n_{\emptyset}^{L,k}} \cdot \omega_{L+1,0}^{n^{L+1,0}}\cdots\omega_{2L,0}^{n^{2L,0}}\end{aligned}$$

Exploiting the prior independence of the two stick-breaking processes, this formula suggests the following two representation:

$$U_{lk}^{\emptyset} \sim Beta\left(1 + n_{\emptyset}^{lk}, \alpha + \sum_{r=l+1}^{L} n_{\emptyset}^{r,k}\right) \quad l = 1,\dots,L, \quad \forall k = 1,\dots,K.$$

and

$$U_l^0 \sim Beta\left(1 + n^{l',0}, \alpha + \sum_{r=l'+1}^{2L} n^{r,0}\right) \quad l' = L+1,\dots,2L.$$

The last two full conditionals for $\boldsymbol{\beta}$ and $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ are easily recovered employing the usual conjugacy properties.

7. Full conditional for $\boldsymbol{\theta} = (\mu, \sigma^2)$:

$$p\left(\boldsymbol{\mu}, \boldsymbol{\sigma^2}\right) \propto p\left(\boldsymbol{\mu}, \boldsymbol{\sigma^2}\right) p\left(\boldsymbol{X^*}|\boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\beta}\right)$$

$$\propto \prod_{lk} \left(\sigma_{lk}^2\right)^{-a-1} \exp\left(-b/\sigma_{lk}^2\right) \frac{\sqrt{\kappa}}{\sqrt{2\pi\sigma_{lk}^2}} \exp\left(-\frac{\kappa\left(\mu_{lk}-m\right)^2}{2\sigma_{lk}^2}\right)$$
$$\times \prod_{ij} \frac{1}{\sqrt{2\pi\sigma_{M_{ij},S_j}^2}} \exp\left(-\frac{\left(x_{ij}^* - \mu_{M_{ij},S_j} - \boldsymbol{Z^j\beta}\right)^2}{2\sigma_{M_{ij},S_j}^2}\right)$$

Let us define $\tilde{x}_{ij} = x_{ij}^* - \boldsymbol{Z^j\beta}$ and n_{lk} as the number of the observations assigned to the observational cluster $l = 1, \ldots, L$ and the distributional cluster $k = 1, \ldots, K$. Moreover, let $\bar{\tilde{x}}_{lk}$ represent the average of the values $\tilde{x}_{lk}$ that have been clustered together and $n_{l'0}$ be the number of the observations that come from the shared distribution assigned to the observational cluster $l' = L+1, \ldots, 2L$. We then obtain, for $l \in \{1, \ldots, L\}$:

$$\begin{aligned} p\left(\mu_{lk}, \sigma_{lk}^2\right) &\propto \left(\sigma_{lk}^2\right)^{-a-1} \exp\left(-b/\sigma_{lk}^2\right) \frac{\sqrt{\kappa}}{\sqrt{2\pi\sigma_{lk}^2}} \exp\left(-\frac{\kappa\left(\mu_{lk}-m\right)^2}{2\sigma_{lk}^2}\right) \\ &\times \prod_{ij:M_{ij}=l,S_j=k} \frac{1}{\sqrt{2\pi\sigma_{lk}^2}} \exp\left(-\frac{\left(\tilde{x}_{ij}-\mu_{lk}\right)^2}{2\sigma_{lk}^2}\right) \\ &\propto \left(\sigma_{lk}^2\right)^{-a-1} \exp\left(-b/\sigma_{lk}^2\right) \frac{\sqrt{\kappa}}{\sqrt{2\pi\sigma_{lk}^2}} \exp\left(-\frac{\kappa\left(\mu_{lk}-m\right)^2}{2\sigma_{lk}^2}\right) \\ &\times \left(2\pi\sigma_{lk}^2\right)^{-n_{lk}/2} \exp\left(-\frac{1}{2\sigma_{lk}^2} \sum_{ij:M_{ij}=l,S_j=k} \left(\tilde{x}_{ij}-\mu_{lk}\right)^2\right) \end{aligned}$$

while for $l' \in \{L+1, \ldots, 2L\}$, similarly we get

$$\begin{aligned} p\left(\mu_{l'0}, \sigma_{l'0}^2\right) &\propto \left(\sigma_{l'0}^2\right)^{-a-1} \exp\left(-b/\sigma_{l'0}^2\right) \frac{\sqrt{\kappa}}{\sqrt{2\pi\sigma_{l'0}^2}} \exp\left(-\frac{\kappa\left(\mu_{l'0}-m\right)^2}{2\sigma_{l'0}^2}\right) \\ &\times \prod_{ij:M_{ij}=l'} \frac{1}{\sqrt{2\pi\sigma_{l'0}^2}} \exp\left(-\frac{\left(\tilde{x}_{ij}-\mu_{l'0}\right)^2}{2\sigma_{l'0}^2}\right) \\ &\propto \left(\sigma_{l'0}^2\right)^{-a-1} \exp\left(-b/\sigma_{l'0}^2\right) \frac{\sqrt{\kappa}}{\sqrt{2\pi\sigma_{l'0}^2}} \exp\left(-\frac{\kappa\left(\mu_{l'0}-m\right)^2}{2\sigma_{l'0}^2}\right) \\ &\times \left(2\pi\sigma_{l'0}^2\right)^{-n_{l'0}/2} \exp\left(-\frac{1}{2\sigma_{l'0}^2} \sum_{ij:M_{ij}=l'} \left(\tilde{x}_{ij}-\mu_{l'0}\right)^2\right) \end{aligned}$$

Following the usual conjugacy reasoning, we obtain

$$\left(\mu_{lk}, \sigma_{lk}^2\right) \sim NIG\left(m^*, \sigma^{2^*}, a^*, b^*\right)$$
$$\left(\mu_{l'0}, \sigma_{l'0}^2\right) \sim NIG\left(m_0^*, \sigma_0^{2^*}, a_0^*, b_0^*\right)$$

where

$$\begin{aligned} m^* &= \frac{\kappa m + n_{lk}\bar{\tilde{x}}_{lk}}{\kappa + n_{lk}} & m_0^* &= \frac{\kappa m + n_{l'0}\bar{\tilde{x}}_{l'0}}{\kappa + n_{l'0}} \\ \kappa^* &= \kappa + n_{lk} & \kappa_0^* &= \kappa + n_{l'0} \\ a^* &= a + n_{lk}/2 & a_0^* &= a + n_{l'0}/2 \end{aligned}$$

and

$$b^* = b + 0.5\left(\sum_{ij:M_{ij}=l,S_j=k}\left(\tilde{x}_{ij}-\bar{\tilde{x}}_{lk}\right)^2 + \left(\frac{\kappa n_{lk}}{\kappa+n_{lk}}\right)\left(\tilde{x}_{lk}-m\right)^2\right)$$

$$b_0^* = b + 0.5\left(\sum_{ij:M_{ij}=l'}\left(\tilde{x}_{ij}-\bar{\tilde{x}}_{l'0}\right)^2 + \left(\frac{\kappa n_{l'0}}{\kappa+n_{l'0}}\right)\left(\tilde{x}_{l'0}-m\right)^2\right)$$

For computational and notational convenience, we can set up the following $N \times p$ matrix, where $N = \sum_{j=1}^{J} n_j$ and n_j is the number of observation inside group j. We can then define

$$\boldsymbol{Z}_t^* = \underbrace{(Z_{t1},\ldots,Z_{t1}}_{n_1\text{times}},\underbrace{Z_{t2},\ldots,Z_{t2}}_{n_2\text{times}},\ldots,\underbrace{Z_{tJ},\ldots,Z_{tJ}}_{n_J\text{times}})'$$

and

$$\boldsymbol{Z}^* = (\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_t,\ldots,\boldsymbol{Z}_p) = (\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_i,\ldots\boldsymbol{Z}_N)'.$$

Define

$$\boldsymbol{\Sigma}_{\boldsymbol{Z}^*} = \begin{pmatrix} \sigma^2_{M_{11},S_1} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma^2_{M_{12},S_1} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_{M_{ij},S_j} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \sigma^2_{M_{n_JJ},S_J} \end{pmatrix},$$

let $d_{ij} = x^*_{ij} - \mu_{M_{ij},S_j}$ and denote by $\boldsymbol{D}$ the correspondent $N \times 1$ vector. Now the distribution $p\left(\boldsymbol{Z}^*|\boldsymbol{M},\boldsymbol{S},\boldsymbol{\mu},\boldsymbol{\sigma^2},\boldsymbol{\beta}\right) = \prod_{ij}\frac{1}{\sqrt{2\pi\sigma^2_{M_{ij},S_j}}}\exp\left(-\frac{\left(y^*_{ij}-\mu_{M_{ij},S_j}-\boldsymbol{Z^j\beta}\right)^2}{2\sigma^2_{M_{ij},S_j}}\right)$ can be rewritten as $\left(\prod_{ij}\frac{1}{\sqrt{2\pi\sigma^2_{M_{ij},S_j}}}\right)\exp\left(-0.5\left(\boldsymbol{D}-\boldsymbol{Z}^*\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{Z}^*}\left(\boldsymbol{D}-\boldsymbol{Z}^*\boldsymbol{\beta}\right)\right)$

8. Finally, the full conditional for $\boldsymbol{\beta}$ is

$$\begin{aligned} p\left(\boldsymbol{\beta}|\cdots\right) &\propto p\left(\boldsymbol{Z}^*|\boldsymbol{M},\boldsymbol{S},\boldsymbol{\mu},\boldsymbol{\sigma^2},\boldsymbol{\beta}\right)p\left(\boldsymbol{\beta}\right) \\ &\propto c_1 exp\left(-0.5\left(\boldsymbol{D}-\boldsymbol{Z}^*\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{Z}^*}\left(\boldsymbol{D}-\boldsymbol{Z}^*\boldsymbol{\beta}\right)\right) \\ &\times c_2\exp\left(-0.5\left(\boldsymbol{\beta}-\boldsymbol{m}_\beta\right)'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}-\boldsymbol{m}_\beta\right)\right) \\ &\propto \exp\left(-0.5\left(\boldsymbol{\beta'Z^{*\prime}\Sigma^{-1}_{Z^*}Z^*\beta} - 2D'\boldsymbol{\Sigma^{-1}_{Z^*}Z^*\beta}\right)\right) \\ &\times \exp\left(-0.5\left(\boldsymbol{\beta'\Sigma^{-1}_{\beta}\beta} - 2\boldsymbol{m'_\beta\Sigma^{-1}_{\beta}\beta}\right)\right) \\ &\propto \exp\left(-0.5\left(\boldsymbol{\beta'}\underbrace{\left(\boldsymbol{Z^{*\prime}\Sigma^{-1}_{Z^*}Z^*}+\boldsymbol{\Sigma^{-1}_{\beta}}\right)}_{\boldsymbol{V^*}}\boldsymbol{\beta} - 2\underbrace{\left(\boldsymbol{m'_\beta\Sigma^{-1}_{\beta}} + D'\boldsymbol{\Sigma^{-1}_{Z^*}Z^*}\right)}_{\boldsymbol{d^*}}\boldsymbol{\beta}\right)\right) \end{aligned}$$

We now employ the ellipsoidal rectification

$$\boldsymbol{u'Au - 2a'u} = \left(\boldsymbol{u}-\boldsymbol{A^{-1}a}\right)'A\left(\boldsymbol{u}-\boldsymbol{A^{-1}a}\right) - \boldsymbol{a'A^{-1}a},$$

obtaining

$$\begin{aligned} p\left(\boldsymbol{\beta}|\cdots\right) &\propto \exp\left(-0.5\left(\boldsymbol{\beta}-\boldsymbol{V^{*-1}d^{*}}\right)'\boldsymbol{V^{*}}\left(\boldsymbol{\beta}-\boldsymbol{V^{*-1}d^{*}}\right)-0.5\boldsymbol{d^{*'}V^{*-1}d^{*}}\right) \\ &\propto \exp\left(-0.5\left(\boldsymbol{\beta}-\boldsymbol{V^{*-1}d^{*}}\right)'\boldsymbol{V^{*}}\left(\boldsymbol{\beta}-\boldsymbol{V^{*-1}d^{*}}\right)\right). \end{aligned}$$

In the last expression we recognize a multivariate normal kernel, so we can write

$$\boldsymbol{\beta}|\cdots \sim MVN\left(\boldsymbol{V^{*-1}d^{*}},\boldsymbol{V^{*-1}}\right)$$

Chapter 2

Bayesian Nonparametric Analysis of Nested Data via Common Atom priors

"Chiunque crede che tutti i frutti maturino contemporaneamente come le fragole, non sa nulla dell'uva."
Paracelso

"Don't mistake motion for progress."
A FB poster hanging inside the Squeeze In, a breakfast place in Menlo park

Abstract

The use of high-dimensional data for targeted therapeutic interventions requires new ways to characterize the heterogeneity observed across subgroups of a specific population. In particular, models for partially exchangeable data are needed for inference on nested datasets, where the observations are assumed to be organized in different units and some sharing of information is required to learn distinctive features of the units. In this manuscript, we propose a nested Common Atoms Model (CAM) that is particularly suited for the analysis of nested data sets where the distributions of the units are expected to differ only over a small fraction of the observations sampled in that unit. The proposed CAM allows a two-layered clustering at the distributional and observational level and is amenable to scalable posterior inference through the use of a computationally efficient nested slice-sampler algorithm. We further discuss how to extend the proposed modeling framework to handle discrete measurements, and discuss posterior inference on a real microbiome dataset from young children in Mali, to investigate how large-scale alterations in intestinal microbiota composition can be associated with diarrhea onset. We further study the performance of our model in capturing true distributional structures in the population by means of a simulation study.

2.1 Introduction

The use of high-dimensional data for targeted therapeutic interventions requires new ways to characterize the heterogeneity observed across subgroups of a specific population. In particular, models for partially exchangeable data are needed for inference on nested datasets, where the observations are assumed to be organized in different units and some sharing of information is required to learn distinctive features of the units. For example, in the application to a microbiome dataset in Section 2.5, we have available sequence count measurements for 212 children (units) in Mali, and the interest is to describe the different patterns of microbial diversity observed in the population of children, since those patterns could be exploited to inform future interventions.

The description of microbial diversity requires investigating the structure, concentration, and richness of microbiota in each subject and how the distributions of microbiota abundances vary across units and subgroups of units. As the groups are typically unknown, they need to be estimated from the data. A few approaches have been proposed in the literature for clustering distributional features directly. For example, Irpino and Verde (2015) have recently proposed clustering methods in symbolic statistics, by employing the Wasserstein distance on histograms treated as units. Similarly, Batagelj et al. (2015) have proposed generalized leaders and Ward's hierarchical methods to cluster modal valued symbolic data. These are exploratory tools, which extend usual multivariate clustering methods to the analysis of (empirical) probability distributions, but they do not allow for a probabilistic assessment of cluster uncertainty.

The nested Dirichlet Process (nDP, Rodríguez et al., 2008) and its extensions have been widely employed to identify distributional groups in Bayesian Nonparametric model-based approaches. For example, Rodriguez and Dunson (2014) have proposed a generalization of the nDP for functional data analysis; Graziani et al. (2015) have investigated how the distribution of a targeted biomarker changes due to treatment and whether it is associated with a clinical outcome; and Zuanetti et al., 2018 have discussed a marginal nDP for clustering genes related to DNA mismatch repair by the distribution of gene-gene interactions with other genes. The nDP leads to a two-layered clustering: first, it allows grouping together similar distributions (distributional clustering), and then it clusters similar observations within each distributional cluster (observational clustering).

However, Camerlenghi et al. (2019b) have recently proved that the inference obtained using the nDP may be affected by a *degeneracy* property: if two distributions share even only one atom in their support, the two distributions are automatically assigned to the same cluster. More precisely, the partially exchangeable partition probability function (pEPPF), i.e. the function which describes the probability of each clustering allocation for partially exchangeable data modeled with a nDP, collapses to a fully exchangeable case when ties are present among the observational atoms. To overcome this drawback, Camerlenghi et al. (2019b) propose a class of latent nested processes, which relies on estimating a latent mixture of shared and idiosyncratic processes across the subgroups. However, the computational complexity of the resulting sampling scheme limits the applicability of the model to relatively small data sets.

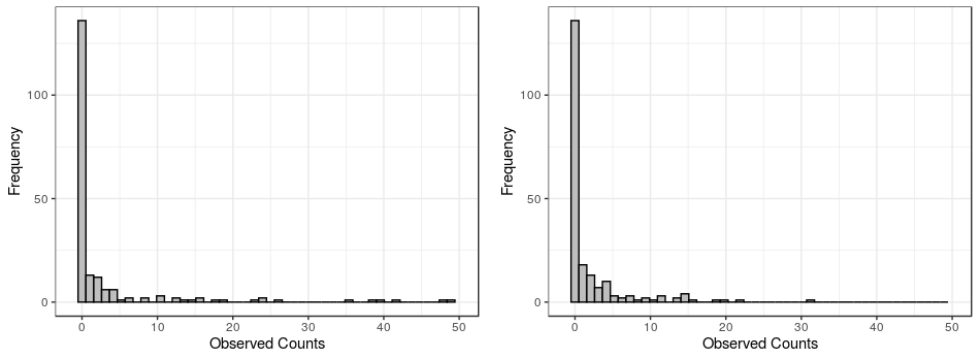
The degeneracy of the nDP is particularly problematic when analyzing high-dimensional data, such as those commonly encountered in genomics and microbiome studies. Here, the distribution profiles of sequencing data are expected to be quite similar across individuals, and to vary only for a small fraction of differentially abundant sequences, which directly intervene to regulate the biological processes and their dysfunctions. Thus, the distributions of genomic sequences across two population-subgroups are expected to be highly overlapping and correlated. In those applications, the nDP may provide unreliable inferences when comparing distributional patterns across individuals.

In this manuscript, we propose a nested Common Atoms Model (CAM) that is particularly suited for the analysis of nested data sets, where the distributions of the units are expected to differ

only over a small fraction of the observations. Although our proposal could be described as a constrained modification of the nDP, where atoms are allowed to be shared across all subgroups, the CAM does not suffer from potential degeneracy issues. One of the consequences of the nDP degeneracy is that unit-level measurements can be clustered together only within units that are assigned to the same group. The within-group clustering still contributes to a compact representation of the data, but unit-level inference across subgroups is precluded. Instead, the proposed CAM framework naturally allows unit-level inference and clustering of observations across groups, since the common atoms structure allows mapping group-specific distributional patterns to a shared support. With respect to the alternative proposal of Camerlenghi et al. (2019b), our proposal is computationally more efficient and it allows to conduct inference on a larger number of observations and population subgroups. We further develop a novel nested slice sampler algorithm (Kalli et al., 2011), which allows sampling directly from the true posterior distribution, without employing the standard truncation-based approximation, which is typically used for posterior inference with nDP models.

We apply the proposed modeling framework to the analysis of a microbiome dataset. Here, a primary goal is to study *microbial diversity*, i.e. how the distribution of microbial units varies across subgroups of a population. A number of diseases have been associated with decreased microbiome diversity (Morgan and Huttenhower, 2012). Typically, summary statistics are used to capture characteristics of species' distributions, e.g. α-diversity and β-diversity metrics such as the Shannon's entropy and Bray-Curtis dissimilarity indexes, respectively (Whittaker, 2006). However, those metrics do not fully capture the complexity of microbiome data, which poses distinctive statistical challenges (Mao et al., 2017). In particular, the data are recorded as counts of the observed microbial genome sequences. The resulting histograms are highly skewed and sparse, due to the many low- or zero- frequency counts and to the presence of a few dominant sequences. Figure 2.1 reports a snapshot of the observed microbial distributions for two representative individuals from the dataset we analyze in Section 2.5. The two subjects have been assigned with high probability to two different population subgroups by the proposed CAM model. In addition to the typical microbiome distributional features discussed above, we note that the two distributions share many common atoms, and they are quite similar except for the presence of a small set of sequences that appear with high frequency. In the microbiome literature, ad-hoc solutions are sometimes adopted to address the challenges put forward by the analysis of microbiome data. For example, when dealing with the excess of zero counts, some authors simply add a small number (e.g. 1) to each count, thus generating "pseudo counts". Here, we propose to embed the proposed CAM framework within a rounded mixture of Gaussians (RGM) model (Canale and Dunson, 2011). In this way, we effortlessly obtain a BNP Nested model for count data that can naturally handle the sparsity and the zero-inflation typical of microbiome abundance tables. The resulting discrete CAM model allows to cluster rows of an abundance table according to their distributional characteristics, providing a partition of patients with similar microbiome distribution.

The remainder of the article is as follows. In Section 2 we introduce our model for continuous measurements, and discuss its properties. In Section 3 we discuss how to adapt the model to count data. In Section 4, we discuss posterior inference and outline the nested version of the slice sampler. Section 5 applies our model to a publicly available microbiome dataset, which contains the OTU counts observed in a sample of children from Mali. Section 6 presents a simulation study to assess the clustering behavior of the model as the number of observations and groups grow in different scenarios. Section 7 summarizes our contributions and discusses some future directions of research.



(A) Subject 1 (B) Subject 2

FIGURE 2.1: Histograms of the microbiome populations of two Mali children. As we can see, the distributions appear very similar and extremely skewed.

2.2 Common Atoms Model for Continuous Measurements

We consider a *nested* dataset, where we have available continuous measurements $\boldsymbol{y}_j = (y_{1,j}, \dots, y_{n_j,j})$ observed over J experimental units. We assume that each observation $y_{i,j}$, $i = 1, \dots, n_j$ and $j = 1, \dots, J$, takes values in a suitable Polish space $\mathbb{X}$ endowed with the respective Borel σ-field $\mathcal{X}$. Similarly as in the nDP (Rodríguez et al., 2008), our goal is to achieve a partition of the vectors $\boldsymbol{y}_1, \dots, \boldsymbol{y}_J$ into few, say $L \leq J$, distributional clusters. However, Camerlenghi et al. (2019b) have shown that the partially exchangeable partition probability function (pEFFP) of the nDP implies that distributions collapse into a common cluster when they share even only one atom. This unappealing behavior can be avoided if the prior explicitly models commonality of atoms between groups. Here, we propose a common atoms model (CAM) such that distributions belonging to different clusters are characterized by specific weights assigned to a common set of atoms. In this section, we define the model and investigate its properties for analyzing high-dimensional data. More specifically, let G_j, $j = 1, \dots, J$ denote the distribution of the j-th experimental unit,

$$y_{i,1}, \dots, y_{i,J} | \, G_1, \dots, G_J \overset{ind.}{\sim} G_1 \times \cdots \times G_J \qquad i = 1, \dots, n_j. \tag{2.1}$$

Then, similarly as in the nested DP formulation, we assume that the G_j's are a sample from an almost surely discrete distribution Q over the space of probability distributions on $\mathcal{X}$, namely

$$G_1, \dots, G_J | Q \overset{i.i.d.}{\sim} Q, \qquad Q = \sum_{k \geq 1} \pi_k \, \delta_{G_k^*}. \tag{2.2}$$

where $G_k^* = \sum_{l \geq 1}, \omega_{lk} \, \delta_{\theta_l}$, $k \geq 1$, and the **common atoms** $\theta_1, \theta_2, \dots$ are drawn from a non-atomic, shared base measure H on $(\mathbb{X}, \mathcal{X})$. We further assume the Griffiths-Engen-McCloskey (GEM) distribution for the weights, which characterizes the stick-breaking (or Sethuraman's) construction of the Dirichlet process (Sethuraman, 1994), i.e. we consider $V_k \sim Beta(1, \alpha)$, $k \geq 1$, and then set $\pi_1 = V_1$, and $\pi_k = V_k \prod_{r=1}^{k-1}(1 - V_r)$, $k > 1$.We indicate this construction writing $\boldsymbol{\pi} = \{\pi_k\}_{k \geq 1} \sim GEM(\alpha)$. Similarly, $\boldsymbol{\omega}_k = \{\omega_{lk}\}_{l \geq 1} \sim GEM(\beta)$ for all $k \geq 1$.

Due to the commonality of the atoms at the unit level, our construction is reminiscent of the Hierarchical Dirichlet process (HDP) of Teh et al. (2006). However, there are crucial differences between the two constructions, and - to the best of our knowledge - the common atoms structure we propose has not been previously investigated in the literature. More specifically, the

HDP does allow a flexible representation of each unit-level distribution G_j, but does not induce distributional clusters among the units. Our formulation preserves a two-layered clustering structure, across units (first layer) and between observations within each unit (second layer). Thus, with reference to this structure, the proposed CAM model is close in spirit to recently developed hierarchical topic models (Paisley et al., 2015; Tekumalla et al., 2015), where a HDP is adopted as a base measure of an (outer) DP, in symbols $Q \sim DP(\alpha, HDP(\beta, H))$. However, these nested HDP formulations aim at describing topic distributions which can be obtained as mixtures of separate topics (i.e. a document may contain words typical of both medicine and sport news), whereas our objective is to cluster individual distributions and the observations therein (a patient-specific distribution is not obtained as a mixture of other patients' distributions). In this regard, our proposal mimics the intended purpose of the original nDP model. We also notice that the prior distribution on Q in (2.2) can be seen as a common atom Dependent Dirichlet process. A similar structure is used in Hatjispyros et al. (2016), where pairwise dependence between m random density functions is induced modeling each of them with a mixture of DPs characterized by common atoms.

2.2.1 Properties of the Common Atoms model

In the following, we investigate the properties of the CAM model. More specifically, we show how the model does not suffer from the theoretical degeneracy of the nDP and, consequently, from the implied dependence between pairs of observations and distributions.

Combinatorial structure. The partitions defined by the model in (2.1)–(2.2) can be described via the so–called partially Exchangeable Partition Probability Function (pEPPF). For notational simplicity, we illustrate the main results by focusing on $J = 2$, but the results easily extend to the general case. We further assume that there are $s > 0$ distinct measurements out of a sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J$, which will be denoted by $y_1^*, \ldots, y_s^*$, with frequencies $\boldsymbol{n}_j = (n_{1,j}, \cdots, n_{s,j})$, and with $n_{i,j}$ indicating the number of times the i-th distinct value y_i^* has been observed in the initial sample from population j, i.e. the absolute frequencies of the distinct values. We indicate by $\mathsf{P}_{\mathbb{X}}$ the space of all random probability measures on $\mathbb{X}$. We first derive a result which is similar to Camerlenghi et al. (2019b, Proposition 2), and characterizes the mixed moments of the random probability measures G_1 and G_2 as a convex combination of fully exchangeable and independent samples. The proof is given in the Appendix.

Proposition 1. *Let f_1 and f_2 be two measurable functions defined on $\mathsf{P}_{\mathbb{X}}$ and taking values in $\mathbb{R}^+$, then*

$$\mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1) f_2(g_2) Q(dg_1) Q(dg_2)\right] = q_1 \mathbb{E}[f_1(G_1^*) f_2(G_1^*)] + (1 - q_1)\mathbb{E}[f_1(G_1^*) f_2(G_2^*)] \tag{2.3}$$

where we have set $q_1 := \mathbb{P}(G_1 = G_2)$.

Following Camerlenghi et al. (2019a), we formally define the pEPPF as the probability of the observed allocation $\{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J\}$ of $s > 0$ distinct observations through the hierarchical structure,

$$\Pi_n^{(s)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) = \mathbb{E}\int_{\mathbb{X}^s} \prod_{j=1}^{J} \prod_{i=1}^{s} G_j^{n_{i,j}}(dy_i^*), \tag{2.4}$$

with $n = \sum_{j=1}^{J} n_j$. The i-th distinct value is shared by any two experimental units j and κ if and only if $n_{i,j}\, n_{i,\kappa} \geq 1$. If $J = 1$ one obtains the usual exchangeable partition probability function (EPPF) for an individual sample, defined by (Pitman, 1995), and denoted here as $\Phi_{n_j}^{(s)}(\boldsymbol{n}_j)$. In the case of the Dirichlet process, this coincides with the well–known Ewens's sampling formula,

$\Phi^{(s)}_{n_j}(\boldsymbol{n_j}) = \frac{\alpha^s\Gamma(\alpha)}{\Gamma(\alpha+n_j)}\prod_{i=1}^{s}(n_{i,j}-1)!$ (Ewens, 1972). The pEPPF for the CAM model is described by the following theorem, for the case $J=2$.

Theorem 1. *Let $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ be samples from $J=2$ experimental units under the CAM model (2.1)–(2.2). Then, the induced random partition of $s>0$ distinct observations may be expressed as*

$$\Pi^{(s)}_n(\boldsymbol{n}_1,\boldsymbol{n}_2) = q_1\Phi^{(s)}_{n_1+n_2}(\boldsymbol{n}_1+\boldsymbol{n_2}) + (1-q_1)\int_{\mathbb{X}^s}\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G^*_j)^{n_{i,j}}(dy^*_i). \tag{2.5}$$

Although a closed form expression is not available, due to the presence of the integral over $\mathbb{X}^s$ on the right hand side, the result is important to show that the proposed CAM model does not reduce to the fully exchangeable case in the presence of common observations across the two samples. Indeed, we can prove the following

Proposition 2. *Assume that two samples $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ share $s_0>0$ distinct observations. Then, necessarily,*

$$\int_{\mathbb{X}^s}\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G^*_j)^{n_{i,j}}(dy^*_i) > 0.$$

In other words, the pEPPF (2.5) does not reduce to the EPPF of the full exchangeable model. The proof of Proposition 2 is deferred to the Appendix, where we also provide an explicit expression for the integral in (2.5) (see Equation 2.18).
Of course, ties among distributions at the outer level are still possible, since

$$\mathbb{P}\left(G_j = G_{j'}|Q\right) = \sum_{k\geq 1}\pi_k^2 > 0, \quad \text{and} \quad \mathbb{P}\left(G_j = G_{j'}\right) = \frac{1}{1+\alpha}. \tag{2.6}$$

Moreover, the probability of a tie between two data points in two separate units can be computed as

$$\mathbb{P}\left[x_{ij} = x_{i'j'}\right] = \frac{1}{1+\alpha}\frac{1}{1+\beta} + \frac{1}{1+\alpha}\frac{1}{2\beta+1}. \tag{2.7}$$

This shows that CAM induces a two-fold clustering structure: it clusters together experimental units characterized by similar distribution profiles, and it also groups together observations, allowing for borrowing information across the two layers.

Correlation between random probability measures. The covariance and correlation are useful quantities to investigate the dependence of random probability measures and are also of paramount importance in applications. For the CAM model, it can be shown that (Appendix (Section 2.A))

$$\begin{aligned} Cov\left(G_j(A), G_{j'}(B)\right) &= H(A\cap B)\left(\frac{q_1}{1+\beta} + \frac{1-q_1}{1+2\beta}\right) \\ &\quad + H(A)H(B)\left(-\frac{q_1}{1+\beta} - \frac{1-q_1}{1+2\beta}\right), \\ \rho_{ij} := Corr(G_j(A), G_{j'}(A)) &= 1 - \frac{\beta}{2\beta+1}\frac{\alpha}{1+\alpha}. \end{aligned} \tag{2.8}$$

where $q_1 = \frac{1}{1+\alpha}$. It is interesting to note that $\rho_{ij} \in \left(\frac{1}{2}, 1\right)$, due to the commonality of the atoms. In many applications, especially in genomics, distribution profiles are expected to be quite similar across experimental units (e.g., subjects), and to vary only for a small fraction of the observations (e.g., genes). For the nDP, $Corr\left(G_j(A), G_{j'}(A)\right) = \frac{1}{1+\alpha} > 0$, where the expression does not depend on β: this is because the nDP assumes independence between atoms in separate distributions.

2.2.2 Common Atoms Mixture Model

The CAM model defined above assumes a.s. discrete distributions. For modeling continuous distributions, one could follow established literature (Ferguson, 1983) and consider a nonparametric mixture model where (2.1) is substituted by

$$\begin{aligned} y_{i,1},\dots,y_{i,J}|\, f_1,\dots,f_J &\overset{ind.}{\sim} f_1\times\cdots\times f_J \qquad i=1,\dots,n_j,\\ f_j(\cdot) &= \int_\Theta p(\cdot|\theta)\, G_j(d\theta), \quad j=1,\dots,J,\\ G_j|Q &\overset{i.i.d.}{\sim} Q, \end{aligned} \tag{2.9}$$

where $p(\cdot|\theta)$ denotes an appropriate parametric continuous kernel density, and $G_j|Q \overset{i.i.d.}{\sim} Q$ as in (2.2). In the rest of the paper, we will adopt Gaussian kernels, i.e. we assume $p(\cdot|\theta)$ to be Normal and $\theta=(\mu,\sigma^2)$ is a vector of location and scale parameters.
To simplify the computational algorithm, we introduce an alternative representation using two sequences of latent variables, $\{S_j\}_{j\geq 1}$ and $\{M_{ij}\}_{i\geq 1,j\geq 1}$, describing – respectively – the clustering process at the distributional level and at the observational level, i.e. $S_j=k$ and $M_{ij}=l$ if the observation i in unit j is assigned to the l-th observational cluster and the k-th distributional cluster:

$$\begin{aligned} y_{ij}|\boldsymbol{M},\boldsymbol{\theta} &\sim N\left(\cdot|\theta_{M_{ij}}\right), & M_{ij}|\boldsymbol{S},\boldsymbol{\omega} &\sim \sum_{l=1}^{\infty}\omega_{l,S_j}\delta_l(\cdot),\\ \boldsymbol{\omega}_k|\boldsymbol{S}=\boldsymbol{\omega}_k &\sim GEM(\alpha), & S_j|\boldsymbol{\pi} &\sim \sum_{k=1}^{\infty}\pi_k\delta_k(\cdot),\\ \boldsymbol{\pi} &\sim GEM(\beta), & \theta_l &\sim \pi(\theta_l). \end{aligned} \tag{2.10}$$

In the following, we assume $\theta_l=(\mu_l,\sigma_l^2)\sim NIG(m_0,\kappa_0,\alpha_0,\beta_0)$, i.e. $\mu_l|\sigma_l^2\sim N\left(m_0,\sigma_l^2/\kappa_0\right)$ and $\sigma_l^2\sim IG\left(\alpha_0,\beta_0\right)$.

2.3 Common Atoms Model for Count Data

In Section 2.5, we consider an application to microbiome data, which can be represented by abundance tables, containing the observed frequency of a particular microbial sequence is in a sample - or subject (unit). In this Section, we describe how the CAM model can be adapted to take into account count data, characterized by skewness and zero-inflation as typically observed in microbiome studies. Let $z_{ij}\in\mathbb{N}$ be the observed count of microbial sequence $i=1,\dots,n$ in subject $j=1,\dots,J$. Consequently, the vector $\boldsymbol{z}_j=(z_{1j},\dots,z_{nj})$ will denote the observed microbiome of individual j. Here, we embed model 2.1–2.2 in the rounded mixture of Gaussians framework of Canale and Dunson (2011). See also Bandyopadhyay and Canale (2016) and Canale and Prünster (2017), where the rounded mixture framework is compared to less flexible nonparametric mixtures of Poisson densities for count data. In order to define a probability mass function for the discrete measurements z, Canale and Dunson (2011) consider a data augmentation framework by latent continuous variables y, such that

$$p_Z(Z=j)=\int_{a_j}^{a_{j+1}} f(y)\,dy, \quad j\in\mathbb{N}$$

for a sequence of thresholds $a_0<a_1<a_2<\ldots<a_\infty$ and for some density function $f(\cdot)$, such that $\int_{a_0}^{a_+\infty} f(y)\,dy=1$. Typically, the sequence of thresholds is set as $\mathbf{a}=\{a_j\}_{j=0}^{+\infty}=\{-\infty,0,1,2,\dots,+\infty\}$ and $f(\cdot)$ is a Dirichlet Process mixture density, to ensure a flexible representation of the table of counts. Here, we propose a novel nested formulation, where f is

modeled as a CAM mixture eq. (2.10) . More specifically, we consider

$$z_{ij}|y_{ij} \sim \sum_{g=0}^{+\infty} \delta_g(\cdot)\mathbf{1}_{[a_g,a_{g+1})}(y_{ij}), \tag{2.11}$$

where y_{ij} is distributed as in (2.10). Thus,

$$\begin{aligned}\pi\left(z|\boldsymbol{M},\boldsymbol{\mu},\boldsymbol{\sigma^2}\right) &= \int_{a_z}^{a_{z+1}} \phi\left(y;\mu_{M_{ij}},\sigma^2_{M_{ij}}\right)dy \\ &= \Phi\left(a_{z+1};\mu_{M_{ij}},\sigma^2_{M_{ij}}\right) - \Phi\left(a_z;\mu_{M_{ij}},\sigma^2_{M_{ij}}\right) \\ &:= \Delta\Phi\left(a_z;\mu_{M_{ij}},\sigma^2_{M_{ij}}\right).\end{aligned}$$

where ϕ and Φ denote the p.d.f. and the c.d.f. of a Gaussian r.v., respectively.
We will refer to this new setting as the Discrete Common Atom Model (DCAM).

2.4 Posterior Inference

Typically, posterior samples for the nDP have been obtained using a truncated version of the Blocked-Gibbs Sampler (Ishwaran and James, 2001), i.e. by choosing proper upper bounds for the infinite-sums. The model representation in eq. (2.10) is useful to obtain an algorithm, which we detail in Appendix B, where we also provide upper bounds for the resulting truncation error. Here, instead, we present a novel nested version of the independent slice-efficient algorithm (Walker, 2007; Kalli et al., 2011). With respect to truncation-based algorithms, the proposed slice sampler has two main advantages: it allows to sample from the true posterior distribution and it considerably decreases the computational time. To the best of our knowledge, no slice sampler has been proposed for nested-type models. The proposed slice sampling scheme can be easily extended to the nDP, and is related to the sampling scheme in Banerjee et al. (2013), although their model is substantially different from ours. In the following, we focus on the Common Atoms Mixture model (2.9), as variations of the algorithm to accommodate count data are straightforward.
Let $f(y_{ij}|\theta)$ denote a generic likelihood function for the observation y_{ij}, let $\boldsymbol{\pi}$ and $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots)$ be the two sets of weights, one referred to the distributional clusters, the other referred to the observational clusters. Then, we can write:

$$p(y_{ij}|\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\pi}) = \sum_{k\geq 1}\pi_k \sum_{l\geq 1}\omega_{lk} f(y_{ij}|\theta_l).$$

We augment the model introducing two sets of latent variables controlling which components of the mixture are "active" and which can be ignored. More specifically, we introduce $\boldsymbol{u^O} = \left(u_1^O,\ldots,u_J^O\right)$ – where the O in the superscript stands for the "outer"– , i.e. the distributional, model – and, within every unit $j = 1,\ldots,J$, we define "inner" sets of latent variables, $\boldsymbol{u_j^I} = \left(u_{1j}^I,\ldots,u_{n_jj}^I\right)$. Moreover, following Kalli et al. (2011) , we also consider the following deterministic sequences: $\boldsymbol{\xi}^O = \left(\xi_1^O,\xi_2^O,\xi_3^O,\ldots,\xi_k^O,\ldots\right)$ and, for every k, $\boldsymbol{\xi_k^I} = \left(\xi_{1k}^I,\xi_{2k}^I,\xi_{3k}^I,\ldots,\xi_{lk}^I,\ldots\right)$. Then the model can be rewritten as

$$p_{\boldsymbol{\xi}^O,\boldsymbol{\xi}^I}\left(Y_{ij},u_j^O,u_{ij}^I|\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\pi}\right) = \sum_{k\geq 1}\mathbb{1}_{\{u_j^O<\xi_k^O\}}\frac{\pi_k}{\xi_k^O}\sum_{l\geq 1}\mathbb{1}_{\{u_{ij}^I<\xi_{lk}^I\}}\frac{\omega_{lk}}{\xi_{lk}^I}f(Y_{ij}|\theta_l). \tag{2.12}$$

Notice that if we assume $\xi_k^O = \pi_k$ and $\xi_{lk}^I = \omega_{lk}$, we recover the nested version of the efficient-dependent slice sampler, as presented in (Kalli et al., 2011; Papaspiliopoulos and Roberts, 2008) Introducing two sets of latent labels that identify the distributional ($\boldsymbol{S}$) and observational ($\boldsymbol{M}$) cluster in which the observation is allocated, allows us to get rid of the infinite sums in the previous equations. The distribution for a single observation becomes

$$p_{\boldsymbol{\xi}^O,\boldsymbol{\xi}^I}\left(Y_{ij}, u_j^O, u_{ij}^I, M_{ij}, S_j|\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\pi}\right) = \mathbb{1}_{\{u_j^O<\xi_{S_j}^O\}}\frac{\pi_{S_j}}{\xi_{S_j}^O}\mathbb{1}_{\{u_{ij}^I<\xi_{M_{ij}S_j}^I\}}\frac{\omega_{M_{ij}S_j}}{\xi_{M_{ij}S_j}^I}f\left(Y_{ij}|\theta_{M_{ij}}\right), \quad (2.13)$$

while the entire sample is modeled by:

$$p_{\boldsymbol{\xi}^O,\boldsymbol{\xi}^I}\left(\boldsymbol{Y},\boldsymbol{u^O},\boldsymbol{u^I},\boldsymbol{M},\boldsymbol{S}|\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\pi}\right) = \prod_{j=1}^{J}\mathbb{1}_{\{u_j^O<\xi_{S_j}^O\}}\frac{\pi_{S_j}}{\xi_{S_j}^O}\prod_{i=1}^{n_j}\mathbb{1}_{\{u_{ij}^I<\xi_{M_{ij}S_j}^I\}}\frac{\omega_{M_{ij}S_j}}{\xi_{M_{ij}S_j}^I}f\left(Y_{ij}|\theta_{M_{ij}}\right). \quad (2.14)$$

If we assume $f\left(Y_{ij}|\theta_l\right) = N(Y_{ij}|\theta_l)$ we recover the CAM model listed in Equation (2.9). Otherwise, if we postulated the existence of a latent variable as in Equation (2.11), we recover the DCAM setting as in $f\left(Y_{ij}|\theta_l\right) = \Delta\Phi\left(a_{y_{ij}};\mu_{M_{ij}},\sigma_{M_{ij}}^2\right)$.

In a general framework, the nested slice sampler is obtained by looping over the following full conditionals:

1. Sample each u_j^O from a continuous uniform distribution $U\left(0,\xi_{S_j}^O\right)$.
2. Sample each u_{ij}^I from a continuous uniform distribution $U\left(0,\xi_{M_{ij}S_j}^I\right)$.
3. The proportions $\boldsymbol{v}$ needed to compute the SB weights $\boldsymbol{\pi}$ are sampled independently: $v_k \sim Beta\left(a_k,b_k\right)$, where $a_k = 1+\sum_{j=1}^{J}\mathbb{1}_{\{S_j=k\}}$ and $b_k = \alpha+\sum_{j=1}^{J}\mathbb{1}_{\{S_j>k\}}$. This full conditional is recovered integrating $\boldsymbol{u^O}$ out.
4. For each k, the proportions $\boldsymbol{u_k}$ needed to compute the SB weights $\boldsymbol{\omega_k}$ are sampled independently: $u_{lk} \sim Beta\left(a_l^k,b_l^k\right)$, where $a_l^k = 1+\sum_{i=1}^{N}\mathbb{1}_{\{M_{ij}=l,S_j=k\}}$ and $b_l^k = \beta+\sum_{i=1}^{N}\mathbb{1}_{\{M_{ij}>l,S_j=k\}}$. This full conditional is recovered collapsing both $\boldsymbol{u^O}$ and $\boldsymbol{u^I}$.
5. Sample the distributional labels from the following full conditional distribution:

$$p\left(S_j=k|\cdots\right) \propto \mathbb{1}_{\{u_j^O<\xi_k^O\}}\frac{\pi_k}{\xi_k^O}\prod_{i=1}^{n_j}\mathbb{1}_{\{u_{ij}^I<\xi_{M_{ij}k}^I\}}\frac{\omega_{M_{ij}k}}{\xi_{M_{ij}k}^I}.$$

   Following Banerjee et al. (2013) and Porteous et al. (2006), we can obtain more efficient updates trough partial collapsing integrating over the inner level slice variables:

$$p\left(S_j=k|\cdots-\boldsymbol{u^I}\right) \propto \mathbb{1}_{\{u_j^O<\xi_k^O\}}\frac{\pi_k}{\xi_k^O}\prod_{i=1}^{n_j}\omega_{M_{ij}k}.$$

6. Sample the observational labels from the following full conditional distribution:

$$p\left(M_{ij}=l|\cdots\right) \propto \mathbb{1}_{\{u_{ij}^I<\xi_{lS_j}^I\}}\frac{\omega_{lS_j}}{\xi_{lS_j}^I}f\left(Y_{ij}|\theta_l\right)$$

7. Sample θ_l from its full conditional, which is a conjugate NIG.

For the DCAM, an additional step (Step 1 of the algorithm in Appendix 1.A) is added in order to update the latent variable. At each step, to make the actual computation of steps 5, 6 and 7

feasible, we need to stochastically truncate the number of mixture components to a sufficiently high integer to ensure that the two steps can be carried out exactly. At each iteration we sample among K^* possible distributional cluster labels and $L^{**} = \max\{L_1^*, \dots, L_{K^*}^*\}$ possible observational labels. If $\xi_k^O = \pi_k$ and $\xi_{lk}^I = \omega_{lk}$, the values are the lowest integers that ensure, respectively, that

$$\sum_{k=1}^{K^*} \pi_k \geq 1 - \min_{j=1,\dots,J} u_j^O \quad \text{and} \quad \sum_{l=1}^{L_k^*} \omega_{lk} \geq 1 - \min_{i=1,\dots,n_j} u_{ij}^I \quad \forall k = 1, \dots, K^*. \tag{2.15}$$

Instead of relying on the efficient-dependent version, according to Kalli et al. (2011) and Hong and Martin (2016), we adopt the following geometric deterministic sequences. In this case, it is sufficient to focus only on one Inner deterministic sequence, say $\boldsymbol{\xi}^I$, being $\boldsymbol{\xi}_k^I$ the same for every k.

$$\xi_k^O = (1 - \kappa_O)\, \kappa_O^{k-1}, \qquad \xi_{lk}^I = \xi_l^I = (1 - \kappa_I)\, \kappa_I^{l-1}.$$

In this case, denoting with $u_{min}^O = \min_j u_j^O$ and $u_{min}^I = \min_{i,j} u_{ij}^I$ we can compute the two thresholds at each MCMC sweep:

$$K^* = \left\lfloor \frac{\log\left(u_{min}^O\right) - \log\left(1 - \kappa_O\right)}{\log\left(\kappa_O\right)} \right\rfloor, \qquad L^* = \left\lfloor \frac{\log\left(u_{min}^I\right) - \log\left(1 - \kappa_I\right)}{\log\left(\kappa_I\right)} \right\rfloor.$$

Again, in case the precision parameters α and β of the two DPs are assumed stochastic, distributed as $Gamma\left(a_\alpha, b_\alpha\right)$ and $Gamma\left(a_\beta, b_\beta\right)$, the full conditionals distributions can be sampled following a neat procedure, as suggested in Walker (2007) and Escobar and West (1995): denote with c^* the number of unique values sampled and with n the number of observations ($n = J$ when the Outer DP is considered, otherwise $n = \sum_{j=1}^J n_j$). Then the precision parameter of the DP γ ($\gamma = \alpha$ when Outer DP, $\gamma = \beta$ otherwise), for both the DPs, can be sampled in two stage, introducing another latent variable η: (a) sample $\eta|\gamma, c^* \sim Beta\left(\gamma + 1, n\right)$ and (b) sample a new γ from the mixture

$$\gamma \sim \pi_\eta G(a + k, b - \log(\eta)) + (1 - \pi_\eta)\, G\left(a + k - 1, b - \log(\eta)\right),$$

where $\pi_\eta = \pi_\eta / \left(1 - \pi_\eta\right) = (a + k - 1)/\{n(b - \log(\eta)\}$.
The exploration of the space of cluster memberships (labels) is a delicate task. Differently from the marginal specification, where simulation methods are devised in a way that the resulting Markov Chain explores the space of the partitions as equivalence classes over cluster values, a conditional/stick-breaking specification operates on the space of the explicit cluster labels (Porteous et al., 2006). In this second scenario, it could easily happen that the chain exploring the cluster membership shows poor mixing, being stuck in one of the local maxima of the posterior. This happens more frequently when a quite large amount of data is available. To overcome this issue, the label switching moves described in Papaspiliopoulos and Roberts (2008) and Hastie et al. (2015) can be added to our setup to improve the mixing.

2.5 Analysis of microbial distributions of infants in low-income countries

The increased availability of high-throughput sequencing techniques has allowed researchers to investigate the impact of the human microbiome and its diversity on our health with increasing detail (see, e.g., Dinan and Cryan, 2013; Ley, 2010; Goodman and Gardner, 2018; Kau et al., 2011). A taxonomical classification of microbial species is typically conducted based on sequence alignments, e.g. through the use of 16S rRNA sequences. More specifically, “practically

identical" sequenced tags (95%, 97% or 99% of degree of similarity) are clustered together into the same *phylotype*, and referred to as an *operational taxonomic unit* (OTU). Thus, for each specimen (e.g. fecal sample) obtained from a particular ecosystem (e.g. the gut), the number of recurrences of each OTU is recorded (Jovel et al., 2016; Kaul et al., 2017). Collecting samples from distinct individuals leads to the construction of an *abundance table*, a matrix formed by the OTU counts (taxa) observed in each sample.

Let $\boldsymbol{Z}$ indicate a $J \times n$ abundance table where each entry $z_{ij} \in \mathbb{N}$ is the frequency of the i-th OTU observed in the j-th subject, $i = 1, \ldots, N$, $j = 1, \ldots, N$. Thus, the vector $\boldsymbol{z}_j = (z_{ij})_{i=1\ldots,n_j} = \left(z_{1j}, \ldots, z_{n_j}\right)$ denotes the observed microbiome sample of individual j, $j = 1, \ldots, J$. Due to the sampling mechanism and the heterogeneity of the microbiome in the population, the observed distribution of the OTU counts is typically skewed and over-dispersed: very few important microbes show a very high frequency, while a vast number of OTUs have been recorded just a few times or have never been observed at all (see Figure 2.1). Indeed, when compared across subjects, microbiota abundance data show a characteristic zero-inflation. Kaul et al. (2017) identifies three possible types of zero values in microbiome samples: *structural* zeros, which record truly absent OTUs; *sampling* zeros, which are due to the sampling depth of the sequencing technique; and *outliers*, which are due to extraneous reasons, independent of the sequencing depth. As a result, only a few major bacterial taxa of the microbiota are shared across samples and the remaining bacteria are detected only in a small percentage of the samples.

In order to understand the varying composition of the microbiome in the population, we apply the DCAM model proposed in Section 2.3 to a publicly available dataset from the study of Pop et al. (2014), which contains the OTU counts of young children from low-income countries. The goal of the study was to understand how large-scale alterations in intestinal microbiota composition can be associated with diarrhea onset. More specifically, we focus here on the 212 records of Mali children. We combine together the OTU at the species level, and subsequently we follow standard preprocessing steps in microbiome analysis, by filtering out the species that have more than 85% of zero entrances across all records in the datasets. This leaves us with 142 taxa from the 212 children, resulting in a total of 30,104 observations.

The varying sequencing depths results also affect the so-called *library size*, i.e. the total frequencies of the observed species. Let $X_j = \sum_{i=1}^{n_j} z_{ij}$ indicate the library size for subject j and let $\gamma_j = X_j / \bar{X}_j$ denote the same quantity divided by the mean of all observed library sizes (Bullard et al., 2010; Witten, 2011). We incorporate the library sizes as a regressor in the latent modeling formulation of the DCAM, i.e. we consider

$$y_{ij} | \boldsymbol{M}, \boldsymbol{\mu}, \boldsymbol{\sigma^2} \sim N\left(\gamma_j \cdot \mu_{M_{ij}}, \sigma^2_{M_{ij}}\right). \tag{2.16}$$

The mean of the latent continuous random variable, which reflects the magnitude of the count of a specific OTU, is decomposed multiplicatively in a term that captures the true intensity underlying the process and a deterministic term that describes the depth of the sequencing.

We adopt standard prior settings for all the hyperparameters $(m, \kappa, \alpha_0, \beta_0, a_\alpha, b_\alpha, a_\beta, b_\beta)$. More specifically, following an empirical Bayes rationale, we set $m = 0$ and κ equal to the inverse of the overall sample variance. According to Rodríguez et al. (2008), we then set $\beta_0 = 1$ and $\alpha_0 = a_\alpha = b_\alpha = a_\beta = b_\beta = 3$. A MCMC sample of 50,000 iterations was collected after a burn in period of the equal length. Convergence of the MCMC chain was assessed based on visual inspection and standard convergence diagnostics (Plummer et al., 2006).

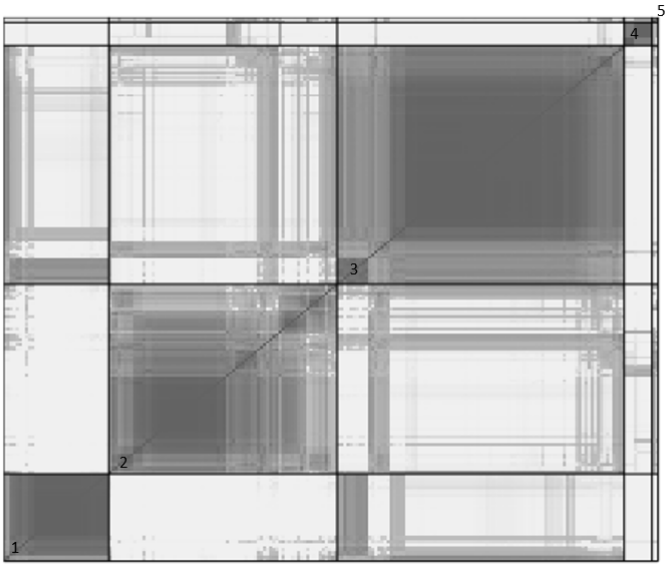
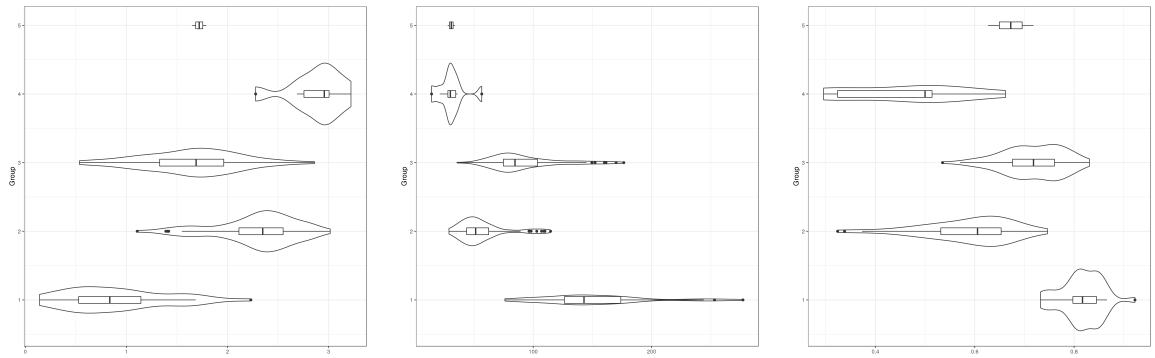


FIGURE 2.1: Pairwise Posterior Probability matrix of coclustering among the 212 subjects. A partition of the subjects' distributions into five clusters is obtained after minimization of the posterior expected Variation of information loss function.

We first start investigating the heterogeneity of the distributions of taxa abundances among the subjects. From the MCMC output, it is possible to obtain the posterior matrix of pairwise co-clusterings, which provides an initial idea of the underlying structure of the data. We can further estimate an optimal partition of the resulting distributions by considering a decision-theoretic approach and minimizing the expected posterior loss under a specific loss function specification. A widely used approach, defined directly on the space of the partitions, relies on the Binder loss function (Binder, 1978; Lau and Green, 2007). However, the Binder loss has been shown to exhibit asymmetrical behaviors, as it leads to split clusters more likely than merging them. This behavior could result in the creation of many small clusters, often containing a single observation. Thus, Wade and Ghahramani (2015) propose to rely instead on the minimization of the Variation of Information loss function developed by Meilă (2007). The Variation of Information distance compares the entropy information in two clusterings with the information shared between the two clusterings. Applying this criterion to our MCMC output, we can find five main clusters among the subjects, of different sizes, which are visualized in Figure 2.1.
The resulting clustering of the subjects' abundance distributions appears to capture relevant distributional characteristics. For example, the Shannon index (Shannon, 1948) is often used to measure the α-diversity of a microbiome community, i.e. the richness (number) and evenness (frequencies' similarity) of the different OTUs observed in a sample. More in detail, the Shannon index for an individual j is defined as $H_j = -\sum_{i=1}^{n_j} p_{ij} \ln p_{ij}$, where p_{ij} is the ratio of the abundance of taxa i over the library size of individual j. Figure 2.2 (a) shows the violin and boxplots of the α-diversity values for the five clusters identified by the optimal partition estimated using the Variation of Information criterion. The identified clusters correspond to marked differences in the α-diversity indexes and their distributions, whereby some clusters are less diverse than others. Figure 2.2 (b) shows the distribution of the within-subject means of all non-zero counts. This measure can be seen as a proxy to evaluate the richness of the distributions. Thus, clusters characterized by high means correspond to low-diversity indices, due to the presence of a few high-abundant microbe sequences. Similarly, Figure 2.2 (c) considers the percentage of non-zero counts in the species, which can also provide some information about the richness and skewness



(A) α-diversity (B) Mean of non-zero abundances (C) % of zero counts

FIGURE 2.2: Violin and boxplot combinations for three distributional summaries stratified by the partition clusters estimated using the Variation of Information criterion: a) α-diversity as measured by the Shannon entropy index; b) the mean of the non-zero entries in the abundance table for each subject; c) the percentage of OTU sequences with zero counts.

of the observed sample counts.

Diarrhea Onset $\sim$ *Cluster Allocation*					*Age* $\sim$ *Cluster Allocation*				
	50%	2.5%	97.5%	$\mathbb{P}(\beta > 0 \vert \text{data}) > 0$		**50%**	2.5%	97.5%	$\mathbb{P}(\beta > 0 \vert \text{data})$
β_0	**1.042**	0.307	1.872	0.998	β_0	**10.883**	7.646	14.116	1.000
β_1	**−1.210**	-2.150	-0.337	0.003	β_1	**9.678**	5.786	13.591	1.000
β_2	**−1.023**	-1.943	-0.175	0.009	β_2	**0.119**	-3.646	3.893	0.524
β_3	**−1.293**	-2.934	0.277	0.053	β_3	**8.653**	1.617	15.703	0.992

TABLE 2.1: Posterior median, 95% credible intervals and posterior probability $\mathbb{P}(\beta > 0 \vert \text{data})$ for each coefficient from a Bayesian regression models, to study the association between diarrhea onset (left) and Age (right) with the estimated cluster allocation.

To further investigate possible interpretations of the estimated distributional clusters, we have fit two simple Bayesian regression models where we estimate the association between additional variables, which are available in the dataset, and the estimated cluster allocations. More specifically, we consider a logistic regression to study the association with the onset of diarrhea, a dichotomous variable indicating if a child has suffered from disrupting diarrhea (*Type*=1) or not (=0), and a linear regression model to study the association with the variable *Age*, which reports the children's age in month. Since there are only two observations clustered in one group, we limited this study only to the four remaining clusters. We treat the cluster indicator variable as a 4-levels factor, from which we recover 3 dummy variables. For both the regression models, we run 5 different chains for 100,000 iterations in `rstan` (Stan Development Team, 2019), discarding the first half as burn-in period.

Table 2.1 reports the posterior medians, 95% credible intervals and the estimated posterior probabilities $\mathbb{P}(\beta > 0 \vert \text{data})$ for each of the coefficients. The first cluster from the left in Figure 2.1 is positively associated with the onset of diarrhea and characterizes children with a mean age of 10.883 months (95% CI: 7.646, 14.116). All the other clusters are characterized by lower odds of diarrhea onset and higher ages with respect to cluster one. However, the first and third clusters are characterized by very similar ages. Albeit limited by the type of additional variables available in the dataset, our exploratory analysis suggests that the distributional clusters may highlight demographic or clinical differences in the studied population.

Finally, we consider inference at the OTU (microbial sequence) level, by investigating the estimated clustering of microbiome sequences, which we have previously referred to as *observational clustering*, and by assessing the variability of the microbiome abundances across subjects. We condition the OTU level inference on the optimal distributional cluster configuration estimated in the previous paragraph. Let $g = 1, \dots, G$ indicate the estimated distributional clusters (here, $G = 5$) and let $S_g = \{1, 2, \dots, n_g\}$ denote a collection of indexes identifying the subjects assigned to each cluster g. In order to investigate how the observed abundance of a given microbial sequence varies within a cluster of subjects, we assess if the abundance of each microbe appears more similar within- than between- the estimated groups. If two microbes i and i' are assigned to the same observational cluster, we can expect $\mu_{M_{ij}} = \mu_{M_{i'j'}}$ with high probability. Let $\mathcal{O}_t^{i,g}$ denote the set of allocation indicators (observational labels) which have been assigned to OTU i across the different subjects that have been assigned to group g at iteration $t = 1, \dots, T$ of the MCMC sampler, for each $g = 1, \dots, G$. Since the allocation indicators are just computational tools that inform about cluster frequencies and not parameter values, we apply a measure of dissimilarity on $\mathcal{O}_t^{i,g}$ to summarize how much the allocations of the OTUs agree between subjects assigned to the same group. More specifically, we use again the Shannon index, by computing $H\left(\mathcal{O}_t^{i,g}, S_g\right) = -\sum_{S_g} p_t^{i,g} \log p_t^{i,g}$, appropriately normalized, where the $p_t^{i,g}$'s denote the relative frequencies of the observational labels of OTU i in cluster g at the t-th iteration. Then, we compute the ergodic means, $\bar{H}^{i,g} = \sum_{t=1}^{T} H\left(\mathcal{O}_t^{i,g}, S_g\right)$, $i = 1, \dots, N$, $g = 1, \dots, G$. The results are reported in Figure 2.3. Two clusters are characterized by low entropy, suggesting that most of the microbial sequences in the groups are characterized by similar abundances. On the other hand, two clusters are characterized by higher entropy, i.e. the abundances of the microbial sequences vary across subjects assigned to the group. Results may vary for specific microbes. For example, the *Faecalibacterium prausnitzii* is characterized by an entropy below 0.5 in three of the five groups considered here. Upon further inspection, in the groups where the entropy is high, the microbe abundances are characterized by high sampling variability, i.e. microbes may have zero counts in some of the samples. We can also investigate the observational clustering structure per se. Table 2.2 reports some summary statistics – average, median, standard deviation and Shannon Index – computed on the OTU counts stratified by different clusters.

Observational Cluster	1	2	3	4	5
n	5586	20349	588	1966	1615
Average	2.460	0.000	1000.447	13.348	87.473
Median	2	0	626	11	66
Std. Dev.	2.003	0.000	980.670	7.801	64.977
Shannon Index	8.368	0.000	6.027	7.436	7.144

TABLE 2.2: Summary statistics stratified by observational cluster, considering the $> 30k$ OTU counts as units.

Among the others, Cluster 2 and 3 are particularly relevant. In fact, they contain the absent and the most present OTUs, respectively. We further investigate the percentage of times a particular OTU species is included in these two clusters, across all the 212 subjects. We report the top three OTU species in terms of presence in Table 2.3.
Our analysis confirms the relevance of studying how microbiome abundances vary across subjects, in order to capture microbial diversity, an important concept in microbiome analysis. Further analysis is needed to associate the abundances of specific microbes to available clinical or demographic variables.

Cluster	% of presence	OTU species
2	92.4	*Bacteroides sp. XO77B42*
2	91.9	*Catenibacterium mitsuokai*
2	91.5	*Prevotella sp. oral clone P4PB_83 P2*
3	47.6	*Escherichia coli*
3	22.6	*Prevotella sp. BI42*
3	13.2	*Prevotella sp. DJF_B112*

TABLE 2.3: Top three OTU species in Observational Clusters 2 and 3 in terms of presence across individuals.

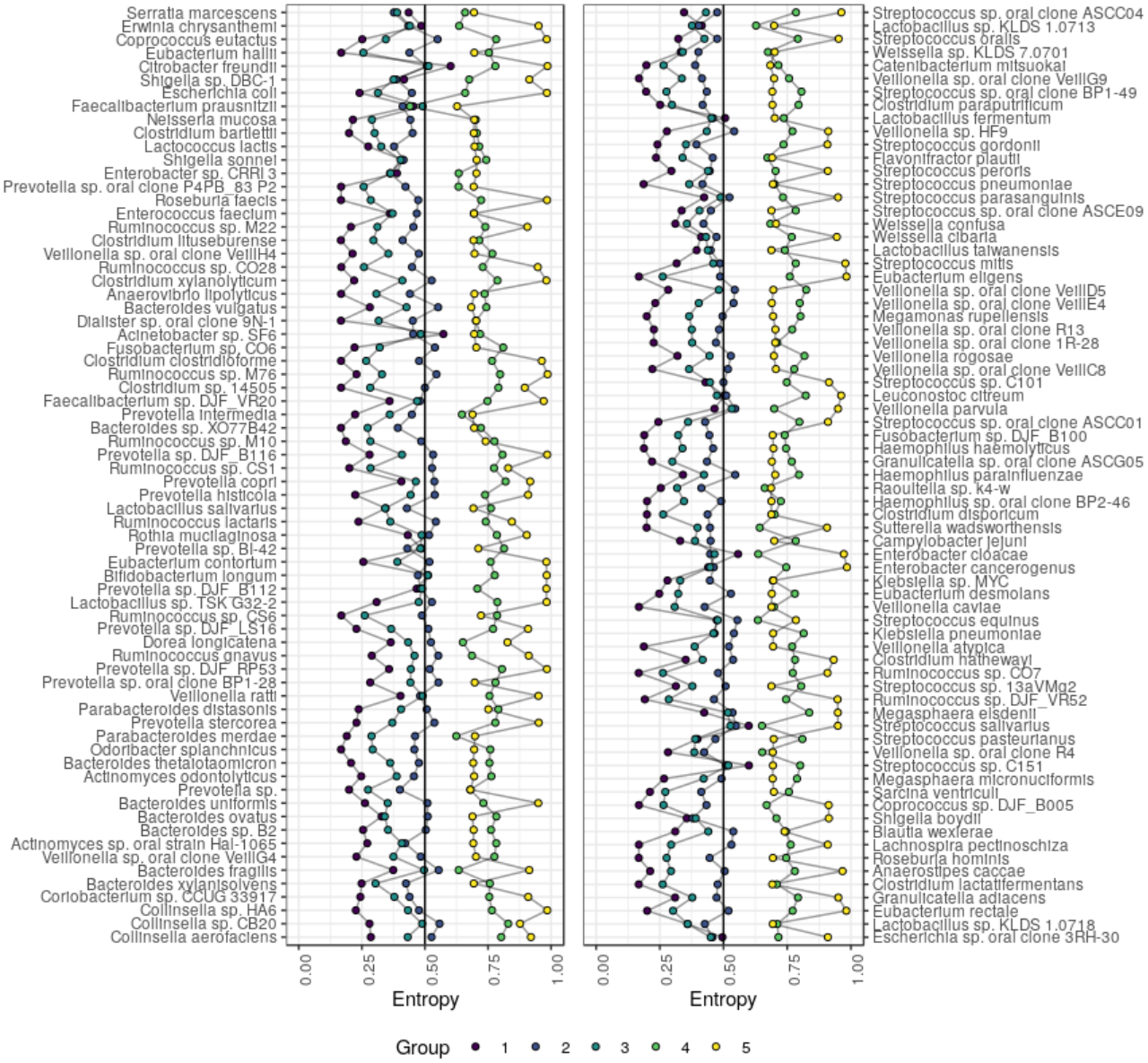


FIGURE 2.3: The Entropy index $\bar{H}^{i,g}$ described in Section 2.5: for each row, representing a specific OTU, we observe the Entropy indexes stratified by the optimal distributional allocation of subjects, which has been estimated using the Variation of Information criterion.

2.6 Simulation study

We study the clustering performance of the proposed methodology for continuous (CAM) and discrete measurements (DCAM) in a simulation study. More specifically, we investigate how our model is able to recover the true distributional clustering structure when the number of observations within each unit, or the number of units itself, increase. The prior specifications are similar as in the case study. We consider the following three scenarios:

- ***Scenario 1 - CAM***. We consider two replications for each of the six units characterized by the following distributions:

$$
\begin{aligned}
&\boldsymbol{Y}_1 \sim N(0,.6), \qquad \boldsymbol{Y}_2 \sim \frac{1}{2}N(0,.6) + \frac{1}{2}N(5,.6),\\
&\boldsymbol{Y}_3 \sim \frac{1}{3}N(0,.6) + \frac{1}{3}N(5,.6) + \frac{1}{3}N(10,.6),\\
&\boldsymbol{Y}_4 \sim \frac{1}{4}N(0,.6) + \frac{1}{4}N(5,.6) + \frac{1}{4}N(10,.6) + \frac{1}{4}N(13,.6),\\
&\boldsymbol{Y}_5 \sim \frac{1}{5}N(0,.6) + \frac{1}{5}N(5,.6) + \frac{1}{5}N(10,.6) + \frac{1}{5}N(13,.6) + \frac{1}{5}N(16,.6),\\
&\boldsymbol{Y}_6 \sim \frac{1}{6}N(0,.6) + \frac{1}{6}N(5,.6) + \frac{1}{6}N(10,.6) + \frac{1}{6}N(13,.6) + \frac{1}{6}N(16,.6) + \frac{1}{6}N(20,.6),
\end{aligned}
$$

  Since every unit is sampled two times, therefore $J = 12$. Notice how the supports of these distributions overlap to a great extent. To assess how the model behaves with asymmetries in the unit sample sizes, we simulate observations from the aforementioned distributions following two different approaches. First, we generate the different units with equal cardinality n_A (Case A), where we consider $n_A \in \{25, 50, 75\}$. Then, we generate inhomogenous units fixing the number of observations per mixture component n_B (Case B). The cardinality of each center is $n_j = \#\boldsymbol{Y}_j = n_B \cdot j$ for $j = 1, \ldots, 6$, where $n_B \in \{5, 10, 20\}$. To assess the distributional clusters (DC), we verify if the couples with identical distributional law are clustered together. We then study how the observations are partitioned (OC) among the different components of the mixtures.

- ***Scenario 2 - CAM***. We test the performance of our model in a challenging framework, where four different distributions for the units are considered:

$$
\begin{aligned}
&\boldsymbol{Y}_1 \sim 0.75N(0,1) + 0.25N(3,0.6),\\
&\boldsymbol{Y}_2 \sim 0.50N(0,1) + 0.50N(3,0.6),\\
&\boldsymbol{Y}_3 \sim 0.33N(0,1) + 0.34N(-2,0.6) + 0.33N(2,1),\\
&\boldsymbol{Y}_4 \sim 0.25N(0,1) + 0.25N(-2,0.6) + 0.25N(2,1) + 0.25N(10,1),
\end{aligned}
$$

  We keep the number of observation per unit constant, equal to $n = 40$. Instead, we vary the number r of replications considered. Therefore, $r = 1, \ldots, 6$ and $J_r = 4 \cdot r$. Therefore, the total number of considered units ranges from $J_1 = 4$ to $J_6 = 24$. In this way, we can investigate the estimated distributional clustering structures as the total number of units increases. Notice that, while the number of true DCs grows, the number of true OCs is fixed to 2 for the first two distributions, 3 and 4 for the last two ones.

- ***Scenario 3 - DCAM***. Let δ_x denote a point mass placed on x and let $UD(q, Q)$ represent a Uniformly Discrete distribution placed on the set of integers $\{q, \ldots, Q\}$, where $q < Q$ and $q, Q \in \mathbb{N}$. We consider three different possible unit distributions to sample $J = 10$ centers:

$$
\boldsymbol{Y}_{r,1} \sim \omega_1\delta_0 + \omega_2\delta_1 + \omega_3 UD[0, 10], \qquad r = 1, 2
$$

$$\boldsymbol{Y}_{r,2} \sim \omega_1\delta_0 + \omega_2\delta_1 + \omega_3 UD\,[0, 50]\,, \qquad r = 3, 4, 5, 6$$
$$\boldsymbol{Y}_{r,3} \sim \omega_1\delta_0 + \omega_2\delta_1 + \omega_3 UD\,[0, 100]\,, \qquad r = 7, 8, 9, 10,$$

where $\omega_1, \omega_2, \omega_3$ are the mixture weights and r denotes the replication. We set $\omega_1 = \omega_2$, generating the 50 observations equal to zero and 50 equal to one to simulate a case of low value inflation. We investigate the performance of the model in 4 cases (I-IV), differentiating by the number of observations assigned to the last mixture component: 10, 15 ,25 and 50 respectively. Notice that the number of true DCs is fixed, equal to 3. However, the assessment of the OCs requires more care since there is ground truth available. We assume four different clusters: a first set of "low-expressed" observations containing only zeros and ones, and three others, partitioning the support in $[2, 10]$,$(11, 50]$ and $(51, 100]$.

We report a picture of the densities of the distributions of each scenario in the Appendix. For each scenario, we also run a nDP mixture model, indicated as nDP, in the case with biggest sample size. We truncate the observational DP at 30 and the distributional DP at 25. In Table 2.1 we assess the goodness of the estimated optimal partition with the ground truth via Adjusted Rand Index (Hubert and Arabie, 1985) and Classification Error. We see how sometimes the model detects the correct number of cluster, but misassigns some of the distributions. We can appreciate how the model is able to recover the observational clusters when they are distinct and separated (Scenario 1). In Scenario 2 and 3 the model correctly identifies the components that are separated from the data while it struggles to distinguish between similar generating distributions, tending to be parsimonious. However, very good results are recovered when we focus on the distributional clustering. In fact, we can also appreciate how the model is able to recover the truth as either n_j or J increase, in some cases even with very small sample sizes. Hence, these results are remarkable and promising: the CAM and the DCAM are able to allocate the various units into the correct distributional clusters, even if the various distributions overlap. Lastly, in each scenario it is evident how the overlap of the data impacts the estimated partition of the vanilla nDP, which always collapses two units.

2.7 Discussion

We have introduced a nested nonparametric model that allows investigating distributional heterogeneity in nested data. The proposed Common Atom Model allows a two-layered clustering at the distributional and observational level, similarly to the nDP of Rodríguez et al. (2008). By construction, our model formulation allows sharing of atoms with different weights across distributions, and thus does not suffer from the degeneracy properties of the nDP whenever there is a tie between atoms. The common atom model specification is appealing and convenient for a variety of reasons: it is simple, allows a more refined description of distributional clusters, and it is computationally efficient. We can extend the methodology to allow the modeling and clustering of discrete distributions, by considering a rounded mixture of Gaussian kernels as in Canale and Dunson (2011). A further contribution of this work is the implementation of a nested version of the independent efficient slice sampler. We applied our methodology to a real microbiome dataset, aiming to cluster patients characterized by similar taxa distributions. Controlling for each subject's library size (the total frequencies of the observed species), we grouped the data minimizing the Variation of Information loss function and we showed how the model is able to detect clusters catching main differences among the distributions. We also assess the performance of our modeling approach by means of a simulation study, where the overlap between different distributions is evident.
The application of the proposed model to the real data set is limited by the type and number of clinical and demographic covariates that are available. If additional covariates were available,

Scenario 1 - DC	$n_A = 25$	$n_A = 50$	$n_A = 75$	$n_B = 5$	$n_B = 10$	$n_B = 20$	nDP
True Clusters	6	6	6	6	6	6	6
Clusters Detected	4	5	6	3	5	6	5
ARI	0.4210	0.7179	1.000	0.3333	0.7179	1.000	0.7179
Class Error	0.3333	0.1667	0.000	0.5000	0.1667	0.000	0.1667
Scenario 1 - OC	$n_A = 25$	$n_A = 50$	$n_A = 75$	$n_B = 5$	$n_B = 10$	$n_B = 20$	-
True Clusters	6	6	6	6	6	6	-
Clusters Detected	5	5	6	5	7	6	-
ARI	0.9772	0.9813	0.9848	0.9617	0.9522	0.9691	-
Class Error	0.0433	0.0316	0.0144	0.0619	0.0261	0.0178	-
Scenario 2 - DC	$J_1 = 4$	$J_3 = 8$	$J_3 = 12$	$J_4 = 16$	$J_5 = 20$	$J_6 = 24$	nDP
True Clusters	4	4	4	4	4	4	4
Clusters Detected	3	3	4	5	4	4	3
ARI	0.0000	0.5882	0.6393	0.6767	0.859	1.0000	0.6849
Class Error	0.2500	0.2500	0.1667	0.1250	0.050	0.0000	0.2500
Scenario 2 - OC	$J_1 = 4$	$J_3 = 8$	$J_3 = 12$	$J_4 = 16$	$J_5 = 20$	$J_6 = 24$	nDP
True Clusters	6	6	6	6	6	6	-
Clusters Detected	4	4	4	4	6	5	-
ARI	0.6340	0.4318	0.5661	0.5347	0.5880	0.6355	-
Class Error	0.2062	0.2875	0.2270	0.2390	0.2187	0.1854	-
Scenario 3 - DC	I	II	III	IV	-	-	nDP
True Clusters	3	3	3	3	-	-	3
Clusters Detected	3	3	3	3	-	-	2
ARI	0.7232	0.8258	0.8258	1.000	-	-	0.6341
Class Error	0.1000	0.1000	0.1000	0.000	-	-	0.2000
Scenario 3 - OC	I	II	III	IV	-	-	-
True Clusters	4	4	4	4	-	-	-
Clusters Detected	4	5	5	5	-	-	-
ARI	0.9452	0.9667	0.9633	0.9592	-	-	-
Class Error	0.0163	0.0260	0.0392	0.0540	-	-	-

TABLE 2.1: True number of clusters, detected number of clusters, Adjusted Rand Index and Classification Error computed in the three simulation scenarios to quantitatively compare the clustering performance

they could be used to model more complex dependencies, e.g. by constructing dependent random measures with covariate-dependent weights as in MacEachern (2000) (see also Barrientos et al., 2012), or to build risk-prediction models. Another interesting extension considers the incorporation of a time dimension, and studying how distributional clusters vary across time. These directions are left for future investigation.

Appendix

2.A Proofs

Proof of Equation 2.6
Let G_i and G_j be two realizations from the random probability measure Q, as defined in eq. (2.1). Then,

$$\begin{aligned}\mathbb{P}\left(G_i = G_j|Q\right) &= \sum_{k\geq 1}\mathbb{P}\left(G_i = G_j = G_k^*|Q\right) = \sum_{k\geq 1}\mathbb{P}\left(G_i = G_k^*, G_j = G_k^*|Q\right)\\ &= \sum_{k\geq 1}\mathbb{P}\left(G_i = G_k^*|Q\right)\mathbb{P}\left(G_j = G_k^*|Q\right) = \sum_{k\geq 1}\pi_k^2 > 0.\end{aligned}$$

So,

$$\mathbb{P}\left(G_i = G_j\right) = \mathbb{E}\left[\mathbb{P}\left(G_i = G_j|Q\right)\right] = \mathbb{E}\left[\sum_{k\geq 1}\pi_k^2\right] = \sum_{k\geq 1}\mathbb{E}\left[\pi_k^2\right] = \frac{1}{1+\alpha}.$$

where the last equality is motivated in one of the next proofs.

Proof of Equation 2.7
Let $x_{ij} \sim G_j$ and $x_{i'j'} \sim G_{j'}$ be two realizations from two probability measures both sampled from Q. Then,

$$\begin{aligned}\mathbb{P}\left[x_{ij} = x_{i'j'}\right] &= \mathbb{E}\left[\mathbb{P}\left[x_{ij} = x_{i'j'}|G_j, G_{j'}\right]\right]\\ &= \mathbb{E}\left[\frac{1}{1+\alpha}\mathbb{P}\left[x_{ij} = x_{i'j'}|G_j = G_{j'}\right] + \frac{\alpha}{1+\alpha}\mathbb{P}\left[x_{ij} = x_{i'j'}|G_j \neq G_{j'}\right]\right]\\ &= \frac{1}{1+\alpha}\mathbb{E}\left[\sum_{r\geq 1}\omega_{rj}^2\right] + \frac{\alpha}{1+\alpha}\mathbb{E}\left[\sum_{r\geq 1}\omega_{rj}\omega_{rj'}\right]\\ &= \frac{1}{1+\alpha}\frac{1}{1+\beta} + \frac{1}{1+\alpha}\frac{1}{2\beta+1} = Corr\left(x_{ij}, x_{i'j'}\right).\end{aligned}$$

Covariance and Correlation among two measures
Suppose G_j's are defined on a Polish space $(\mathbb{X}, \mathcal{X})$. Let us consider $A, B \in \mathcal{X}$. Recall that $G_j, G_{j'}|Q \overset{i.i.d.}{\sim} Q$. In the following, to ease the notation, we assume without loss of generality that $j = 1$ and $j' = 2$. Denote with G_k^* a realization of a Dirichlet Process and we will differentiate among different realizations calling them $G_{k_1}^*$ and $G_{k_2}^*$. Then

$$\begin{aligned}\mathbb{E}\left[G_1(A)\cdot G_2(B)\right] =& \mathbb{E}\left[\mathbb{E}\left[G_1(A)\cdot G_2(B)|Q\right]\right] =\\ &\mathbb{P}\left[G_{k_1}^* = G_{k_2}^* = G_k^*\right]\mathbb{E}\left[G_k^*(A)\cdot G_k^*(B)\right] + \mathbb{P}\left[G_{k_1}^* \neq G_{k_2}^*\right]\mathbb{E}\left[G_{k_1}^*(A)\cdot G_{k_2}^*(B)\right].\end{aligned}$$

We then notice that $\mathbb{P}\left[G^*_{k_1} = G^*_{k_2} = G^*_k\right] = \mathbb{E}\left[\sum_{l\geq 1}\pi_l^2\right] = \frac{1}{\alpha+1}$, so the previous equation becomes

$$\mathbb{E}\left[G_1(A)\cdot G_2(B)\right] = \frac{1}{\alpha+1}\mathbb{E}\left[G^*_k(A)\cdot G^*_k(B)\right] + \frac{\alpha}{\alpha+1}\mathbb{E}\left[G^*_{k_1}(A)\cdot G^*_{k_2}(B)\right].$$

The first expected value can be expressed as

$$\begin{aligned}
\mathbb{E}\left[G^*_k(A)\cdot G^*_k(B)\right] =& \mathbb{E}\left[\sum_{l\geq 1}\omega_{l,k}\delta_{x_l}(A)\cdot\sum_{l\geq 1}\omega_{l,k}\delta_{x_l}(B)\right]\\
=& \mathbb{E}\left[\sum_{l\geq 1}\omega_{l,k}^2\delta_{x_k}(A\cap B)\right] + \mathbb{E}\left[\sum_{l\geq 1}\sum_{r\neq l}\omega_{l,k}\omega_{r,k}\delta_{x_l}(A)\delta_{x_r}(B)\right]\\
=& \mathbb{E}\left[\sum_{l\geq 1}\omega_{l,k}^2\right]H(A\cap B) + \left(1-\sum_{l\geq 1}\mathbb{E}\left[\omega_{l,k}^2\right]\right)H(A)H(B)\\
=& \frac{1}{\beta+1}H(A\cap B) + \frac{\beta}{\beta+1}H(A)H(B),
\end{aligned}$$

while the second one

$$\begin{aligned}
\mathbb{E}\left[G^*_{k_1}(A)\cdot G^*_{k_2}(B)\right] =& \mathbb{E}\left[\sum_{r\geq 1}\omega_{r,k_1}\delta_{x_r}(A)\cdot\sum_{l\geq 1}\omega_{l,k_2}\delta_{x_l}(B)\right]\\
=& \mathbb{E}\left[\sum_{r\geq 1}\omega_{r,k_1}\omega_{r,k_2}\delta_{x_r}(A\cap B)\right] + \mathbb{E}\left[\sum_{r\neq l}\omega_{r,k_1}\omega_{l,k_2}\delta_{x_r}(A)\delta_{x_l}(B)\right]\\
=& \mathbb{E}\left[\sum_{r\geq 1}\omega_{r,k_1}\omega_{r,k_2}\right]H(A\cap B) + \mathbb{E}\left[\sum_{r\neq l}\omega_{r,k_1}\omega_{l,k_2}\right]H(A)H(B).
\end{aligned}$$

Then, we notice that the previous needs to hold also when $A = B = \mathbb{X}$. In that case

$$\begin{aligned}
1 = \mathbb{E}\left[G^*_{k_1}(\mathbb{X})\cdot G^*_{k_2}(\mathbb{X})\right] &= \sum_{r\geq 1}\mathbb{E}\left[\omega_{r,k_1}\right]\mathbb{E}\left[\omega_{r,k_2}\right]H(\mathbb{X}) + \sum_{r\neq l}\mathbb{E}\left[\omega_{r,k_1}\right]\mathbb{E}\left[\omega_{l,k_2}\right]H(\mathbb{X})H(\mathbb{X})\\
&\iff \left(1-\sum_{r\geq 1}\mathbb{E}\left[\omega_{r,k_1}\right]^2\right) = \sum_{r\neq l}\mathbb{E}\left[\omega_{r,k_1}\right]\mathbb{E}\left[\omega_{l,k_2}\right].
\end{aligned}$$

So we can conclude that

$$\begin{aligned}
\mathbb{E}\left[G^*_{k_1}(A)\cdot G^*_{k_2}(B)\right] &= \sum_{r\geq 1}\mathbb{E}\left[\omega_{r,k_1}\right]\mathbb{E}\left[\omega_{r,k_2}\right]H(A\cap B) + \sum_{r\neq l}\mathbb{E}\left[\omega_{r,k_1}\right]\mathbb{E}\left[\omega_{l,k_2}\right]H(A)H(B)\\
&\overset{i.i.d.}{=} \sum_{r\geq 1}\{\mathbb{E}\left[\omega_{r,k_1}\right]\}^2 H(A\cap B) + \left(1-\sum_{r\geq 1}\{\mathbb{E}\left[\omega_{r,k_1}\right]\}^2\right)H(A)H(B).
\end{aligned}$$

We then notice that

$$\sum_{r\geq 1}\{\mathbb{E}\left[\omega_{r,k_1}\right]\}^2 = \sum_{r\geq 1}\left\{\mathbb{E}\left[v_r\prod_{q=1}^{r-1}(1-v_q)\right]\right\}^2 =$$

$$= \sum_{r\geq 1} \left[\frac{1}{(1+\beta)^2} \left(\frac{\beta}{1+\beta} \right)^{2(r-1)} \right] = \frac{1}{2\beta+1}.$$

So we can complete the derivation of $\mathbb{E}\left[G^*_{k_1}(A) \cdot G^*_{k_2}(B)\right]$:

$$\mathbb{E}\left[G^*_{k_1}(A) \cdot G^*_{k_2}(B)\right] = \frac{1}{1+2\beta} H(A \cap B) + \frac{2\beta}{1+2\beta} H(A)H(B).$$

Now let us define as $q_1 := \frac{1}{\alpha+1}$. Then we can write

$$\mathbb{E}\left[G_1(A) \cdot G_2(B)\right] = H(A \cap B) \left(\frac{q_1}{1+\beta} + \frac{1-q_1}{1+2\beta} \right) + H(A)H(B) \left(q_1 \frac{\beta}{1+\beta} + (1-q_1) \frac{2\beta}{1+2\beta} \right).$$

So we derive for two general indexes j and j'

$$\begin{aligned} Cov\left(G_j(A), G_{j'}(B)\right) &= H(A \cap B) \left(\frac{q_1}{1+\beta} + \frac{1-q_1}{1+2\beta} \right) + H(A)H(B) \left(q_1 \frac{\beta}{1+\beta} + (1-q_1) \frac{2\beta}{1+2\beta} - 1 \right) \\ &= H(A \cap B) \left(\frac{q_1}{1+\beta} + \frac{1-q_1}{1+2\beta} \right) + H(A)H(B) \left(-\frac{q_1}{1+\beta} - \frac{1-q_1}{1+2\beta} \right), \end{aligned}$$

and in the case $A = B$, we have

$$Cov\left(G_j(A), G_{j'}(A)\right) = \left(\frac{q_1}{1+\beta} + \frac{1-q_1}{1+2\beta} \right) H(A)(1-H(A)).$$

Moreover, let us find $Corr\left(G_j(A), G_{j'}(A)\right) = \frac{Cov\left(G_j(A), G_{j'}(A)\right)}{\sqrt{Var(G_j(A)) \cdot Var(G_{j'}(A))}}$. We compute

$$\begin{aligned} Var(G_j(A)) =& \mathbb{E}\left[G_j(A)^2\right] - \mathbb{E}\left[G_j(A)\right]^2 \\ =& \mathbb{E}\left[\mathbb{E}\left[G_j(A)^2 | Q\right]\right] - \mathbb{E}\left[G_j(A)\right]^2 = \mathbb{E}\left[G^*_j(A)^2\right] - \mathbb{E}\left[G_j(A)\right]^2 \\ =& \frac{1}{\beta+1} H(A) + \frac{\beta}{1+\beta} H(A)^2 - H(A)^2 \\ =& \frac{1}{\beta+1} H(A)\left(1 - H(A)\right). \end{aligned}$$

Finally,

$$\begin{aligned} \rho_{ij} := Corr(G_j(A), G_{j'}(A)) =& \left(\frac{q_1}{\beta+1} + \frac{1-q_1}{2\beta+1} \right) \Big/ \frac{1}{\beta+1} \\ =& q_1 + \frac{\beta+1}{2\beta+1}(1-q_1) = 1 - \frac{\beta}{2\beta+1}(1-q_1) \\ =& 1 - \frac{\beta}{2\beta+1} \cdot \frac{\alpha}{1+\alpha}. \end{aligned}$$

We underline that $\rho_{ij} \in \left(\frac{1}{2}, 1\right)$.

Proof of Proposition 1
Recalling the CAM model (2.1)–(2.2), we get

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)Q(dg_1)Q(dg_2)\right] \\
&\quad = \mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)\sum_{k_1\geq 1}\pi_{k_1}\delta_{G^*_{k_1}}(dg_1)\sum_{k_2\geq 1}\pi_{k_2}\delta_{G^*_{k_2}}(dg_2)\right] \\
&\quad = \mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)\sum_{k\geq 1}\pi_k^2\delta_{G^*_k}(dg_1)\delta_{G^*_k}(dg_2)\right] \\
&\qquad + \mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)\sum_{k_1\neq k_2}\pi_{k_1}\pi_{k_2}\delta_{G^*_{k_1}}(dg_1)\delta_{G^*_{k_2}}(dg_2)\right].
\end{aligned}$$

Observe that the G^*_k's are all Dirichlet processes having the same law on the space $\mathsf{P}_{\mathbb{X}}$, which will be denoted by $\mathcal{P}$, depending on the total mass α and the base measure H. We also point out that the G^*_k's are not independent random elements for different values of k, indeed they share the same random atoms $(\theta_l)_{l\geq 1}$, nevertheless if $k_1 \neq k_2$, the distribution of $(G^*_{k_1}, G^*_{k_2})$ equals the distribution of (G^*_1, G^*_2), which will be denoted by $\mathcal{P}_{[2]}$. Therefore, by applying the Tonelli–Fubini Theorem, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)Q(dg_1)Q(dg_2)\right] \\
&\quad = \sum_{k\geq 1}\pi_k^2\mathbb{E}\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)\delta_{G^*_k}(dg_1)\delta_{G^*_k}(dg_2) \\
&\qquad + \sum_{k_1\neq k_2}\pi_{k_1}\pi_{k_2}\mathbb{E}\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)\delta_{G^*_{k_1}}(dg_1)\delta_{G^*_{k_2}}(dg_2) \\
&\quad = q_1\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g)f_2(g)\mathcal{P}(dg) + (1-q_1)\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)\mathcal{P}_{[2]}(dg_1, dg_2),
\end{aligned}$$

and then the thesis follows.

Proof of Theorem 1
We first evaluate the expected value in the definition of pEPPF (2.4), for $J = 2$,

$$\begin{aligned}
\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s} G_j^{n_{i,j}}(dy_i^*) &= \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{2}\prod_{i=1}^{s} G_j^{n_{i,j}}(dy_i^*)\Big|Q\right]\right] \\
&= \mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2}\prod_{j=1}^{2}\prod_{i=1}^{s} g_j^{n_{i,j}}(dy_i^*)Q(dg_1)Q(dg_2)\right].
\end{aligned}$$

Now we apply Equation (2.3) to the previous integral where the functions f_j, as $j = 1, 2$, are defined by

$$f_j(g_j) := \prod_{i=1}^{s} g_j^{n_{i,j}}(dy_i^*),$$

and then we get

$$\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s} G_j^{n_{i,j}}(dy_i^*) = q_1\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G_1^*)^{n_{i,j}}(dy_i^*) + (1-q_1)\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G_j^*)^{n_{i,j}}(dy_i^*). \tag{2.17}$$

We finally integrate over the space $\mathbb{X}^s$ to get the result, i.e. (2.5).

Proof of Proposition 2

Assume that the two samples $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ share $s_0 > 0$ distinct values denoted here as $y^*_{1,0}, \dots y^*_{s_0,0}$ with frequencies $(q_{1,j}, \dots, q_{s_0,j})$ in the j-th sample, as $j = 1, 2$. We further suppose that the j-th sample contains exactly s_j distinct observations not shared with the other one, and denoted here by $y^*_{1,j}, \dots, y^*_{s_j,j}$, for $j = 1, 2$; besides the vector of corresponding frequencies will be denoted as $(r_{1,j}, \dots, r_{s_j,j})$. We obviously have that $s = s_0 + s_1 + s_2$.

Using the representation of the G^*_k's in the CAM model (2.1)–(2.2), we get

$$\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G^*_j)^{n_{i,j}}(dy^*_i) = \mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}\left(\sum_{l\geq 1}\omega_{l,j}\delta_{\theta_l}(dy^*_i)\right)^{n_{i,j}}.$$

Exploiting the partition of the data described at the beginning of the proof, we obtain

$$\begin{aligned}
&\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G^*_j)^{n_{i,j}}(dy^*_i) \\
&\quad= \mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s_j}\left(\sum_{l\geq 1}\omega_{l,j}^{r_{i,j}}\delta_{\theta_l}(dy^*_{i,j})\right)\prod_{i=1}^{s_0}\left(\sum_{l\geq 1}\omega_{l,1}^{q_{i,1}}\omega_{l,2}^{q_{i,2}}\delta_{\theta_l}(dy^*_{i,0})\right) + o\left(\prod_{j=0}^{2}\prod_{i=1}^{s_j}H(dy^*_{i,j})\right) \\
&\quad= \sum_{\neq}\mathbb{E}\left[\prod_{j=1}^{2}\prod_{i=1}^{s_j}\omega_{l_{i,j},j}^{r_{i,j}}\prod_{i=1}^{s_0}\omega_{l_{i,0},1}^{q_{i,1}}\omega_{l_{i,0},2}^{q_{i,2}}\right]\prod_{j=0}^{2}\prod_{i=1}^{s_j}H(dy^*_{i,j}) + o\left(\prod_{j=0}^{2}\prod_{i=1}^{s_j}H(dy^*_{i,j})\right).
\end{aligned}$$

where the sum $\sum_{\neq}$ is extended over all possible vectors of distinct natural numbers $(l_{i,j})_{i=1,\dots,s_j,\ j=0,1,2} \in \mathbb{N}^3$. Integrating over $\mathbb{X}^s$ we get that

$$\int_{\mathbb{X}^s}\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G^*_j)^{n_{i,j}}(dy^*_i) = \sum_{\neq}\mathbb{E}\left[\prod_{j=1}^{2}\prod_{i=1}^{s_j}\omega_{l_{i,j},j}^{r_{i,j}}\prod_{i=1}^{s_0}\omega_{l_{i,0},1}^{q_{i,1}}\omega_{l_{i,0},2}^{q_{i,2}}\right] \tag{2.18}$$

which is positive whenever $s_0 > 0$.

2.B Truncated Blocked Gibbs Sampler for CAM

The posterior distribution is analytically intractable, which forces us to develop sampling algorithms to simulate from it. A Polya Urn representation would be too expensive in computational cost. Instead, we provide two different algorithms: a Blocked Gibbs sampler (Ishwaran and James, 2001), mimicking the one proposed in (Rodríguez et al., 2008) and a nested slice sampler (Damien et al., 1999; Walker, 2007; Kalli et al., 2011). Here we discuss the former one. The Truncated CAM model has the following form:

$$\begin{aligned}
y_{ij}|\boldsymbol{M},\boldsymbol{\theta} &\sim N\left(\cdot|\theta_{M_{ij}}\right), & M_{ij}|\boldsymbol{S},\boldsymbol{\omega} &\sim \sum_{l=1}^{L}\omega_{l,S_j}\delta_l(\cdot), \\
\boldsymbol{\omega}_k|\boldsymbol{S}=\boldsymbol{\omega}_k &\sim GEM(\alpha), & S_j|\boldsymbol{\pi} &\sim \sum_{k=1}^{K}\pi_k\delta_k(\cdot), \\
\boldsymbol{\pi} &\sim GEM(\beta), & \theta_l &\sim \pi(\theta_l).
\end{aligned} \tag{2.19}$$

The Truncated version of CAM (TCAM) (2.19) can be extended to a Truncated version of DCAM (TDCAM) once the likelihood is modified according to (2.11), In the following we report the Gibbs Sampler for the TDCAM, the extension of the sampler to accomodate the presence of a

covariate linearly introduced. Notice that some of the conditioning variables are collapsed (Liu, 1994), to enhance the speed of convergence and the mixing of the chains.

TDCAM: Gibbs Sampler
The steps of the MCMC are the following:

1. The full conditional for each $y_{ij}, \forall i, j$ is Truncated Normal, with support $\left[a_{z_{ij}}, a_{z_{ij}+1}\right)$:

$$p\left(y_{ij}|z_{ij}, \cdots\right) \sim TN(\mu_{M_{ij}}, \sigma^2_{M_{ij}}; a_{z_{ij}}, a_{z_{ij}+1}).$$

   This can be done easily with the help of the R package `TruncatedNormal`, which relies on a recently improved algorithm exploiting minmax tilting (Botev, 2017).

2. The full conditional for the observational cluster labels M_{ij}, once the latent variable $\boldsymbol{y}$ is integrated out, is a discrete distribution, given by $\forall i, j$

$$p\left(M_{ij} = l|y_{ij}, S_j = k, \cdots - \boldsymbol{y_{ij}}\right) \propto \omega_{l,S_j} \Delta\Phi\left(a_{z_{ij}}; \mu_{M_{ij}}, \sigma^2_{M_{ij}}\right).$$

3. The full conditional for the distributional cluster labels S_j is given by, $\forall j$:

$$p\left(S_j = k|\cdots - \boldsymbol{y}^*_j, -\boldsymbol{M}_j\right) \propto \pi_k \prod_i \left(\sum_{m=1}^{L} \omega_{m,k} \Delta\Phi\left(a_{y_{ij}}; \mu_m, \sigma^2_m\right)\right).$$

4. To sample the full conditional of the weights $\boldsymbol{\pi}$ at the distributional level, we first need to define m^*_k as the number of units assigned to the same distributional cluster k, where $\sum_{k=1}^{K} m^*_k = J$ the total number of observed units. Then,

$$p\left(\boldsymbol{\pi}|\cdots\right) \propto p\left(\boldsymbol{S}|\boldsymbol{\pi}\right) p\left(\boldsymbol{\pi}\right) \propto p\left(\boldsymbol{\pi}\right) \pi_1^{m^*_1} \cdots \pi_K^{m^*_K}.$$

   Referring to the Stick Breaking representation, we can define the full conditional of the various sticks v_k, $\forall k = 1, \ldots, K$ as:

$$v_k \sim Beta\left(1 + m^*_k, \beta + \sum_{s=k+1}^{K} m^*_k\right).$$

5. The derivation of the full conditional for $\boldsymbol{\omega}$ is similar, even it requires more care. We have

$$p\left(\boldsymbol{\omega}|\cdots\right) \propto p\left(\boldsymbol{M}|\boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\xi_0}\right) p\left(\boldsymbol{\omega}\right) \propto \prod_{k=1}^{K} p\left(\boldsymbol{\omega_k}\right) \prod_{i,j} \left(\sum_{l=1}^{L} \omega_{l,S_j} \delta_l(\cdot)\right).$$

   The previous formula can be decomposed into the product of K elements and we can focus only on the case $S_j = k$. Let us define n^{lk} as the total number of observations assigned to the distributional cluster k in the observational unit l. The full conditional has this Stick-Breaking representation for $u_{lk}, \forall k$:

$$u_{lk} \sim Beta\left(1 + n^{lk}, \alpha + \sum_{r=l+1}^{L} n^{rk}\right), \quad l = 1, \ldots, L.$$

6. Let us define $n^{l\cdot} = \sum_{k=1}^{K} n^{lk}$ and denote with $\bar{y}_{l\cdot} = \frac{1}{n^{l\cdot}} \sum_{i,j:M_{i,j}=l} y_{ij}$. Exploiting the conjugacy property, we obtain the full conditional for $\theta_l = (\mu_l, \sigma_l^2)$:

$$\left(\mu_l, \sigma_l^2\right) | \cdots \sim NIG\left(m_0^*, \kappa_0^*, \alpha_0^*, \beta_0^*\right).$$

where

$$m_0^* = \frac{\kappa_0 m_0 + n^{l\cdot}\bar{y}_{l\cdot}}{\kappa_0 + n^{l\cdot}} \quad \kappa_0^* = \kappa_0 + n^{l\cdot} \quad \alpha_0^* = \alpha_0 + n^{l\cdot}/2$$

and

$$\beta_0^* = \beta + 0.5\left(\sum_{ij:M_{ij}=l} (y_{ij} - \bar{y}_{l\cdot})^2 + \left(\frac{\kappa_0 n^{l\cdot}}{\kappa_0 + n^{l\cdot}}\right)(y_{lk} - m_0)^2\right).$$

7. In case the precision parameters α and β of the two DPs are assumed stochastic, distributed as $Gamma\left(a_\alpha, b_\alpha\right)$ and $Gamma\left(a_\beta, b_\beta\right)$, we can still exploiting conjugacy. The full conditionals distributions are:

$$\alpha| \cdots \sim Gamma\left(a_\alpha + (K-1), b_\alpha - \sum_{k=1}^{K-1} \log(1 - v_k)\right),$$
$$\beta| \cdots \sim Gamma\left(a_\beta + K \cdot (L-1), b_\beta - \sum_{l=1}^{L-1}\sum_{k=1}^{K} \log(1 - u_{lk})\right).$$

Notice that we naturally set $a_{y_{ij}} = y_{ij}$. As suggested in Rodríguez et al. (2008), each step of this algorithm can be parallelized, in order to gain computational speed.

Linearly incorporating a covariate in the Likelihood

If we want to linearly add regressor to the mean, we update Equation (2.19) simply assuming:

$$z_{ij}|y_{ij} \sim \sum_{g=0}^{+\infty} \delta_g(\cdot)\mathbf{1}_{[a_g, a_{g+1})}\left(y_{ij}\right) \qquad y_{ij}|\boldsymbol{M}, \boldsymbol{\mu}, \boldsymbol{\sigma^2} \sim N\left(\mu_{M_{ij}} + \beta X_j, \sigma_{M_{ij}}^2\right). \tag{2.20}$$

We espouse such representation because of its interpretability: the latent continuous random variable y_{ij} can be decomposed as $y_{ij} = \mu_{M_{ij}} + \beta X_j + \varepsilon_{ij}$, where $\varepsilon \sim N\left(0, \sigma_{M_{ij}}^2\right)$. In other words, we model the every single latent value as the sum of an effect specific for each observational cluster, an effect due to the regressor value of each individual multiplied by a overall coefficient and a completely random effect, whose entity still depends on the observational cluster. This choice does not complicate the algorithm presented in the previous section: the full conditionals 1-3 are preserved if the mean is modified accordingly, switching from $\mu_{M_{ij}}$ to $\mu_{M_{ij}} + \beta X_j$. Step 6 remains the same once we substitute y_{ij} with $d_{ij} = y_{ij} - \beta X_j$. Steps 4, 5 and 7 are not affected by this change.
Finally, if we assume $\beta \sim N\left(m_\beta, \frac{1}{\kappa_\beta}\right)$, we can perform inference on the introduced coefficient. Define $R^1 = \sum_{i,j} \frac{X_j^2}{\sigma_{M_{ij}}^2}$ and $R^2 = \sum_{i,j} \frac{d_{ij} \cdot X_j}{\sigma_{M_{ij}}^2}$. The full conditional for β is:

$$\beta| \cdots \sim N\left(\frac{m_\beta \kappa_\beta + R^2}{\kappa_\beta + R^1}, \frac{1}{\kappa_\beta + R^1}\right).$$

This framework can be easily extended to accommodate for the presence of multiple covariates.

2.B.1 Error bounds estimation

Total Variation Distance (TVD). Let P, Q probability measures defined on the space $(\mathbb{X}, \mathcal{X})$. We define

$$d_{TV}(P,Q) = \sup_{A \in \mathcal{X}} |P(A) - Q(A)|.$$

If P, Q are absolutely continuous w.r.t. μ then

$$d_{TV}(P,Q) = \frac{1}{2}\int_{\mathcal{X}} \left| \frac{dP}{d\mu} - \frac{dQ}{d\mu} \right| d\mu.$$

Moreover, if $\mathcal{X}$ is a discrete space or if P and Q are concentrated on $\Omega \subset \mathcal{X}$ countable then

$$d_{TV}(P,Q) = \frac{1}{2}\sum_{x \in \Omega} \Big| P(x) - Q(x) \Big|.$$

Consider then two sequences a_i and b_i such that $|a_i| \leq 1$ and $|b_i| \leq 1$ $\forall i$. Then the following inequality holds:

$$(*) \quad \left| \prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i \right| \leq \sum_{i=1}^{n} |a_i - b_i|.$$

Approximation Error in CAM
Let us now consider

$$Q = \sum_{k \geq 1} \tilde{\pi}_k \delta_{\left(\sum_{l \geq 1} \omega_{lk} \delta_{\tilde{x}_l}\right)}$$

and its truncated version

$$Q^{(K,L)} = \sum_{k=1}^{K} \tilde{\pi}_k^{(K,L)} \delta_{\left(\sum_{l=1}^{L} \omega_{lk}^{(K,L)} \delta_{\tilde{x}_l}\right)}.$$

This will induce truncated probability measures $G_1^{(K,L)}, \ldots, G_J^{(K,L)}$, defined as follows. Consider $\xi_j | Q \overset{iid}{\sim} \sum_{k=1}^{+\infty} \tilde{\pi}_k \delta_k$, a latent random variable that identifies the mixture component from which G_j is generated, conditionally on Q.
So we define

$$G_j^{(K,L)} = \begin{cases} \sum_{l=1}^{L} \omega_{lk}^{(K,L)} \delta_{\tilde{x}_l} & \text{if } \xi_j \leq K \\ \sum_{l=1}^{L} \omega_{lK}^{(K,L)} \delta_{\tilde{x}_l} & \text{if } \xi_j > K \end{cases}$$

with

$$\begin{aligned} \omega_{lk}^{(K,L)} &= \omega_{lk} && \text{if } l \leq L-1 \\ \omega_{lK}^{(K,L)} &= 1 - \omega_{l1} - \ldots - \omega_{l(K-1)} \\ \tilde{\pi}_k^{(K,L)} &= \tilde{\pi}_k && \text{if } k \leq K-1 \\ \tilde{\pi}_K^{(K,L)} &= 1 - \tilde{\pi}_1 - \ldots - \tilde{\pi}_{K-1} \end{aligned}$$

In this way, we can write $G_1^{(K,L)}, \ldots, G_J^{(K,L)} | Q^{(K,L)} \overset{iid}{\sim} Q^{(K,L)}$.

We adopt a fixed j and, **conditioning on** $\xi_j = k$, we compute the TVD between G_j and $G_j^{(K,L)}$.

Case 1.1: $\xi_j \leq K$. We compare $G_j = \sum_{l\geq 1} \tilde{\omega}_{lk}\delta_{\tilde{x}_l}$ and $G_j^{(K,L)} = \sum_{l=1}^{L} \tilde{\omega}_{lk}^{(K,L)}\delta_{\tilde{x}_l}$. We get:

$$\begin{aligned} d_{TV}(G_j, G_j^{(K,L)}) &= \frac{1}{2}\left(\sum_{l=1}^{L} |\tilde{\omega}_{lk} - \tilde{\omega}_{lk}^{(K,L)}| + |\tilde{\omega}_{lk} - \tilde{\omega}_{lk}^{(K,L)}| + \sum_{l\geq L+1} |\tilde{\omega}_{lk} - 0|\right) \\ &= \frac{1}{2}\left(|\tilde{\omega}_{Lk} - 1 + \tilde{\omega}_{1k} + \ldots + \tilde{\omega}_{1(K-1)}| + \sum_{l\geq L+1} \tilde{\omega}_{lk}\right) \\ &= \frac{1}{2}\left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk} + \sum_{l\geq L+1} \tilde{\omega}_{lk}\right) = \left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk}\right). \end{aligned}$$

Case 1.2: if $\xi_j = k > K$ We compute

$$\begin{aligned} &\mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)|\xi, Q\right]\right] \\ =&\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\pi_k}\mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)|\xi = k, Q\right] + \sum_{k=K+1}^{+\infty} \tilde{\pi_k}\mathbb{E}\left[\underbrace{d_{TV}\left(G_j, G_j^{(K,L)}\right)}_{\leq 1}|\xi = k, Q\right]\right] \\ \leq&\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\pi_k}\left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk}\right) + \sum_{k=K+1}^{+\infty} \tilde{\pi_k}\right] = (indep + linearity) \\ =&\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\pi_k}\right] \cdot \mathbb{E}\left[\left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk}\right)\right] + \mathbb{E}\left[\sum_{k=K+1}^{+\infty} \tilde{\pi_k}\right] \\ \leq&1 \cdot \mathbb{E}\left[\left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk}\right)\right] + \mathbb{E}\left[1 - \sum_{k=1}^{K} \tilde{\pi}_l\right] \\ =&\left(1 - \left(\frac{\alpha}{1+\alpha}\right)^K\right)\left(\frac{\beta}{1+\beta}\right)^L + \left(\frac{\alpha}{1+\alpha}\right)^K. \end{aligned}$$

Se we can conclude that

$$\mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)\right] \leq \left(1 - \left(\frac{\alpha}{1+\alpha}\right)^K\right)\left(\frac{\beta}{1+\beta}\right)^L + \left(\frac{\alpha}{1+\alpha}\right)^K.$$

Approximation Error in Mixture Models (CAMM)
Consider J units, each of them containing n_j observations, $j = 1, \ldots, J$. Denote with $\boldsymbol{X} = \left(x_{1j}, \ldots, x_{n_j j}\right)$ for $j = 1, \ldots, J$ the observations from the $j-$th component of the mixture model $x_{ij}|\theta_{ij} \sim f(\cdot|\theta_{ij})$ with $\theta_{ij}|G_1, \ldots, G_J \sim G_j$ where the G_j's are generated according to a CAM.
The law of the data is given by:

$$\pi\left(\boldsymbol{X}\right) = \mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{\Theta} f\left(x_{ij}|\theta_{ij}\right) G_j(d\theta_{ij})\right],$$

and its truncated version is

$$\pi^{(K,L)}(\boldsymbol{X}) = \mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{\Theta} f(x_{ij}|\theta_{ij})\, G_j^{(K,L)}(d\theta_{ij})\right].$$

The next step is to evaluate $d_{TV}\left(\pi, \pi^{(K,L)}\right)$. Let us define $n := n_1 + \cdots + n_J$. We obtain:

$$\begin{aligned}
d_{TV}\left(\pi, \pi^{(K,L)}\right) =& \frac{1}{2}\int_{\mathbb{X}^n}\left|\frac{d\pi}{d\boldsymbol{X}} - \frac{d\pi^{K,L}}{d\boldsymbol{X}}\right| d\boldsymbol{X} \\
=& \frac{1}{2}\int_{\mathbb{X}^n}\left|\mathbb{E}\left[\prod_{i=1}^{d}\prod_{j=1}^{n_j}\int_{\Theta} f(x_{ij}|\theta_{ij})\, G_j(d\theta_{ij}) - \prod_{i=1}^{d}\prod_{j=1}^{n_j}\int_{\Theta} f(x_{ij}|\theta_{ij})\, G_j^{(K,L)}(d\theta_{ij})\right]\right| d\boldsymbol{X} \\
=& \frac{1}{2}\int_{\mathbb{X}^n}\left|\mathbb{E}\left[\int_{\Theta^n}\prod_{i=1}^{d}\prod_{j=1}^{n_j} f(x_{ij}|\theta_{ij})\prod_{i=1}^{d}\prod_{j=1}^{n_j} G_j(d\theta_{ij}) - \int_{\Theta^n}\prod_{i=1}^{d}\prod_{j=1}^{n_j} f(x_{ij}|\theta_{ij})\prod_{i=1}^{d}\prod_{j=1}^{n_j} G_j^{(K,L)}(d\theta_{ij})\right]\right| d\boldsymbol{X}.
\end{aligned}$$

First we apply Fubini, then we take advantage of the fact that the absolute value of an integral in lower or equal to the integral of expected value

$$\begin{aligned}
=& \frac{1}{2}\int_{\mathbb{X}^n}\left|\int_{\Theta^n}\prod_{i=1}^{d}\prod_{j=1}^{n_j} f(x_{ij}|\theta_{ij})\,\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{ij}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{ij})\right]\right| d\boldsymbol{X} \\
\leq& \frac{1}{2}\int_{\mathbb{X}^n}\int_{\Theta^n}\prod_{j=1}^{J}\prod_{i=1}^{n_j} f(x_{ij}|\theta_{ij})\, d\boldsymbol{X}\left|\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{ij}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{ij})\right]\right| \\
=& \frac{1}{2}\int_{\Theta^n}\underbrace{\int_{\mathbb{X}^n}\prod_{j=1}^{J}\prod_{i=1}^{n_j} f(x_{ij}|\theta_{ij})\, d\boldsymbol{X}}_{=1}\left|\mathbb{E}\left[\underbrace{\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{ij})}_{m} - \underbrace{\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{ij})}_{m^{(K,L)}}\right]\right| \\
=& \frac{1}{2}\int_{\Theta^n}\left|\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{ij}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{ij})\right]\right| = d_{TV}\left(m, m^{(K,L)}\right) \\
=& \sup_{A_{ij}\in\mathbb{X}^n}|m(A_{ij}) - m^{(K,L)}(A_{ij})| = \sup_{A_{ij}\in\mathbb{X}^n}\left|\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{ij}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{ij})\right]\right|.
\end{aligned}$$

Now notice that, in general, $\sup_A f(A,x) \geq f(A,x) \implies \int \sup_A f(A,x)dx \geq \int f(A,x)dx \implies \int \sup_A f(A,x)dx \geq \sup_A \int f(A,x)dx$. Then

$$\begin{aligned}
&\leq \sup_{A_{ij}\in\mathbb{X}^n}\mathbb{E}\left[\left|\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{ij}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{ij})\right|\right] \\
&\leq \mathbb{E}\left[\sup_{A_{ij}\in\mathbb{X}^n}\left|\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{ij}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{ij})\right|\right].
\end{aligned}$$

Then we apply (*), obtaining

$$\leq \mathbb{E}\left[\sup_{A_{ij}\in\mathbb{X}^n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left|G_j(d\theta_{ij}) - G_j^{(K,L)}(d\theta_{ij})\right|\right] \leq \mathbb{E}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sup_{A_{ij}\in\mathbb{X}^n}\left|G_j(d\theta_{ij}) - G_j^{(K,L)}(d\theta_{ij})\right|\right]$$

where we recognize $\sup_{A_{ij}\in\mathbb{X}^n}\left|G_j(d\theta_{ij})-G_j^{(K,L)}(d\theta_{ij})\right|=d_{TV}\left(G_j,G_j^{(K,L)}\right)$, so

$$
\begin{aligned}
=&\mathbb{E}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}d_{TV}\left(G_j,G_j^{(K,L)}\right)\right]=\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbb{E}\left[d_{TV}\left(G_j,G_j^{(K,L)}\right)\right]\\
\leq& n\left[\left(\frac{\beta}{1+\beta}\right)^L+\left(\frac{\alpha}{1+\alpha}\right)^K\right].
\end{aligned}
$$

2.C Summary of Differences between nDP, HDP and CAM

Level	HDP	nDP	CAM
Data	$F_j(\cdot\|\phi)=\int_\Theta p(\cdot\|\theta,\phi)G_j(d\theta)$	$F_j(\cdot\|\phi)=\int_\Theta p(\cdot\|\theta,\phi)G_j(d\theta)$	$F_j(\cdot\|\phi)=\int_\Theta p(\cdot\|\theta,\phi)G_j(d\theta)$
I-LL	$G_j(\cdot)\sim\mathrm{DP}\left(\alpha G_0\right)$ $G_j(\cdot)=\sum_l w_l^*(\alpha)\delta_{\theta_l^*}(\cdot)$ $\theta_l^*\sim G_0$	$G_j(\cdot)\sim Q$	$G_j(\cdot)\sim Q$
II-L	$G_0=\sum_k w_k^*(\beta)\delta_{\theta_k^*}(\cdot)$ $\theta_k^*\sim H$	$Q=\sum_k \pi_k^*(\alpha)\delta_{G_k^*}(\cdot)$ $G_k^*(\cdot)=\sum_l w_{lk}^*(\beta)\delta_{\theta_{lk}}^{*(\cdot)}$ $\theta_{lk}\sim H$	$Q=\sum_k \pi_k^*(\alpha)\delta_{G_k^*}(\cdot)$ $G_k^*(\cdot)=\sum_l w_{lk}^*(\beta)\delta_{\theta_l}^{*(\cdot)}$ $\theta_l\sim H$

2.D Densities of the three scenarios considered in the simulation study

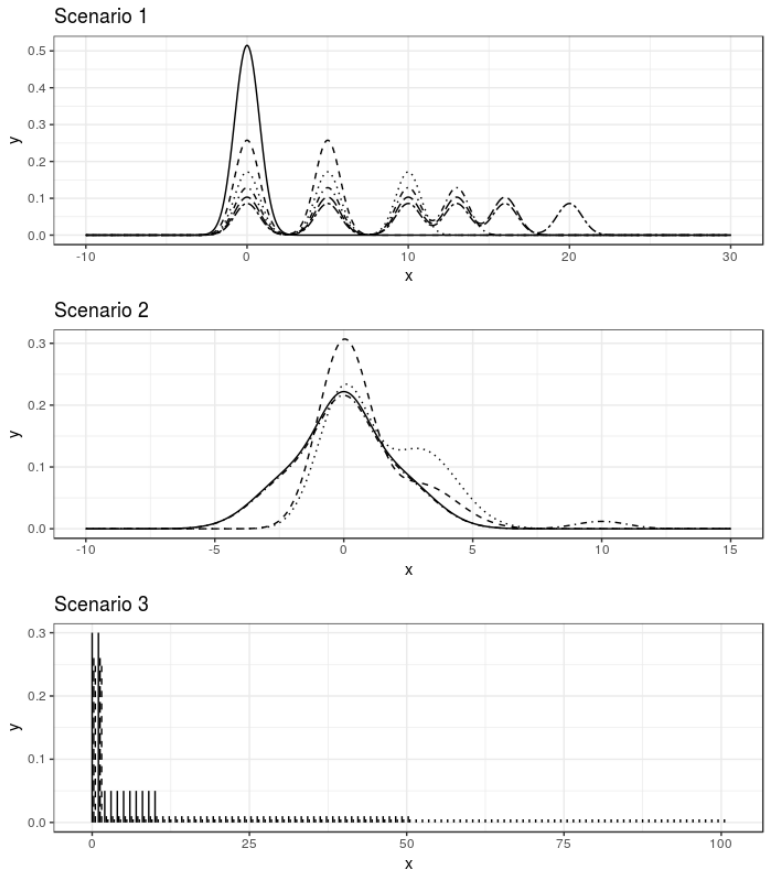


FIGURE 2.D.1: The densities distributions of each unit in three scenarios considered.

Chapter 3

Two-group Poisson-Dirichlet mixtures for multiple testing

"Ogni volta che spiego la geometrica agli studenti, dico sempre che è come la vita: il numero di possibili fallimenti è infinito."
"Sì, ma lo è anche il numero di prove!"
L'ottimo Bvignone, saggio collega

"People love to say, 'Give a man a fish, and he'll eat for a day. Teach a man to fish, and he'll eat for a lifetime.' What they don't say is, 'And it would be nice if you gave him a fishing rod.' That's the part of the analogy that's missing."
Born a crime – Trevor Noah

Abstract

The simultaneous testing of multiple hypotheses is common to the analysis of high-dimensional data sets. The two-group model, first proposed in Efron (2004), identifies significant comparisons by allocating observations to a mixture of an empirical null and an alternative distribution. In the Bayesian nonparametrics literature, many approaches have suggested using mixtures of Dirichlet Processes in the two-group model framework. Here, we investigate employing instead mixtures of two-parameter Poisson Dirichlet Processes (2PPD), and show how they provide a more flexible and effective tool for large-scale hypothesis testing. Our model further employs non-local prior densities to allow separation between the two mixture components. We obtain a closed form expression for the exchangeable partition probability function of the two-group model, which leads to a straightforward MCMC implementation. We compare the performances of our method for large-scale inference in a simulation study and illustrate its use on a case-control microbiome study of the gastrointestinal tracts in children from underdeveloped countries who have been recently diagnosed with moderate to severe diarrhea.

3.1 Introduction

The availability of high-dimensional data in domains as diverse as genomics, imaging, and astronomy, has brought the necessity to screen a large number of hypotheses simultaneously. Here, we focus on the two-group modeling framework (Efron, 2004; Efron, 2008). To illustrate, we assume that the observations are suitably defined difference scores $z_i, i = 1, \ldots, n$ over a large number of distinct hypotheses. The two-group model assumes that the z_i's are drawn either from a null (f_0) or a non-null (f_1) distribution, i.e., each score is drawn from a mixture,

$$z_i \sim f = (1 - \rho) f_0 + \rho f_1, \tag{3.1}$$

for some weight $\rho \in (0, 1)$, and some probability (density) functions f_0 and f_1. The null component is typically assumed standard normal; however, the true null distribution may differ from the theoretical null, e.g., due to limited sample size or unaccounted correlation. Thus, Efron proposes the estimation of an "empirical null" distribution to adequately capture the range of parameter values coherent with the null hypothesis and accordingly evaluate each testing decision.
In Bayesian nonparametrics, the Dirichlet process (DP) has been extensively used to provide flexible estimates of f_0, or f_1, or both, as well as for clustering the z_i's into common "expression" levels (Do et al., 2005; Dahl and Newton, 2007; Kim et al., 2009; Kottas and Fellingham, 2012). Martin and Tokdar (2012) develop a flexible hierarchical nonparametric approach where f_0 is assigned a Normal distribution with unknown mean and variance, whereas f_1 is a location mixture of normals. One appealing feature of the two-group model is that the resulting inference is immediately amenable to interpretation in a decision theoretic framework. For example, Efron (2004) describes a local version of the false discovery rate (*local fdr*), which represents the posterior probability that a difference score z_i is generated according to the null hypothesis, $fdr(z_i) = (1 - \rho)\, f_0(z_i)/f(z_i)$. The selection of interesting scores is conducted by flagging all z_i's such that $fdr(z_i) < \alpha$, $\alpha \in (0, 1)$, allowing control of the Benjamini–Hochberg FDR (Benajmini and Hochberg, 1995) at level α. More generally, the decision problem could minimize loss functions that compound expected false positive and false negative decisions. The optimal decision would then lead to thresholding the posterior probability of the alternative (e.g., see Muller et al., 2006).

In this manuscript, we investigate the use of a mixture prior of two–parameter Poisson–Dirichlet (2PPD) processes (Pitman, 1996) in lieu of the commonly used DPs. The 2PPD process, also known as the *Pitman-Yor* process, is a generalization of the DP and is characterized by two parameters: a "concentration" parameter θ (analogous to the single parameter of the DP), and a "discount" parameter σ. The additional parameter allows for more flexible clustering behavior than the DP and can be used to tune the reinforcement mechanism of large clusters (Lijoi et al., 2007). We show how the proper choice of σ can be used to model the empirical null distribution f_0 and the uncertainty related to the non-null distribution in the two-group model, leading to improved testing procedures. Our modeling framework further employs non-local prior densities for the base measure of the random probability measures under the alternative hypothesis to allow better separation between the two mixture components. We derive the expression of the exchangeable partition probability function (EPPF), induced by the proposed two-group 2PPD mixture process and observe that, conditional on the assignment of the observations to the null or the alternative hypothesis, the respective random partitions are independent. This property conveniently facilitates posterior inference obtained via MCMC algorithms, which take into account the conditional independence of the partitions. By means of a simulation study, we discuss the performances of our method with respect to the commonly used mixture of DPs and existing state-of-the-art approaches for large-scale multiple comparison problems. We also

illustrate the use of the proposed 2PPD processes mixture model on a publicly available dataset from a microbiome study, where the aim was to characterize the microbial composition of the gastrointestinal tracts of children from underdeveloped countries who have been diagnosed with moderate to severe diarrhea. Our study suggests that mixture of DPs should be used with some caution in large scale multiple-testing, and that the use of 2PPD processes could lead to improved operating characteristics.

3.2 A review of the 2PPD process

In this Section we provide an overview of the 2PPD process with particular regard to its use for density estimation and its clustering properties. Let $Z_1, \ldots, Z_n$ be a sample of n data measurements (e.g. raw observations or summary statistics), drawn from a sequence of exchangeable random elements $Z_1, Z_2, \ldots$, taking value in a complete and separable metric space $\mathbb{Z}$ endowed with its Borel σ-algebra $\mathcal{Z}$. By virtue of the de Finetti representation theorem,

$$\begin{aligned} Z_i \mid \tilde{p} &\sim \tilde{p} \qquad i = 1, \ldots, n, \\ \tilde{p} &\sim Q, \end{aligned} \tag{3.2}$$

for any $n \geq 1$, and for $\tilde{p}$, a random probability measure, with distribution Q defined on the space $\mathcal{P}(\mathbb{Z})$ of probability measures on $\mathbb{Z}$. In a Bayesian framework, Q represents the prior distribution and the model is said to be parametric whenever Q degenerates on a finite dimensional subspace of $\mathcal{P}(\mathbb{Z})$; otherwise, the model is denoted as nonparametric.

Here, we consider the 2PPD process for the random probability measure $\tilde{p}$, which can be represented almost surely as an infinite mixture, i.e.,

$$\tilde{p} = \sum_{k=1}^{\infty} \tilde{w}_k \, \delta_{Y_k},$$

where δ_c denotes the point mass at c, the $\tilde{w}_k$'s are random weights obtained as $\tilde{w}_1 = V_1$ and $\tilde{w}_k = V_k \prod_{j=1}^{k-1} (1 - V_j)$, $k \geq 2$ with $V_j \overset{ind}{\sim} \text{Beta}(1-\sigma, \theta + j\sigma)$, $j \geq 1$ (Pitman, 1995; Pitman, 1996), for some $\sigma \in [0, 1)$ and $\theta > -\sigma$. The Y_k's are random locations in $\mathbb{Z}$, independent of the weights $\tilde{w}_k$'s, and assumed as random draws from a non-atomic *base measure* P^*, i.e., $Y_k \overset{i.i.d.}{\sim} P^*$, $k \geq 1$, which represents the prior expected value of the random distribution $\tilde{p}$, i.e., $\mathbb{E}[\tilde{p}(A)] = P^*(A)$ for any $A \in \mathcal{Z}$. We should note that the 2PPD process is also well defined for $\sigma < 0$ and $\theta = r|\sigma|$, with r being an integer; however, in such case the process reduces to the parametric Fisher model (Ghosal and Vaart, 2017). Hereafter, we will use $Z_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$, with $\tilde{p} \overset{d}{=} \text{2PPD}(\sigma, \theta, P^*)$, $i = 1, \ldots, n$ to indicate a sample from a 2PPD with parameters σ and θ, and base measure P^*. If $Z_1, \ldots, Z_n$ is a realization from an exchangeable sequence driven by a 2PPD process, there is a positive probability of ties, i.e., $\mathbb{P}[Z_i = Z_j] > 0$ for any $i \neq j$. This *clustering* property often motivates the use of the 2PPD process in statistical applications, e.g. to model data from heterogeneous populations.

The clustering behavior of the 2PPD process can also be investigated by considering the exchangeable partition probability function (EPPF), which defines the probability that $Z_1, \ldots, Z_n$ are partitioned into K distinct clusters with respective sizes $n_1, \ldots, n_K$. For the 2PPD process,

such probability is

$$\Pi_K^{(n)}(n_1,\dots,n_K) = \frac{\prod_{j=1}^{K-1}(\theta+j\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{K}(1-\sigma)_{n_j-1} \tag{3.3}$$

for any choice of positive integers $n_1,\dots,n_K$ such that $\sum_{i=1}^K n_i = n$, with $K \in \{1,\dots,n\}$ and $(a)_q = \Gamma(a+q)/\Gamma(a)$, for any non-negative integer q. The expression highlights how the values of the parameters σ and θ affect the clustering structure induced by the 2PPD process. It is well-known that if K_n denotes the number of distinct values recorded in a sample $Z_1,\dots,Z_n$ of an exchangeable sequence drawn according to a 2PPD(σ,θ) process, then $K_n/n^\sigma \to S_{\sigma,\theta}$ as $n\to\infty$ (almost surely) for some positive random variable $S_{\sigma,\theta}$ when $\sigma \in (0,1)$(see Theorem 3.8 in Dümbgen, 1994). When $\sigma = 0$, we recover the clustering behavior of the Dirichlet process, obtaining $K_n/\log n \to \theta$ as $n\to\infty$ (almost surely). Hence, the larger σ is, the larger the number of clusters. Moreover, σ controls the reinforcement of the partition, i.e., the ability of big clusters to attract even more observations, as highlighted by the predictive distribution of the 2PPD process,

$$\mathbb{P}[Z_{n+1}\in A \mid Z_1,\dots,Z_n] = \frac{\theta+\sigma K_n}{\theta+n-1}P^* + \sum_{j=1}^{K_n}\frac{n_j-\sigma}{\theta+n-1}\delta_{Z_j^*}(A), \tag{3.4}$$

where the probability that a new observation is assigned to an existing cluster, and assumes value Z_j^*, $j=1,\dots,K_n$, is proportional to $n_j-\sigma$. Therefore, values of σ close to 1 favor the formation of a large number of clusters, most of which are singletons (Lijoi et al., 2007).
Finally, we consider the variability of realizations from a 2PPD process around the base measure P^*. The variance of the process is $\mathrm{Var}[\tilde{p}(A)] = \frac{1-\sigma}{\theta+1}P^*(A)[1-P^*(A)]$, for any $A\in\mathcal{Z}$ and $j=0,1$. Large values of σ correspond to random probability measures which are more concentrated around the base measure P^*. Therefore, one should expect that the empirical distribution function of any sample $Z_1,\dots,Z_n$ drawn from a 2PPD process with high values of σ, $F_n(b) = \tilde{p}(\infty,b] = \sum_{k=1}^\infty \tilde{w}_k\,\delta_{Z_k^*}(\infty,b]$, would be characterized by a large number of weights $\tilde{w}_k$ of similar size. In the next Sections we will exploit these properties to guide the use of the 2PPD process in the two-group model for multiple testing.

3.3 Methods

3.3.1 A two-group 2PPD model

The different clustering behavior that the 2PPD process exhibits as a function of σ can be exploited for distinguishing between the null and alternative distributions in the two-group model. More precisely, we first rewrite model (3.2) as the two–component mixture,

$$\tilde{p} = (1-\rho)\,\tilde{p}_0 + \rho\,\tilde{p}_1, \tag{3.5}$$

where $\tilde{p}_j \sim$ 2PPD$(\sigma_j,\theta_j,P_j^*)$ represents the unknown distribution under the null and the alternative hypotheses, for $j=0$ and $j=1$, respectively. Similarly as in (3.1), the mixture weight ρ is a random variable independent of the $\tilde{p}_j$'s and takes values in $[0,1]$. We further introduce an auxiliary binary random variable γ_i, $i=1,\dots,n$, such that $Z_i\sim\tilde{p}_0$ if $\gamma_i=0$ and $Z_i\sim\tilde{p}_1$ if

$\gamma_i = 1$. Thus, conditionally on the $\gamma_i's$, we can rewrite (3.2)–(3.5) as

$$\begin{aligned} Z_i \mid \gamma_i &\overset{\text{ind}}{\sim} \tilde{p}_{\gamma_i}, \qquad i = 1, \dots, n, \\ \gamma_i \mid \rho &\overset{\text{iid}}{\sim} \text{Bernoulli}(\rho), \\ \rho &\sim \text{Beta}(a, b), \\ \tilde{p}_{\gamma_i} &\sim \text{2PPD}(\sigma_{\gamma_i}, \theta_{\gamma_i}, P^*_{\gamma_i}), \end{aligned} \tag{3.6}$$

with $\tilde{p}_0$ and $\tilde{p}_1$ independent, and assuming a Beta distribution on ρ.

We exploit the properties of the 2PPD process discussed in Section 3.3.1 and propose to specify the hyper-parameters of the null and non-null random probability measures in (3.6) as follows. In accordance with Efron's idea that the empirical null distribution should capture only small departures from the theoretical null, we let $\tilde{p}_0$ concentrate around the theoretical null. Furthermore, we assume that there's no good model *a priori* for the non-null distribution. Therefore, $\tilde{p}_1$ is allowed to vary more freely on the space of the alternative distributions. Under the null distribution, the process should encourage the creation of a large number of clusters each composed by few observations, so that the empirical distribution well approximates the theoretical null. For the non-null distribution, we should expect a more uneven distribution of the realizations. Based on those considerations, we propose to set $\sigma_0 > \sigma_1$. We will discuss how such a choice might help discriminating between the null and the alternative distribution in the multi-comparison problem.

We conclude this Section by considering the joint partition structure induced by model (3.6) for a sample $Z_1, \dots, Z_n \mid \tilde{p} \sim \tilde{p}$. Let $\Pi^{(n)}_{K,j}(n_1, \dots, n_K)$ denote the EPPF of process $\tilde{p}_j$, $j = 0, 1$, that is the probability that n observations are assigned to K different clusters of sizes $(n_1, \dots, n_K)$. For notational simplicity, we assume that

$$\Pi^{(n)}_{K+1,j}(n_1, \dots, n_{i-1}, 0, n_{i+1}, \dots, n_K, n_{K+1}) \equiv \Pi^{(n)}_{K,j}(n_1, \dots, n_K),$$

for any $j = 0, 1$ and $n_1, \dots, n_K \geq 1$ such that $\sum_{i=1}^K n_i = n$. Then the following result provides the EPPF of the mixture of 2PPD processes as below:

Proposition 1. *The EPPF associated to the mixture of 2PPD processes in* (3.6) *is given by:*

$$\begin{aligned} \Pi^{(n)}_K(n_1, \dots, n_K) = \frac{1}{(a+b)_n} \sum_{\boldsymbol{i} \in \times_{j=1}^K \{0, n_j\}} (a)_{|\boldsymbol{i}|} (b)_{n-|\boldsymbol{i}|} \times \\ \Pi^{(|\boldsymbol{i}|)}_{K_0,0}(i_1, \dots, i_K)\, \Pi^{(n-|\boldsymbol{i}|)}_{K_1,1}(n_1 - i_1, \dots, n_K - i_K) \end{aligned} \tag{3.7}$$

where $\boldsymbol{i} = (i_1, \dots, i_K)$, $|\boldsymbol{i}| = i_1 + \cdots + i_K$, $K_0 = card\{j : i_j = n_j\}$ *and* $K_1 = K - K_0$. *In case* $i_k = n_k$ *or* $i_k = 0\ \forall k$, *we assume* $\Pi^{(n)}_K(i_1, \dots, i_K) = 1$.

Direct use of (3.7) is far from trivial. Nonetheless, the expression lends itself to an interesting interpretation: conditional on the assignment of the clusters to either $\tilde{p}_0$ or $\tilde{p}_1$, the respective random partitions are still independent. This remark is useful for devising a suitable computational algorithm for posterior inference.

3.3.2 Bayesian hierarchical two-group mixture model

In many applications, the discreteness of the realizations of the 2PPD process may be considered inadequate. Thus, in lieu of (3.6), it is often common to assume for a sample $Z_1, \dots, Z_n$ a

hierarchical mixture model with continuous components, i.e.

$$Z_i|\tilde{p} \overset{iid}{\sim} \tilde{p}, \quad \text{with} \tilde{p} = (1-\rho) \int k_0(Z_i, \vartheta)\, \tilde{p}_0(d\vartheta) + \rho \int k_1(Z_i, \vartheta)\, \tilde{p}_1(d\vartheta), \tag{3.8}$$

that is the two-group model is characterized by a null and non-null distributions which are each defined as a 2PPD process mixture. Here, $f_{\tilde{p}}(Z_i)$ is the random density induced by the random probability measure $\tilde{p}$, while $k_j : \mathbb{Z} \times \Theta \to \mathbb{R}^+$, $j = 0, 1$ are general kernels such that for $\boldsymbol{\vartheta} \in \Theta$ and some σ–finite measure λ on $(\mathbb{Z}, \mathcal{Z})$ one has $\int_{\mathbb{Z}} k_j(x, \boldsymbol{\vartheta})\, \lambda(\mathrm{d}x) = 1$, $j = 0, 1$. For our purposes, it is convenient to set $\mathbb{Z} = \mathbb{R}$ and let λ coincide with the Lebesgue measure on $\mathbb{R}$ so that the previous model defines a prior on the space of density functions on $\mathbb{R}$. By conditioning on the auxiliary group indicator variables γ_i, $i = 1, \ldots, n$, we can rewrite model (3.8) as a hierarchical Bayes *two-group 2PPD process mixture*,

$$\begin{aligned} Z_i \mid \boldsymbol{\vartheta_i}, \gamma_i &\overset{\text{ind}}{\sim} k_{\gamma_i}(Z_i \mid \boldsymbol{\vartheta}_i), \qquad i = 1, \ldots, n \\ \boldsymbol{\vartheta}_i \mid \gamma_i, \tilde{p} &\overset{\text{ind}}{\sim} \tilde{p}_{\gamma_i}, \\ \gamma_i \mid \rho &\overset{\text{iid}}{\sim} \text{Bernoulli}(\rho), \\ \rho &\sim \text{Beta}(a, b) \\ \tilde{p}_{\gamma_i} &\sim \text{2PPD}(\sigma_{\gamma_i}, \theta_{\gamma_i}, P^*_{\gamma_i}), \end{aligned} \tag{3.9}$$

where $\boldsymbol{\vartheta}_i$ may indicate either a scalar or a vector parameter. In general, $k_0(\cdot)$ and $k_1(\cdot)$ could be different. Here, we assume $k_{\gamma_0}(\cdot) = k_{\gamma_1}(\cdot) = k(\cdot)$ to be a Normal kernel and set $\boldsymbol{\vartheta_i} = (\mu_i, \tau_i^2)$. For notational simplicity, in (3.9) we have omitted additional hyper-parameters which may feature in the kernel function $k(\cdot)$ but are not relevant for the decision problem and thus are assigned separate priors.

We conclude the specification of the two-group model (3.9) by discussing the choice of the base measures P_0^* and P_1^*. On the one hand, we achieve flexible estimation of the so-called "empirical null" distribution by setting

$$P_0^*\left(\mu, \tau^2\right) = \pi(\mu) \times \pi(\tau) = N(0, 1) \times IG(a_0, b_0).$$

where the parameters of the IG on τ^2 are chosen so to allow relatively small deviations from the theoretical null distribution. For example, by assuming $a_0 = 5$, $b_0 = .2$, the induced marginal distribution on Z_i has only slightly fatter tails than the standard normal.

Moreover, P_0^* and P_1^* should not have significantly overlapping supports, i.e. they should assign high probability to regions of the parameter space that are consistent with the null and the alternative hypotheses, respectively. In the Bayesian multiple hypotheses testing framework, this requirement has sometimes been advocated to ensure enough *separation* between the null and the alternative models. Thus, we first model P_1^* as a symmetric bimodal mixture of Normal-Inverse Gamma (NIG) distributions, as $P_1^* = \frac{1}{2}\text{NIG}(-|m_1|, k_1, \alpha_1, \beta_1) + \frac{1}{2}\text{NIG}(|m_1|, k_1, \alpha_1, \beta_1)$, with $m_1 \in \mathbb{R}$, and $k_1, \alpha_1, \beta_1 \in \mathbb{R}^+$. Marginally,

$$\pi(\mu_1|m_1) = \frac{1}{2}\left[\sqrt{\frac{\beta_1}{\alpha_1}} t_{2\alpha_1} - |m_1|\right] + \frac{1}{2}\left[\sqrt{\frac{\beta_1}{\alpha_1}} t_{2\alpha_1} + |m_1|\right].$$

We further achieve separation in the multiple hypotheses testing problem by modeling the location parameter m_1 with a non-local prior (NLP), i.e. a prior that assigns vanishing density to small neighborhoods of the null hypothesis (Johnson and Rossell, 2010). Several types of NLP have been proposed in the literature. See, for instance, Johnson and Rossell (2012) and Rossell

and Telesca (2017). Here, we adopt a r-th moment prior for m_1, with

$$\pi_{MOM}\left(m_1; 0, \kappa^2, r\right) = \frac{m_1^{2r}}{\xi} \frac{e^{-m_1^2/2\kappa^2}}{\sqrt{2\pi\kappa^2}}, \tag{3.10}$$

where ξ is the normalizing constant, and we write $m_1 \sim NLP_{MOM}\left(0, \kappa^2, r\right)$. Specific hyper-parameter specifications will be detailed in Section 3.4. Here, we only note that our choice of base measures favors large effect sizes for the alternative hypothesis. The non-local prior specification in P_1^* should provide enough separation at the origin to ensure good estimation of the posterior probability of the alternative.
Finally, the other parameters of the 2PPD processes are set such that $\theta_0 = \theta_1$ and $\sigma_0 > \sigma_1$. In general, θ_0 and θ_1 are chosen relatively small, in order to enforce coarser clustering structures, especially under the alternative hypothesis. Typically, in Dirichlet-Process two-groups models, $\theta_0 = \theta_1 = 1$ (see, e.g. Do et al., 2005). From the discussion at the end of Section 3.2, it follows that realizations of the 2PPD null process are expected to be more concentrated around the base measure. In the next Sections we will investigate the effect of different choices for the parameter values of the 2PPD processes for the multiple comparison problem.

3.3.3 Posterior inference

Posterior inference for model (3.6) or (3.9) relies on Markov Chain Monte Carlo techniques, since the posterior distributions are not available in closed form. Our primary interest is in the group indicators γ_i's, which uniquely identify the random probability measure from which the data Z_i's were generated, and, correspondingly, the probability of group membership, ρ. For the sampling of the γ_i's, we exploit the independence of the random partitions implied by the EPPF (3.7) of the proposed mixture of 2PPD processes. More specifically, if $Z_1, \ldots, Z_n$ are a random sample from (3.6) and P_j^*, $j = 0, 1$ are non-atomic base measures with common support, then $\mathbb{P}[Z_i = Z_j \mid \gamma_i \neq \gamma_j] = 0$ for $i \neq j$. Thus, all the Z_i's in a cluster are generated by the same 2PPD process. The details of the MCMC algorithms are provided in the Appendix. In particular, we employ a split-merge move to speed up computations for large sample sizes (Dahl, 2003; Dahl, 2005). The computational burden of the MCMC algorithm increases for higher values of either θ_0, θ_1, σ_0 or σ_1 due to the increased number of latent clusters generated by the 2PPD process. A discussion of the computational efficiency of a plain Polya-Urn sampler versus the split-merge implementation is also provided in the Appendix.
Posterior inference on the weight ρ in (3.6) is conducted by means of post-MCMC analysis, by approximating the posterior expected value $\mathbb{E}[\rho \mid \text{data}]$ using auxiliary indicators, say $\boldsymbol{\gamma}_t^* = (\gamma_{1,t}^*, \ldots, \gamma_{K^{(t)},t}^*)$, which denote if cluster $k \in 1, \ldots K^{(t)}$ at iteration $t = 1, \ldots, T$ is a realization from $\tilde{p}_0$ or $\tilde{p}_1$. More precisely, if we denote by $B < T$ the burn-in period of the chain, we can compute the following Monte Carlo approximation of the posterior expected value $\mathbb{E}[\rho \mid \text{data}] \approx \frac{1}{T-B} \sum_{t=B+1}^{T} \frac{a + \sum_{k=1}^{K} n_{k,t}(1-\gamma_{k,t}^*)}{a+b+n}$.
Similarly, the posterior probability that an observation belongs to the non-null group can be obtained from the MCMC output as $PP_i^1 = p(\gamma_i = 1 \mid \text{data}) \approx \frac{1}{T-B} \sum_{t=B+1}^{T} \gamma_{i,t}$, where the $\gamma_{i,t}$'s indicate the MCMC draws of the component indicators γ_i's. Then, a score Z_i's is considered significant if the corresponding PPI_i^1 is larger than a threshold, say κ, chosen to control the Bayesian FDR at a pre-assigned $\alpha \times 100\%$ level , $BFDR(\kappa) = \frac{\sum_{\nu=1}^{V}(1-\mathrm{PP}_{\mathrm{i}}^1) I(\mathrm{PP}_{\mathrm{i}}^1 > \kappa)}{\sum_{\nu=1}^{V} I(\mathrm{PP}_{\mathrm{i}}^1 > \kappa)} < \alpha$ (Newton et al., 2004; Muller et al., 2006).

3.4 Applications

3.4.1 Simulation study

We investigate the performances of the Bayesian hierarchical 2PPD mixture modeling framework described in (3.8)–(3.9) for large-scale multiple hypothesis testing by means of a simulation study under $S = 5$ scenarios. More specifically, we simulate z-scores from mixture (3.1), where $f_0(x) = N(x \mid 0, \sigma_s^2)$. We set $\sigma_s^2 = 1$ for $s = 1, \dots 4$. For the fifth scenario, we set $\sigma_5^2 = 1.5$ to model the effect of hidden correlation among observations and association with unobserved covariates, that may lead to departures from standard gaussianity. For f_1 we choose:

- **Scenario 1**: $f_1(x) = 0.67 \cdot \mathcal{N}(x| -3, 2) + 0.33 \cdot \mathrm{N}(x|3, 2)$,
- **Scenario 2**: $f_1(x) = \mathcal{N}(x \mid u, 1)$ with $u \sim \mathrm{Uniform}(2, 4)$,
- **Scenario 3**: $f_1(x) = \mathcal{N}(x \mid u, 1)$ with $u \sim \mathrm{Uniform}([-4, -2] \cup [2, 4])$,
- **Scenario 4**: $f_1(x) = \mathrm{Gamma}((-1)^v \cdot x \mid a, b)$ with $a = 4$, $b = 1$ and $v \sim \mathrm{Bernoulli}(0.5)$,
- **Scenario 5**: $f_1(x) = 0.5 \cdot \mathcal{N}(5, 1) + 0.5 \cdot \mathcal{N}(-5, 1)$.

i.e. f_1 is assumed asymmetric unimodal (scenario 1), symmetric bimodal (scenarios 2 and 5), asymmetric bimodal (scenario 3) and symmetric bimodal with fat tails (scenario 4), thus mimicking typical high-dimensional testing situations. An illustrative plot of data generated under the five scenarios is provided in the Appendix. In all scenarios, we set $\rho = 0.05$, since typically only a small proportion of the comparisons is expected to be significant in large-scale inference hypothesis testing. Each simulation includes $n = 1,000$ simulated scores and is replicated 30 times to allow quantification of posterior uncertainty and of the frequentist operating characteristics of the testing procedures.

For model fitting, we employ the mixture model (3.8)–(3.9), where we assume $k(\cdot \mid \theta_i) = \mathrm{Normal}(\cdot \mid \boldsymbol{\vartheta}_i)$, with $\boldsymbol{\vartheta}_i = (\mu_i, \tau_i^2)$. The base measure of the 2PPD process $\tilde{p}_0$ is chosen as described in Section 3.3.2, with $a_0 = 5$, $b_0 = .2$. For P_1^*, we set $k_1 = 1/3$, $\alpha_1 = 1$, $\beta_1 = 1$. A NLP_{MOM} prior is assumed for m_1, with $r = 3$ and $\kappa = 2$. For the parameters characterizing the clustering behavior of the 2PPD process priors, we investigate the effect of different choices of (σ_0, σ_1) on the inference, with $\sigma_0 > \sigma_1$. More specifically, here we report the inference for the following values for the pair (σ_0, σ_1): $(0.75, 0)$, which corresponds to assuming a DP on the non-null component; in addition to $(0.75, 0.1)$, $(0.75, 0.25)$, $(0.9, 0.25)$ to investigate the effect of decreased prior uncertainty, $Var(\tilde{p})$, on the components of the two-group 2PPD mixture. We further set the concentration parameters $\theta_0 = \theta_1 = 1$ (Do et al., 2005). For the Beta prior on ρ, we set $a = 1$ and $b = 9$. For each dataset, the MCMC algorithm was run for 2,500 iterations after a 2,500 iterations burn-in period. The evaluation of posterior convergence was conducted using standard Bayesian convergence diagnostics on the chains of the traceable parameters, m_1 and ρ, by monitoring the number of group components and by inspecting the estimated densities of the null and non-null processes.

We compare the performance of our modeling approach with five alternative methods for large-scale hypothesis testing: (a) a two-group DP mixture model, which can be seen as a special case of the modeling framework proposed here, obtained by setting $\sigma_0 = \sigma_1 = 0$, with a non-local prior on the base measure for the alternative distribution (b) the local false discovery rate of Efron (2004); (c) the Benjamini and Hochberg procedure (BH, Benajmini and Hochberg, 1995); (d) the empirical Bayes mixture model of Muralidharan (2012), which allows simultaneous estimation of the effect size and of the local false discovery rate, and (e) the empirical Bayes semi-parametric approach of Martin and Tokdar (2012).

TABLE 3.1: Simulation study: sensitivity results across different settings for σ_0 and σ_1 for the five simulation scenarios considered in Section 3.4.1 ($\rho = 0.05$). The values in the table represent the average MCC and F_1 scores, the average precision (PRE), specificity (SPEC), accuracy (ACC) and the area under the curve (AUC) of the corresponding receiver operating characteristic curve, over 30 replicates with corresponding standard deviations between brackets.

	$\sigma_0 = 0.75$						$\sigma_0 = 0.9$	
	$\sigma_1 = 0$		$\sigma_1 = 0.1$		$\sigma_1 = 0.25$		$\sigma_1 = 0.25$	
Scenario 1								
AUC	0.9095	(0.0245)	0.9143	(0.0224)	0.9201	(0.0253)	0.9200	(0.0207)
PRE	0.9775	(0.0357)	0.9773	(0.0333)	0.9713	(0.0346)	0.9776	(0.0328)
SPEC	0.9995	(0.0007)	0.9995	(0.0007)	0.9994	(0.0008)	0.9995	(0.0007)
ACC	0.9676	(0.0049)	0.9680	(0.0052)	0.9690	(0.0048)	0.9683	(0.0049)
MCC	0.5777	(0.0903)	0.5833	(0.0940)	0.6020	(0.0835)	0.5893	(0.0876)
F1	0.5197	(0.1125)	0.5269	(0.1169)	0.5520	(0.1051)	0.5342	(0.1095)
Scenario 2								
AUC	0.9526	(0.0218)	0.9563	(0.0185)	0.9581	(0.0170)	0.9523	(0.0183)
PRE	0.9710	(0.0338)	0.9725	(0.0373)	0.9680	(0.0396)	0.9712	(0.0384)
SPEC	0.9993	(0.0008)	0.9994	(0.0009)	0.9993	(0.0009)	0.9993	(0.0009)
ACC	0.9703	(0.0043)	0.9703	(0.0044)	0.9701	(0.0043)	0.9696	(0.0040)
MCC	0.6249	(0.0673)	0.6242	(0.0695)	0.6212	(0.0681)	0.6135	(0.0644)
F1	0.5809	(0.0831)	0.5796	(0.0855)	0.5771	(0.0835)	0.5665	(0.0808)
Scenario 3								
AUC	0.9335	(0.0235)	0.9401	(0.0238)	0.9477	(0.0180)	0.9452	(0.0209)
PRE	0.9721	(0.0438)	0.9714	(0.0425)	0.9682	(0.0402)	0.9772	(0.0360)
SPEC	0.9995	(0.0007)	0.9995	(0.0007)	0.9994	(0.0007)	0.9996	(0.0006)
ACC	0.9624	(0.0044)	0.9624	(0.0045)	0.9638	(0.0045)	0.9631	(0.0043)
MCC	0.5081	(0.0842)	0.5080	(0.0847)	0.5340	(0.0797)	0.5224	(0.0808)
F1	0.4320	(0.1053)	0.4320	(0.1056)	0.4659	(0.1001)	0.4489	(0.1018)
Scenario 4								
AUC	0.9552	(0.0162)	0.9627	(0.0119)	0.9685	(0.0107)	0.9661	(0.0087)
PRE	0.9787	(0.0264)	0.9736	(0.0284)	0.9532	(0.0312)	0.9657	(0.0289)
SPEC	0.9993	(0.0009)	0.9991	(0.001)	0.9984	(0.0013)	0.9988	(0.0010)
ACC	0.9789	(0.0034)	0.9792	(0.0035)	0.9797	(0.0035)	0.9794	(0.0036)
MCC	0.7513	(0.0462)	0.7554	(0.0462)	0.7625	(0.0461)	0.7572	(0.0478)
F1	0.7354	(0.0538)	0.7413	(0.0528)	0.7535	(0.0518)	0.7449	(0.0542)
Scenario 5								
AUC	0.9985	(0.0010)	0.9985	(0.0010)	0.9985	(0.0011)	0.9985	(0.0011)
PRE	0.8346	(0.0300)	0.8170	(0.0334)	0.7560	(0.0330)	0.7856	(0.0326)
SPEC	0.9898	(0.0021)	0.9885	(0.0025)	0.9832	(0.0029)	0.9859	(0.0026)
ACC	0.9886	(0.0020)	0.9875	(0.0028)	0.9831	(0.0030)	0.9855	(0.0029)
MCC	0.8951	(0.0219)	0.8832	(0.0249)	0.8529	(0.0232)	0.8694	(0.0241)
F1	0.8920	(0.0229)	0.8860	(0.0241)	0.8534	(0.0235)	0.8710	(0.0238)

For each simulation replicate, results were compared using several performance measures: the Matthews Correlation Coefficient (MCC), which can be computed from a confusion matrix as $\text{MCC} = (TP \times TN - FP \times FN)/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$, where TP,

TN, FP, and FN are the number of true positive, true negative, false positive and false negative results, respectively; the F1 score, $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$; as well as precision, specificity, accuracy and the area under the curve (AUC) of the corresponding receiver operating characteristic curve. For each simulation, we identify significant scores by controlling the Bayesian false discovery rate (Newton et al., 2004), the local false discovery rate (Efron, 2004) and the frequentist false discovery rate (Benajmini and Hochberg, 1995) at the 10% level.

TABLE 3.2: Simulation study: performance metrics for five other multiple comparison methods in the five simulation scenarios considered in Section 3.4.1 ($\rho = 0.05$). The values in the table represent the average MCC and F_1 scores, the average precision (PRE), specificity (SPEC), accuracy (ACC) and the area under the curve (AUC) of the corresponding receiver operating characteristic curve, over 30 replicates with corresponding standard deviations between brackets.

	DPmix		*local fdr*		(Benajmini and Hochberg, 1995)		(Muralidharan, 2012)		(Martin and Tokdar, 2012)	
Scenario 1										
AUC	0.9053	(0.0301)	0.8869	(0.0365)	0.9237	(0.0205)	0.9242	(0.0230)	0.9230	(0.0216)
PRE	0.1113	(0.0092)	0.9897	(0.0216)	0.9974	(0.0019)	0.9915	(0.0227)	0.9825	(0.0284)
SPEC	0.6150	(0.0432)	0.9998	(0.0004)	0.9726	(0.0044)	0.9999	(0.0004)	0.9996	(0.0006)
ACC	0.6297	(0.0396)	0.9671	(0.0042)	0.9141	(0.0603)	0.9653	(0.0043)	0.9679	(0.0044)
MCC	0.2329	(0.0209)	0.5708	(0.0721)	0.6629	(0.0648)	0.5379	(0.0804)	0.5835	(0.0757)
F1	0.1980	(0.0145)	0.5067	(0.0932)	0.6427	(0.0736)	0.4643	(0.1023)	0.5251	(0.0962)
Scenario 2										
AUC	0.9524	(0.0218)	0.9231	(0.0428)	0.9544	(0.0169)	0.9668	(0.0174)	0.9762	(0.0087)
PRE	0.1140	(0.0087)	0.9796	(0.0304)	0.9974	(0.0018)	0.9895	(0.0216)	0.9698	(0.0332)
SPEC	0.6033	(0.0364)	0.9995	(0.0008)	0.9729	(0.0042)	0.9998	(0.0004)	0.9993	(0.0008)
ACC	0.6213	(0.0339)	0.9694	(0.0048)	0.9129	(0.0556)	0.9676	(0.0040)	0.9715	(0.0044)
MCC	0.2506	(0.0180)	0.6088	(0.0773)	0.6674	(0.0638)	0.5805	(0.0667)	0.6435	(0.0686)
F1	0.2037	(0.0138)	0.5578	(0.0991)	0.6486	(0.0721)	0.5194	(0.0857)	0.6048	(0.0846)
Scenario 3										
AUC	0.9400	(0.0206)	0.9069	(0.0359)	0.9500	(0.0191)	0.9481	(0.0197)	0.9481	(0.0182)
PRE	0.1085	(0.0079)	0.9759	(0.0424)	0.9972	(0.0020)	0.9901	(0.0328)	0.9707	(0.0437)
SPEC	0.5679	(0.0349)	0.9995	(0.0008)	0.9710	(0.0040)	0.9999	(0.0005)	0.9994	(0.0010)
ACC	0.5880	(0.0330)	0.9641	(0.0050)	0.9089	(0.0561)	0.9611	(0.0043)	0.9652	(0.0044)
MCC	0.2337	(0.0195)	0.5397	(0.0854)	0.6544	(0.0578)	0.4840	(0.0883)	0.5591	(0.0740)
F1	0.1948	(0.0129)	0.4708	(0.1087)	0.6342	(0.0670)	0.3973	(0.1080)	0.4970	(0.0948)
Scenario 4										
AUC	0.9612	(0.0156)	0.9406	(0.0246)	0.9709	(0.0085)	0.9627	(0.0159)	0.9658	(0.0139)
PRE	0.1217	(0.0099)	0.9919	(0.0172)	0.9965	(0.0023)	0.9972	(0.0106)	0.9953	(0.0123)
SPEC	0.6288	(0.0345)	0.9998	(0.0005)	0.9812	(0.0036)	0.9999	(0.0003)	0.9999	(0.0004)
ACC	0.6459	(0.0325)	0.9758	(0.0033)	0.9136	(0.0502)	0.9741	(0.0034)	0.9741	(0.0032)
MCC	0.2671	(0.0194)	0.7080	(0.0474)	0.7849	(0.0443)	0.6840	(0.0486)	0.6831	(0.0470)
F1	0.2161	(0.0156)	0.6801	(0.0586)	0.7853	(0.0448)	0.6492	(0.0602)	0.6485	(0.0595)
Scenario 5										
AUC	0.9980	(0.0012)	0.7885	(0.0449)	0.9986	(0.0010)	0.8128	(0.0430)	0.8185	(0.0364)
PRE	0.3622	(0.0163)	0.9998	(0.0005)	0.6433	(0.0531)	0.9999	(0.0003)	0.9999	(0.0003)
SPEC	0.9058	(0.0155)	0.9652	(0.0036)	0.9705	(0.0067)	0.9645	(0.0037)	0.9644	(0.0036)
ACC	0.9104	(0.0385)	0.9879	(0.0258)	0.9710	(0.0064)	0.9954	(0.0185)	0.9926	(0.0234)
MCC	0.5715	(0.0361)	0.5379	(0.0661)	0.7861	(0.0361)	0.5240	(0.0685)	0.5236	(0.0667)
F1	0.5303	(0.0420)	0.4650	(0.0870)	0.7792	(0.0395)	0.4451	(0.0877)	0.4453	(0.0845)

In Table 3.1 we report the performance metrics achieved in the different simulation scenarios as a function of the combinations of hyper-parameters of the 2PPD process. Overall, the performances of the proposed 2PPD process are similar, as long as $\sigma_1 < \sigma_0$. Higher values of σ_0 lead to draw samples from f_0 which are closer to the theoretical null, but the implied tighter control of the variance of the null process may lead to a slightly decreased performance in some scenarios. If $\sigma_1 > \sigma_0$, the performances can deteriorate considerably (see Appendix).
Table 3.2 reports the results from the comparison with alternative multiple testing methods. The method of Martin and Tokdar (2012) performs quite well in all scenarios, comparably with our method, except in the fat tails Scenario 4, where our 2PPD model outperforms the competitors. The BH procedure also performs quite well, although with slightly lower specificity, in the first four scenarios. However, small departures from the standard Gaussian null assumption (scenario

5) considerably affect the performance of the BH procedure. The performance of two-group DP mixtures is impacted by the flexible modeling of both the null and alternative distribution, which leads to a relatively high number of false assignments. The result is remarkable as various types of mixture of DP processes have been often proposed for hypothesis testing in the two-group modeling framework.

3.4.2 Case study: Microbiome data

We illustrate the applicability of the proposed two-group 2PPD process model for large-scale inference on a publicly available dataset of microbial abundances from a case-controlled study on post-diarrheal disruption in children from low-income countries. Microbial community compositional data are obtained by sequencing highly variable regions of the 16S rRNA gene. Most bacterial species have specific instances of this marker gene. Therefore, the 16S rRNA gene can be used to map, at least approximately, individual sequences obtained from a sample into an individual bacterium of a given Operational Taxonomic Unit (OTU, Morgan and Huttenhower, 2012). Thus, the observed sequenced abundance is used as a proxy for OTU abundance.
The purpose of the study was to identify potential microbiota which may be responsible for exacerbating clinical conditions by showing positive associations with presence of moderate-to-severe diarrhea (MSD) in the case group. Taxa with negative associations are also of interest since they may indicate potential target treatments for recovery from dysbiosis.
Stool samples were obtained from 992 children between the ages of 0 and 59 months, 508 of whom had recently suffered from moderate to severe diarrhea, with the remaining 484 children acting as age-matched controls. The samples were obtained in Mali (M), the Gambia (G), Kenya (K), and Bangladesh (B) and case/control proportions were approximately equal from each country. We aggregate OTUs at the Species Level. In our dataset, 535 taxa have been measured and the number of non-null taxa per sample ranged from 14 to 184, with a median of 64 and an average of 64.53.
Due to the nature of the sampling mechanism, the distribution of the microbiome counts is highly skewed, i.e., there are few taxa present in very high abundances and many taxa with low frequencies (Chen and Li, 2016; Shi et al., 2016). Here, we are interested in evaluating the ability of our model to identify taxa which may differently abundant in healthy and MSD subjects. Therefore, we employ a Negative-Binomial regression model on the taxonomic abundances y_{ij}, where $j = 1, \dots, J_i$ indexes the taxa, and $i = 1, \dots, n$ indexes the observations. As it is typical when dealing with sequencing data (Anders and Huber, 2010; Witten, 2011; Love et al., 2014), we let s_i denote an estimate of a sample-specific size factor, in order to take into account the different sequencing depths of the samples. Also, we let x_{ij}^{case}, x_{ij}^{age} and $x_{ij}^{country}$ denote three available covariates for the MSD status, age and country. More specifically, $x_{ij}^{case} = 1$ for cases and $x_{ij}^{case} = 0$ for the matched controls. In this analysis, we adopt Gambia as the reference value for the other countries., and indicate with x_{ij}^{K}, x_{ij}^{B}, and x_{ij}^{M} the dummy variables for the other countries. Then, we consider the following model:

$$\begin{aligned} y_{ij} &\sim NB(\mu_{ij}, \alpha_j), \qquad j = 1, \dots, J_i;\ i = 1, \dots, n, \\ \log(\mu_{ij}) &= \log(s_i) + \beta_{0,j} + \beta_{1,j}\, x_{ij}^{case} + \beta_{2,j}\, x_{ij}^{age} + \beta_{3,j}\, x_{ij}^{M} + \beta_{4,j}\, x_{ij}^{B} + \beta_{5,j}\, x_{ij}^{K} + \epsilon_{ij}, \end{aligned}$$

where α_j represents a taxon-specific dispersion parameter, and $\beta_{0,j}$ represents a taxon-specific effect, which captures the abundance of taxon j in the control group, and the $\beta_{k,j}$'s represent the effects of each covariate on the taxon abundance. The Negative Binomial distribution was chosen due to its flexibility over the Poisson alternative. The model was fitted using the `glmmTMB` (Brooks et al., 2017) package. To illustrate our multiple testing procedure, we consider the fixed case-control effect captured by the estimates of the coefficients $\beta_{1,j}$'s, which provide the z-scores for testing the differences in abundance between healthy and MSD subjects. A histogram of the

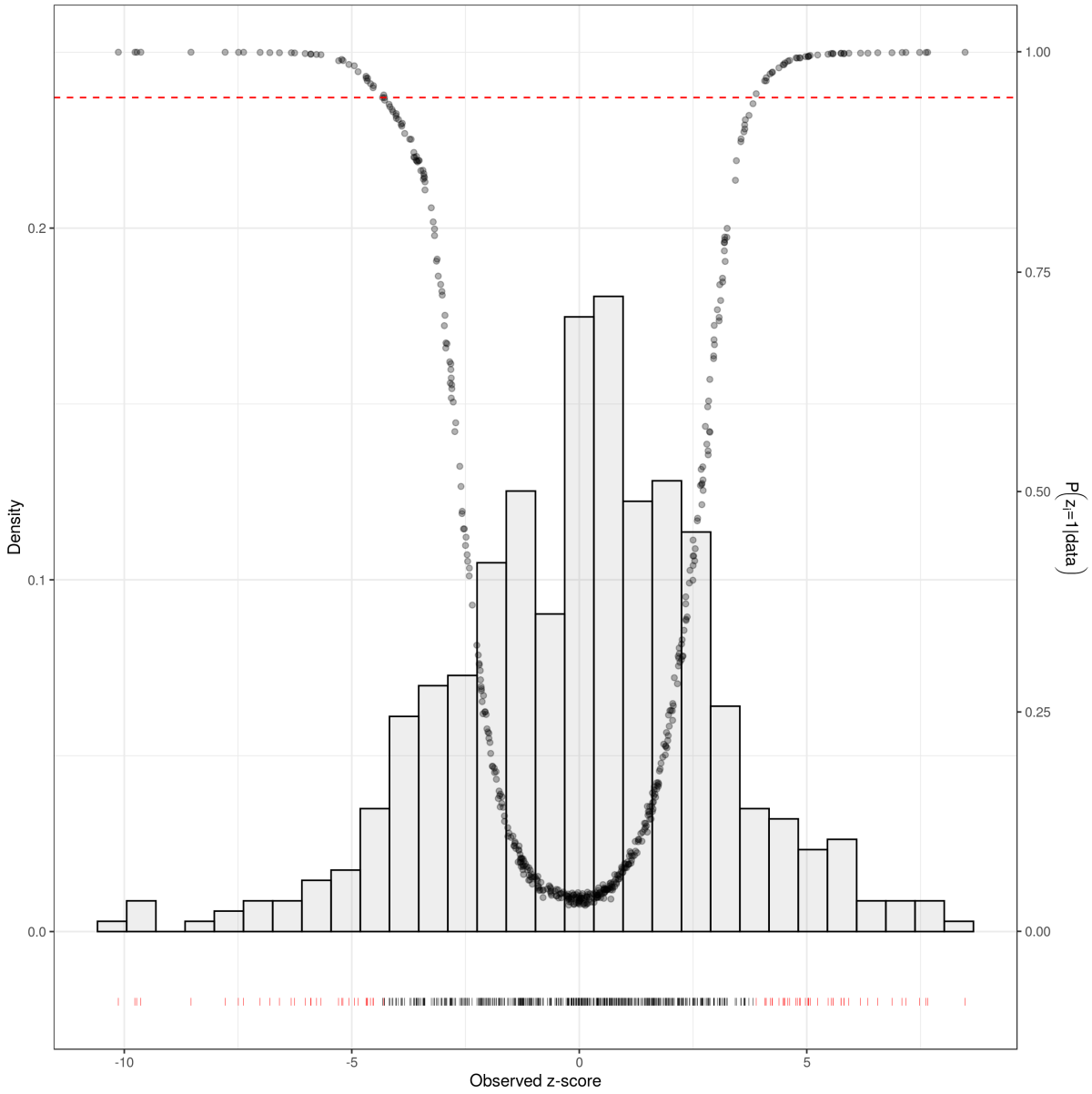


FIGURE 3.1: Microbiome data case study: Histogram of 535 z-scores obtained from the case term (β_1) in the Negative Binomial generalized linear mixed effects model. We superimpose the posterior probabilities of the events $\{\gamma_i = 1|\boldsymbol{z}\}$ and the threshold corresponding to a Bayesian FDR of 1%.

535 z-scores from the data is given in Figure 3.1. It is important to note that since the estimated coefficients are a function of the original data, the independence assumption may not be satisfied if the original taxonomic abundances are correlated. Indeed, the presence of hidden correlation among the observables, or unknown associations with unobserved covariates, has been one of the major motivations for the two-group model formulation in (Efron, 2004).

In the two-group model (3.8)–(3.9), we fix the hyper-parameters for the prior processes as $\theta_0 = \theta_1 = 1$, $\sigma_0 = 0.75$, $\sigma_1 = 0.10$ and $a = 1$ and $b = 99$. This choice allows to describe small departures from the theoretical null, while maintaining computational feasibility in the generation of the latent clusters from the null. The hyper-parameters of the base measures were set to the same values as in Section 3.4.1. For the results provided here, we run the

MCMC algorithm described in Section 3.3.3 for 20,000 iterations after a 20,000 iteration burn-in period. No evidence of lack of convergence was observed by running standard diagnostic procedures. Figure 3.1 overlays the Monte Carlo estimates of the posterior probability of each taxon belonging to the non-null distribution to the histogram of the z-scores. By thresholding the Monte Carlo estimate of posterior probability of the non-null process at a value corresponding to a Bayesian false discovery rate (Newton et al., 2004) of 1%, we identify a total of 74 non-null taxa. The *locfdr* model detects as relevant only 6 taxa. On the contrary, the BH procedure leads to 143 significant microbes, when controlling the FDR at the 1% level.Tables 1 and 2 in the Appendix report the taxa with the highest discovery probabilities, separately for positive and negative z-scores.
A close inspection of our results reveals some interesting biological findings. Among the species which were identified by our method as having significantly less abundance in MSD children, we found *Prevotella* species and *Clostridium* species (see Table 1 in the Appendix). *Prevotella* spp. are common bacteria in the gut and are commonly found in children from rural and underdeveloped areas (Di Paola et al., 2010) as well as children whose diets predominantly consist of carbohydrates and fiber (Chen et al., 2011). Thus the severely decreased abundance of the *Prevotella* spp. is reasonable in light of the gastrointestinal disruption the children experienced. As for the *Clostridium* spp., it is well-known that *C. difficile* is a toxigenic bacteria in adults but it is also found asymptomatically in large proportions in infants and neonates (Jangi and Lamont, 2010). Another interesting species is *Megasphaera*, which was recently suggested for reclassification to *Clostridium* (Yutin and Galperin, 2013). Finally, *Eubacterium rectale*, *Bacteroides*, and *Faecalibacterium prausnitzii*, have all been shown a marked reduction in concentrations in patients affected by chronic idiopathic diarrhea (Swidsinski et al., 2008).
Of the species identified as significantly more abundant in the MSD children, many of them belong to the *Streptococcus* species. Some *Streptococcus* species are well known human pathogens causing conjunctivitis, respiratory infections and urinary tract infections. Other species in the genus are opportunistic pathogens, meaning they are asymptomatically present in healthy individuals but will flourish in individuals with weakened immune systems such as the patients in this dataset. The pathogenic genus *Shigella* is present and is well-known for causing dysentery. It has been suggested that the *Shigella* spp. are closely related to another well-known pathogen, *Escherichia coli* (Lan and Reeves, 2002) which is also differentially abundant in these patients. A *Granulicatella* species has also been identified as differentially abundant. However, these bacteria are usually implicated in childhood infective endocarditis or infection of the heart.

3.4.3 Case study: Prostate Cancer Dataset

To assess how our model performs in large-sample cases, we apply our methodology also to the widely known *Prostate* dataset of Singh et al. (2002). See also Efron (2009). We introduce a split-merge move in the MCMC to reduce the computational cost. The dataset is composed of 6,033 genes for 102 observations from 52 prostate cancer patients and 50 healthy men. We adopt the same prior specification as in the microbiome case study, with the exception of choosing $b = 9$ to further encourage sparsity of discoveries. Figure 3.2 reports the posterior probabilities of discovery for this dataset. When thresholding the BFDR at the 20% level, our method flags only 18 genes as relevant. Similarly, the *locfdr* procedure flags 19 genes. On the contrary, the BH procedures identifies 60 genes as significant, when thresholding the FDR at the 10% level.

3.5 Discussion and Conclusion

We have considered the two-group model by Efron (2004) for multiple hypotheses testing and we have proposed the use of a mixture prior of two–parameter Poisson–Dirichlet processes as a flexible class of prior processes in that framework. In particular, an appropriate choice of the

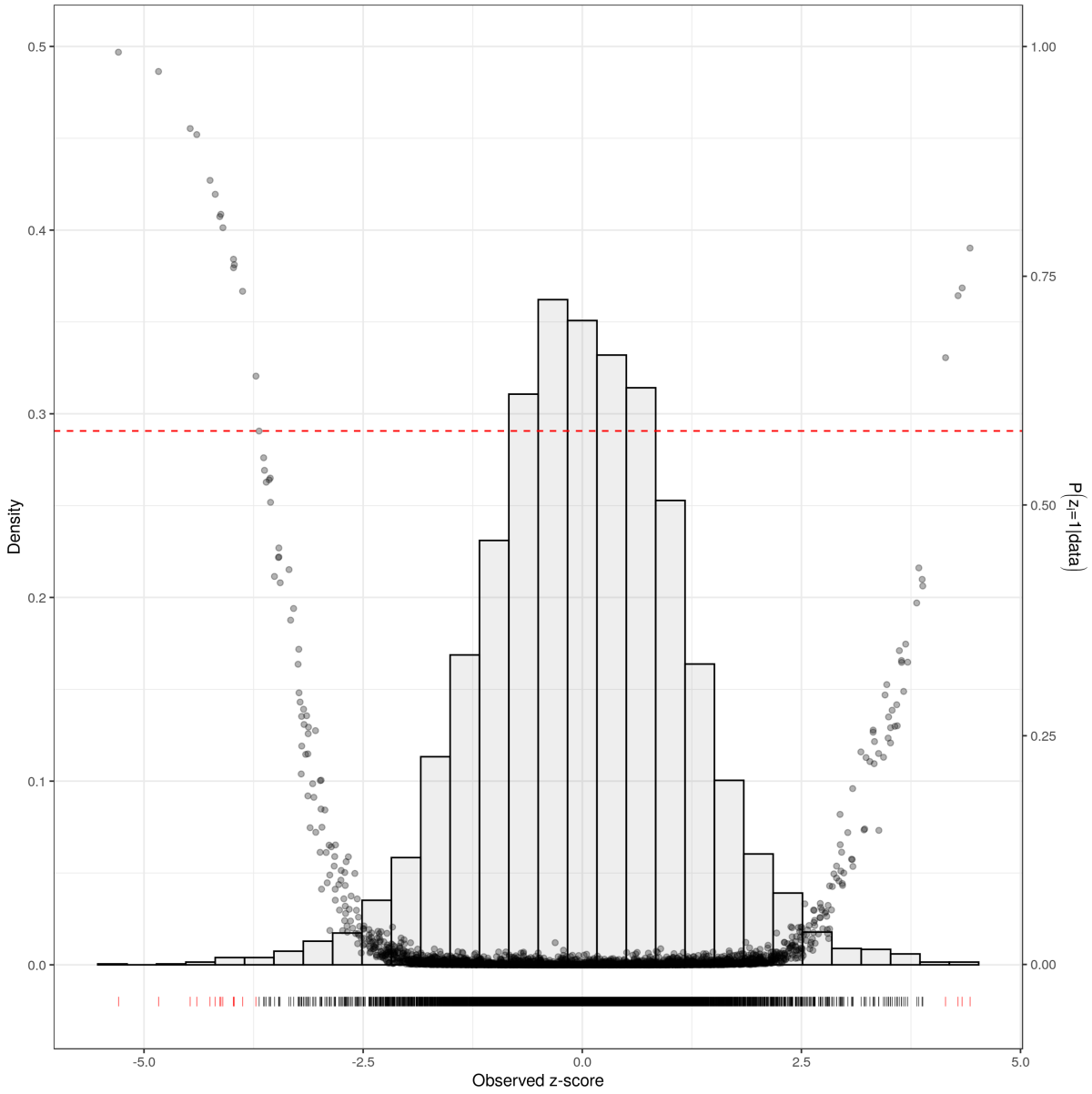


FIGURE 3.2: Prostate dataset: Histogram of 6033 z-scores obtained from a two-groups comparison. We superimpose the posterior probabilities of the events $\{\gamma_i = 1|\boldsymbol{z}\}$ and the threshold corresponding to a Bayesian FDR of 20%.

hyper-parameters of the 2PPD processes allows the characterization of small departures from the theoretical null in the estimation of the empirical null distribution, while leaving flexibility in the modeling of the non-null distribution. We have also employed a mixture of non-local prior densities as base measure for the alternative distribution, to improve separation and facilitate the estimation and identifiability of the mixture components. We have characterized the behavior of the 2PPD mixture prior by deriving its exchangeable partition probability function, and highlighted the conditional independence assumptions in the clustering of the mixture process. The proposed approach has been shown to provide a robust testing procedure, which compares favorably with recently proposed methods for estimating the components of the two-group model, including the widely-used DP mixture models. Finally, we have illustrated the use of the proposed 2PPD process mixture model on two publicly available datasets, one from a benchmark microarray study and one from a more recent microbiome study.

One limitation of our analysis of the microbiome data is related to the fact that many of the species which can now be sequenced using 16S technologies remain relatively unknown in the microbiology literature, since they cannot be cultured in the lab. Indeed, it is common for microbiome studies to assign taxa to the genus or even family level and the interpretation is often conducted at those taxonomic levels. Since the specific functions of many taxa are still poorly understood, the interpretation of findings at the species level remains difficult.
Another limitation is related to the computing effort, since Markov chain Monte Carlo algorithms for Bayesian nonparametric models typically require considerable computational time for posterior inference. While the data sizes considered in this manuscript are typical of current microbiome studies, a full MCMC approximation of the posterior distributions may become less appealing as sequencing techniques improve. Metagenomics shotgun sequencing has been increasingly adopted in lieu of 16S rRNA sequencing in human microbiome studies, thus requiring software able to handle the large amount of genomic information being sampled (Sharpton, 2014). To provide an illustration of current capabilities, in the analysis of the Prostate cancer dataset of Section 3.4.3, it took approximately 56 hours to run 20,000 MCMC iterations on a Xeon(R) E5-2640 v4, 2.40GHz Linux sever, with the computational bottleneck being represented by the iterations requiring a full Polya-Urn sampling. Variational Bayes techniques have been developed for many Bayesian nonparametric models, including the 2PPD process (see, e.g. Jordan and Blei, 2006; Sudderth and Jordan, 2009). However, the speed up of MCMC algorithms for Bayesian nonparametric models in high-dimensional settings is still a topic of ongoing research (see, e.g., Canale et al., 2019).
A careful choice of the hyperparameters of the two-group 2PPD model is essential to ensure good operating characteristics of the testing procedures. We have followed prevailing practices and set $\theta_0 = \theta_1 = 1$ in both the simulations and the data analyses. Of course, larger values of θ_0 could also be considered to improve the identification of the null distribution. However, the specification of the parameter $\theta_0 > 0$ to that purpose is not straightforward and in our experience the performances may vary considerably for different choices of the parameters. Similarly, priors on θ_0 and θ_1 would need to incorporate constraints to facilitate the identification of the two-group components.
Finally, in our data analyses, we have proposed a two-group model for the analysis of data observed under two conditions. However, often the interest is in studying longitudinal changes of repeated measurements within a subject. Therefore, models that take into account the temporal dependence of the hypotheses are required. Our model could also be employed as an alternative to Bayesian nonparametric DP mixtures for finding subsets of variables discriminating between two groups of observations (Kim et al., 2006). These directions will be explored in future research.

Appendix

3.A Posterior inference

In this section, we detail the MCMC algorithm for posterior inference for model (3.6) and model (3.8)–(3.9), with particular regards to inference on the auxiliary variable indicators γ_i's and the two-group components' weight ρ.
Before deriving the full conditional distributions, we first need to outline a few properties of the clustering implied by the EPPF (3.7). More specifically, an underlying assumption of the two-group model is that if two observations assume the same value, they should not be assigned to different groups. For the 2PPD processes, the following result holds

LEMMA 1. *If $Z_1, \dots, Z_n$ are a random sample from an exchangeable sequence $\boldsymbol{Z}$ governed by a random probability measure as defined in (6), with P_j^*, $j = 0, 1$, being non-atomic base measures with common support, then*

$$\mathbb{P}[Z_i = Z_j \mid \gamma_i \neq \gamma_j] = 0$$

for $i \neq j$.

The result follows directly from the characterization of the 2PPD process as an infinite mixture and the fact that P_j^* is non-atomic. A notable consequence of this result is that Z_i's belonging to the same cluster are generated by the same PD process and distinct clusters are generated by distinct random probability measures. In the case of the 2PPD mixture models (9), Lemma 1 applies to the atoms generated by the nonparametric priors. Given the same set of hypotheses, we have that

$$\mathbb{P}[\boldsymbol{\vartheta}_i = \boldsymbol{\vartheta}_j \mid \gamma_i \neq \gamma_j] = 0.$$

As we outline below, this result is crucial for a proper characterization of the full conditionals of each model.

- **MCMC for model (6):** At any iteration of the MCMC algorithm, the vector of observations $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ is partitioned into K separate clusters, $K \geq 1$. Let $Z_1^*, \dots, Z_K^*$ denote the $K \leq n$ unique values in $\boldsymbol{Z}$. We denote the corresponding partition sets by $C_{k,n} = \{i : Z_i = Z_k^*\}$, $k = 1, \dots, K$, and by $n_k = |C_{k,n}|$, the cardinality of each set. By virtue of Lemma 1, two observations assigned to the same cluster are also assigned to the same random probability measure. Therefore, let γ_k^* be an auxiliary random variable such that $\gamma_k^* = 0$ if the partition set $C_{k,n}$ contains draws from $\tilde{p}_0$ and $\gamma_k^* = 1$ otherwise. Then, for any $i \in C_{k,n}$, one has $\gamma_i = \gamma_k^*$, and, conditional on the partition sets $C_{k,n}$, $k = 1, \dots, K$, the K-tuple $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_K^*) \in \{0, 1\}^K$ describes the solution of the multiple testing problem, analogously to the vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n) \in \{0, 1\}^n$. Then, posterior samples for $\boldsymbol{\gamma}$ can be immediately derived from the posterior samples of the vector $\boldsymbol{\gamma}^*$ and the configuration of the partition sets $C_{k,n}$, $k = 1, \dots, K$, which can be obtained by means of a Gibbs sampling scheme.

More specifically, let us consider the joint probability distribution of the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$, and the partition $\boldsymbol{C}_n = \{C_{1,n}, \ldots, C_{K,n}\}$,

$$\mathcal{L}(\boldsymbol{\gamma}, C_{1,n}, \ldots, C_{K,n}) = \mathcal{L}(\boldsymbol{\gamma})\, \mathcal{L}(C_{1,n}, \ldots, C_{K,n} \mid \boldsymbol{\gamma}).$$

The joint distribution of $\boldsymbol{\gamma}$ can be obtained as

$$\mathcal{L}(\boldsymbol{\gamma}) = \frac{(a)_{|\boldsymbol{\gamma}|}(b)_{n-|\boldsymbol{\gamma}|}}{(a+b)_n}, \tag{3.11}$$

where $|\boldsymbol{\gamma}| = \sum \gamma_i$. Here, $|\boldsymbol{\gamma}|$ indicates the number of observation currently assigned to the non-null process. By Lemma 1 and Proposition 1, the conditional distribution of $C_{1,n}, \ldots, C_{K,n}$ given $\boldsymbol{\gamma}$ can be written as

$$\begin{aligned} \mathcal{L}(C_{1,n}, \ldots, C_{K,n} \mid \boldsymbol{\gamma}) &= \mathcal{L}(C_{1,n}, \ldots, C_{K,n} \mid \boldsymbol{\gamma}^*) \\ &= \Pi_{K_0,0}^{(n-|\boldsymbol{\gamma}|)}(n_1(1-\gamma_1^*), \ldots, n_K(1-\gamma_K^*)) \\ &\times\ \Pi_{K_1,1}^{(|\boldsymbol{\gamma}|)}(n_1\gamma_1^*, \ldots, n_K\gamma_K^*) \prod_{k=1}^{K} \prod_{i \in C_{k,n}} \mathbb{1}_{\{\gamma_k^*\}}(K_i), \end{aligned} \tag{3.12}$$

where $K_1 = \sum_{k=1}^{K} \gamma_k^*$ and $K_0 = K - K_1$ indicate the number of clusters belonging to $\tilde{p}_1$ and to $\tilde{p}_0$, respectively. Expression (3.11) and (3.12) emphasize that it is sufficient to consider only the cluster-based vector of indicators $\boldsymbol{\gamma}^*$ in order to determine the joint probability distribution of the $\boldsymbol{\gamma}$ and the partition $\boldsymbol{C}_n$. Therefore, in order to obtain the full conditional distribution of the γ_i's, we may focus only on the γ_k^*'s and the vector of $K \leq n$ observations in $\boldsymbol{Z}^*$. Let $\boldsymbol{\gamma}_{-k}^*$ denote the $\boldsymbol{\gamma}^*$ vector with the kth entry removed. Similarly, we define $\boldsymbol{\gamma_{-i}}$ as the vector where the ith entry is deleted. This implies that $\boldsymbol{\gamma}_{-k}^* \in \{0,1\}^{K-1}$ and that $\boldsymbol{\gamma}_{-i} \in \{0,1\}^{n-1}$. Also let $n_{-k,1} = \sum_{l \neq k}^{K} n_l \gamma_l^*$ denote the number of observations not included in $C_{k,n}$ which at the same time come from the non-null process. Furthermore, for any $k = 1, \ldots, K$, let $K_{-k,1} = |\boldsymbol{\gamma}_{-i}^*| = \sum_{l \neq k}^{K} \gamma_l^*$ indicate the number of clusters assigned to the non-null distribution $\tilde{p}_1$ after removing cluster $C_{k,n}$.

We can now write the full conditional of the γ_k^*'s. For notational simplicity, let $\gamma_k^* = \xi$, where $\xi \in \{0,1\}$. Then, the full conditional $\mathcal{L}_k(\xi \mid \boldsymbol{\gamma}_{-k}^*, \boldsymbol{Z}^*, \boldsymbol{C}_n) \propto p_{k,\xi}$ where

$$\begin{aligned} p_{k,\xi} &= (a)_{n_{-k,1}+n_k\xi}(b)_{n-n_{-k,1}-n_k\xi}\, P_\xi^*(\mathrm{d}Z_k^*) \prod_{j \neq k} P_{\gamma_j^*}^*(\mathrm{d}Z_j^*) \prod_{l \in C_{k,n}} \mathbb{1}_{\{\gamma_k^*\}}(\gamma_l) \\ &\times\ \Pi_{K-K_{-k,1}-\xi,0}^{(n-n_{-k,1}-n_k\xi)}(n_1(1-\gamma_1^*), \ldots, n_k(1-\xi), \ldots, n_K(1-\gamma_K^*)) \\ &\times\ \Pi_{K_{-k,1}+\xi,1}^{(n_{-k,1}+n_k\xi)}(n_1\gamma_1^*, \ldots, n_k\xi, \ldots, n_K\gamma_K^*). \end{aligned}$$

In particular, the probability $P(\gamma_k^* = 1 \mid \boldsymbol{\gamma}_{-k}^*, \boldsymbol{Z}^*, \boldsymbol{C}_n)$ determines the probability that for any $i \in C_{k,n}$, the observations $Z_i = Z_k^*$ are assigned to the non-null distribution, and can be obtained as

$$\mathcal{L}_k(\xi = 1 \mid \boldsymbol{\gamma}_{-k}^*, \boldsymbol{Z}^*, \boldsymbol{C}_n) = \frac{1}{1 + (p_{k,0}/p_{k,1})},$$

where the ratio

$$\frac{p_{k,0}}{p_{k,1}} = \frac{P_0^*(\mathrm{d}Z_k^*)}{P_1^*(\mathrm{d}Z_k^*)} \frac{\Gamma(a + n_{-k,1})}{\Gamma(a + n_{-k,1} + n_k)} \frac{\Gamma(b + n - n_{-k,1})}{\Gamma(b + n - n_{-k,1} - n_k)}$$

$$\times \frac{\Pi^{(n-n_{-k,1})}_{K-K_{-k,1},0}(n_1(1-\gamma^*_1),\ldots,n_k,\ldots,n_K(1-\gamma^*_K))}{\Pi^{(n-n_{-k,1}-n_k)}_{K-K_{-k,1}-1,0}(n_1(1-\gamma^*_1),\ldots,n_K(1-\gamma^*_K))} \times \frac{\Pi^{(n_{-k,1})}_{K_{-k,1},1}(n_1\gamma^*_1,\ldots,n_K\gamma^*_K)}{\Pi^{(n_{-k,1}+n_k)}_{K_{-k,1}+1,1}(n_1\gamma^*_1,\ldots,n_k,\ldots,n_K\gamma^*_K)}.$$

The previous expression is valid for all normalized random measures with independent increments. If $\tilde{p}_j$ is a 2PPD$(\sigma_j, \theta_j, P^*_j)$, $j = 0, 1$ process, then the above ratio further simplifies as:

$$\frac{p_{k,0}}{p_{k,1}} = \frac{P^*_0(\mathrm{d}Z^*_k)}{P^*_1(\mathrm{d}Z^*_k)} \frac{(b+n-n_{-k,1}-n_k)_{n_k}}{(a+n_{-k,1})_{n_k}} \frac{(1-\sigma_0)_{n_k-1}}{(1-\sigma_1)_{n_k-1}} \times \frac{\theta_0+(K-K_{-k,1}-1)\,\sigma_0}{\theta_1+K_{-k,1}\,\sigma_1} \frac{(\theta_1+n_{-k,1})_{n_k}}{(\theta_0+n-n_{-k,1}-n_k)_{n_k}}. \quad (3.13)$$

It is worth noting that if the two base measures coincide, i.e., $P^*_0 = P^*_1$, then the full conditional does not depend on Z^*_k; therefore, the probability that Z^*_k is a draw from the non-null depends only on the clustering behavior of the 2PPD process implied by the parameters characterizing $\tilde{p}_0$ and $\tilde{p}_1$.

To summarize, in order to implement a Gibbs sampler for sampling the auxiliary indicators in (6), at each iteration $t = 1, \ldots, T$, we draw each γ^*_k, from the full conditional $\mathcal{L}_k(\xi = 1 \mid \boldsymbol{\gamma}^*_{-k}, \boldsymbol{Z}^*, \boldsymbol{C}_n)$. The vectors $\boldsymbol{\gamma}^*_t$, $t = 1, \ldots, T$, can then be mapped to the vector $\boldsymbol{\gamma}$ using the partition sets $\boldsymbol{C}_n$.

- **MCMC for model (8)–(9):** The full conditionals for the vectors γ^*_k are derived in a similar way as above. More specifically, for $\gamma^*_k = \xi$, let $\vartheta^*_{k,\xi} \sim P^*_\xi$, indicate an atom of $\tilde{p}_\xi$, $\xi \in \{0, 1\}$, $k = 1, \ldots K$. Then,

$$\begin{aligned}
&\mathcal{L}[\xi \mid \boldsymbol{\gamma}^*_{-k}, C_{1,n}, \ldots, C_{K,n}, \boldsymbol{Z}] \\
&\quad \propto \left\{ \int \prod_{l \in C_{k,n}} k_\xi\left(\boldsymbol{Z}_l | \boldsymbol{\vartheta}^*_{k,\xi}\right) P^*_\xi(\mathrm{d}\boldsymbol{\vartheta}^*_{k,\xi}) \right\} \Gamma(a+n_{-k,1}+n_k\xi)\Gamma(b+n-n_{-k,1}-n_k\xi) \\
&\qquad \times \Pi^{(n-n_{-k,1}-n_k\xi)}_{K-K_{-k,1}-\xi,0}(n_1(1-\gamma^*_1),\ldots,n_k(1-\xi),\ldots,n_K(1-\gamma^*_K)) \\
&\qquad\quad \times \Pi^{(n_{-k,1}+n_k\xi)}_{K_{-k,1}+\xi,1}(n_1\gamma^*_1,\ldots,n_k\xi,\ldots,n_K\gamma^*_K) \prod_{l \in C_{k,n}} \mathbb{1}_{\{\xi\}}(\gamma_l),
\end{aligned}$$

If we denote with $f_{k,\xi}$ the marginal likelihoods

$$\pi\left(\boldsymbol{Z}_{l \in C_{k,n}}\right) = \left\{ \int \prod_{l \in C_{k,n}} k_\xi\left(Z_l | \boldsymbol{\vartheta}^*_{k,\xi}\right) P^*_\xi(\mathrm{d}\boldsymbol{\vartheta}^*_{k,\xi}) \right\}$$

we can recover an expression similar to (3.13):

$$\frac{p_{k,0}}{p_{k,1}} = \frac{f_{k,0}}{f_{k,1}} \frac{(b+n-n_{-k,1}-n_k)_{n_k}}{(a+n_{-k,1})_{n_k}} \frac{(1-\sigma_0)_{n_k-1}}{(1-\sigma_1)_{n_k-1}}$$

$$\times \frac{\theta_0 + (K - K_{-k,1} - 1)\sigma_0}{\theta_1 + K_{-k,1}\,\sigma_1} \frac{(\theta_1 + n_{-k,1})_{n_k}}{(\theta_0 + n - n_{-k,1} - n_k)_{n_k}}. \quad (3.14)$$

Since ρ is conditionally independent from the observations Z_i's given the γ_i's and the parameter $\boldsymbol{\vartheta}_i$, we can obtain the full conditional distribution of ρ as

$$\mathcal{L}[\mathrm{d}\rho \mid Z_1, \ldots, Z_n, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_n, \gamma_1, \ldots, \gamma_n] = \frac{\Gamma(a+b+n)}{\Gamma(a+n^*)\Gamma(b+n-n^*)} \rho^{a+n^*-1}(1-\rho)^{b+n-n^*-1}\, \mathbb{1}_{(0,1)}(\rho),$$

that is, as a sample from a Beta$(a + n^*, b + n - n^*)$, with $n^* = \sum_{j=1}^{K} n_j \gamma_j^*$. The sampling algorithm is then completed by drawing samples of the $\boldsymbol{\vartheta}_i^*$'s from the respective full conditionals.

The full conditional of $\boldsymbol{\vartheta}_i$ is obtained as

$$\mathcal{L}[\mathrm{d}\boldsymbol{\vartheta}_{i,\gamma_i} \mid \boldsymbol{\vartheta}_{-i}, \boldsymbol{\gamma}^*, \boldsymbol{C}_n, \boldsymbol{Z}] \propto q_0^{(i)} \frac{k_{\gamma_i}(Z_i \mid \boldsymbol{\vartheta}_{i,\gamma_i})\, P^*_{\gamma_i}(\mathrm{d}\boldsymbol{\vartheta}_{i,\gamma_i})}{\int_{\mathbb{R}\times\mathbb{R}^+} k_{\gamma_i}(Z_i \mid \boldsymbol{\vartheta}_{\boldsymbol{i},\boldsymbol{\gamma_i}})\, P^*_{\gamma_i}(\mathrm{d}\boldsymbol{\vartheta}_{\boldsymbol{i},\boldsymbol{\gamma_i}})} + \sum_{j=1}^{\kappa^{(-i)}_{\gamma_i}} q_j^{(i)}\, \delta_{\boldsymbol{\vartheta}^*_{j,\gamma_i}}(\mathrm{d}\boldsymbol{\vartheta}_{i,\gamma_i})$$

where $\boldsymbol{\vartheta}_{-i} = \{(\mu_j, \tau_j^2)_{\gamma_j} : j \neq i\}$ and $\kappa^{(-i)}_{\gamma_i}$ is the number of unique values in $\boldsymbol{\vartheta}_{-i}$ that share the same generating random probability measure $\tilde{p}_{\gamma_i}$ with $\boldsymbol{\vartheta}_{i,\gamma_i}$. Correspondingly, the respective frequencies are denoted as $\boldsymbol{n}^{(-i)}_{\gamma_i} = \left(n^{(-i)}_{1,\gamma_i}, \ldots, n^{(-i)}_{\kappa^{(-i)}_{\gamma_i},\gamma_i}\right)$ with weights $q_0^{(i)}$ and $q_j^{(i)}$ as follows

$$q_0^{(i)} \propto \Pi^{(|n^{(-i)}_{\gamma_i}|+1)}_{\kappa^{(-i)}_{\gamma_i}+1,\gamma_i}(n^{(-i)}_{1,\gamma_i}, \ldots, n^{(-i)}_{\kappa^{(-i)}_{\gamma_i},\gamma_i}, 1)\, k_{\gamma_i}(Z_i \mid \boldsymbol{\vartheta}_i)$$
$$q_j^{(i)} \propto \Pi^{(|n^{(-i)}_{\gamma_i}|+1)}_{\kappa^{(-i)}_{\gamma_i},\gamma_i}(n^{(-i)}_{1,\gamma_i}, \ldots, n^{(-i)}_{j,\gamma_i} + 1, \ldots, n^{(-i)}_{\kappa^{(-i)}_{\gamma_i},\gamma_i})$$

Specializing the previous formula to the 2PPD process, we obtain:

$$q_0^{(i)} \propto \left(\theta_{\gamma_i} + \kappa^{(-i)}_{\gamma_i}\sigma_{\gamma_i}\right) \int k_{\gamma_i}(Z_i \mid \boldsymbol{\vartheta}_i)\, P^*_{\gamma_i}(d\boldsymbol{\vartheta}_i),$$
$$q_j^{(i)} \propto \left(n^{(-i)}_{j,\gamma_i} - \sigma_{\gamma_i}\right) k_{\gamma_i}\left(Z_i \mid \vartheta^*_{j,\gamma_i}\right).$$

where $\int k_{\gamma_i}(Z_i \mid \boldsymbol{\vartheta}_i)\, P^*(d\boldsymbol{\vartheta}_i)$ is the marginal likelihood based only on the observation Z_i

Finally, the full conditional for m_1 is sampled with a simple adaptive Metropolis Hasting step (Roberts and Rosenthal, 2007; Roberts and Rosenthal, 2009). Here, we can take advantage of the conjugacy properties of the model and employ a marginal Gibbs sampler as discussed in Ishwaran and James (2001). Alternatively, for non-conjugate models, a Metropolis-Hastings algorithm would be equivalently straightforward to implement, and mimic widely used algorithms for Dirichlet Process mixture models (Neal, 2000a).

3.B Proof of Proposition 1.

We first evaluate the probability distribution of the Z_i's, partitioned into K distinct clusters with representatives located at infinitesimal intervals $\mathrm{d}z_1, \ldots, \mathrm{d}z_K$, around points $z_1, \ldots, z_K$, with respective multiplicities $n_1, n_2, \ldots, n_K$.

$$\begin{aligned}
&\mathbb{P}[Z_1^* \in \mathrm{d}z_1, \ldots, Z_K^* \in \mathrm{d}z_K, n_1, \ldots, n_K] \\
&= \mathbb{E}\left[\prod_{j=1}^{K}\left\{w\frac{\mu_0(\mathrm{d}z_j)}{\mu_0(\mathbb{Z})} + (1-w)\frac{\mu_1(\mathrm{d}z_j)}{\mu_1(\mathbb{Z})}\right\}^{n_j}\right] \\
&= \sum_{i_1=0}^{n_1}\cdots\sum_{i_K=0}^{n_K}\binom{n_1}{i_1}\cdots\binom{n_K}{i_K} w^{i_1+\cdots+i_K}(1-w)^{n-(i_1+\cdots+i_K)} \times \\
&\qquad\qquad \mathbb{E}\left[\prod_{j=1}^{K}\left(\frac{\mu_0(\mathrm{d}z_j)}{\mu_0(\mathbb{Z})}\right)^{i_j}\left(\frac{\mu_1(\mathrm{d}z_j)}{\mu_1(\mathbb{Z})}\right)^{n_j-i_j}\right]
\end{aligned}$$

From Lemma 1, it follows that for any $i_j \notin \{0, n_j\}$ the expected value above vanishes. Hence, for $i_j \in \{0, n_j\}$ one has

$$\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{K}\left(\frac{\mu_0(\mathrm{d}z_j)}{\mu_0(\mathbb{Z})}\right)^{i_j}\left(\frac{\mu_1(\mathrm{d}z_j)}{\mu_1(\mathbb{Z})}\right)^{n_j-i_j}\right] &= \prod_{j=1}^{K}[P_0^*(\mathrm{d}z_j)]^{\frac{i_j}{n_j}}[P_1^*(\mathrm{d}z_j)]^{\frac{n_j-i_j}{n_j}} \\
&\quad \Pi_{|\boldsymbol{i}|,0}^{(\boldsymbol{n}'\boldsymbol{i})}(i_1, \ldots, i_K)\, \Pi_{K-|\boldsymbol{i}|,1}^{(n-\boldsymbol{n}'\boldsymbol{i})}(n_1 - i_1, \ldots, n_K - i_K)
\end{aligned}$$

The representation in Proposition 1, then, follows when integrating out with respect to ρ and $z_1, \ldots, z_k$.

3.C Split-Merge move for the two-group 2PPD model

In the cases of Dirichlet and Pitman-Yor process mixture models, posterior sampling is often performed via Gibbs sampler. Unfortunately, especially in marginal models, the Gibbs sampler explores the state space by means of conditional updates cycling through all observations and can get stuck in local modes. Thus, it can mix poorly across modes that have high probability (Dahl, 2003). To address the deficiencies of the Gibbs sampler, Jain and Neal (2004) and Jain and Neal (2005) propose a split-merge (SM) algorithm which greatly improves the mixing. To obtain the best performance, they recommend to cycle between the usual Gibbs sampler and their SM proposal. Nevertheless, their sampler is known to be computationally demanding. To obviate this problem, Dahl proposes an enhanced versions of the first split-merge sampler, called SAMS, that can be used in conjugate cases (Dahl, 2003) and in non-conjugate cases (Dahl, 2005) for Dirichlet Process mixture models.

In the following, we adapt the conjugate version of Dahl's SM sampler to be employed in Pitman-Yor mixture models. In particular, according to Lemma 1, given the process allocation variables $\boldsymbol{\gamma}$, the two processes are independent. This allows us to perform two separate Split and Merge steps, one per process. In the following, we present the sampler for the generic process γ', $\gamma' = 0, 1$. Let us denote the partition observed among the observations with $\eta = \{S_1, S_2, \ldots, S_{K^*}\}$, where S_l denotes the subset of indexes assigned to the l-th cluster

and K^* denotes the number of subsets in the partition.

Steps

1. Among the observations assigned to the same process γ', uniformly select a pair of distinct indices i and j.

2. If i and j belong to the same component in η , propose $\eta^\star$ by attempting a split move:

   S1 For convenience, denote the common component containing indexes i and j as S. Remove the indexes i and j from S and form singleton sets $S_i = \{i\}$ and $S_j = \{j\}$.

   S2 Letting k be successive values in a **uniformly-selected permutation** of the indexes in S, add k to S_i with probability

   $$Pr\left(k \in S_i | S_i, S_j, y\right) = \frac{\left(n_{i,\gamma} - \sigma_\gamma\right) \int F\left(y_k; \vartheta\right) dH_{S_i}(\vartheta)}{\left(n_{i,\gamma} - \sigma_\gamma\right) \int F\left(y_k; \vartheta\right) dH_{S_i}(\vartheta) + \left(n_{j,\gamma} - \sigma_\gamma\right) \int F\left(y_k; \vartheta\right) dH_{S_j}(\vartheta)}$$

   where H_S is the posterior distribution of a component location ϑ based on the prior $P^*_{\gamma'}$ and the data corresponding to the indices in S . Otherwise, add k to S_j . Note that, at each iteration above, either S_i or S_j gains an index resulting in $n_{i,\gamma'}$ or $n_{j,\gamma'}$ increasing by 1 . Further, H_{S_i} and H_{S_j} evolve to account for each additional index. We remark how our model employs non-conjugate base measures for the Null process. However, it is straightforward to perform precise numerical integration over the parameters' space. Working with marginal distributions $m_{\gamma'}\left(S_l\right)$ of all the observations in the l-th cluster, we can rewrite

   $$\int F\left(y_k; \vartheta\right) dH_{S_i}(\vartheta) = \int F\left(y_k; \vartheta\right) \frac{F\left(S_i; \vartheta\right) P^*_{\gamma'}(\vartheta) d\vartheta}{m_{\gamma'}\left(S_i\right)} = \frac{m_{\gamma'}\left(y_k, S_i\right)}{m_{\gamma'}\left(S_i\right)}$$

   and consequently

   $$Pr\left(k \in S_i | S_i, S_j, y\right) = \frac{\left(n_{i,\gamma'} - \sigma_{\gamma'}\right) \frac{m_{\gamma'}(y_k, S_i)}{m_{\gamma'}(S_i)}}{\left(n_{i,\gamma'} - \sigma_{\gamma'}\right) \frac{m_{\gamma'}(y_k, S_i)}{m_{\gamma'}(S_i)} + \left(n_{j,\gamma'} - \sigma_{\gamma'}\right) \frac{m_{\gamma'}(y_k, S_j)}{m_{\gamma'}(S_j)}} \tag{3.15}$$

   Alternatively, one can employ the steps presented in Dahl (2005) for non-conjugate cases.

   S3 Compute the Metropolis-Hastings ratio and accept η^* as the current state η with probability given by this ratio. The calculation of the Metropolis-Hastings ratio is discussed below.

3. Otherwise, i and j belong to different components in η. Propose $\eta^\star$ by attempting a merge move:

   M1 For convenience, let S_i and S_j denote the components in η containing i and j, respectively.

   M2 Form a merged component $S = S_i \cup S_j$.

   M3 Propose the following set partition: $\eta^\star = \eta \cup \{S\} \backslash \{S_i, S_j\}$.

   M4 Compute the Metropolis-Hastings ratio and accept $\eta^\star$ as the current state η with probability given by this ratio. Again, the calculation of the Metropolis-Hastings ratio is discussed below.

Computing the Metropolis Ratio The MH ratio for the SAMS sampling algorithm is given as:

$$a\left(\eta^{\star}|\eta\right)=\min\left[1,\frac{p\left(\eta^{\star}|y\right)}{p(\eta|y)}\frac{Pr\left(\eta|\eta^{\star}\right)}{Pr\left(\eta^{\star}|\eta\right)}\right]$$

where $p\left(\eta^{\star}|y\right)$ is the partition posterior distribution evaluated at $\eta^{\star}$ and $Pr\left(\eta^{\star}|\eta\right)$ is the probability of proposing $\eta^{\star}$ from the state η.

We have that $p(\eta|y)\propto p(y|\eta)p(\eta)$, where $p(\eta)=\frac{\prod_{j=1}^{K-1}(\theta_{\gamma'}+j\sigma_{\gamma'})}{(\theta_{\gamma'}+1)_{n_{\gamma'}-1}}\prod_{j=1}^{K}(1-\sigma_{\gamma'})_{n_{j,\gamma'}-1}$ is the 2PPD process EPPF and

$$p(y|\eta)=\prod_{l=1}^{K}p\left(S_l\right)=\prod_{l=1}^{K}\int\prod_{k\in S_l}F\left(y_k;\vartheta\right)P^{*}_{\gamma'}(\vartheta)d\vartheta=\prod_{l=1}^{K}m\left(S_l\right)$$

.

Finally, let us focus on $p(\eta^{\star}|\eta)$. When the proposal $\eta^{\star}$ is a split update, $Pr(\eta^{\star}|\eta)$ is merely the product of the probabilities in (3.15) associated with the chosen allocations. Since these two split components could only be merged in one way, $Pr(\eta|\eta^{\star})=1$. Conversely, when the proposal $\eta^{\star}$ is a merge update, $Pr(\eta^{\star}|\eta)$ is 1, but $Pr(\eta|\eta^{\star})$ is the product of the probabilities in (3.15) associated with the allocation choices that would need to be made to obtain the split partition η, although no actually splitting is performed. Dahl underlines that it is critical that a random permutation of the indexes is used when performing this imaginary split.

3.D Microbiome data case study: list of differentially abundant taxa

TABLE 3.D.1: Microbiome data case study: differentially abundant taxa with negative z-scores indicating less abundance in the children with moderate to severe diarrhea. Most are well known commensal bacteria, e.g. *Prevotella* spp. and *Clostridium* spp. Posterior probability that the z-score belongs to the non-null group is given for each taxa. The dotted line highlights the difference between the genes flagged as relevant by our method and the ones found with the *locfdr* model.

Taxon	z-score	$p(K_i = 1\|\text{data})$	Efron LocFdr
Prevotella copri	-10.13	1.00	0.99
Prevotella sp. DJF_RP53	-9.76	1.00	0.98
Prevotella sp. BI-42	-9.72	1.00	0.98
Prevotella sp. DJF_B112	-9.64	1.00	0.98
Clostridium lituseburense	-8.53	1.00	0.86
Clostridium paraputrificum	-7.78	1.00	0.62
Faecalibacterium prausnitzii	-7.49	1.00	0.50
Prevotella sp. oral clone BP1-28	-7.38	1.00	0.46
Clostridium bartlettii	-7.01	1.00	0.33
Clostridium sp. FRC_C11	-6.80	1.00	0.27
Faecalibacterium sp. DJF_VR20	-6.59	1.00	0.23
Clostridium disporicum	-6.33	1.00	0.19
Collinsella sp. CB20	-6.26	1.00	0.19
Ruminococcus gnavus	-6.02	1.00	0.18
Bacteroides fragilis	-5.90	1.00	0.18
Clostridium butyricum	-5.90	1.00	0.18
Enterococcus sp. L2	-5.78	1.00	0.18
Prevotella intermedia	-5.68	1.00	0.18
Clostridium glycolicum	-5.29	0.99	0.20
Bacteroides sp. CJ78	-5.22	0.99	0.21
Collinsella aerofaciens	-5.19	0.99	0.21
Eubacterium rectale	-5.06	0.99	0.21
Bacteroides xylanisolvens	-4.95	0.98	0.21
Clostridium hathewayi	-4.86	0.98	0.22
Collinsella sp. HA6	-4.69	0.97	0.21
Turicibacter sanguinis	-4.68	0.97	0.21
Clostridium sp. CJ66	-4.66	0.97	0.21
Prevotella sp. oral clone AO009	-4.65	0.97	0.21
Enterococcus gallinarum	-4.59	0.96	0.21
Megasphaera sp. TrE9262	-4.53	0.96	0.20
Bacteroides ovatus	-4.52	0.96	0.20
Clostridium difficile	-4.30	0.95	0.19

TABLE 3.D.2: Microbiome data case study: differentially abundant taxa with positive z-scores indicating greater abundance in the children with moderate to severe diarrhea. Most are well known pathogenic bacteria, e.g. *Shigella* spp. and *E. coli*. Posterior probability that the z-score belongs to the non-null group is given for each taxa. The dotted line highlights the difference between the genes flagged as relevant by our method and the ones found with the *locfdr* model.

Taxon	z-score	$p(K_i = 1\|\text{data})$	Efron LocFdr
Escherichia coli	8.48	1.00	0.83
Streptococcus sp. C101	7.66	1.00	0.70
Haemophilus haemolyticus	7.62	1.00	0.69
Streptococcus mitis	7.48	1.00	0.65
Erwinia chrysanthemi	7.18	1.00	0.58
Streptococcus sp. oral clone ASCE09	7.10	1.00	0.55
Enterobacter cloacae	6.87	1.00	0.49
Acinetobacter sp. SF6	6.56	1.00	0.40
Granulicatella sp. oral clone ASCG05	6.34	1.00	0.33
Streptococcus sp. oral clone ASCC04	6.18	1.00	0.28
Shigella boydii	5.93	1.00	0.20
Streptococcus sp. oral clone ASCC01	5.82	1.00	0.17
Streptococcus peroris	5.81	1.00	0.17
Rothia mucilaginosa	5.76	1.00	0.15
Streptococcus oralis	5.75	1.00	0.15
Escherichia sp. oral clone 3RH-30	5.58	1.00	0.11
Citrobacter freundii	5.58	1.00	0.11
Granulicatella adiacens	5.54	1.00	0.10
Streptococcus sanguinis	5.47	1.00	0.08
Escherichia albertii	5.24	1.00	0.03
Escherichia sp. EMB 210	5.08	1.00	0.01
Granulicatella elegans	5.04	1.00	0.00
Streptococcus pneumoniae	5.03	0.99	0.00
Fusobacterium nucleatum	5.03	1.00	0.00
Serratia marcescens	4.97	1.00	0.00
Streptococcus sp. oral strain T4-E3	4.86	0.99	0.00
Streptococcus sp. oral clone DP009	4.84	0.99	0.00
Shigella sonnei	4.79	0.99	0.00
Fusobacterium periodonticum	4.76	0.99	0.00
Neisseria sp. oral clone BP2-82	4.62	0.99	0.00
Actinobacillus pleuropneumoniae	4.59	0.99	0.00
Streptococcus parasanguinis	4.52	0.99	0.00
Streptococcus sp. C163	4.50	0.99	0.00
Fusobacterium sp. oral clone BS011	4.48	0.99	0.00
Haemophilus sp. oral clone BJ021	4.39	0.98	0.00
Streptococcus sp. oral clone BP1-49	4.25	0.98	0.00
Abiotrophia defectiva	4.24	0.98	0.00
Streptococcus sp. oral clone MCE7_144	4.19	0.98	0.00
Haemophilus influenzae	4.11	0.97	0.00
Campylobacter jejuni	4.11	0.97	0.00
Citrobacter sp. SVUB3	4.07	0.97	0.00
Enterobacter sp. CRRI 3	3.89	0.95	0.00

3.E Plots of the five scenarios considered in the simulation study

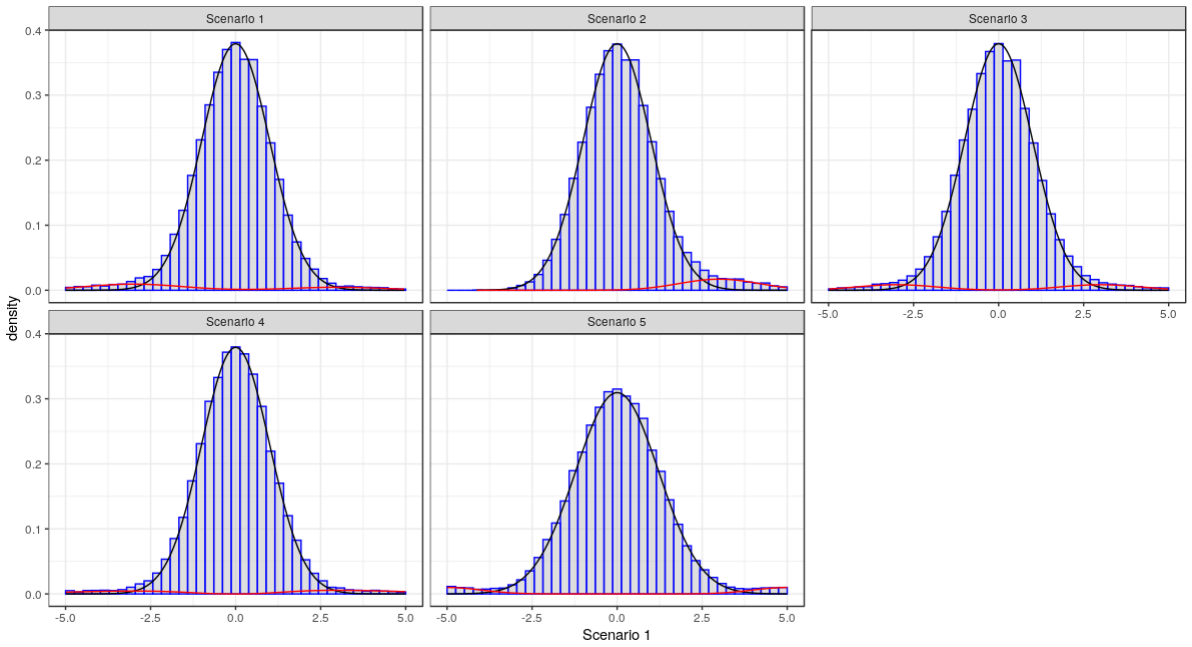


FIGURE 3.E.1: The five Scenarios considered in the simulation study in Section 4.1. We plot the histograms of the simulated data and superimpose the null (blue) and alternative (red) density functions.

We recall here the five scenarios that we investigate in our simulation study:
We simulate z-scores from mixture (1), where $f_0(z) = N(z \mid 0, 1)$ and for f_1 we choose:

- **Scenario 1**: $f_1(z) = 0.67 \cdot \mathcal{N}(z| -3, 2) + 0.33 \cdot \mathcal{N}(z|3, 2)$,
- **Scenario 2**: $f_1(z) = \mathcal{N}(z \mid u, 1)$ with $u \sim \text{Uniform}(2, 4)$,
- **Scenario 3**: $f_1(z) = \mathcal{N}(z \mid u,)$ with $u \sim \text{Uniform}([-4, -2] \cup [2, 4])$,
- **Scenario 4**: $f_1(z) = \text{Gamma}((-1)^v \cdot z \mid a, b)$ with $a = 4$, $b = 1$ and $v \sim \text{Bernoulli}(0.5)$,

i.e. f_1 is assumed asymmetric unimodal (scenario 1), symmetric bimodal (scenario 2), asymmetric bimodal (scenario 3) and symmetric bimodal with fat tails (scenario 4), thus mimicking typical high-dimensional testing situations. Moreover, we also considered the following case

- **Scenario S5**: $f(z) = 0.95\mathcal{N}(z|0, 1.5) + 0.025\mathcal{N}(z|5, 1) + 0.025\mathcal{N}(z| -5, 1)$,

where the null distribution departs from the theoretical standard Gaussian and the alternative distribution is chosen to be easily detectable, being very separated from the null one. Figure 3.E.1 shows the histograms of data simulated under the five scenarios.

3.F Computational Burden

We perform a simulation study where we investigate the computational time needed by the algorithm in its original specification (Polya Urn scheme - PUS) compared to the Split-Merge (SM) alternative proposed in the previous subsection. More precisely, we keep track of the time in seconds that the model needs to run 100 iterations. For the SM case, we perform 10 SM moves per iterations, and full sweep is executed every 10 steps.

We study how the computational time varies as the sample size and the discount parameter of the null process change. These two quantities play a pivotal role in the sampler efficiency, since the expected number of clusters grows as both n and σ_0 increase. This in turn means more clusters to update at each iteration. Table 3.F.1 reports the results in seconds. We can appreciate how the burden becomes more consistent with the growth of the sample size. The impact of σ_0 on the computational time is amplified with the sample size. We see how, for sample sizes beyond 1,000 observations, the PUS sampler becomes extremely inefficient, making the inference infeasible. Relying on the SM moves seems to solve this issue, consistently speeding up the algorithm. We also notice how the value of σ_0 has a small impact on the SM sampler.

		$\sigma_0 = 0.5$	$\sigma_0 = 0.75$	$\sigma_0 = 0.9$
$n = 100$	PUS	5.480	5.916	6.007
	SM	6.373	5.625	5.614
$n = 200$	PUS	12.396	14.498	15.009
	SM	10.448	10.625	10.696
$n = 500$	PUS	51.319	58.532	61.201
	SM	22.625	23.585	18.599
$n = 1,000$	PUS	186.226	221.853	236.200
	SM	54.390	52.042	42.521
$n = 2,000$	PUS	914.227	1137.689	1160.732
	SM	145.904	146.850	144.341

TABLE 3.F.1: Computational time in seconds for 100 iterations with the Polya Urn Scheme (PUS) sampler, compared with the Split-Merge (SM) scheme. We investigate how the time varies as the sample size n and the discount parameter of the null process change.

3.G Additional simulation study

In this Section, we study the effect of considering discount parameters σ_0, σ_1 with $\sigma_0 < \sigma_1$. More specifically, we assume $\sigma_0 = 0.1$ and $\sigma_1 = 0.75$. The remaining hyper-parameters are set according to the simulation scenarios in Section 3.4.1. When thresholding the BFDR at 10%, all the observations are flagged as interesting, since all the posterior probability of inclusion too high. To exemplify, Figure 3.G.1 reports the posterior probability of inclusion for all the observations of one of the datasets in the five different scenarios. To obtain the results reported in Table 3.G.1, we set the threshold at the 1% level. We see how the model fails to recognize the two underlying distributions, since the null distribution is free to vary and even the values around zero are taken over by the alternative.

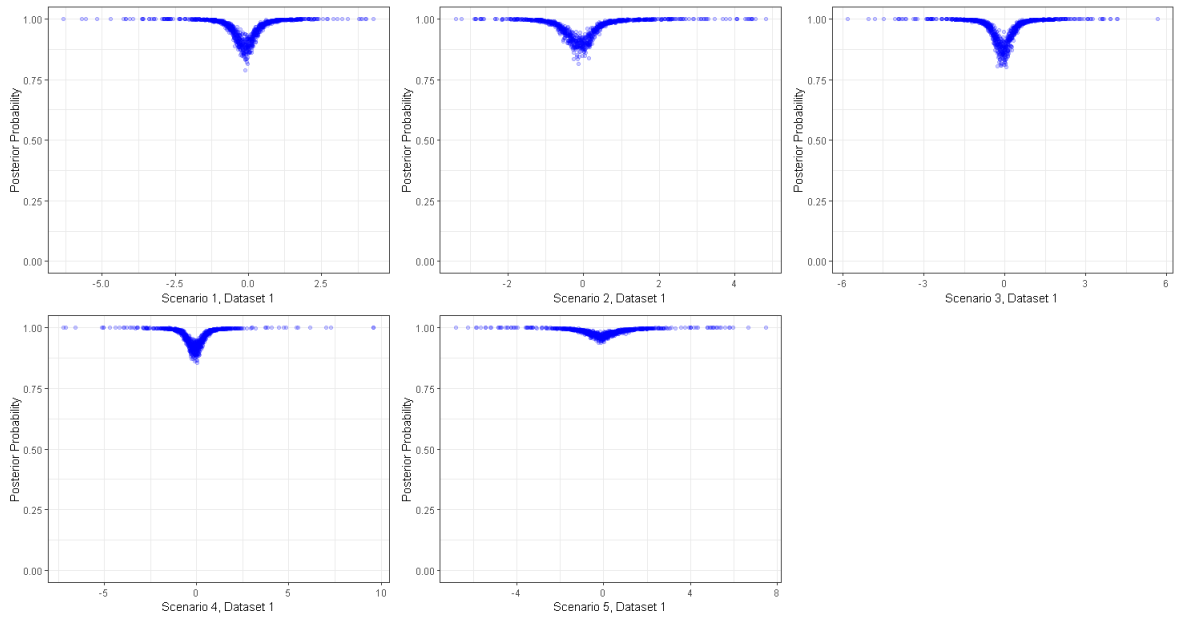


FIGURE 3.G.1: Posterior probability of inclusion for observations sampled in the first dataset of the five different scenarios, estimated adopting $\sigma_0 = 0.1$ and $\sigma_1 = 0.75$

Scenario	MCC		F1		SPEC		ACC		PRE		AUC	
1	0.1569	(0.0147)	0.1409	(0.0062)	0.3823	(0.0272)	0.4114	(0.0258)	0.0760	(0.0035)	0.9189	(0.0213)
2	0.1668	(0.0136)	0.1436	(0.0066)	0.3792	(0.0312)	0.4096	(0.0296)	0.0774	(0.0038)	0.9477	(0.0168)
3	0.1664	(0.0131)	0.1468	(0.0064)	0.3724	(0.0320)	0.4042	(0.0302)	0.0793	(0.0037)	0.9448	(0.0178)
4	0.1697	(0.0168)	0.1445	(0.0094)	0.3765	(0.0459)	0.4075	(0.0436)	0.0779	(0.0055)	0.9656	(0.0100)
5	0.1157	(0.0125)	0.1183	(0.0048)	0.2138	(0.0363)	0.2531	(0.0345)	0.0629	(0.0027)	0.9867	(0.0026)

TABLE 3.G.1: Simulation study: sensitivity results across the five simulation scenarios considered in Section 4.1 ($\rho = 0.05$), modeled with PYs processes characterized by $\theta_0 = \theta_1 = 1$ and $\sigma_0 = 0.1 < \sigma_1 = 0.75$. The values in the table represent the average MCC and F_1 scores, the average precision (PRE), specificity (SPEC), accuracy (ACC) and the area under the curve (AUC) of the corresponding receiver operating characteristic curve, over 30 replicates with corresponding standard deviations between brackets.

Chapter 4

Bayesian Two-Group Model: a Non-Local Likelihood Approach

"Non aver paura di invecchiare.
L'età che si dimostra attrae, vedi Persepoli.
Non avere paura di stonare.
Non le popstar, le campane suonano per secoli."
Migliora La Tua Memoria Con Un Click – Michele Salvemini feat. Max Gazzé

"Stay out of my territory."
Breaking Bad s2e10 – ~~The null distribution~~ Heisenberg

Abstract

The two-group model of Efron (2004) discriminates among interesting and uninteresting test statistics by assigning instances to either a null or an alternative density, whose mixture describes the data distribution. The amount of separation between these two competing densities can crucially affect the performance of the classification, and the assumed mixture distribution poses no control on a possible detrimental overlap. In this work, we employ non-local distributions (Johnson and Rossell, 2010) as an important device to model the non-null density in the two-group model, with the purpose of reducing the overlap with the null distribution. We provide a theoretical justification in terms of improved false discovery rate, false negative rate, and statistical power. Moreover, we develop efficient Gibbs Sampler algorithms for both parametric and nonparametric model specifications. We illustrate the performance of our methodology in an extensive simulation study and employ it on several publicly available genomics datasets including the analysis of microarray data, microbiome abundance tables, and proteomics measurements. In one of these applications, we extend the model to account for data grouped by different experiment conditions.

4.1 Introduction

With the increasing availability of large datasets, many researchers have addressed the problem of multiple hypotheses testing (MHT), both in a classical setting as well as in a Bayesian framework. An important branch of the literature has been developed starting from the seminal paper of Benajmini and Hochberg (1995), where the concept of False Discovery Rate (FDR) was introduced to control the Type-I error when conducting multiple comparisons. Numerous authors devoted to this topic, see Goeman and Solari (2014) for a review. In particular, Efron extended the idea of FDR in an Empirical Bayes setting, introducing the concepts of Local FDR ($lfdr$, Efron 2004), then formalized in his famous two-group model (Efron, 2004; Efron, 2007; Efron, 2008), a milestone in the MHT literature. The two-group model assumes that the distribution of the data - which are often in the form of test statistics properly standardized- is a mixture between the *null* distribution, f_0 and the *alternative* distribution, f_1. The observations (estimated to be) generated from the null distribution (namely, under H_0) are then labeled as "irrelevant", whilst those attributed to the alternative (under H_1) are regarded as "relevant". Identifying the relevant observations by deconvoluting the two-group model is the main goal of this methodology.
Numerous efforts have been dedicated to providing reliable estimates for f_0 and f_1. Some authors opted for finite mixtures defined directly on p-values (Pounds and Morris, 2003; Liao et al., 2004; Allison et al., 2002). Martin and Tokdar (2012) develop a likelihood-based analysis of the two-group model, imposing regularization on the parameters of the empirical null distribution and the mixture weight, and by modeling semiparametrically the alternative density f_1. Muralidharan (2012) proposes a hierarchical model, estimated in an empirical Bayes framework, to describe simultaneously the effect size and the local or tail-area false discovery rate for each observed test statistic. Do et al. (2005) employ Dirichlet Processes Mixture Models to estimate separately f_0 and f_1, adopting a Standard Gaussian and a mixture of two Gaussians centered in ± 2 as base measures, respectively.
The idea of employing mixture models for the two unknown densities is appealing, but it may cause a lack of separation between the null and the alternative model: without any constraint, the two distributions might freely overlap, jeopardizing the classification. In fact, we should require f_1 to be longer-tailed than f_0, with the non-null z_i's tending to occur far away from the origin (Efron, 2007). Generally, if the sample space $\mathcal{X}$ is partitioned into two complementary parts, say $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$, relative to two different scenarios $\mathcal{S}_0$ and $\mathcal{S}_1$, we propose to achieve the desired separation between the two competing distributions by imposing what we call a non-local density (NLD) on f_1, having the property of assigning very low or null probability to the portion of the sample space $\mathcal{X}_0$ under the scenario $\mathcal{S}_1$.
The concept of a non-local version of distributions was first proposed by Johnson and Rossell (2010) for the prior of the parameters of interest, introducing the so-called non-local Prior (NLP), with the scope of performing (non-multiple) hypothesis testing and variable selection. The properties of NLDs have been studied in numerous works, e.g. Johnson and Rossell (2012), Consonni et al. (2012), Zhu et al. (2013), and Rossell and Telesca (2017). In particular, Johnson and Rossell (2010) introduce the Moment (MOM) and Inverse Moment (iMOM) NLD specifications, while Johnson and Rossell (2012) prove the consistency of a model selection procedure for linear model settings when the number of possible covariates p is bounded by the number of observations n, once NLDs are imposed on the model parameters. Consonni et al. (2012) adopt NLD to perform model selection for DAGs: starting from a MOM prior under the full, unconstrained model, they use the fractional Bayes factor for comparisons with restricted nested models. More recently, Shin et al. (2015) employ NLDs for variable selection in high dimensional settings and Zhu et al. (2013) use NLDs on the parameters of a generalized linear model in the context of admixture mapping. Finally, Rossell and Telesca (2017) develop a NLD-based Bayesian model averaging framework, where NLD is expressed as a mixture of truncated distributions

that facilitates posterior sampling.
Generalizing the structure of the MOM density we propose a new class of NLDs as members of a family of weighted densities that encompasses numerous other distributions. We employ our new class of NLDs in the definition of f_1 for the two-group model, to perform MHT with an appropriate degree of separation among the two densities under the two competing hypotheses. For the problem of detecting relevant z-scores, we fix in the sequel $\mathcal{X}_0$ to be a region of the sample space around zero, implying that the scenario $\mathcal{S}_0$ coincides with the null hypothesis, i.e. $\mathcal{S}_0 = H_0$, and consequently assuming $\mathcal{S}_1 = H_1$. However, the proposed framework does not prevent the inclusion of data-driven external information for the choice of a different sample space partition.
We then develop a parametric, interpretable, yet flexible Bayesian two-group model and provide an estimate of $lfdr$ naturally constrained in $[0, 1]$. Following the same rationale, we also propose a nonparametric alternative. By choosing a threshold on the posterior probability of inclusion in the alternative scenario $(1 - lfdr)$ that ensures a Bayesian FDR bounded above by a user-defined constant α (Newton et al., 2004; Do et al., 2005), a critical region $[z_1, z_2]$ is derived, outside of which a z-score is labeled as "relevant". To conduct posterior inference, we propose a collapsed Gibbs sampler, and compare our methodology in simulated data in four different scenarios against established alternatives.
The article proceeds as follows: in Section 2 we introduce the NLDs as members of the weighted distributions family and develop the modeling framework.
Section 3 discusses posterior inference building the required algorithm, whilst Section 4 contains a simulation study and the real-data applications on three freely available biostatistical datasets: a microarray experiment, a microbiome abundance table, and a grouped proteomics data frame. Section 5 discusses and concludes.

4.2 Non-local Likelihood

4.2.1 Weighted densities and Non-local Distributions

A density $\pi_{NL}(x)$ is a NLD with respect to the sample space partition $\{\mathcal{X}_0, \mathcal{X}_1\}$ if, for some $\varepsilon, \zeta > 0$,

$$\pi_{NL}(x) < \varepsilon \quad \text{for all } x \in \mathcal{X}_0 : \inf_{x \in \mathcal{X}_0} |x - x_0| < \zeta, \tag{4.1}$$

therefore attributing low density to the sample subspace corresponding to the null scenario (Johnson and Rossell, 2010).A density that does not satisfy (4.1) is referred to as *local density*. Consider an univariate random variable X with density function $\pi(x; \lambda)$ parameterized by some λ. A MOM distribution (Johnson and Rossell, 2010) has density

$$\pi_{MOM}(x, \lambda, x_0, k) \propto (x - x_0)^{2k} \pi(x; \lambda), \tag{4.2}$$

a weighted version of $\pi(x; \lambda)$ (Consonni et al., 2012; Rossell and Telesca, 2017). It is evident that the MOM is an NLD, where a quadratic weight constrains the density to be close to zero in a neighborhood of x_0 which we name the origin.
In the same way, starting from a generic density and a suitable weight function w, we can define a family of weighted distributions:

Definition 2 (Weighted Density)**.** Consider a random variable X, with support $\mathcal{X}$ and (proper) probability density function $\pi(x; \xi)$. For a non-negative function $w(x; \xi)$, with $\mathbb{E}_\pi[w(x; \xi)] < \infty$, the corresponding weighted density is

$$\pi_W(x; \xi, \lambda) = \frac{w(x; \xi)}{\int w(x; \xi)\pi(x; \lambda)dx}\pi(x; \lambda) = \frac{w(x; \xi)}{\mathbb{E}_{\pi(x;\lambda)}[w(x; \xi)]}\pi(x; \lambda). \tag{4.3}$$

Most NLDs in the literature can be viewed as weighted densities characterized by specific weight functions. For example, if $w(x; x_0, k) = (x - x_0)^{2k}$ we obtain the MOM distribution, while if we set $w(x; x_0, \phi, \tau) = \exp\left(\sqrt{2} - \frac{\tau\phi}{(x-x_0)^2}\right)$ we recover the eMOM (Rossell et al., 2013). More generally, following Rossell and Telesca (2017) we can obtain a non-local Distribution around x_0 imposing that $w(x; \xi) \to 0$ as $x \to x_0$, regardless the form of $\pi(x; \lambda)$.
The family of weighted distributions defined in (4.3) is very general and encompasses many known statistical distributions besides the non-local ones. For instance, the choice $w(x; \xi) = \mathbb{I}_{\{x \in [a,b]\}}$ or the sum of indicator functions on disjoints sets recovers all the truncated distributions; using a Gaussian c.d.f. as weight for a Gaussian density results in a Skew Normal distribution (Capitanio, 2011; O'Hagan and Leonard, 1976); multivariate repulsive priors of Petralia et al. (2012) can also be shown to be in this family.
Furthermore, the broad class of weighted distributions can also be linked to regularized methods. Writing the log-posterior of the model highlights how the prior can be seen again as a penalty term: $\log \pi(\theta|\boldsymbol{x}) \propto \log \mathcal{L}(\boldsymbol{x}|\theta) + \log \pi(\theta)$. Adopting a weighted density $\pi_W(\theta)$ adds an extra term in the log-posterior, introducing extra flexibility in the way the shrinkage is performed: $\log \pi(\theta|\boldsymbol{x}) \propto \log \mathcal{L}(\boldsymbol{x}|\theta) + \log \pi(\theta) + \log w(\theta; \xi)$. Then specifying an observations-dependent or group-dependent weight function would be an interesting way to improve penalization methods, achieving a combination between local and global shrinkage, in the spirit of Carvalho et al. (2009). We leave this open issue for future research and, for the rest of the paper, we focus on non-local distributions, that can be seen as a particular subset of the weighted density family, and their use for likelihood specification.

4.2.2 Non-local two-group Model

In a multiple hypotheses testing framework, let us denote the set of N hypotheses with $\boldsymbol{H} = \left(H^{(1)}, \dots, H^{(N)}\right)$. Often, we reduce the evaluation of each hypothesis $H^{(i)}$, $i = 1, \dots, N$, to a corresponding test statistic z_i, sometimes opportunely transformed in a way that

$$H_0^{(i)} : z_i \sim f_0 \quad \text{vs} \quad H_1^{(i)} : z_i \sim f_1. \tag{4.4}$$

where f_0 and f_1 are referred as the *null distribution* ad the *alternative distribution*, respectively. Assuming the exchangeability of the test statistics, the two-group model rephrases (4.4) into the following mixture:

$$z_i|\rho, f_0, f_1 \overset{iid}{\sim} f = (1 - \rho)f_0 + \rho f_1 \tag{4.5}$$

where $\rho \in (0, 1)$ is the mixture proportion. Theoretically, if all the common Normal sampling assumptions are met, the distribution f_0 should coincide with a $N(0, 1)$ under the null scenario. However, a *theoretical null* density $\phi_0(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ could be too restrictive in practice, because of failed model assumptions, unobserved covariates, correlation of the measurements across subjects and within the same subject (Efron, 2007). Thus, we estimate an *empirical null* distribution, which should be "close" to a Gaussian but with estimated mean and variance. From now on, to ease the notation, we will drop the z variable when referring to the Normal density: $\phi(z; \mu, \sigma^2) = \phi(\mu, \sigma^2)$.
We model f_0 as a Normal distribution $\phi_0(\mu_0, \sigma_0^2)$, with Normal-Inverse Gamma prior concentrated around $(0, 1)$ for (μ_0, σ_0^2). In contrast, we model f_1 with a non-local distribution of the form $\pi_W(z; \lambda) = w(z; \xi)\pi(z; \lambda)$, employing a weighted density properly chosen to induce null mass in the origin to ensure separation from f_0. For the cases under study, we noticed that parsimonious models are sufficient to ensure the desired amount of flexibility. To balance flexibility and tractability, we then adopt $\pi(z; \lambda) = (1 - \alpha)\phi_1(z; \mu_1, \sigma_1^2) + \alpha\phi_2(z; \mu_2, \sigma_2^2)$, with $\alpha \in (0, 1)$ to be a mixture of two different Normals, in order to account for asymmetries in the tails of f. But in principle, nothing prevents from resorting to more flexible priors if suggested by the

application, for instance with $\pi(z;\lambda)$ modeled by Dirichlet Process Mixture Model (Escobar and West, 1995; Antoniak, 1974), since whichever distribution we choose, the weight will ensure separation in a neighborhood of the origin as long as $w \to 0$ if $z \to 0$.

Let $\boldsymbol{z} = (z_1, \dots, z_N)$ be a collection of test statistics. For exchangeable data, the likelihood can be written as $\mathcal{L}\left(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}\right) = \prod_{i=1}^{N} \mathcal{L}\left(z_i|\tilde{\boldsymbol{\theta}}\right)$, where

$$\mathcal{L}\left(z_i|\tilde{\boldsymbol{\theta}}\right) = (1-\rho)\,\phi_0 + \rho \underbrace{\left[\frac{w(z;\xi)}{\tilde{\mathcal{K}}(\tilde{\boldsymbol{\theta}}_1)}\left[(1-\alpha)\phi_1 + \alpha\phi_2\right]\right]}_{\pi_W} \tag{4.6}$$

where $\tilde{\boldsymbol{\theta}} = \left(\rho, \alpha, \{\mu_j, \sigma_j^2\}_{j=0}^2, \xi\right)$ is the vector that collects all the parameters of the model, $\tilde{\boldsymbol{\theta}}_1 = \left(\alpha, \{\mu_j, \sigma_j^2\}_{j=1}^2, \xi\right)$ is the subvector that contains all the parameters that pertain to the alternative distribution, and $\mathcal{K}$ is the normalizing constant of the latter, given by

$$\tilde{\mathcal{K}} = \tilde{\mathcal{K}}(\xi, \alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \int_{\mathbb{R}} w(z;\xi)\left[(1-\alpha)\phi_1 + \alpha\phi_2\right] dz.$$

The proportion parameters ρ and α belong to $(0,1)$. To proceed in a fully Bayesian setting, we need to provide prior distributions for all the parameters contained in $\tilde{\boldsymbol{\theta}}$. We introduce, for inference and computational advantages, the following latent variables: $\boldsymbol{\Lambda} = (\lambda_1, \dots, \lambda_N)$ and $\boldsymbol{\Gamma} = (\gamma_1, \dots, \gamma_N)$, where each λ_i is binary and $\gamma_i \in \{-1, 0, 1\}$. For all $i = 1, \dots, N$, these variables indicate from which model the observation is extracted namely: from ϕ_0 when $(\lambda_i = 0, \gamma_i = -1)$, from ϕ_1 when $(\lambda_i = 1, \gamma_i = 0)$ or from ϕ_2 when $(\lambda_i = 1, \gamma_i = 1)$. The remaining configurations are assumed to have null probability. Consequently we redefine $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}_1$ as $\boldsymbol{\theta} = \left(\{\mu_j, \sigma_j^2\}_{j=0}^2, \xi, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}\right)$ and $\boldsymbol{\theta}_1 = \left(\{\mu_j, \sigma_j^2\}_{j=1}^2, \xi, \boldsymbol{\Gamma}\right)$, respectively. The likelihood can be rewritten as

$$\mathcal{L}(z_i|\boldsymbol{\theta}) = (1-\lambda_i)\,\phi_0 + \lambda_i \left[\frac{w(z;\xi)}{\mathcal{K}_{\gamma_i}(\boldsymbol{\theta}_1)}\left[\phi_1\delta_0(\gamma_i) + \phi_2\delta_1(\gamma_i)\right]\right], \tag{4.7}$$

where $\delta_x(y)$ is the usual Dirac delta, non-null and equal to one only when $y = x$. We call the distribution stated in Equation (4.7) non-local likelihood (NoLLik). Thanks to the introduction of the latent variables γ_i, we are able to simplify the expression of the normalizing constant as $\mathcal{K}_{\gamma_i}(\xi, \mu_j, \sigma_j^2) = \mathbb{E}_{\phi\left(\mu_j, \sigma_j^2\right)}\left[w(z;\xi)\right]$, where $j = 1$ if $\gamma_i = 0$ and $j = 2$ if $\gamma_i = 1$. The model is formulated as follows:

$$\begin{aligned}
z_i|\boldsymbol{\theta} &\overset{i.i.d}{\sim} NoLLik\left(\cdot|\lambda_i, \gamma_i, \{\mu_j, \sigma_j^2\}_{j=0}^2, \xi\right) \\
\pi(\gamma_i|\lambda_i, \alpha) &\overset{i.i.d}{=} \delta_{-1}(\cdot)\,\delta_0(\lambda_i) + \delta_0(\cdot)\,\delta_1(\lambda_i)\,(1-\alpha) + \delta_1(\cdot)\,\delta_1(\lambda_i)\,\alpha; \\
\lambda_i|\rho &\overset{i.i.d}{\sim} Bern(\rho); \\
\left(\mu_j, \sigma_j^2\right) &\sim NIG(m_j, \kappa_j, a_j, b_j)\,\mathbb{I}_{\mu_j \in \mathcal{M}_j}, \qquad j = 0, 1, 2 \\
\rho &\sim Beta(a_\rho, b_\rho), \qquad \alpha \sim Beta(a_\alpha, b_\alpha), \qquad \xi \sim \Xi
\end{aligned} \tag{4.8}$$

where $\mathcal{M}_0 = \mathbb{R}$, $\mathcal{M}_1 = \mathbb{R}^-$, $\mathcal{M}_2 = \mathbb{R}^+$ and Ξ is the law that describes the distribution of the parameters in the weight function. The indicator functions over $\{\mathcal{M}_j\}_{j=0}^2$ for μ_j separate the location parameters of the mixture of Normals in f_1: one component on the negative semi-axis, one component on the positive one, avoiding identifiability problems.
Let us underline the hierarchical dependence among the latent indicators $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$. According to the data generating mechanism, the observation i is assigned either to the null ($\lambda_i = 0$) or to

the alternative distribution ($\lambda_i = 1$). If it is assigned to the null, the second indicator loses its meaning, so we set $\gamma_i = -1$. On the other hand, if $\lambda_i = 1$, then observation i con be allocated to the positive ($\gamma_i = 1$) or to the negative ($\gamma_i = 0$) component of the mixture. Figure 4.1 helps visualize this scheme, reporting the a priori joint probabilities of each couple of values.

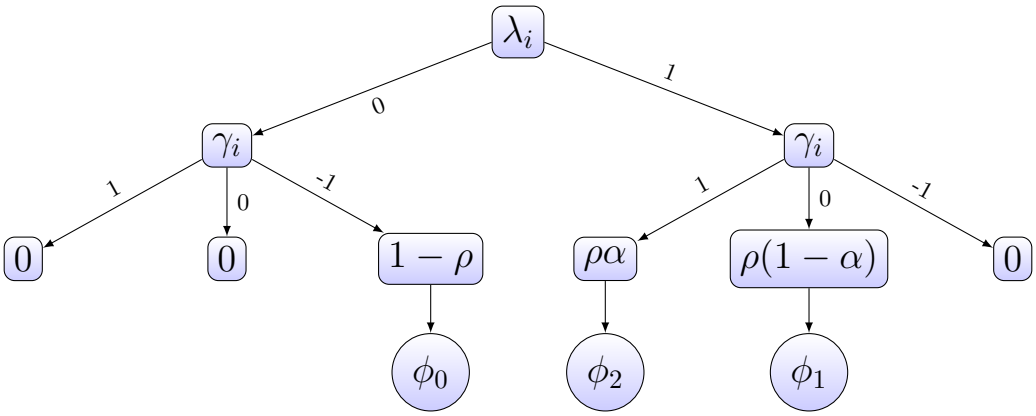


FIGURE 4.1: Visualization of the dependence between the latent variables. The rectangular leaves of the tree report the variables of the model. The bottom line of rectangular leaves contains the r.v.s representing probabilities: connected to them we find circular leaves, containing the corresponding densities for each scenario.

A Bayesian Nonparametric Alternative

The model in eq. (4.8) is efficient and parsimonious. However, one could argue that the alternative distribution a priori should be unknown and thus estimated in a more flexible way. We can employ a weighted Dirichlet Process Mixture Model (DPMM) for the distribution of f_1. To build the model we first consider a classical DP prior and then apply our weight function, properly normalized:

$$f_1 = \frac{w(z;\xi)}{\tilde{\mathcal{K}}(\tilde{\boldsymbol{\theta}}_1)} \int \phi(\vartheta) P(d\vartheta), \qquad P \sim DP(a, G)$$

where $\vartheta = (\mu, \sigma^2)$ and DP is the Dirichlet Process with concentration parameter a and base measure G (Ferguson, 1973). Adopting the DP Stick Breaking representation of Sethuraman (1994), we can extend the likelihood in (4.6) writing

$$\mathcal{L}\left(z_i|\tilde{\boldsymbol{\theta}}\right) = (1-\rho)\,\phi_0 + \rho \left[\frac{w(z;\xi)}{\tilde{\mathcal{K}}(\tilde{\boldsymbol{\theta}}_1)} \left[\sum_{j=1}^{+\infty} \pi_j \phi_j\right]\right] \tag{4.9}$$

where $\tilde{\boldsymbol{\theta}} = \left(\rho, \boldsymbol{\pi} = \{\pi_j\}_{j\geq 1}, \vartheta_j = \{\mu_j, \sigma_j^2\}_{j=0}^{+\infty}, \xi\right)$, $\tilde{\boldsymbol{\theta}}_1$ contains only the parameters relative to f_1, $\boldsymbol{\pi} \sim SB(a)$, meaning that $\pi_j = u_j \prod_{l=1}^{j-1}(1-u_l)$ with $u_l \sim Beta(1, a)$ for $l \leq J$, and $\vartheta_j \sim G$. The normalizing constant can be rewritten as

$$\tilde{\mathcal{K}} = \int_{\mathbb{R}} w(z;\xi) \sum_{j=1}^{+\infty} \pi_j \phi_j dz = \sum_{j=1}^{+\infty} \pi_j \int_{\mathbb{R}} w(z;\xi)\phi_j dz = \sum_{j=1}^{+\infty} \pi_j \mathbb{E}_{\phi_j}\left[w(z,\xi)\right].$$

applying the Dominated Convergence Theorem for series of probability densities, assuming boundedness of the weight function.

We remark that the model still consists of two nested mixtures: one between f_0 and f_1, the other one defining f_1. Again we introduce latent allocation variables to ease the computational aspects of the model. The only difference with eq. (4.7) is the support of the variables γ_i, that we opportunely change in $\gamma_i \in \{0, 1, 2, \dots\}$, being zero whenever $\lambda_i = 0$. Then the likelihood becomes

$$\mathcal{L}(z_i|\boldsymbol{\theta}) = (1-\lambda_i)\,\phi_0 + \lambda_i \left[\frac{w(z;\xi)}{\mathcal{K}_{\gamma_i}(\boldsymbol{\theta}_1)}\left[\sum_{j=1}^{+\infty}\phi_j\delta_j(\gamma_i)\right]\right] \tag{4.10}$$

or, shortly, $\mathcal{L}(z_i|\boldsymbol{\theta}) = (1-\lambda_i)\,\phi_0 + \lambda_i\left[\frac{w(z_i;\xi)}{\mathcal{K}_{\gamma_i}(\boldsymbol{\theta}_1)}\phi_{\gamma_j}\right]$. We name this density BNP-Nollik. The formulation in (4.10) allows us to rephrase the normalizing constant $\mathcal{K}$ in simple terms. In fact

$$\tilde{\mathcal{K}}_{\gamma_i}(\boldsymbol{\theta}_1) = \int w(z;\xi)\sum_{j=1}^{+\infty}\phi_j\delta_j(\gamma_i)dz = (\boldsymbol{\theta}_1) = \int w(z;\xi)\phi_j dz$$

letting us integrate one mixture component at a time, greatly simplifying the posterior simulation. Model eq. (4.8) becomes

$$\begin{aligned}
z_i|\boldsymbol{\theta} &\overset{i.i.d}{\sim} \text{BNP-}NoLLik\left(\cdot|\lambda_i,\gamma_i,\{\mu_j,\sigma_j^2\}_{j=0}^2,\xi\right),\\
\pi(\gamma_i|\lambda_i,\alpha) &\overset{i.i.d}{=} \delta_0(\cdot)\,\delta_0(\lambda_i) + \delta_1(\lambda_i)\left[\sum_{j\geq 1}\pi_j\delta_j(\cdot)\right],\\
\lambda_i|\rho &\overset{i.i.d}{\sim} Bern(\rho), \qquad \rho \sim Beta(a_\rho, b_\rho), \qquad \boldsymbol{\pi} \sim SB(a), \qquad \xi \sim \Xi\\
\left(\mu_0,\sigma_0^2\right) &\sim NIG(m_0,\kappa_0,a_0,b_0), \qquad \left(\mu_j,\sigma_j^2\right) \sim G = NIG(m_G,\kappa_G,a_G,b_G),
\end{aligned} \tag{4.11}$$

where m_G, κ_G, a_G, b_G denote the hyperparameters of the base measure.

Extensions to a covariate-based MHT framework can be achieved without increasing the complexity of the model, introducing $\boldsymbol{\Lambda}$ dependent on some covariates $\boldsymbol{X} = (X_1, \dots, X_p)$. Specifying for every i:

$$\lambda_i \sim Bern(p_i), \quad p_i = g(\boldsymbol{X}_i, \boldsymbol{\eta}), \tag{4.12}$$

where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})$ is the vector of covariate values for individual i. This formulation has two main advantages. First, the tractability of the MCMC is not altered, being $\boldsymbol{\Lambda}$ separated from the other parameters in the hierarchical structure. Second, with this formulation, the covariates impact directly the parameters that drive the allocation to the two latent classes. The function g can be assumed to be a Gaussian process accordingly constrained (Riihimäki and Vehtari, 2010; Tolvanen, 2014) or, more commonly, a link function as the Logistic or the Probit (Normal c.d.f., denoted with $\Phi(x)$). The latter case is particularly appealing, since it has been recently proven in Durante (2019) that for Probit regression $g = \Phi(\boldsymbol{X}_i'\boldsymbol{\eta})$, if we assume Gaussian distribution for every component of $\boldsymbol{\eta}$, the posterior has a closed form and belongs to the Unified Skew-Normal (SUN) family.

4.2.3 On the choice of the weight function

Quite generally, to impose non-locality around the origin, we only need: (i) $w(0;\xi) = 0$; (ii) $w(z,\xi)$ non-decreasing for $z > 0$ and non-increasing for $z < 0$; (iii) weight that shrinks the density towards zero in its neighborhood, without over-inflating the density far away from zero: weight function whose growing rate to infinity is higher than the decreasing rate to zero of the local distribution should be avoided; (iv) w bounded by a constant K, even if not strictly necessary. Without loss of generality, we assume $K = 1$ (the difference will be embodied in the normalizing constant $\mathcal{K}$); (v) $w(z;\xi)$ continuous, to avoid discontinuity points in the resulting

likelihood; (vi) w symmetric w.r.t. 0, in the case we have no a priori information, avoiding to favor deviations of one direction over the other. We call a weight function satisfying properties (i)-(v) a *smooth* weight function, and *smooth symmetric* if (vi) is also satisfied.
Following the above requirements, we will consider and suggest smooth symmetric weight functions, bounded between $[0, 1]$. Two functions appear particularly well behaved, and we will compare their performance in the simulation study below:

$$w_1(z;\xi,k) = 1 - \exp\left[-\left(\frac{z}{\xi}\right)^{2k}\right], \qquad w_2(z;\xi,k) = \exp\left[-\left(\frac{z}{\xi}\right)^{-2k}\right]. \tag{4.13}$$

Apparently similar, the two functions behave differently in the way they approach the origin: w_1 has the same behavior of a Gaussian density, whilst w_2 mimics an Inverse Gamma. The latter approaches faster zero, leading to bigger areas of low density for the same value of k and ξ. Notice that the second weight is exactly the eMOM term, up to a proportionality constant $c = e^{\sqrt{2}}$ that cancels out with the normalizing constant.
Both ξ and k are parameters useful to tune the amount of separation. As they increase, the weight function decreases faster towards zero. Examples are reported in Figure 4.2, where different behaviors corresponding to different values of ξ and k are reported. The first two panels show the weight function w_1 when the values of ξ (I), and k (II) vary between 1 and 4, fixing the other parameter equal to 2. The second two panels show the same for the weight function w_2. We can appreciate the different effects the two parameters induce on the chosen functions: while ξ (left column) affects the functions globally imposing a milder growth as the parameter increases, k (right column) instead affects the function only in a neighborhood of the origin.

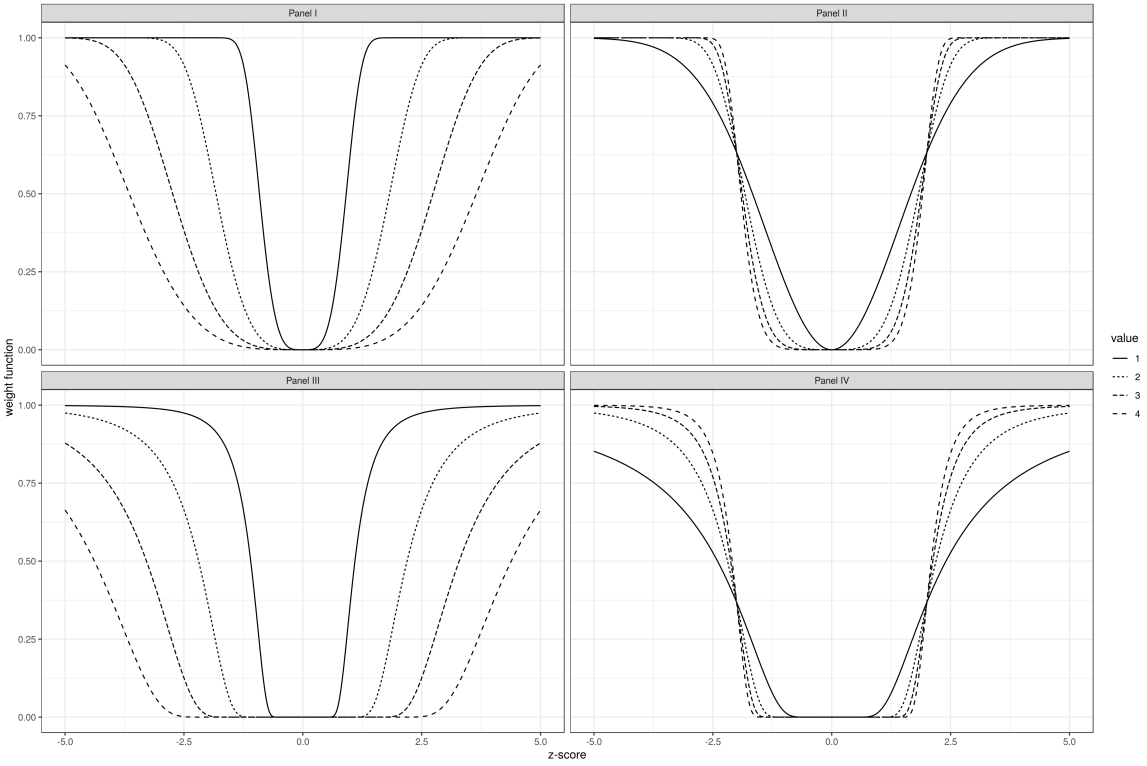


FIGURE 4.2: The panel in the first row report the different behaviours of the weight function w_1 when the *value* of ξ (I) and k (II) changes, keeping the other parameter fixed equal to 2. The second row shows the same for the weight function w_2.

Outside the multiple hypotheses testing framework, a NLD can be employed to model multivariate data. Because of their quadratic structure, the weight functions can be readily extended to the multivariate case, following Johnson and Rossell (2010). Given a $d-$dimensional vector $\boldsymbol{z}$, we can define the quantity

$$Q(\boldsymbol{z}) = \frac{(\boldsymbol{z} - \boldsymbol{z}_0)' \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{z}_0)}{n\xi\sigma^2},$$

where Σ is a positive definite matrix and σ^2 and ξ are scalars. The latter controls the dispersion around $\boldsymbol{z}_0$. For example, it is straightforward to extend the weight functions in (4.13) to the multivariate case:

$$w_1(\boldsymbol{z};\xi) = 1 - \exp\left[-Q(\boldsymbol{z})^k\right], \qquad w_2(\boldsymbol{z};\xi) = \exp\left[-Q(\boldsymbol{z})^{-k}\right]. \tag{4.14}$$

4.3 Theoretical justifications

In the present section we show that the NLD under the alternative hypotheses improves the performance of the two-group model, in terms of lower False Discovery Rate (FDR), lower False Negative Rate (FNR), and lower Type II Error (β), relative to the same model with unweighted alternative hypotheses. This is always true for a smooth symmetric alternative multiplied by a symmetric weight function, and we provide guidance and conditions for the more general asymmetric case.

Consider a generic density - also called *local density* as opposed to NLD - $f_1(z)$ for the alternative distribution, and its weighted counterpart $f_1^W(z) = w(z;\xi)f_1(z)/\mathcal{K}$, where $\mathcal{K} = \int_{-\infty}^{\infty} w(z;\xi)f_1(z)dz$ and $w(z;\xi)$ is a non-local weight function. Let $F_1(z)$ and $F_1^W(z)$ the corresponding c.d.f.s. The decision process reduces to the specification of an interval $\mathcal{A} = [z_1, z_2]$ outside which z-scores are flagged as interesting. Without loss of generality, we assume $z_2 > 0$ and $z_1 < 0$. We call $\mathcal{A}$ the *acceptance region* of the decision process.

Recall that $FDR(\mathcal{A}) = \mathbb{P}[H_0|z \notin \mathcal{A}]$, $FNR(\mathcal{A}) = \mathbb{P}[H_1|z \in \mathcal{A}]$ and $\beta(\mathcal{A}) = \mathbb{P}[z \in \mathcal{A}|H_1]$. Define the following differences: $\Delta FDR = FDR(\mathcal{A}) - FDR^W(\mathcal{A})$, $\Delta FNR = FNR(\mathcal{A}) - FNR^W(\mathcal{A})$, and $\Delta\beta = \beta(\mathcal{A}) - \beta^W(\mathcal{A})$, i.e. the discrepancies between the aforementioned indexes in their unweighted and weighted versions. In the Appendix, we show that all these differences simplify to the sum of two discrepancies among the alternative c.d.f. F_1 and the weighted alternative c.d.f. F_1^W:

$$\begin{aligned} & FDR(\mathcal{A}) - FDR^W(\mathcal{A}) \geq 0 \\ \iff \quad & FNR(\mathcal{A}) - FNR^W(\mathcal{A}) \geq 0 \\ \iff \quad & \beta(\mathcal{A}) - \beta^W(\mathcal{A}) \geq 0 \\ \iff \quad & F_1(z_2) - F_1^W(z_2) + F_1^W(z_1) - F_1(z_1) \geq 0. \end{aligned} \tag{4.15}$$

Then a sufficient condition for Equation (4.15) to hold is that the weighted c.d.f. has to be lower than its unweighted counterpart in z_2, and higher in z_1.

Proposition 3. *Given a smooth symmetric weight function w, a symmetric density f_1, an acceptance region $\mathcal{A}$, and a null hypothesis $H_0 : z \sim f_0$, the test $H_1^W : z \sim f_1^W$ has higher power, lower FDR and lower FNR than the test $H_1 : z \sim f_1$.*

Proof. Consider a generic random variable Z, characterized by a local density $f(z)$ and its weighted, non-local version Z_W, with $f_W(z)$. Let $Z^T = |Z|$ and $Z_W^T = |Z_W|$ denote the truncations of the r.v.s on the positive semi-axis. Thanks to the symmetry of the distributions, we can state that $f^T(z) = 2f(z)\mathbb{I}_{[0,+\infty)}$ and let $F^T(z) = 2F(z) - 1$ for $z > 0$ be its c.d.f. The same can be said about $f_W^T(z)$. Applying Lemma 2 in Dharmadhikari and Joag-Dev (1983) –

reported in the Appendix – we want to conclude that Z_W^T stochastically dominates (I order) Z^T, i.e. $\forall z\,[0,+\infty)$, meaning that $F^T(z) \geq F_W^T(z)$ and $F^T(z) > F_W^T(z)$ for at least one z.
To verify condition (a) of the Lemma, we need to study the sign of $\Delta(z) = f^T(z) - f_W^T(z) = (1 - w(z;\xi)/\mathcal{K})f^T(z)$. The function $g(z) = 1 - w(z,\xi)/\mathcal{K}$ is monotone decreasing, given the monotonicity of $w(z,\xi)$ on the positive semi-axis. Moreover, $\mathcal{K} \leq K$. Thus, $g(z)$ is 1 in zero and for $z \to +\infty$ it tends to $1 - K/\mathcal{K} \leq 0$ and admits an unique root z^*. The monotonicity of g and the positivity of f^T imply that $\Delta(z) = g(z) \cdot f^T(z)$ has only one zero, occurring in z^* as well. Since $\lim_{z\to z^{*}+} \Delta(z) = 0^-$ and $\lim_{z\to z^{*}-} \Delta(z) = 0^+$, we can conclude that f crosses f^T just once and from above, so the condition is satisfied.
This in turn implies that, $\forall z \geq 0$, $F^T(z) \geq F_W^T(z) \iff 2F(z) - 1 \geq 2F_W(z) - 1 \iff F(z) \geq F_W(z)$, showing that, on the positive semi-axis, the c.d.f. of Z is always greater than its weighted counterpart. Exploiting the symmetry of the densities of the two random variables Z and Z_W, the converse holds on the negative semi-axis, implying (4.15).

□

Note that for this proof to hold, the conditions (iv) and (v) stated in Section 4.2.3 are not strictly necessary, and can be dropped as long as the resulting weighted density is proper.
When the alternative density is regular but asymmetric, we need to introduce further assumptions. The following proposition provides guidance in this case.

Proposition 4. *Consider a symmetric weight function, monotone on both the semi-axes and bounded by a positive constant K. Denote with $g(z) = 1 - w(z;\xi)/\mathcal{K}$ and with $\pm z^*$ the solutions of the equation $g(z) = 0$. Moreover, let*

$$H(z) = F_1(z) - F_1^W(z) = \int_{-\infty}^{z} g(x) f_1(x) dx$$

and call $\hat{z}$ the unique point where $H(\hat{z}) = 0$. Then (4.15) holds if $\mathcal{A}$ is such that such that $z_1 < \hat{z} < z_2$. A stronger condition for (4.15) to hold is $z_1 < -z^*$ and $z^* < z_2$.

Proof. We have that $\lim_{x\to\pm\infty} w(x;\xi) = K$. Let $\pm z^* = \pm \sup\{z \geq 0 : g(z) > 0\}$. If w is monotone on both the semi-axes, then $\pm z^* = \pm w^{-1}(\mathcal{K})$. Now consider the function $H(z) = F_1(z) - F_1^W(z) = \int_{-\infty}^{z} g(x) f_1(x) dx$. In general, $\lim_{x\to-\infty} H(z) = 0^-$ and $\lim_{x\to+\infty} H(z) = 0^+$. We can study the sign of the derivative $H'(z) = g(z) f_1(z)$: H starts negative, decreases until its point of global minimum $-z^*$, then increases until its point of global maximum z^* and finally decreases towards zero from above as $z \to +\infty$. Let us call $\hat{z}$ the *unique* point where $H(\hat{z}) = 0$. It must be that $\hat{z} \in [-z^*, z^*]$. Since (4.15) can be reprhased as $H(z_2) - H(z_1)$, the previous conditions ensure that this difference is positive.

□

For (4.15) to hold, we can assess, a posteriori, if the obtained $\mathcal{A}$ satisfies the following condition: $z_1 < \hat{z} < z_2$. Or, if more caution is needed, we can tune the weight function so that $z_1 < -z^*$ and $z^* < z_2$, an even more conservative requirement .
We remark that these guidelines are extremely general, holding every time a two-tailed test is adopted: given a fixed acceptance region, the NLD performs better.

In Bayesian MHT, we recover $\mathcal{A}$ thresholding the a posteriori probability of inclusion in the alternative distribution, which, in the two-group model, is equivalent to thresholding the posterior $lfdr$, defined as

$$lfdr(z) = \frac{(1-\rho)f_0(z)}{f(z)} = \frac{(1-\rho)f_0(z)}{(1-\rho)f_0(z) + \rho f_1(z)}. \tag{4.16}$$

In detail, $\mathcal{A}$ is defined as

$$\mathcal{A} = \left\{ z \in \mathbb{R} : lfdr(z) = \frac{(1-\rho) f_0(z)}{f(z)} \leq \nu^* \right\} = \left\{ z \in \mathbb{R} : P_1(z) = \frac{\rho f_1(z)}{f(z)} \geq \nu \right\}, \tag{4.17}$$

where ν^* and ν are two suitable thresholds on $(0,1)$ such that $\nu = 1-\nu^*$ and $P_1(z)$ is the (local) posterior probability of inclusion in the alternative scenario when the value z is observed.

Example. Consider a generic alternative density f_1 with c.d.f. F_1 and the weight function $w(z;\xi) = \mathbb{1}_R$, where $R = (-\infty, -\delta) \cup (\delta, +\infty)$. This gives us $\mathcal{K} = F_1(-\delta) + 1 - F_1(\delta)$ and $z^* = \pm\delta$. We can derive a close expression for $H(z)$:

$$H(z) = F_1(z) - \frac{1}{\mathcal{K}} \left[\mathbb{1}_{z<-\delta} F_1(z) + \mathbb{1}_{z\in(-\delta,\delta)} F_1(-\delta) + \mathbb{1}_{z>\delta} \left(F_1(z) - F_1(\delta) + F_1(-\delta) \right) \right].$$

Knowing that the only root $\hat{z}$ can only be in $(-\delta, \delta)$, we can compute

$$\hat{z} = F_1^{-1} \left(\frac{F_1(-\delta)}{\mathcal{K}} \right).$$

In this particular example, applying the criterion showed in (4.17) we conclude that $\forall z \in R$ we have that $P_1^W = 0$. Thus, we are sure that, for $\nu > 0$, $\mathcal{A}$ can be either $(-\delta, \delta)$ or wider. This means that condition (4.15) is respected and that FNR, FDR and β are lower in this weighted case.

4.4 Posterior Computation

The posterior distribution $\pi(\boldsymbol{\theta}|z)$ for model (4.8) is not analytically tractable and we need to rely on MCMC sampling techniques. To this extent, we propose a Gibbs sampler (Gelfand et al., 1990). The details of the full conditional distributions are reported in section 4.4.2. For the parametric model, the full conditional distributions for ξ and $\left(\mu_j, \sigma_j^2\right)$ $j = 1, 2$ require a Metropolis step (Metropolis et al., 1953). In particular, we employ a random walk Metropolis algorithm (RWM), with Gaussian proposal distribution, for the means and the log-variances. We also adopt an adaptive strategy, to better tune the variance of the proposal during the run of the algorithm, as suggested in Roberts and Rosenthal (2007). Starting from covariance matrices $\Sigma_{(\mu_1,\sigma_1^2)}$,$\Sigma_{(\mu_2,\sigma_2^2)}$ and the scalar σ_ξ^2, every n_{batch} MCMC samples the values are updated in the following way: at the t-th iteration if the acceptance rate in the last examined batch is lower than the optimal rate of 0.44, the logarithm of the standard deviation is lowered by the quantity $\delta(t) = \min\left(0.01, t^{-1/2}\right)$, otherwise it is increased of the same quantity. Notice that the adaptive term is vanishing, so the convergence to the desired target distribution is preserved (Roberts and Rosenthal, 2007; Roberts and Rosenthal, 2009).

4.4.1 Inference on f_0 and f_1

The fully Bayesian specification of the model allows the estimation of the parameters and their functions, along with their uncertainty quantification. In particular, we are interested in P_1, the a posteriori probability that an observation is marked as relevant, i.e. generated from f_1. A simple way to estimate P_1 exploits the presence of latent variables in our MCMC scheme and permits the computation of an estimate for each of the observations, namely $\hat{P}_{1i}$. We only need to evaluate the ergodic mean $\hat{P}_{1i} = \sum_{t=1}^{T} \lambda_{it}/T$, where T is the number of iterations and λ_{it} is the value of the chain for the parameter λ_i at the t-th MCMC sweep. Another solution involves the local FDR ($lfdr$). Once f_0 and f_1 are given after imputing the posterior means of

the parameters, we can compute the $lfdr$. It is easy to show that, applying Bayes theorem to eq. (4.5), it holds that $1 - P_1(z) = lfdr(z)$. This means that we are able to estimate the $lfdr$ as a function of the z-score and, more importantly, we can effortlessly derive $P_1(z)$, the probability of the alternative hypothesis, also as a function of z. Moreover, we underline that the adoption of a non-local likelihood makes the function in (1.22) assume all the values in $[0, 1)$ reflecting the rationale that if $z = 0$ is observed, we want to, almost surely, mark it as "irrelevant". Do et al. (2005) argue that this second estimate is preferable since it is the Rao-Blackwellized version (where the λ_i's are integrating out) of the previous one (Casella and Robert, 1996).

The next step is to classify the observations according to their distribution of origin. We can accomplish this task thresholding the point-wise estimates for P_{i1}. Thresholding the function $P_1(z)$ we are able to derive the corresponding critical values (z_1, z_2) on the z-scores domain. We choose a threshold that guarantees to control, at a given level α, the Bayesian FDR (BFDR, Newton et al. 2004), defined, as a function of the threshold ν, as

$$\text{BFDR} = E(\text{FDR}|\text{Y}) = \frac{\sum_{i=1}^{N} (1 - P(z_i)) \, \mathbb{I}_{\{P(z_i) > \nu\}}}{\sum_{i=1}^{N} \mathbb{I}_{\{P(z_i) > \nu\}}}. \tag{4.18}$$

Given a specified level of α, we obtain the threshold ν^* solving the inequality $\text{BDFR}(\nu) < \alpha$ for ν. Thresholding $P_1(z)$ is equivalent to thresholding the $lfdr$, a procedure that mimics the Bayes oracle rule (Efron, 2008; Muralidharan, 2012; Sun and Cai, 2007; Martin and Tokdar, 2012).

4.4.2 Gibbs Sampler

Parametric model specification

In what follows, we detail the steps of the Gibbs sampler for the Nollik model in (4.8). The algorithm proceeds iteratively sampling from the following full conditionals:

1. The full conditional of ρ is $Beta\left(a_\rho + \sum_{i=1}^{N} \lambda_i, b_\rho + N - \sum_{i=1}^{N} \lambda_i\right)$, due to conjugacy.

2. Given their dependence, we sample together $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ from their joint full conditional, given by:

$$\pi\left((\boldsymbol{\Lambda}, \boldsymbol{\Gamma}) \,|\, \cdots\right) = \prod_{i=1}^{N} Bern(\lambda_i; \rho) \cdot \pi\left(\gamma_i | \lambda_i, \alpha\right) \times \left((1 - \lambda_i)\, \phi_0 + \lambda_i \left[\frac{w(z;\xi)}{\mathcal{K}_{\gamma_i}(\boldsymbol{\theta}_1)} \left[\phi_1 \delta_0\left(\gamma_i\right) + \phi_2 \delta_1\left(\gamma_i\right)\right] \right] \right)$$

   We can update each component of $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ individually, rewriting:

$$\pi\left((\lambda_i, \gamma_i) \,|\, \cdots\right) \propto \rho^{\lambda_i} (1 - \rho)^{1 - \lambda_i} \pi\left(\gamma_i | \lambda_i, \alpha\right) \cdot \left(\phi_0^{1 - \lambda_i} \cdot \left[\frac{w(z;\xi)}{\mathcal{K}_{\gamma_i}(\boldsymbol{\theta}_1)} \left[\phi_1^{\delta_0(\gamma_i)} \cdot \phi_2^{\delta_1(\gamma_i)}\right] \right]^{\lambda_i} \right).$$

   In particular, the only scenarios with non-null probabilities are:

$$\begin{aligned}
\pi\left(\lambda_i = 0, \gamma_i = -1 | \cdots\right) &\propto (1 - \rho)\phi_0 \\
\pi\left(\lambda_i = 1, \gamma_i = 0 | \cdots\right) &\propto (1 - \alpha) \cdot \rho \left[\frac{w(z;\xi)}{\mathcal{K}_0(\boldsymbol{\theta}_1)} \phi_1 \right] \\
\pi\left(\lambda_i = 1, \gamma_i = 1 | \cdots\right) &\propto \alpha \cdot \rho \left[\frac{w(z;\xi)}{\mathcal{K}_1(\boldsymbol{\theta}_1)} \phi_2 \right]
\end{aligned}$$

3. Let $n_{0,1} = \sum_{i=1}^{N} \mathbb{I}_{\{\gamma_i \neq -1\}}$. The full conditional of α is $Beta\left(a_\alpha + \sum_{i=1}^{N} \gamma_i, b_\alpha + n_{0,1} - \sum_{i=1}^{N} \gamma_i\right)$, due to conjugacy.

4. Let $n_0 = \sum_{i=1}^{N} \mathbb{I}_{\lambda_i=0}$. Let us define $\bar{z}_0 = \frac{\sum_{i=1}^{N} z_i \cdot \mathbb{I}_{\lambda_i=0}}{n_0}$ and $SQ_0^2 = \sum_{i=1}^{N} (z_i - \bar{z}_0)^2 \mathbb{I}_{\lambda_i=0}$. The full conditional for (μ_0, σ_0^2), is given by

$$\pi\left(\left(\mu_0, \sigma_0^2\right) | \cdots\right) \sim NIG(m_0^*, \kappa_0^*, a_0^*, b_0^*)$$

where $m_0^* = \frac{\kappa_0 m + n_0 \bar{z}_0}{\kappa + n_0}$, $\kappa^* = \kappa_0 + n_0$, $a_0^* = a_0 + \frac{1}{2} n_0$
and $b_0^* = b_0 + \frac{1}{2} SQ_0^2 + \frac{n_0 \kappa_0}{n_0 + \kappa_0} \frac{(\bar{z}_0 - m_0)^2}{2}$.

5. Let $n_{1j} = \sum_{i=1}^{N} \mathbb{I}_{\lambda_i=1} \cdot \mathbb{I}_{\gamma_i=j-1}$.
Define $\bar{z}_{1j} = \frac{\sum_{i=1}^{N} z_i \cdot \mathbb{I}_{\lambda_i=1} \cdot \mathbb{I}_{\gamma_i=j-1}}{n_{1j}}$ and $SQ_{1j}^2 = \sum_{i=1}^{N} (z_i - \bar{z}_{1j})^2 \mathbb{I}_{\lambda_i=1} \cdot \mathbb{I}_{\gamma_i=j-1}$
The full conditional for $\left(\mu_j, \sigma_j^2\right)$, for $j = 1, 2$ is given by

$$\begin{aligned}
\pi\left(\left(\mu_j, \sigma_j^2\right) | \cdots\right) &\propto NIG(m_j, \kappa_j, a_j, b_j) \cdot \mathbb{I}_{(-1)^j \mu_j > 0} \cdot \prod_{\lambda_i=1, \gamma_i=j-1} \frac{\phi_j}{\mathcal{K}_i(\boldsymbol{\theta}_1)} \\
&\propto NIG(m_j^*, \kappa_j^*, a_j^*, b_j^*) \cdot \mathbb{I}_{\{(-1)^j \mu_j > 0\}} \cdot \frac{1}{\mathcal{K}_{j-1}(\boldsymbol{\theta}_1)^{n_{1j}}}
\end{aligned}$$

where $m_j^* = \frac{\kappa_j m + n_{1j} \bar{z}_{1j}}{\kappa_j + n_{1j}}$, $\kappa_j^* = \kappa_j + n_{1j}$, $a_j^* = a_j + \frac{1}{2} n_{1j}$
and $b_j^* = b_j + \frac{1}{2} SQ_{1j}^2 + \frac{n_{1j} \kappa_j}{n_{1j} + \kappa_j} \frac{(\bar{z}_{1j} - m_j)^2}{2}$.

6. The full conditional of ξ is given by:

$$\pi\left(\xi | \cdots\right) \propto \pi\left(\xi\right) \prod_{\lambda_i=1} \frac{w(z_i; \xi)}{\mathcal{K}_{\gamma_i}(\boldsymbol{\theta}_1)} = \frac{\pi\left(\xi\right) \cdot \prod_{\lambda_i=1} w(z_i; \xi)}{\mathcal{K}_0(\boldsymbol{\theta}_1)^{n_{11}} \cdot \mathcal{K}_1(\boldsymbol{\theta}_1)^{n_{12}}}.$$

Notice that Steps 2 and 5 are embarrassingly parallelizable. We observed that, in some real applications, this Gibbs sampler shows poor mixing, especially when the tails of the distribution f are light. To improve the mixing and, at the same time, to provide an alternative for practitioners, we also implement Nollik with `Stan`, the probabilistic programming language which resorts on a no-U-turn sampler (NUTS) Hamiltonian Monte Carlo algorithm (Hoffman and Gelman, 2014; Carpenter et al., 2017).

Nonparametric model specification

To implement the sampling algorithm for the Bayesian nonparametric version of Nollik model, we use the truncated representation of Ishwaran and James (2001), where the infinite sum in (4.6) is substituted with a big enough number of mixture components J. The collapsed Gibbs sampler we employ mimics the finite-dimensional case with few modifications. Recall that now $\gamma_i \in \{0, 1, 2, \ldots\}$. Steps 1 and 4 are unchanged, whilst the others become:

2. The non-null scenarios for $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ are

$$\begin{aligned}
&\pi\left(\lambda_i = 0, \gamma_i = 0\right) \propto (1 - \rho)\phi_0, \\
&\pi\left(\lambda_i = 1, \gamma_i = j\right) \propto \rho \pi_j \left[\frac{w(z; \xi)}{\mathcal{K}_{\gamma_i}} \phi(\mu_{\gamma_i}, \sigma_{\gamma_i}^2)\right] \text{ for } j \in \{1, 2, \ldots, J\}.
\end{aligned}$$

3. Let $n_{1j} = \sum_{i=1}^{N} \mathbb{I}_{\lambda_i=1}\mathbb{I}_{\gamma_i=j}$. The Stick-Breaking weights are constructed with the auxiliary variables u_j, which in turn have full conditionals of the form

$$u_j \sim Beta\left(1 + n_{1j}, a + \textstyle\sum_{l<j} n_{1l}\right) \text{ for } j \in \{1, 2, \ldots, J\}.$$

5. With reference to step 5. in the previous algorithm, now we have:

$$\pi\left(\left(\mu_j, \sigma_j^2\right) | \cdots\right) \propto NIG(m_j^*, \kappa_j^*, a_j^*, b_j^*) \cdot \frac{1}{\mathcal{K}_{\gamma_i}(\boldsymbol{\theta}_1)^{n_{1j}}}.$$

6. Lastly,

$$\pi\left(\xi | \cdots\right) \propto \frac{\pi\left(\xi\right) \cdot \prod_{\lambda_i=1} w(z_i; \xi)}{\mathcal{K}_1(\boldsymbol{\theta}_1)^{n_{11}} \cdots \mathcal{K}_J(\boldsymbol{\theta}_1)^{n_{1J}}}.$$

7. We can place a conjugate Gamma prior, $Gamma(\alpha_a, \beta_a)$, on the concentration parameter a, obtaining $a| \cdots \sim Gamma\left(\alpha_a + (J-1), \beta_a + \sum_{j=1}^{J-1} \log(1-u_j)\right)$.

The code of the Gibbs samplers and the Stan model are available at the this page.
The proposed algorithms are efficient and able to handle large datasets. To provide evidence for this claim, Table 4.1 reports the mean and the standard deviation of the running times (in seconds) that the model takes to complete 1,000 iterations for different sample sizes on an i7-5500U – 2.40GHz laptop, computed over 20 different runs. To compare with, we report the running times in seconds of the BNP version of the model, truncated ad $J = 20$. The parametric model is from 6 to 10 times faster than the BNP version.

	$n = 100$	$n = 500$	$n = 1,000$	$n = 5,000$	$n = 10,000$	$n = 50,000$
Nollik	1.038	1.4444	1.8858	5.9271	10.2036	45.8604
	(0.088)	(0.2031)	(0.1942)	(0.7047)	(0.8043)	(1.8636)
BNP-*Nollik*	6.5350	10.4658	15.2201	50.8832	96.2305	451.3572
	(0.1187)	(0.1242)	(0.1542)	(0.2324)	(0.5391)	(2.8465)

TABLE 4.1: Elapsed time in seconds of the Nollik parametric and nonparametric models using w_1 as weight function to obtain 1,000 iterations for different values of the sample size.

4.5 Applications

4.5.1 Simulated Data

We test the model on 50 datasets generated under 4 scenarios. Each dataset contains 1,000 observations: 90% of the sample is drawn from f_0, the remaining 10% from f_1. The four scenarios distributions are:

- **Scenario S1**: $z_i \sim 0.90\mathcal{N}(0, 1.5) + 0.05\mathcal{N}(5, 1) + 0.05\mathcal{N}(-5, 1)$.
- **Scenario S2**: $z_i \sim 0.90\mathcal{N}(0, 1) + 0.05\mathcal{N}(3, 1) + 0.05\mathcal{N}(-5, 1.5)$.
- **Scenario S3**: each $z_i \sim \mathcal{N}(\gamma_i, 1)$, where γ_i is sampled from the mixture

$$0.90\delta_0 + 0.1\mathcal{N}(-3, 1).$$

This scenario was proposed in Efron (2008), equations (6.1) and (7.1).

- **Scenario S4**: $z_i \sim \mathcal{N}(\gamma_i, 1)$, where γ_i is sampled from the mixture:

$$0.90\delta_0 + 0.10\left(0.5\mathcal{U}_{[-4,-2]} + 0.5\mathcal{U}_{[2,4]}\right).$$

  This scenario is similar to the one proposed in Muralidharan (2012).

The hyperprior parameters' specification is described in what follows.
Parametric case. Regarding ρ and α: on one hand, we first set $a_\rho = 1$ and $b_\rho = 9$ to follow the rationale that only a small proportion of the observations is of interest. On the other hand, we have no a priori information about the inner mixture proportions, so we set $a_\alpha = b_\alpha = 1$. For simplicity, we fix $k = 2$, instead of letting it be random since the fourth power $2k$ provides a good reduction of the weight in a reasonably large neighborhood of the origin. We then assume ξ to be Inverse gamma with hyperparameters $a_\xi = 20$ and $b_\xi = 57$. This choice, a priori, ensures $\mathbb{E}[\xi] = 3$, while the variance is reasonably low and equal to 0.5: in this manner, we enforce a decent level of "repulsion" from the origin. Regarding (μ_i, σ_i^2), for $i = 1, 2$, we set $\kappa_i = 1$, $a_i = 2$, $b_i = 5$. This implies that $\mathbb{E}[\sigma_i^2] \approx 1.67$ and $Var[\sigma_i^2] = 6.25$. In this way we are fairly uninformative while keeping the values of the variances on reasonable levels helping the stability of the simulations. Moreover, we adopt $m_1 = 3$ and $m_2 = -3$. For the Empirical Null distribution, we need to be informative: we set $a_0 = b_0 = 10$ to induce a density for σ_0^2 peaked around 1. We finally set $\kappa_0 = 100$ and $m_0 = 0$. The initial covariance matrices of the jumps or the random walk Metropolis steps are all fixed equal to $\Sigma_{(\mu_1,\sigma_1^2)} = \Sigma_{(\mu_2,\sigma_2^2)} = diag(0.5, 0.5)$ and $\sigma_c^2 = 0.5$. n_{batch} is fixed to 50 iterations.
Nonparametric case. We truncate the SB process at $J = 30$, we fix the concentration parameter to 1 and as base measure for DP we choose $NIG(0, 100, 3, 1)$. All the other specifications are equal to the parametric case.

We ran 50,000 iterations as a burn-in period. We then run the algorithm for 150,000 sweeps and thin the output every 30 iterations to annihilate the autocorrelation, obtaining a sample of 5,000 MCMC sweeps. Visual inspection of the traceplots reveal good mixing and the convergence of the chains was tested by visual inspection and with the help of the usual diagnostics (Plummer et al., 2006). In each scenario, we compute the mean $lfdr$ as a function of the z-score. To evaluate this function in the nonparametric case, instead of saving all the parameters estimated during the MCMC iterations, we evaluate the posterior densities f_0 and f_1 on a grid of points and then we consider their pointwise mean: for T MCMC sweeps, then

$$\hat{f}_0^{post}(x) = \frac{1}{T}\sum_{t=1}^{T}\phi_0(x;\mu_{0,t},\sigma_{0,t}^2) \quad \hat{f}_1^{post}(x) = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{J}\pi_j\phi_j(x;\mu_{j,t},\sigma_{j,t}^2)\right)$$

and we recover all other values by interpolation.

We classify the observations as "interesting" vs "uninteresting" thresholding the posterior probability of inclusion with a value that controls the BFDR (4.18) at a level of 0.05. We compare our method, adopting three different weight functions (quadratic (MOM), w_1, and w_2 as in (4.13)), with the Mixfdr model (Muralidharan, 2012), the LocFdr model (Efron, 2004) and the Benjamini-Hochberg (BH) procedure for adjusting p-values (Benajmini and Hochberg, 1995). For the first two competitors, we threshold the estimated $lfdr$ at 0.20, as suggested in the respective papers. We threshold the BH adjusted p-values at 0.05.
To assess the performance we compute several indexes from the confusion matrix between the predicted and actual classes. Denoting the number of false positive with FP, of false negative with FN, etc, we report the model's Accuracy (ACC), Specificity (SPE), Precision (PRE), and AUC. Moreover, we compare Matthew's Correlation Coefficient (MCC) and the F_1 score, defined

as

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad F_1 = \left(\frac{2}{\text{REC}^{-1} + \text{PRE}^{-1}} \right),$$

since more comprehensive measures of the overall binary classification performance. All the indexes are then synthesized by their mean and standard deviation across the 50 repetitions. Table 4.1 reports the results. In the same way, we report the average of the posterior means of the parameters ρ, α and ξ.

TABLE 4.1: Simulation study. Posterior Probability of Inclusion thresholded for BFDR=0.05. The *locfdr* and *MixFDR* provide estimates for the $lfdr$, with threshold at 0.2 as suggested in Efron (2007) and Muralidharan (2012). BH adjusted p-values thresholded at 5%. The table shows the performance of 7 different models: Nollik with quadratic (MOM), w_1 and w_2 - as in (4.13) -as weight functions, the BNP Nollik with w_1, the MixFDR of Muralidharan (2012), the LocFDR of Efron (2007) and the classical BH procedure. The performances have been measured in terms of Accuracy (ACC), Specificity (SPE), Precision (PRE), and AUC. Moreover, we compare the Matthew's Correlation Coefficient (MCC) and the F_1 score. The highest MCC and F_1 scores among the *Nollik* models and among the competitors are highlighted.

	Nollik: MOM	*Nollik*: w_1	*Nollik*: w_2	BNP-*Nollik*: w_1	MixFDR	LocFDR	BH
Scenario 1							
MCC	0.9327 (0.0168)	0.9348 (0.0161)	0.9359 (0.0184)	**0.9367** (0.0173)	0.9163 (0.0266)	**0.9281** (0.0224)	0.8846 (0.0228)
F1	0.9392 (0.0152)	0.9411 (0.0145)	0.9421 (0.0166)	**0.9428** (0.0157)	0.9219 (0.0256)	**0.9340** (0.0211)	0.8925 (0.0221)
SPEC	0.9907 (0.0033)	0.9916 (0.0030)	0.9932 (0.0029)	0.9937 (0.0028)	0.9983 (0.0015)	0.9968 (0.0023)	0.9749 (0.0064)
ACC	0.9876 (0.0032)	0.9880 (0.0030)	0.9884 (0.0033)	0.9886 (0.0031)	0.9854 (0.0045)	0.9873 (0.0039)	0.9761 (0.0055)
PRE	0.9206 (0.0255)	0.9271 (0.0242)	0.9403 (0.0239)	0.9441 (0.0230)	0.9830 (0.0149)	0.9693 (0.0212)	0.8156 (0.0377)
AUC	0.9985 (0.0007)	0.9985 (0.0007)	0.9985 (0.0007)	0.9983 (0.0009)	0.9985 (0.0007)	0.9968 (0.0031)	—
$\hat{\rho}$	0.1105 (0.0053)	0.1062 (0.0043)	0.0991 (0.0046)	0.1150 (0.0102)	—	—	—
$\hat{\alpha}$	0.4993 (0.0239)	0.4991 (0.0207)	0.5013 (0.1900)	—	—	—	—
$\hat{\xi}$	—	3.9882 (0.1810)	3.9346 (0.5068)	3.0076 (0.0386)	—	—	—
Scenario 2							
MCC	**0.8130** (0.0414)	0.8117 (0.0423)	0.7977 (0.0410)	0.7899 (0.0436)	0.7591 (0.0513)	0.7104 (0.0462)	**0.8207** (0.0309)
F1	**0.8180** (0.0453)	0.8167 (0.0457)	0.8001 (0.0455)	0.7920 (0.0496)	0.7519 (0.0605)	0.6937 (0.0563)	**0.8280** (0.0313)
SPEC	0.9966 (0.0024)	0.9967 (0.0021)	0.9975 (0.0019)	0.9971 (0.0021)	0.9993 (0.0010)	0.9997 (0.0006)	0.9959 (0.0024)
ACC	0.9685 (0.0064)	0.9684 (0.0065)	0.9662 (0.0062)	0.9650 (0.0066)	0.9604 (0.0076)	0.9533 (0.0065)	0.9697 (0.0048)
PRE	0.9600 (0.0247)	0.9605 (0.0225)	0.9695 (0.0215)	0.9650 (0.0234)	0.9911 (0.0126)	0.9959 (0.0088)	0.9530 (0.0260)
AUC	0.9827 (0.0061)	0.9827 (0.0061)	0.9785 (0.0081)	0.9787 (0.0068)	0.9833 (0.0063)	0.9683 (0.0138)	—
$\hat{\rho}$	0.1019 (0.0087)	0.0926 (0.0075)	0.0849 (0.0069)	0.0869 (0.0096)	—	—	—
$\hat{\alpha}$	0.5004 (0.0371)	0.4752 (0.0356)	0.4440 (0.0356)	—	—	—	—
$\hat{\xi}$	—	3.0292 (0.2023)	2.6954 (0.2357)	2.9065 (0.2391)	—	—	—
Scenario 3							
MCC	**0.7124** (0.0475)	0.7111 (0.0462)	0.7039 (0.0479)	0.6965 (0.0541)	0.6428 (0.0461)	**0.6827** (0.0468)	0.6686 (0.0356)
F1	**0.7049** (0.0569)	0.7039 (0.0549)	0.6935 (0.0577)	0.6836 (0.0669)	0.6125 (0.0593)	**0.6656** (0.0591)	0.6538 (0.0443)
SPEC	0.9972 (0.0020)	0.9972 (0.0019)	0.9977 (0.0017)	0.9979 (0.0017)	0.9994 (0.0009)	0.9984 (0.0016)	0.9974 (0.0018)
ACC	0.9536 (0.0078)	0.9534 (0.0075)	0.9524 (0.0078)	0.9515 (0.0082)	0.9440 (0.0078)	0.9494 (0.0079)	0.9475 (0.0058)
PRE	0.9596 (0.0282)	0.9582 (0.0269)	0.9655 (0.0252)	0.9681 (0.0244)	0.9893 (0.0156)	0.9752 (0.0240)	0.9566 (0.0288)
AUC	0.9503 (0.0154)	0.9466 (0.0159)	0.9337 (0.0163)	0.9401 (0.0159)	0.9535 (0.0180)	0.9059 (0.0250)	—
$\hat{\rho}$	0.0851 (0.0118)	0.0790 (0.0110)	0.0726 (0.0105)	0.0823 (0.0123)	—	—	—
$\hat{\alpha}$	0.0304 (0.0109)	0.0321 (0.0109)	0.0278 (0.0081)	—	—	—	—
$\hat{\xi}$	—	2.6198 (0.2256)	2.3663 (0.1832)	2.3845 (0.1294)	—	—	—
Scenario 4							
MCC	**0.6641** (0.0585)	0.6538 (0.0559)	0.6314 (0.0599)	0.6149 (0.074)	0.5602 (0.0726)	0.6349 (0.0581)	**0.6635** (0.0424)
F1	**0.6490** (0.0716)	0.6339 (0.0708)	0.6042 (0.0762)	0.5816 (0.0966)	0.5059 (0.0940)	0.6093 (0.0736)	**0.6471** (0.0513)
SPEC	0.9968 (0.0027)	0.9974 (0.0023)	0.9980 (0.0020)	0.9982 (0.0022)	0.9996 (0.0008)	0.9979 (0.0019)	0.9974 (0.0022)
ACC	0.9477 (0.0093)	0.9463 (0.0090)	0.9435 (0.0094)	0.9416 (0.0105)	0.9349 (0.0111)	0.9439 (0.0094)	0.9475 (0.0075)
PRE	0.9476 (0.0383)	0.9570 (0.0336)	0.9643 (0.0333)	0.9686 (0.0343)	0.9905 (0.0172)	0.9622 (0.0326)	0.9555 (0.0347)
AUC	0.9565 (0.0118)	0.9566 (0.0119)	0.9510 (0.0148)	0.9568 (0.0121)	0.9556 (0.0118)	0.9199 (0.0209)	—
$\hat{\rho}$	0.0960 (0.0145)	0.0816 (0.0118)	0.0705 (0.0106)	0.0801 (0.0126)	—	—	—
$\hat{\alpha}$	0.4946 (0.0695)	0.4946 (0.0627)	0.4992 (0.0604)	—	—	—	—
$\hat{\xi}$	—	2.8631 (0.1689)	2.6092 (0.2119)	2.521 (0.110)	—	—	—

The highest MCC and F_1 scores among the *Nollik* models and among the competitors are highlighted. All the considered weight functions in the simulated scenarios obtain similar results that are favorable to our proposal. This is interesting because, even if an unbounded weight

function as $w_{MOM} = x^{2k}$ is not optimal for density estimation (since it tends to over-inflate the mass far away from the origin), the posterior probability of inclusion is correctly estimated. Nollik performs better than its direct competitor, the two-group model `locfdr`. That being said, we have to underline that Efron's method, employing an empirical Bayes methodology requires a very short time for the estimation. Also, Nollik performs better than the `MixFDRF` when the performance is assessed with more complete indexes like the MCC and the F_1 score. The only alternative that obtains comparable results with the Nollik models is the BH procedure. However, the output of our procedure is richer, being able to provide uncertainty estimation for each observation. Moreover, the BH procedure performs well only when the null distribution resembles the theoretical one, the Standard Gaussian. Whenever the null distribution departs from the theoretical case, BH method struggles, as we can see in Scenario 1. The nonparametric model seems to perform (slightly) worse than the Nollik parametric models. This is due to the fact that the scenarios in use are suited for the parametric model, while the BNP alternative is more useful in cases where the tails of the distribution in use are more problematic (multimodal/heavy). The posterior mean of α reveals that the model is capable to correctly estimate the proportion of the inner mixture. The same can be said about ρ, but this was expected since we placed a slightly informative prior on that parameter. We can also appreciate how $\hat{\xi}$ varies to accommodate the differences in the distributions of the data.

4.5.2 Gene Expression Case Studies

We apply our model to three different measurements of gene expressions, using the weight function w_1 (4.13). The runs where we employed the weight function w_2 are not reported since we obtained similar results. We compare the results with the `locfdr` model and the BH procedure. We adopt the same hyperparameters already discussed in Section 4.5.1 and collect a sample of 10,000 MCMC sweeps thinned every 50 iterations, after a burn-in period of 50,000 iterations, if not differently reported. In all the applications, N gene expressions between two groups are compared: the case group, composed of n_1 units, and the control group, composed of n_2 units. In the first example, the statistics are the result of a two-sample t-test $t_1, \dots, t_N$ and we map them with an inverse c.d.f. transformation to obtain Normality: $z_i = \Phi^{-1}\left(F^T_{n_1+n_2-2}(t_i)\right)$. In the second case, the statistics come from a series of Wald tests that are asymptotically Normal, so we do not perform any transformation. In the last example, we again apply the inverse c.d.f. transform to *moderated* t-statistics, characterized by a non-integer value of degrees of freedom (Smyth, 2004).

Alon Microarray Data

A classical example of gene expressions dataset is given by the `Alon` microarray matrix, which was collected originally for the study of Alon et al. (1999). We consider the publicly available version of the dataset contained in the `dglars` R package. The data consist of gene expression values of 2,000 microarray genes from 62 patients. Forty patients are diagnosed with colon cancer, and the remaining twenty-two serve as the control group. These 62 samples from colon-cancer patients were analyzed with an Affymetrix oligonucleotide Hum6000 array. Two-thousand out of around 6500 genes were selected based on the confidence in the measured expression levels. Microarray data are continuous, so for each gene, we compute the corresponding t-statistics to test the difference of expression among the two-group. We transform the data using the c.d.f. of a Student T with 60 degrees of freedom. Our model estimates $\hat{\rho} = 0.116$ ($s.e.$ 0.024) to be the proportion of genes flagged as interesting. $\hat{\alpha} = 0.91$ ($s.e.$ 0.082) is the estimated proportion of the over-expressed genes among the flagged ones. The parameter of the weight function is estimated as $\hat{\xi} = 3.158$ ($s.e.$ 0.63).

The number of selected genes varies between the methods. On one hand, thresholding the adjusted p-values with the BH procedure at 0.05 gives us 365 flagged genes. On the other hand, Efron's $lfdr$ methods (thresholded at 0.20 as suggested) finds no interesting genes. Our methods, controlling the Bayesian FDR at the level of 0.05 agrees with Efron. If instead we allow for a Bayesian FDR at the level 0.15, we obtain a threshold on the posterior probability of inclusion at 0.8092, flagging 87 genes as interesting. They are all over-expressed genes.

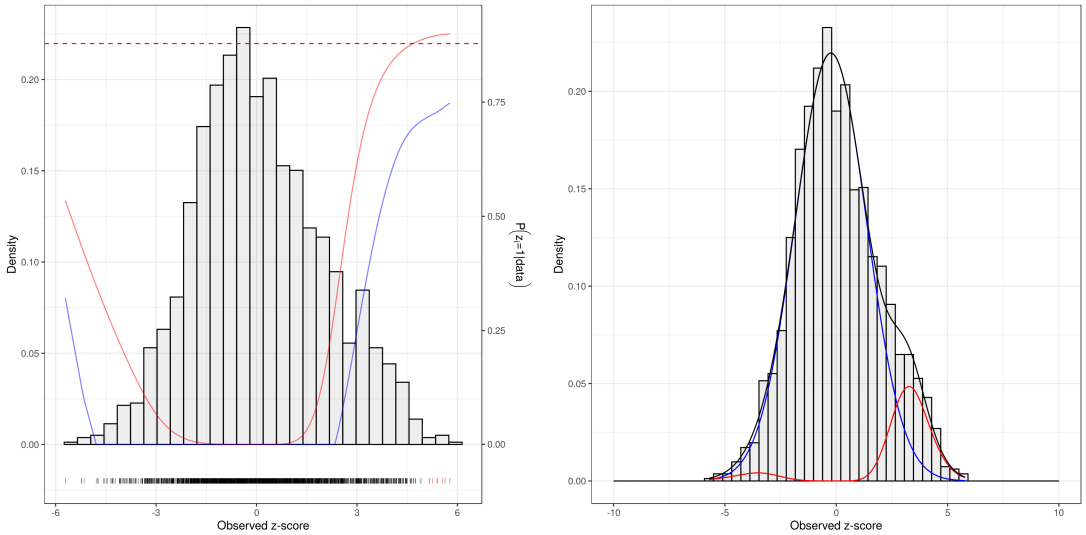


FIGURE 4.1: Alon Dataset. Left panel: Density estimation a posteriori of the global density f (black), the null density f_0 (blue) and the alternative density f_1 (red). Right panel: Histogram of the data with the Posterior Probability of Inclusion function $(1 - locfdr(z))$ superimposed, for both Efron's `locfdr` (blue) and Nollik (red). The dashed line represents the threshold controlling for a BFDR of 5%.

Microbiome Abundance table: Kostic Dataset

Many different models have been developed by bioinformaticians to address the challenges that a differential expression study can raise in this setting (see, for example, *edgeR* and *baySEQ* (Hardcastle and Kelly, 2010; Robinson and Smyth, 2007)). Recently, Love et al. (2014) propose *Deseq2*, a method for differential analysis of count data. In particular, Deseq2 models the count data, K_{ij}, with negative binomial distribution with mean μ_{ij} and dispersion α_i. The mean is taken as a quantity proportional to the concentration of cDNA fragments from the gene in the sample q_{ij}, scaled by a normalization factor. Then, a logarithmic link is used to model q_{ij}: $\log_2 q_{ij} = \sum_r x_{jr}\beta_{ir}$, where x_{jr} are the design matrix elements and β_{ir} is the coefficients. In the simplest case of a comparison between two groups, such as treated and control samples, the design matrix elements indicate whether a sample j is treated or not, and the GLM fit returns coefficients indicating the overall expression strength of the gene and the log-2 fold change between treatment and control. In this framework, it is common to fit the model, collect the test statistics on the coefficients and apply a Bonferroni correction or the BH procedure to the derived adjusted p-values. Instead, we propose to fit our model to the estimated test statistics. The Wald statistics provided as output are known to be asymptotically Normal. For this reason, we do not employ Efron's transform.

The `kostic` dataset is an abundance table with frequency counts of 2505 taxa that appeared in 190 different samples, originally used in Baselga et al. (2012). Out of the 190 samples, 95 are labeled as *Healthy*, 90 as *Tumoral* and 5 are not labeled at all. We first removed the five samples with missing labels, along with the ones that contain more than 95% zero entrances. Moreover,

we prune the taxa that were observed less than 50 times in the entire dataset. We remain with 708 taxa and 184 samples (94 vs 90). We then apply the Deseq2 model to this count table, obtaining a vector of 708 test statistics.
Given the peculiar shape of the tails of the data (see Figure 4.2) we use the BNP version of our model, with weight function w_1. We found that $\hat{\rho} = 0.0276 (s.e.\ 0.0288)$ is the proportion of interesting genes, while $\hat{\xi} = 3.001$ $(s.e.\ 0.71)$. The number of genes marked as interesting by our method with a threshold of 0.68 is 170, Efron's locfdr found 76 interesting genes while BH only found 26. This is due to the peaked shape of the null distribution, which is ignored by BH procedure and that misleads the estimation of Efron's $lfdr$, as we can appreciate in the left panel of Figure 4.2. In fact, the competitor's $lfdr$ function (in blue) is not monotone.

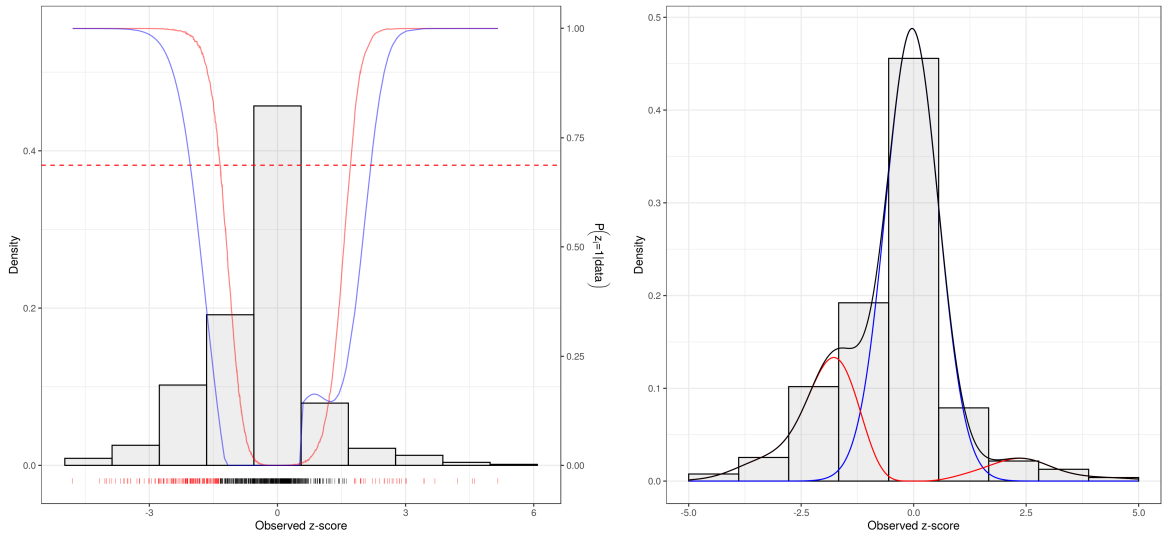


FIGURE 4.2: Kostic Dataset. Left panel: Density estimation a posteriori of the global density f (black), the null density f_0 (blue) and the alternative density f_1 (red). Right panel: Histogram of the data with the Posterior Probability of Inclusion function $(1 - locfdr(z))$ superimposed, for both Efron's `locfdr` (blue) and Nollik (red). The dashed line represents the threshold controlling for a BFDR of 5%.

Grouped Proteomics Data: Ubiquitin-protein interactors

Sometimes in real-case scenarios the differential expression analysis may be conducted separately, stratifying the subjects for some characteristics. For example, the data at hand could be the results of different models fitted using genes belonging to subgroups of subjects characterized by different age, gender, tumor gravity, etc. We can rephrase the model Equation (4.8) to accommodate this grouped setting, allowing the sharing of information across the various groups. The idea is to use all the observations to estimate the distributions f_0 and f_1 while taking into account the grouped structure in the parameters of main interest for us, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$.
Let us denote with z_{ij} the statistics related to the i-th measurement $(i = 1, \dots, N_j)$ found the j-th group $(j = 1, \dots, J)$. We can rewrite the likelihood as:

$$\mathcal{L}(\boldsymbol{z}|\boldsymbol{\theta}) = \prod_{i,j} (1 - \lambda_{ij})\, \phi_0(z_{ij}; \mu_0, \sigma_0^2) + \lambda_{ij} \frac{w(z;\xi)}{\mathcal{K}(\boldsymbol{\theta}_1)} \left[\delta_0(\gamma_{ij}) \phi_1(z_{ij}; \mu_1, \sigma_1^2) + \delta_1(\gamma_{ij}) \phi_2(z_{ij}; \mu_2, \sigma_2^2)\right]. \tag{4.19}$$

Consequently, new priors distributions for $\boldsymbol{\Lambda} = (\lambda_{11}, \dots, \lambda_{N_J,J})$, $\boldsymbol{\Gamma} = (\gamma_{11}, \dots, \gamma_{N_J,J})$, $\boldsymbol{\rho} = (\rho_1, \dots, \rho_J)$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_J)$ are needed. We adopt:

$$\begin{aligned}
&\lambda_{ij} \sim Bern\left(\rho_j\right), \\
&\gamma_{ij} \sim \delta_{-1}\left(\cdot\right)\delta_0\left(\lambda_{ij}\right) + \delta_0\left(\cdot\right)\delta_1\left(\lambda_{ij}\right)\left(1-\alpha_j\right) + \delta_1\left(\cdot\right)\delta_1\left(\lambda_{ij}\right)\alpha_j, \\
&\alpha_j \sim Beta(a_{\alpha_j}, b_{\alpha_j}), \qquad \rho_j \sim Beta(a_{\rho_j}, b_{\rho_j}).
\end{aligned} \tag{4.20}$$

This means that we will be able to estimate a $lfdr$ function, and consequently a posterior probability of inclusion, for each group. Doing so, we can assess the proportion of "significant" observation in each group.

We analyze a proteomics dataset where Ubiquitin-protein interactors with different Ubiquitin lengths are characterize in three different experimental conditions: `Ubi1`, `Ubi4`, and `Ubi6`. See Zhang et al. (2017) for more details on the data. The raw mass spectrometry data were first analyzed using MaxQuant (Cox and Mann, 2008). Then we follow the same pre-processing pipeline indicated in the manual of the R package `DEP` (Smits and Huber, 2017), remaining with 1899 values of genes for the three experimental conditions and a control group. The package uses the model *Limma*, an Empirical Bayes procedure that produces *moderate t-statistics*, in the form $t^{g,mod} = \frac{d}{s+s_0}$, where d is the difference in the sample means, s is the pooled standard deviation and s_0 is a small constant, added to avoid divisions by an extremely small variance estimate (Ritchie et al., 2015). For each experimental condition, we run the model and collect the statistics for $j = 1, 2, 3$, computing

$$z_{ij} = \Phi^{-1}\left(F_g^T(t_{ij}^{g,mod})\right),$$

where $t_{ij}^{g,mod}$ and the estimate of g are provided by the `DEP` methodology. We adopt $a_{\alpha_j} = b_{\alpha_j} = 1$, $a_{\rho_j} = 1$ and $b_{\rho_j} = 9$ for all $j = 1, 2, 3$. We also fix $a = 2.5$ to improve the mixing. Differently from what stated before, we run 400,000 MCMC iterations and, after discarding the first 100,000 as burn-in period, we thin the remaining chain every 30 iterations. Figure 4.3 shows the three histograms of the data under the different conditions and the non-null posterior probability functions. We can see how `Ubi4` and `Ubi6` generated similar statistics, while the observations under `Ubi1` are less spread away from zero. In fact, the three estimated proportions of non-null statistics are 0.104, 0.484, and 0.456 respectively. The three thresholds are all very close to each other, being 0.8357, 0.859, and 0.862. The numbers of non-null protein genes are 95, 492 and 445 respectively.

4.6 Discussion

In the Bayesian MHT, the two-group model of Efron (2004) is one of the most diffused approaches for distinguishing interesting, non-null observations, from the null ones. It postulates that the data generating process is a mixture of two distributions: the null, centered in zero, and the alternative on the tails. Multiple hypotheses testing without further constraints can lead to wide overlaps between the two distributions, compromising data classification. Inspired by the MOM non-local priors (Johnson and Rossell, 2010), we derive a class of weighted distributions that adjusts the likelihood with appropriate weight functions, with the purpose of better discriminating between interesting and uninteresting instances. We propose to model the alternative distribution of the two-group model as a weighted distribution, and we show that under reasonable assumptions the weighted version of the model leads to better results in terms of false discovery rate, false negative rate, and statistical power.

Future research will be focused on how this general class of weighted densities can be employed as prior distributions, contributing to the current literature of regularization methods. Guidance is provided for the choice of the weight function, and parametric and nonparametric sampling schemes are developed for posterior computation. The parametric case turns out to be an efficient, reliable and flexible enough specification. A possible immediate extension, to let the

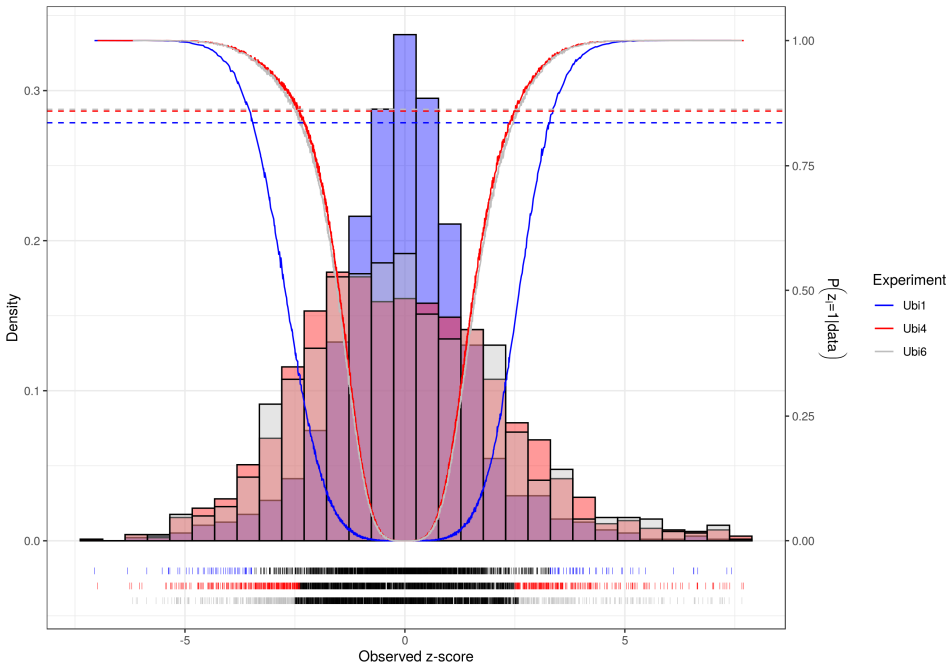


FIGURE 4.3: Ubiquitin-protein interactors dataset. Histograms of the three groups present in the data with the Posterior Probability of Inclusion function $(1 - locfdr(z))$ superimposed, one for each experiment. The three dashed lines represent the thresholds for each group, controlling for a BFDR of 5%.

parametric model handle more complex tails, would be to use a mixture of Gamma densities or Skewed Normals for modeling f_1.

We assessed the performance of our model on four different scenarios in an extensive simulation study, where it is clear that the model performs comparably or even better than the alternatives in the literature and efficiently. We finally apply our model to open source genomics datasets, and extend the methodology to also account for grouped data.

We finally want to underline that a weighted likelihood, especially a non-local one, can be useful also in contexts outside the MHT. The weight function can be properly chosen to reflect external information on the data, like spatial boundaries or penalize for implausible values, and the duality between null and non-null distributions can be used to perform classification tasks.

Appendix

4.A Sampling from Weighted Distributions

We can take advantage of the product form of the NLD to introduce a general Slice Sampler to generate random variates from the distribution of interest with bounded weight. Rossell and Telesca (2017), showing that every NLD can be seen as a mixture of truncated distributions, propose intuitive sampling schemes (Algorithm 1 and 2) which make simple the posterior simulation in a wide variety of cases. However, their result can be seen as a particular version of the Slice Sampler (Damien et al., 1999; Neal, 2003), exploited also in Petralia et al. (2012). We rephrase and adapt the algorithm in our framework, whenever a bounded weight function $w : \mathbb{R} \to [0, K]$ is assumed. Without loss of generality, we set $K = 1$. Using the idea of data augmentation, we introduce a Uniform latent variable in the weighted density, obtaining:

$$\pi_W(\theta, u; \xi, \lambda) \propto \pi(\theta; \lambda) \mathbb{I}_{\{w(\theta;\xi)>u\}}. \tag{4.21}$$

Notice that $\int_0^1 \pi_W(\theta, u; \xi, \lambda)\, du = \pi_W(\theta; \xi, \lambda)$.
Let us denote with (θ_0, u_0) the current values for the parameters of interest and let (θ_*, u_*) their updated version. The Slice Sampler algorithm for the NLD is composed by two steps:

1. Sample u_* from a $U(0, w(\theta_0; \xi))$
2. Sample θ_* from $\pi(\theta; \lambda) \mathbb{I}_{\{A_*\}}$, i.e. sample the new value from the distribution $\pi(\theta; \lambda)$ truncated on $A_* = \{\theta : w(\theta; \xi) > u_*\}$

This algorithm is trivial to implement every time the weight function w is invertible and a sampler for a truncated version of the local density $\pi(\theta; \lambda)$ is available. If a NLD is used as a prior, as long as the local distribution is conjugate with the likelihood distribution $f(\boldsymbol{z}; \theta)$, the derivation of a sampler for the posterior is immediate. In fact, we can recover the same structure of Equation (4.21) writing

$$\pi_W(\theta, u|\boldsymbol{z}; \xi, \lambda) \propto \pi(\theta; \lambda) \mathbb{I}_{\{w(\theta;\xi)>u\}} f(\boldsymbol{z}; \theta) = \pi(\theta|\boldsymbol{z}; \lambda) \mathbb{I}_{\{w(\theta;\xi)>u\}}$$

and then applying the algorithm using $\pi(\theta|\boldsymbol{z}; \lambda)$ as new local distribution.

As an example, consider these two different weighted distributions: a NLD defined by the product of a Standard Gaussian with the weight function w_1 and a Skew-Normal(α):

$$(S1) \quad \pi_W(\theta) = w_1(\theta; a, k)\, N(\theta; 0, 1) \qquad (S2) \quad \pi_W(\theta) = 2\Phi(\alpha\theta)\, N(\theta; 0, 1)$$

To implement the algorithm, we just need to compute the set A_* for both cases. Simple algebra provides the answer:

$$(S1) \quad A_* = \{\theta : |\theta| > a \sqrt[2k]{-\log(1-u)} \cdot\} \qquad (S2) \quad A_* = \{\theta : \theta > \frac{1}{a}\Phi^{-1}\left(\frac{u}{2}\right) \cdot\}$$

Both of the scenarios involve sampling from a Truncated Normal distribution. A recent R library, `TruncatedNormal` Botev (2017) makes this operation extremely smooth. To actually

simulate the values, we assumed $a = 5, k = 1$ and $\alpha = 2$. Figure 4.A.1 shows the histograms of 10,000 random instances sampled with the described algorithm, where the true density has been superimposed.

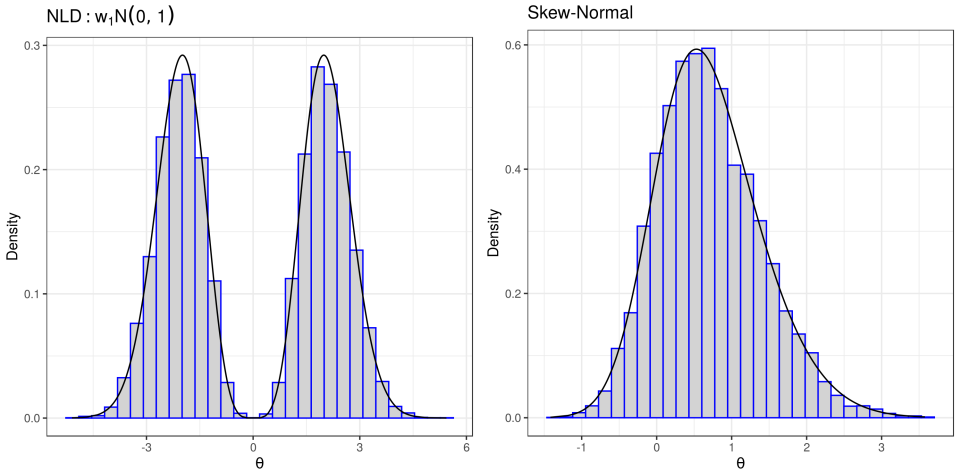


FIGURE 4.A.1: Histograms referring to the two distributions adopted in (S1) and (S2)

4.B FDR, FNR, Type II error as function the Acceptance Region

Let $\mathcal{A} = (z_1, z_2)$. Recall that $FDR(\mathcal{A}) = \mathbb{P}[H_0|z \notin \mathcal{A}]$, $FNR(\mathcal{A}) = \mathbb{P}[H_1|z \in \mathcal{A}]$ and $\beta(\mathcal{A}) = \mathbb{P}[z \in \mathcal{A}|H_1]$. Let also $\mathbb{P}[H_0] = (1-\rho)$, $F(z) = (1-\rho)F_0(z)+\rho F_1(z)$ and $F^W(z) = (1-\rho)F_0(z)+\rho F_1^W(z)$. Then,

$$
\begin{aligned}
FDR(\mathcal{A}) - FDR^W(\mathcal{A}) \geq 0 \iff \\
\frac{\mathbb{P}[z \notin \mathcal{A}|H_0](1-\rho)}{\mathbb{P}[z \notin \mathcal{A}]} - \frac{\mathbb{P}^W[z \notin \mathcal{A}|H_0](1-\rho)}{\mathbb{P}^W[z \notin \mathcal{A}]} \geq 0 \iff \\
\frac{F_0(z_1)+1-F_0(z_2)(1-\rho)}{F(z_1)+1-F(z_2)} - \frac{F_0(z_1)+1-F_0(z_2)(1-\rho)}{F^W(z_1)+1-F^W(z_2)} \geq 0 \iff \\
F^W(z_1)+1-F^W(z_2) - (F(z_1)+1-F(z_2)) \geq 0 \iff \\
(1-\rho)F_0(z_1)+\rho F_1^W(z_1)+1-(1-\rho)F_0(z_2)-\rho F_1^W(z_2)- \\
(1-\rho)F_0(z_1)-\rho F_1(z_1)-1+(1-\rho)F_0(z_2)+\rho F_1(z_2) \geq 0 \iff \\
F_1(z_2)-F_1^W(z_2)+F_1^W(z_1)-F_1(z_1) \geq 0
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
FNR(\mathcal{A}) - FNR^W(\mathcal{A}) \geq 0 \iff \\
\frac{\mathbb{P}[z \in \mathcal{A}|H_1](\rho)}{\mathbb{P}[z \in \mathcal{A}]} - \frac{\mathbb{P}^W[z \in \mathcal{A}|H_1](\rho)}{\mathbb{P}^W[z \in \mathcal{A}]} \geq 0 \iff \\
\frac{F_1(z_2)-F_1(z_1)}{F(z_2)-F(z_1)} - \frac{F_1^W(z_2)-F_1^W(z_1)}{F^W(z_2)-F^W(z_1)} \geq 0 \iff
\end{aligned}
$$

$$
\begin{aligned}
1 - \frac{F_0(z_2) - F_0(z_1)}{F(z_2) - F(z_1)} - 1 + \frac{F_0(z_2) - F_0(z_1)}{F^W(z_2) - F^W(z_1)} \geq 0 \iff \\
F(z_2) - F(z_1) - F^W(z_2) + F^W(z_1) \geq 0 \iff \\
F_1(z_2) - F_1^W(z_2) + F_1^W(z_1) - F_1(z_1) \geq 0
\end{aligned}
$$

and

$$
\begin{aligned}
\beta(\mathcal{A}) - \beta^W(\mathcal{A}) \geq 0 \iff \\
\mathbb{P}\left[z \in \mathcal{A} | H_1\right] - \mathbb{P}^W\left[z \in \mathcal{A} | H_1\right] \geq 0 \iff \\
F(z_2) - F(z_1) - F^W(z_2) + F^W(z_1) \geq 0 \iff \\
F_1(z_2) - F_1^W(z_2) + F_1^W(z_1) - F_1(z_1) \geq 0
\end{aligned}
$$

So we showed as the conditions that the discrepancies in the FDR, FNR and β between unweighted and weighted case all simplify into this expression $F_1(z_2) - F_1^W(z_2) + F_1^W(z_1) - F_1(z_1)$, which is the difference in the areas under the densities f_1 and f_1^W computed over $\mathcal{A}$.

Lemma 1 (Lemma 2 in Dharmadhikari and Joag-Dev (1983))**.** *Let X and Y be real random variables with c.d.f.'s F and G and densities f and g, respectively. Either of the following conditions imply that F stochastically dominates by G, i.e. $\forall x, F(x) \leqslant G(x)$:*
a. The density g crosses f only once and from above.
b. For all $t \in (F(0), 1)$, $\frac{\mathrm{d}}{\mathrm{d}t}\left\{F^{-1}(t) - G^{-1}(t)\right\} \geqslant 0$ or, equivalently $f\left[F^{-1}(t)\right] \leqslant g\left[G^{-1}(t)\right]$.

Chapter 5

Bayesian Mixture Models for Intrinsic Dimension Estimation

"Sono piccolissimi i conigli quando nascono
era semplicissima la via quando la trovi
Non fa più freddissimo quando hai vestiti nuovi
I superlativi, banalissimi quando li scrivi."
Depressissimo – Rancore

"If it's bad, then I hate it because I hate bad writing.
If it's good, then I'll be envious and I'll hate it all the more."
Midnight in Paris – ~~Hemingway~~ Woody Allen

Abstract

Even if they are defined on a space with a large dimension D, data points usually lie onto a hypersurface, or manifold, with a much smaller intrinsic dimension (ID). Data points with ID d can be deemed as a configuration of a Poisson Process (PP) with an intensity proportional to the true underlying density. Postulating the local homogeneity of the PP on the scale of the second nearest neighbor (NN), the ratio of the distances between the second and first NNs follows a Pareto distribution parametrized only by d. Following this rationale, the recent TWO-NN method (Facco et al., 2017) and Hidalgo model (Allegra et al., 2019) allow for estimating the ID when all points lie onto a single subspace or K manifolds, respectively. In particular, the latter setting employs a Bayesian finite mixture of K components. In this paper we extend this theoretical framework, obtaining a closed-form density for the ratio of the distances of L consecutive NNs under the assumption that the homogeneity of the density holds on the scale defined by the distance of the L-th NN of a point. More generally, we are able to provide a distribution for the ratio of the distances of two NNs of general order. These extensions lead to a refined estimator of the data intrinsic dimension, which we name CRIME (Consecutive Ratios Intrinsic Manifold Estimator). Hidalgo obtains remarkable results, but its limitation consists in fixing a priori the number of components in the mixture. To adopt a fully Bayesian nonparametric approach, we let $K \to +\infty$, using a Dirichlet Process Mixture Model as an infinite mixture of Pareto distributions. Since the posterior distribution has no closed-form expression, to sample from it we rely on the slice sampler algorithm (Kalli et al., 2011). From preliminary analyses on simulated and well-known datasets, our method provides promising results allowing us to uncover a rich data structure starting from the intrinsic dimension, a pure geometric data feature, and only requires the definition of a distance measure.

5.1 Introduction

In recent years we have witnessed an unimaginable growth in the production of data: from genomics to personalized medicine, from sports to finance, datasets of important dimensions are now widely available. This poses new interesting challenges for the statistical community, which is called to devise new techniques to analyze this type of data in a reasonable amount of time. The collective behavior of a dataset can be often described by a handful of variables. Thus, one key aspect of statistical analyses nowadays is the estimation of the Intrinsic Dimension (ID) of a dataset, which can be seen as the number of relevant variables that are needed to completely describe entirely the data-generating process, i.e. the dimension of all the non-redundant information contained in a table. The ID can be also seen as the minimum number of parameters needed to accurately describe the important characteristics of a system (Fukanaga, 1972). More formally, the ID d is defined as the dimension of the subspace of $\mathbb{R}^D$ where the data entirely lies, without information loss (Bishop, 1995). The literature regarding methods for ID estimation is vast. We refer to Campadelli et al. (2015) for an extensive review. In the same paper, the authors provide another useful interpretation of the ID in a context of pattern recognition: a point set is viewed as a sample set uniformly drawn from an unknown smooth (or locally smooth) manifold structure, eventually embedded in a higher-dimensional space through a non-linear smooth mapping; in this case, the ID to be estimated is the manifold's topological dimension.

In this framework, many modeling and exploratory techniques are based on some concept of distance between the data. Recently, Duan and Dunson (2018) have propose to model in a Bayesian setting the pairwise distances among distributions to coherently estimate a clustering structure. One drawback of this method is that it involves the computation of each pairwise distance among the data points, which can be extremely computationally expensive. Amsaleg et al. (2015), exploiting results from Houle (2013), suggest modeling a distance random variable using a Generalized Pareto Distribution (Coles and Davison, 2008), a milestone of Extreme Value Theory, since they show that the ID can be recovered, asymptotically, as a function of its parameter. Classical projective methods, such as Principal component analysis (PCA – Jolliffe, 2002), postulate that the sub-manifold where the data lie is linear, but this can be a simplistic hypothesis. To capture the non-linearity of the subspaces - also called manifold learning, Granata and Carnevale (2016) provide a method to estimate a global ID starting from the distribution of distances on graphs, or geodesic distance, since it is being "shape-aware", i.e. capable of measuring the length of paths completely contained in the manifold, and to analyze the scaling behavior of the distance probability distribution at intermediate length-scales. Li et al. (2017) propose a Spherical version of the PCA, where the latent manifolds are locally approximated with the use of Spherelets (pieces of spheres). This idea has been extended in other works: Mukhopadhyay et al. (2019) use a Fisher-Gaussian Kernel to estimate densities on highly non-linear supports, Li and Dunson (2019a) extend the idea to clustering methods and Li and Dunson (2019b) use the Spherical PCA to efficiently estimate the geodesic distances among the points.

The aforementioned methods are based on the assumption that the ID d is constant throughout the entire dataset. However, recent literature discusses that the observations of complex datasets could show more than one ID. We build on the idea of the TWO-NN estimator and its Bayesian version, Hidalgo (Facco et al., 2017; Allegra et al., 2019), extending both the underlying theoretical framework and modeling approach. In detail, we first propose to adopt different parametric prior specifications to improve Hidalgo's performance. Then, we extend the existing theoretical methodology, by deriving closed-form expressions for the distribution for more then one ratio of distances between NNs. Finally, we propose a novel Bayesian nonparametric model

that is able to estimate the ID for each data point.

The paper is organized as follows. In Section 2 we introduce the original estimators TWO-NN and Hidalgo, revisiting some of the theoretical results found in Allegra et al. (2019), and we propose two simple but effective modifications. In Section 3, we extend Hidalgo's theoretical framework and propose a Bayesian Nonparametric model. Section 4 presents an efficient MCMC algorithm for posterior simulation and discusses methods for postprocessing the MCMC output. Section 5 contains applications to both simulated data and real datasets. Section 6 presents future research directions and concludes.

5.2 Background: the TWO-NN estimator and Hidalgo

Consider a dataset $\boldsymbol{X}$, composed of n D-dimensional observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i \in \mathbb{R}^D$. Let us assume that the data are realization from a Poisson point process, characterized by a density $\rho(\boldsymbol{x})$, defined on a manifold of unknown intrinsic dimension $d \leq D$. In general, we expect $d << D$.

For any point $\boldsymbol{x}_i$, we can define $v_{i,l} = \omega_d \left(r_{i,l}^d - r_{i,l-1}^d \right)$ as the volume of the spherical shell enclosed between two successive neighbors, where d is the dimensionality of the space in which the points are embedded, ω_d is the volume of the d-dimensional sphere with unitary radius, and $r_{i,j}$ defines the value of a distance in $\mathbb{R}^+$ between observation i and its j-th nearest neighbor (NN). Note that in the univariate case, $v_{i,l}$ is usually referred to as *inter-arrival time*.

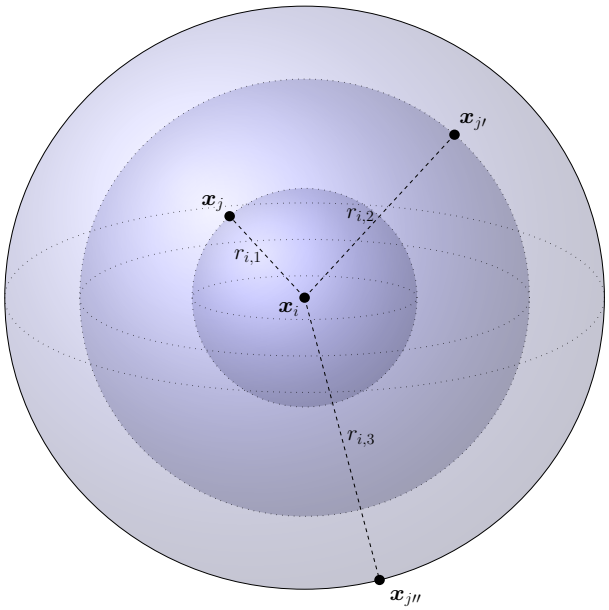


FIGURE 5.1: A pictorial example in $\mathbb{R}^3$ of the quantities involved. The points represent the data. The selected observation, $\boldsymbol{x}_i$ is connected by dashed lines representing the distances $r_{i,j}$, $j = 1, 2, 3$ to its first three NNs. The different spherical shells, characterized by different colors, have area $v_{i,j}$, $j = 1, 2, 3$.

In the univariate case it is well known that, if the density is constant (i.e. $\rho(\boldsymbol{x}) = \rho \ \forall \boldsymbol{x}$), all the v_l's are independently drawn from an Exponential distribution with rate equal to the density ρ (Kingman, 1992). Building on the work of Moltchanov (2012), Facco et al. (2017) have extended this result to the multivariate case, where hyperspherical shells are the proper extension of the inter-arrival times. Then, the following Lemma holds:

Lemma 2. *Consider a distance taking values in $\mathbb{R}^+$ defined among the data points and let $r_{i,j}$ be the value of this distance between observation i and its j-th NN. If we assume that the density ρ is approximately locally constant on the scale defined by the distance of the first two neighbors $r_{i,2}$, it follows that*

$$\mu_i = \frac{r_{i,2}}{r_{i,1}} \sim Pareto(1, d). \tag{5.1}$$

Proof. A detailed proof of this result is contained in Facco and Laio (2017). In the Appendix, we provide an equivalent but simper proof based on the properties of distributions of random variables. We underline that

$$X \sim Pareto(1, \alpha) \iff Y = X^q \sim Pareto(1, \alpha/q).$$

This property, used in the proof, is straightforward to prove. However, to the best of our knowledge, no reference in the literature mentioned it. □

In other words, using only basic properties of the Homogeneous Poisson point process, Facco et al. (2017) showed that the ratio of the distances between a point and, respectively, its first and second NN is Pareto distributed, with scale parameter equal to 1 and shape parameter d, i.e. the Intrinsic Dimension of the dataset. We remark that this result holds every time a distance taking values in $\mathbb{R}^+$ is adopted.
The TWO-NN estimator treats the ratios μ_i's as independent, $i = 1, \dots, n$, and estimates an overall d on the entire dataset following a least-squared approach, linearizing the Pareto c.d.f. (Facco et al., 2017). However, assuming d to be unique on the entire dataset could be limiting. Moreover, the independence assumption is clearly too restrictive. To overcome these issues, Allegra et al. (2019) propose Hidalgo (Heterogeneous Intrinsic Dimension Algorithm), a Bayesian finite mixture model, where the data, treated as exchangeable, have density ρ that is assumed to arise as a convex linear combination of different densities with support on different manifolds of heterogeneous dimension: $\rho(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \rho_k(\boldsymbol{x})$, where the k-th manifold has weight π_k. To reflect this assumption in the data, we can extend eq. (5.1) writing μ_i as a mixture of $Pareto(1, d)$, distributions:

$$\mathcal{L}(\mu_i | \boldsymbol{d}) = \sum_{k=1}^{K} \pi_k \mathcal{P}(\mu_i | d_k) \tag{5.2}$$

where $\mathcal{P}(\mu_i|d) = d\mu_i^{-(d+1)}$ is the usual Pareto density, $\boldsymbol{d} = (d_1, \dots, d_K)$ are the intrinsic dimensions and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ are the mixture weights. The model is completed assuming independent Gamma priors for each d_k and a Dirichlet prior for the mixture weights $\boldsymbol{\pi}$. To ease the posterior computation, it is common to augment the model introducing latent indicators $\boldsymbol{z} = \{z\}_{i=1}^n$ which denote the mixture component for μ_i, i.e. the lower-dimensional manifold to which observation $\boldsymbol{x}_i$ has been assigned. Conditionally on the $\boldsymbol{z}$, the likelihood (5.2) can be re-expressed as $f(\mu_i | \boldsymbol{d}, \boldsymbol{z}) = \mathcal{P}(\mu_i | d_{z_i})$ and $z_i \sim \sum_{k=1}^{K} \pi_k \delta_k(\cdot)$. However, in this classical formulation the estimation of the parameters $\boldsymbol{d}$ is difficult and the presented model can be inaccurate since there is no clear separation between different Pareto distributions. In fact, even when considering very different shape parameters, Pareto distributions overlap in the right-hand side tail to a great extent. Figure 5.2 provides an example.

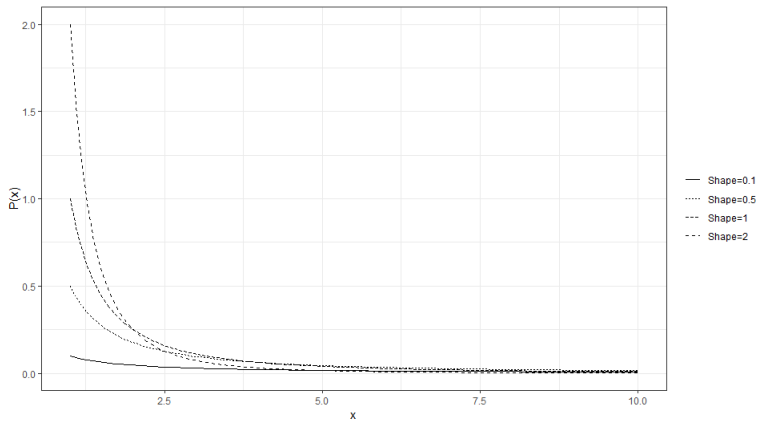


FIGURE 5.2: The graphs reports four Pareto densities characterized by the same scale parameter equal to 1 and different shape parameters, ranging from 0.1 to 2. Even for very different shapes, the densities overlap to a great extent.

In other words, the clustering induced in the data by the latent variables $\boldsymbol{z}$ is of paramount importance: the observations within each group concur to the estimation of a different value of d_{z_i}. If the clustering is inaccurate, so is the estimate. It is crucial to include a source of local homogeneity in the model, and this can be obtained via the following

Assumption: the different manifolds are separated in the space, and the neighborhood of a point should be more likely to contain points sampled from the same manifold than points sampled from a different manifold.

Therefore, Allegra et al. (2019) propose to extract from the original data $\boldsymbol{x}$ another source of information that can be used to penalize for local inhomogeneities: the $n \times n$ proximity matrix $\mathcal{N}^{(q)}$, that we now introduce. The (i, j) entry $\mathcal{N}_{ij}^{(q)}$ of this binary matrix is 1 only if the observation j is one of the first q NNs of observation i, 0 otherwise. Notice that $\sum_j \mathcal{N}_{ij}^{(q)} = q$. To induce local uniformity, we can model $f\left(\mathcal{N}_{ij}^{(q)} = 1 | z_i = z_j\right) = \zeta_0$, and that $f\left(\mathcal{N}_{ij}^{(q)} = 1 | z_i \neq z_j\right) = \zeta_1$, where the probabilities ζ_0, ζ_1 are such that $\zeta_0 > 0.5$ and $\zeta_1 < 0.5$. These inequalities imply that points assigned to the same manifold have more chances to be neighbors.
We underline that the events $\{\mathcal{N}_{ij}^{(q)} = 1 | z_i = z_j\}$ and $\{\mathcal{N}_{ij}^{(q)} = 1 | z_i \neq z_j\}$ are not complementary, therefore in general $\zeta_0 \neq 1 - \zeta_1$. However, for simplicity, the authors propose to set $\zeta_0 = \zeta$ and $\zeta_1 = 1 - \zeta$. Denote with $\mathcal{N}_i^{(q)}$ the i-th row of the adjacency matrix. We can regard the rows of $\mathcal{N}^{(q)}$ as independent and adopt the following distribution:

$$f\left(\mathcal{N}^{(q)} | \mathbf{z}, \zeta\right) = \prod_i f\left(\mathcal{N}_i^{(q)} | \mathbf{z}, \zeta\right) = \prod_i \frac{\zeta^{n_i^{in}(\boldsymbol{z})} (1-\zeta)^{q - n_i^{in}(\boldsymbol{z})}}{\mathcal{Z}\left(\zeta, N_{z_i}\right)} \tag{5.3}$$

with $\zeta \in (0.5, 1)$ is the parameter enforcing uniformity between neighbors ($\zeta = 0.5$ implies no additional term in the likelihood), $n_i^{in}(\mathbf{z}) = \sum_j n_{ij} \mathbb{I}_{z_j = z_i}$ is the number of the q NNs of $\boldsymbol{x}_i$ that are clustered together with observation i, N_{z_i} is the cardinality of cluster of instances grouped with $\boldsymbol{x}_i$ and

$$\mathcal{Z}\left(\zeta, N_{z_i}\right) = (1-\zeta)^q \begin{pmatrix} n - N_{z_i} \\ q \end{pmatrix} {}_2F_1\left(-q, 1 - N_{z_i}, n - N_{z_i} - q, \frac{\zeta}{1-\zeta}\right)$$

is the normalization constant, involving the ordinary hypergeometric function ${}_2F_1$. Details of its derivation can be found in Facco and Laio (2017). The assumption of independence among the

rows of the matrix $\mathcal{N}^q$ is convenient from a computational point of view. However, it may be not completely satisfactory. We leave this research question open for future investigation, noticing that models for graphs and networks can be employed to better represent the aggregating process behind the data. In any case, this additional term removes the independence between the cluster labels and helps to recover a more precise estimate of the ID. The resulting likelihood for $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ is

$$\mathcal{L}\left(\boldsymbol{\mu}, \mathcal{N}^{(q)} | \boldsymbol{d}, \boldsymbol{z}, \zeta\right) = \prod_{i=1}^{n} \mathcal{P}\left(\mu_i | d_{z_i}\right) \times f\left(\mathcal{N}_i^{(q)} | \mathbf{z}, \zeta\right). \tag{5.4}$$

The number of mixture components is chosen ex-post, adopting some postprocessing procedure. Allegra et al. (2019) compare the average likelihood estimated over the MCMC sweeps. Another approach could be comparing the average posterior values for different K. Alternatively, one could use more complete measures of model comparisons, like DIC, BIC, AICm, BICm, or WAIC (Spiegelhalter et al., 2002; Raftery et al., 2007; Hjort et al., 2010; Watanabe, 2013).

5.2.1 Alternative parametric prior specifications for d

Various immediate and simple extensions for improving the prior on the intrinsic dimension parameters d are possible. We discuss their introduction with two examples.

Truncated Support. First, we consider the `iris` dataset, where $D = 4$ measurements of $n = 150$ iris flowers are recorded. The flowers belong to three species: `setosa`, `versicolor` and `virginica`. We remove one observation, being the exact replica of another: the model cannot handle cases where two points coincide, resulting in a distance equal to zero in the denominator of (5.1). After 10000 iterations used as burn-in period and we collect 5000 MCMC samples. According to BICm and AICm, the best value for the number of mixture components is $K = 3$. We derive each observation Intrinsic Dimension using the following estimators, where T denotes the number of MCMC sweeps:

$$\hat{d}_i = \frac{1}{T} \sum_{t=1}^{T} d_{z_i^t}, \qquad \hat{d}_i = \text{median } \{d_{z_i^t}\}_{t=1}^{T}. \tag{5.5}$$

Tracking the chains of parameters actually assigned to each observation via z_i^t is a simple way to deal with the label switching problem. More refined methodologies have been proposed for handling this non-identifiability issue (Robert, 2010; Rodríguez and Walker, 2014; Celeux, 1998; Sperrin et al., 2010). However, for the current illustration, the solution we adopt suffices. Similarly, we derive the two quantiles of order 0.05 and 0.95, to provide Bayesian credible sets. The top left panel of Figure 5.3 reports the estimated median IDs for the iris dataset and the corresponding credible sets. It is interesting how the three different Species of flowers show different intrinsic dimensions. However, an interpretation problem is clear: some of the upper bounds of the credible sets for the IDs are above the maximum dimension D. To obviate this issue, we propose to substitute the Gamma prior on d_k with a Uniform distribution or a Truncated Gamma over $(0, D)$. Alternatively, if one wants to include the case where $d = D$, we can employ the following density, with mixture proportion $\hat{\rho}$:

$$\pi(d_k) \propto \hat{\rho} \frac{b^a}{\Gamma(a)} d_k^{a-1} \exp\{-b d_k\} \mathbb{I}_{(0,D)} + (1 - \hat{\rho}) \delta_D(d_k) \quad \forall k. \tag{5.6}$$

The results with these new priors are reported in Figure 5.3. We see how the estimates are now coherent with the theoretical framework, while the differences between groups are preserved. In detail, the right-hand panels show that adopting a truncated Gamma or a Uniform on the interval $(0, D)$ leads to similar results. The bottom-left panel presents the estimated median ID

when a point mass is placed on D and $\hat{\rho}$ is set to 0.9. In this case, all the `setosa` flowers are estimated to be characterized by an ID equal to the upper-bound dimension of the dataset D. Another simulation study, involving Uniform distributions, is reported in the Appendix.

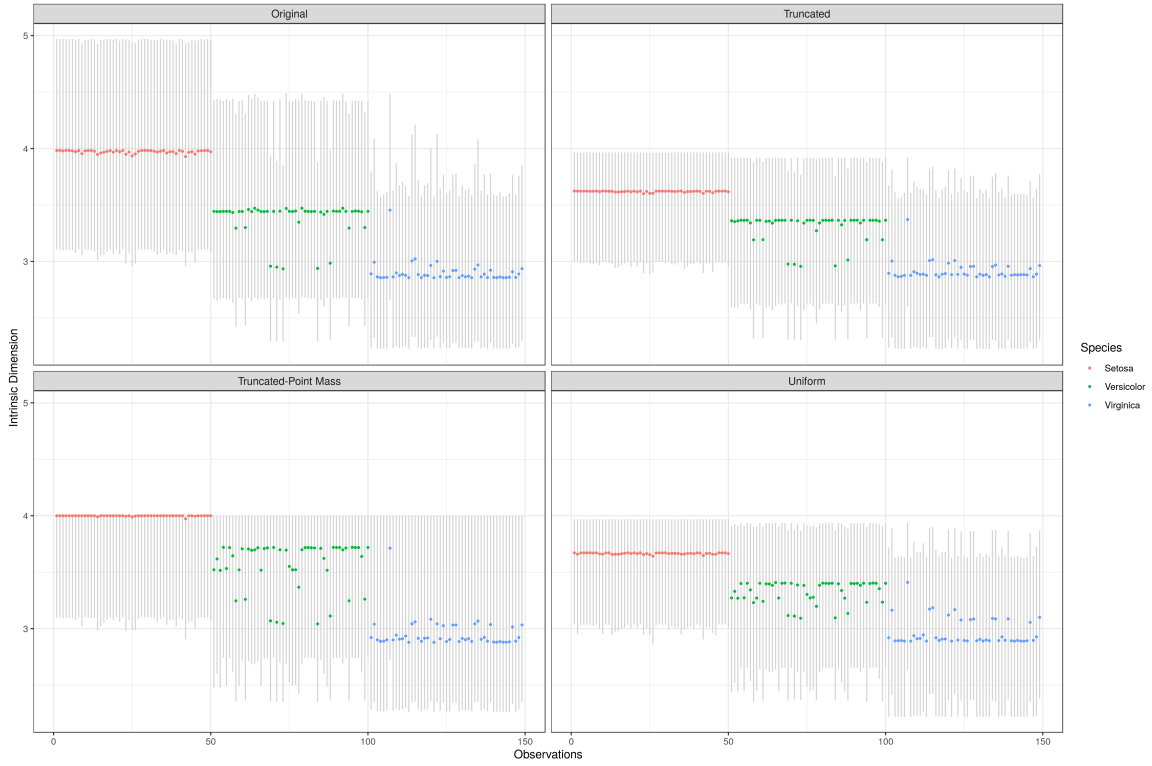


FIGURE 5.3: `iris` dataset. The top-left panel reports the posterior means and the 90% credible sets for each observation, after applying Hidalgo with K=3. The panels on the right show the same output when a truncated Gamma (top) or a Uniform (bottom) prior on d is adopted. The bottom-left panel shows the results when a Truncated Gamma is mixed with a point mass in D, with prior mixture proportion equal to $\hat{\rho} = 0.9$

Repulsive Distribution. Dealing with mixture models could lead to overfitting, in the sense that the model tends to create more components than the ones that are actually needed. Then, in some application one may observe different clusters of observations characterized by very similar ID. This distinction, instead of reflecting a real difference in the latent manifold dimensions, could be simply due to noise in the observed data or small curvatures in the latent geometry. To avoid the creations of redundant components and shrink to zero the fluctuations in the estimation, we employ a repulsive density of the following form as in Petralia et al. (2012):

$$\pi(\boldsymbol{d}) = c_1 \left(\prod_{k=1}^{K} g_0(d_k) \right) h(\boldsymbol{d}), \qquad h(\boldsymbol{d}) = \min_{\{(s,j)\in A\}} g\left(\Delta\left(d_s, d_j\right)\right) \tag{5.7}$$

where Δ is a suitable distance in $\mathcal{R}^+$, g_0 is a univariate density function for d_k, and $A = \{(s,j) : s = 1,\dots,K; j < s\}$. Instead of specifying the function g as in the aforementioned paper, i.e.$g(\Delta) = \exp\left[-\tau(\Delta)^{-\nu}\right]$ with $\tau, \nu > 0$, we adopt the following sigmoidal function:

$$g(\Delta) = \frac{1}{1 + \exp\left[-\frac{\Delta - \tau}{\nu}\right]} \quad \tau, \nu > 0. \tag{5.8}$$

This sigmoidal function is convenient because it allows to directly specify the magnitude of the

repulsion. For $\nu \to 0$ the sigmoid approaches a step function, where the jump is exactly at τ. In other words, choosing a parameter ν small enough, we can induce a distance of at least τ between the realizations of the vector $\boldsymbol{d}$. Sampling from the full conditional induced by the repulsive prior requires to be able to sample from truncated distributions. To this extent, we implement a simple, general algorithm that allows us to sample from interval-truncated random variables, starting from the inverse c.d.f. method. The details are reported in figure 5.C.1 in the Appendix.

To show how this prior works, we apply Hidalgo to the famous `growth` dataset, a collection of 93 growth curves that track children's heights over time (Tuddenham and Snyder, 1954). As usually done in functional data analysis, we employ a basis function representation to smooth the data. In this case, we use $B = 50$ b-splines bases (Ramsay and Silverman, 2005). The smoothed functions are reported in the left panel of Figure 5.4. The dataset contains growth curves of 54 females and 39 males. Once we smooth the functions, we are left with a dataset of 50 spline coefficients for 93 individuals. We apply Hidalgo with $K = 2$ to this dataset, looking for the ID of the spline coefficients. As we can see from the right panel of Figure 5.4 two groups, reflecting the gender partition are evident, but not very well separated: the males are characterized by an ID roughly equal to 4.75 while the ID is approximately 4.9 for the females. When the repulsive prior is applied with $\tau = 0.5$ and $\nu = 0.001$, the two clusters are "pushed apart" from each other favoring more distant realizations of the ID. Namely, in this particular case, the two repulsive IDs are around 4.5 an 5.1.

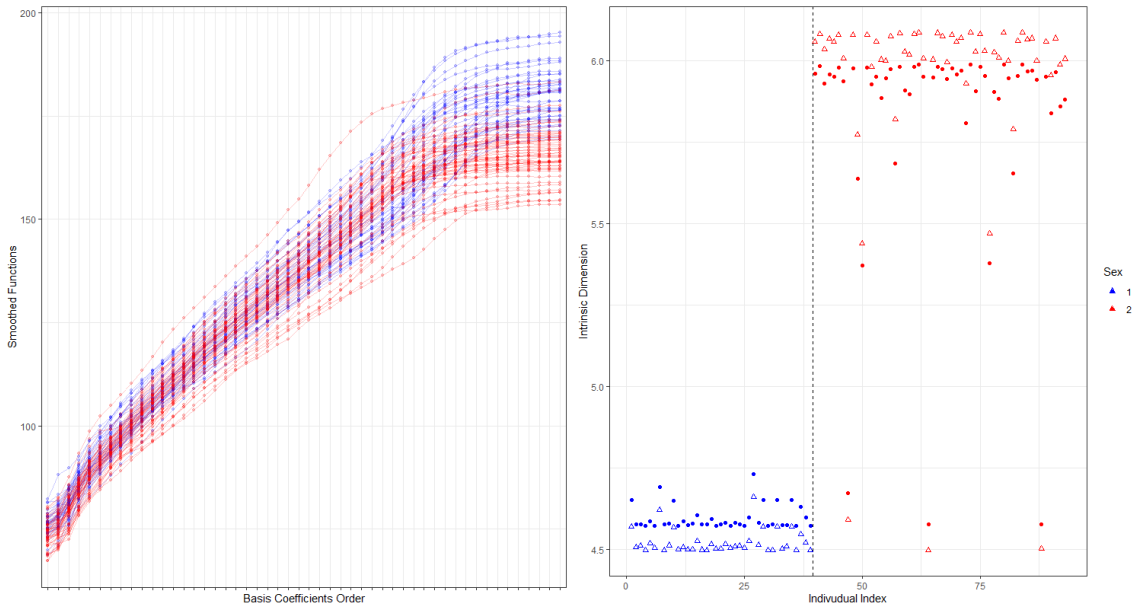


FIGURE 5.4: `growth` dataset. The left panel shows the smoothed growth curves for both male (1, in blue) and female (2, in red). The right panel reports the posterior median of the IDs obtained with Hidalgo, where $K = 2$, comparing the output when a classical independent prior (circles) is adopted against when d is modeled with a repulsive prior (triangles).

5.3 Extending Hidalgo to Consecutive Ratios (CRIME)

We now show how the theoretical results of the previous section can be extended, under the assumption that the homogeneity of the density holds on the scale defined by the distance of the first L neighbors of a point $\boldsymbol{x}_i$. Let us define $V_{i,l} = \omega_d r_{i,l}^d$ as the volume of the hypersphere centered in $\boldsymbol{x}_i$ with radius equal to the distance to its l-th NN. Notice that, for $l = 2, \dots, L$, $v_{i,l}$

and $V_{i,l-1} = v_{i,1} + \cdots + v_{i,l-1}$ are independent. Moreover, $V_{i,l} \sim Erlang(1, l-1)$. The essential point is to notice that we can write

$$\frac{v_{i,l}}{V_{i,l-1}} = \frac{\omega_d \left(r_{i,l}^d - r_{i,l-1}^d\right)}{\omega_d r_{i,l-1}^d} = \left(\frac{r_{i,l}}{r_{i,l-1}}\right)^d - 1$$

which means

$$\mu_{i,l} = \frac{r_{i,l}}{r_{i,l-1}} = \left(\frac{v_{i,l}}{V_{i,l-1}} + 1\right)^{1/d} \sim Pareto(1, (l-1)d)) \tag{5.9}$$

or equivalently

$$\gamma_{i,l} = \log\left(\frac{r_{i,l}}{r_{i,l-1}}\right) \sim Exp((l-1)d). \tag{5.10}$$

Given these premises, the following Lemma holds.

Lemma 3. *Consider a suitable distance in $\mathbb{R}^+$ between the data points and let $r_{i,j}$ be the value of this distance between observation i and its j-th NN. Assume that the density ρ is approximately locally constant on the scale defined by the distance of the first L neighbors, and let $\mu_{i,l} = \frac{r_{i,l}}{r_{i,l-1}}$ and $\gamma_{i,l} = \log(\mu_{i,l})$, for an integer $l \in \{2, \ldots, L\}$. It follows that*

$$\begin{aligned} \boldsymbol{\gamma}_{i,L} &= (\gamma_{i,2}, \ldots, \gamma_{i,L}) \sim Exp(d) \times \cdots \times Exp\left((L-1)d\right), \\ \boldsymbol{\mu}_{i,L} &= (\mu_{i,2}, \ldots, \mu_{i,L}) \sim Pareto(1, d) \times \cdots \times Pareto(1, (L-1)d). \end{aligned} \tag{5.11}$$

Proof. See Appendix. □

In this way, we are able to characterize the distributions of the ratio of consecutive distances. All of them depend on the same parameter d, which can now be estimated using more information extracted from the data. Another important advantage is that now we have a way to justify the choice of the hyperparameter q in $\left(\mathcal{N}_i^{(q)} | \mathbf{z}, \zeta\right)$: if we set $q = L$, it can be interpreted as our prior guess about how diffuse (in terms of number of neighbors) is the postulated local uniformity of the Poisson Process density $\rho(\boldsymbol{x})$.
Alternatively to the equations in (5.11), that suggests a multivariate model, one could apply the following transformation to recover an univariate distribution:

$$\Gamma_{i,L} = \sum_{l=1}^{L-1} l \cdot \gamma_{i,l+1} \sim Erlang\left(\frac{(L-1)L}{2}, d\right), \quad i = 1, \ldots, n. \tag{5.12}$$

Notice that many other distributions can be employed using the properties of the Exponential random variables. For a generic observation i and a generic ratio of order l, the following statements are equivalent

$$\begin{aligned} &(I)\ \gamma_{i,l} \sim Exp\left((l-1)d\right), \qquad (II)\ \exp\left(-\gamma_{i,l}\right) \sim Beta\left(1, (l-1)d\right), \\ &(III)\ \gamma_{i,l}^2 \sim Weibull\left(\frac{1}{2}, \frac{1}{(l-1)^2 d^2}\right), \\ &(IV)\ \mu - \sigma \log\left((l-1)d\gamma_{i,l}\right)\right) \sim GEV\left(\mu, \sigma, 0\right) (Gumbel), \end{aligned} \tag{5.13}$$

where GEV indicates the Generalized Extreme Values distribution (McFadden, 1978). Equations (5.12)-(5.13) provide alternatives to the modeling and the estimate of the ID. Distributions (III) and (IV) are linked to the Extreme Value Theory (EVT). Other authors have recently developed an ID estimator in an EVT framework (Amsaleg et al., 2015; Houle, 2013): we leave for future research the investigation of potential connections among the two fields. From now on,

we proceed considering only the vectors $\boldsymbol{\mu}_{i,L}$, $i = 1, \ldots, n$, i.e. working with Pareto distributions.

Deriving a least-square estimator from (5.11) as in Facco et al. (2017) is not immediate. However, it is straightforward to derive a Maximum likelihood Estimator (MLE) for the parameter d:

$$\hat{d}_{MLE} = \frac{n(L-1)}{\sum_{i=1}^{n} \sum_{l=2}^{L} \log\left(\mu_{i,l}\right)}. \tag{5.14}$$

The theoretical results rely on the fact that the different consecutive ratios are independent of each other. In real applications, this is hardly the case. To assess if considering more than one ratio provides an actual improvement, we simulate 50 independent datasets from Gaussian random variables of dimensions 2, 5, 10, 15, and 20. For each of these 5 simulation scenarios, we consider 4 different sample sizes: (A) $n = 22$, (B) $n = 100$, (C) $n = 500$,(D) $n = 1000$. We are then left with 20 datasets replicated 50 times. For each of the 50 replicates, we compute $\hat{d}_{MLE}$ and then we pool the results, computing the average estimates and the 10th and 90th quantile, to provide a measure of uncertainty estimation.

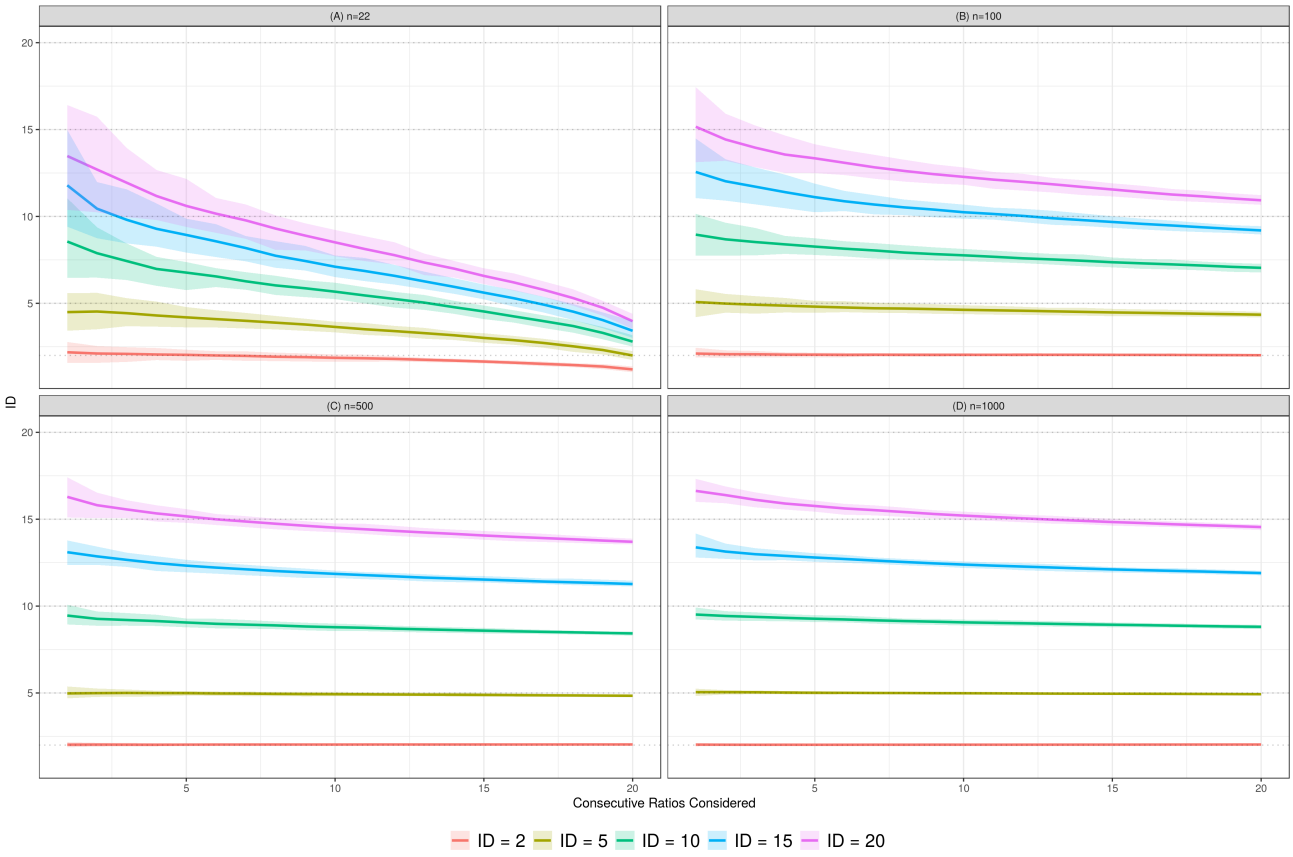


FIGURE 5.1: Each panel shows the MLE of the ID for an increasing number of consecutive ratios $\mu_{l,i}$ considered in the estimation. The five lines correspond to the average $\hat{d}_{MLE}$, while the shaded area highlights the interval between the 10th and 90th quantile.

Figure 5.1 shows some important insights. The lower the sample size, the greater is the underestimation of the true ID. There are multiple reasons for this behavior.
First, since (5.11) is a result based on local homogeneity, we need to consider the curse of dimensionality and the frontier effects that can affect the most extreme points. In fact, it is known that the number of points missing from the neighborhood of the frontier scales exponentially

with the ID (Smith, 1988; Granata and Carnevale, 2016). These are two intrinsic limitations of general ID estimation. In general, the model is capable of correctly estimate the true values the IDs when dealing with low-dimensional manifolds. Also, as n grows, the underestimation of higher IDs reduces.
A second reason is given by the model formulation. The fewer the data points, the less realistic is the hypothesis of independence of the observations.Introducing more than one ratio when the sample size is limited is counterproductive when the ID is high. However, even if considering more consecutive ratios does not improve the point estimation, it is effective in reducing the estimator variability, as shown by the shaded bands that shrink as L grows. To better investigate this behavior, we run a more complete simulation study. We consider 50 replications of independent gaussian datasets in dimension $d = 2, 5, 10, 15, 20$, respectively. The ID is estimated considering a varying number of consecutive ratios $L = 1, 2, 3, 5, 10$ 5 different sample sizes (ranging from $n_1 = 50$ to $n_5 = 10000$ instances). For each case, we compute the median and the width of the interval given between the 10th and the 90th percentile of the MLEs obtained on 50 different sample replicates. The results are reported in Table 5.1. The values confirm our claim: for low ID and high sample size, considering a higher number of consecutive ratios is beneficial, leading to more precise estimates. When the ID is higher (10 or above) increasing L induces a greater underestimation, while increasing n fights this tendency. The study shows the presence of a trade-off between precision in the estimates and underestimation. In general, considering values of $L > 1$ can be beneficial but, especially for small datasets, we suggest to fix its value not above 5. We finally mention that Facco et al., 2017 propose to trim the 10% of the most extreme values of the vector of ratios $\boldsymbol{\mu}$ to recover an estimate less prone to fluctuations given by the presence of outliers. We do not perform trimming, since it is not immediate how to perform it in multidimensional frameworks.

Finally, we derive the distribution of the ratio of the distances between two nearest neighbors of generic order n_1 and n_2, where $n_2 > n_1$ are integers. This is particularly useful to investigate how the ID changes as different scales are considered when studying the data.

Lemma 4. *Consider a suitable distance in $\mathbb{R}^+$ between the data points and let $r_{i,j}$ be the value of this distance between observation i and its j-th NN. Assume that the density ρ is approximately locally constant on the scale defined by the distance of the first n_2 neighbors, and let $x = \mu_{i,n_1,n_2} = \frac{r_{i,n_2}}{r_{i,n_1}}$. It follows that*

$$f_{\mu_{i,n_1,n_2}}(x) = (n_2 - n_1)\binom{n_2 - 1}{n_1 - 1}\frac{d(x^d - 1)^{n_2 - n_1 - 1}}{x^{(n_2-1)d+1}}, \quad x > 1. \tag{5.15}$$

If we set $n_1 = n_0$ and $n_2 = 2n_0$ we obtain

$$f_{\mu_{i,n_0,2n_0}}(x) = \frac{(2n_0 - 1)!}{(n_0 - 1)!^2} \cdot \frac{d(x^d - 1)^{n_0 - 1}}{x^{(2n_0-1)d+1}}, \quad x > 1. \tag{5.16}$$

Proof. See Appendix. □

5.3.1 Extension to the Nonparametric Case

One of the major limitations of Hidalgo is that the number of components K must be specified beforehand, or estimated via cross-validation or by using some information criteria (e.g. BICm) However, this approach does not take into account the uncertainty in the number of mixture components which corresponds to the number of sub-manifolds. A way to overcome this issue would be placing a prior on K: however, having a random dimension would require a Reversible Jump Step (Green, 1995) in the Gibbs Sampler, which is known to be sophisticated

		$d = 2$	$d = 5$	$d = 10$	$d = 15$	$d = 20$
n_1	$L = 1$	2.1875 (0.9885)	4.9243 (2.0925)	8.7124 (2.3071)	12.0778 (4.1211)	14.3320 (4.5960)
	$L = 2$	2.1005 (0.6013)	4.8265 (1.3165)	8.3014 (1.5790)	11.4484 (3.4586)	13.5916 (3.4765)
	$L = 3$	2.1119 (0.6011)	4.7786 (0.9773)	8.0889 (1.5058)	10.9846 (2.5534)	13.1396 (2.4620)
	$L = 5$	2.0702 (0.3993)	4.6742 (0.6537)	7.7210 (1.0096)	10.2892 (1.8154)	12.3440 (1.7075)
	$L = 10$	1.9996 (0.2820)	4.2974 (0.5527)	7.0115 (0.5632)	9.2108 (1.0515)	10.7797 (1.0844)
n_2	$L = 1$	2.0148 (0.2953)	4.9410 (0.6809)	9.4148 (1.1357)	13.1085 (1.4120)	16.2261 (2.3103)
	$L = 2$	2.0241 (0.2154)	5.0049 (0.4823)	9.2663 (0.8204)	12.7952 (1.0518)	15.7747 (1.4639)
	$L = 3$	2.0259 (0.1996)	5.0065 (0.4242)	9.1604 (0.7242)	12.6396 (0.8196)	15.5256 (1.0295)
	$L = 5$	2.0268 (0.1332)	5.0018 (0.2702)	9.0978 (0.5042)	12.3173 (0.7196)	15.1243 (0.7873)
	$L = 10$	2.0375 (0.0977)	4.9261 (0.2510)	8.8030 (0.4324)	11.8869 (0.4216)	14.4834 (0.5023)
n_3	$L = 1$	1.9920 (0.1487)	5.0505 (0.3154)	9.7230 (0.5976)	13.8044 (0.7947)	17.2314 (0.8853)
	$L = 2$	2.0045 (0.0940)	5.0544 (0.2052)	9.6435 (0.3509)	13.5955 (0.4831)	16.9371 (0.5743)
	$L = 3$	2.0017 (0.0690)	5.0516 (0.1936)	9.5906 (0.3281)	13.4748 (0.4386)	16.7916 (0.5515)
	$L = 5$	2.0055 (0.0516)	5.0424 (0.1463)	9.5001 (0.2200)	13.2479 (0.3287)	16.5205 (0.3409)
	$L = 10$	2.0116 (0.0376)	5.0477 (0.1058)	9.3639 (0.1703)	12.9496 (0.2454)	16.0101 (0.2764)
n_4	$L = 1$	2.0138 (0.0896)	5.0502 (0.2165)	9.7686 (0.4388)	13.9549 (0.5241)	17.5438 (0.6003)
	$L = 2$	2.0079 (0.0750)	5.0496 (0.1724)	9.7262 (0.2913)	13.8367 (0.3740)	17.2710 (0.4460)
	$L = 3$	2.0045 (0.0566)	5.0445 (0.1595)	9.6866 (0.2519)	13.7021 (0.3087)	17.1169 (0.4554)
	$L = 5$	2.0053 (0.0343)	5.0454 (0.1214)	9.6145 (0.1851)	13.5431 (0.2686)	16.8715 (0.3377)
	$L = 10$	2.0112 (0.0250)	5.0439 (0.0702)	9.5053 (0.1275)	13.2669 (0.1743)	16.4809 (0.2204)
n_5	$L = 1$	1.9981 (0.0522)	5.0628 (0.1518)	9.8562 (0.3239)	14.1140 (0.4256)	17.8752 (0.5633)
	$L = 2$	1.9988 (0.0385)	5.0598 (0.0970)	9.8045 (0.2179)	13.9763 (0.3282)	17.6310 (0.3447)
	$L = 3$	2.0005 (0.0373)	5.0586 (0.0812)	9.7758 (0.1795)	13.8842 (0.2766)	17.4826 (0.2885)
	$L = 5$	2.0024 (0.0283)	5.0559 (0.0612)	9.7402 (0.1261)	13.7624 (0.1996)	17.2772 (0.1995)
	$L = 10$	2.0062 (0.0205)	5.0568 (0.0459)	9.6624 (0.0716)	13.5372 (0.1363)	16.9074 (0.1537)

TABLE 5.1: Median values of the estimated ID and corresponding width of the credible set between the 10th and 90th quantile across 50 different samples, varying the sample size from $n = 50$ to $n = 10000$. In particular, $n_1 = 50$, $n_2 = 500$, $n_3 = 2500$, $n_4 = 5000$, and $n_5 = 10000$. As expected, the underestimation typical of higher dimensionality is less evident as more data are used. Using more than one ratio helps lowering the variance of the estimates.

and computationally expensive (Bhattacharya, 2008). Instead, we propose a Bayesian nonparametric Intrinsic Dimensions Estimator, adopting a Dirichlet Process Mixture Model to describe the distribution of the transformed data as an infinite components mixture. Letting $K \to \infty$ introduces more flexibility avoiding limiting parametric assumptions and makes the estimation algorithm feasible. Additionally, it is reasonable to think that the number of hidden manifold in the dataset may increase with the advent of new data points.
Let us denote with $\hat{\mathcal{P}}\left(\boldsymbol{\mu}_{i,L}|d\right)$ the density of the vector $\boldsymbol{\mu}_{i,L}$, containing L consecutive distance ratios for observation i. The classical DPMM (Dirichlet Process Mixture Model) (Antoniak, 1974; Lo, 1984) is defined as, for $i = 1, \ldots, n$:

$$f(\mu_i|G) = \int \hat{\mathcal{P}}\left(\boldsymbol{\mu}_{i,L}|d\right) \mathrm{d}G(d), \qquad G \sim DP\left(\alpha, G_0\right), \tag{5.17}$$

where $DP\left(\alpha, G_0\right)$ is the usual Dirichlet Process, a random probability measure characterized by a base measure G_0 and concentration parameter α. Following Sethuraman's stick-breaking representation (Sethuraman, 1994), we can write $G = \sum_{k=1}^{+\infty} \pi_k \delta_{d_k}$, where d_k's are i.i.d. from G_0 and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ is obtained as

$$\pi_k = u_k \prod_{l=1}^{k-1} (1 - u_l) \qquad u_k \sim Beta(1, \alpha) \qquad k \geq 1.$$

We will refer to this representation writing $\boldsymbol{\pi} \sim SB(\alpha)$. Given this characterization, we can write model eq. (5.17) as

$$f(\mu_i|G) = \sum_{k=1}^{+\infty} \pi_k \hat{\mathcal{P}}\left(\boldsymbol{\mu}_{i,L}|d_k\right), \qquad d_k \overset{i.i.d.}{\sim} G_0, \qquad \boldsymbol{\pi} \sim SB(\alpha). \tag{5.18}$$

As in the finite case, we introduce a set of latent allocation variables to ease the computational part. Likewise, the problem of the overlapping densities is still present, therefore we need to also introduce the homogeneity-inducing term in the likelihood $f\left(\mathcal{N}_i^{(q)}|\mathbf{z}\right)$. So the previous parametric-single ratio model in (5.4) is extended to a nonparametric Consecutive Ratios Intrinsic Manifolds Estimator (CRIME) as:

$$\begin{aligned} &\mathcal{L}\left(\boldsymbol{\mu}_{i,L}, \mathcal{N}_i^{(q)}|\mathbf{z}, \mathbf{d}\right) = \hat{\mathcal{P}}\left(\boldsymbol{\mu}_{i,L}|d_{z_i}\right) \times f\left(\mathcal{N}_i^{(q)}|\mathbf{z}\right), \\ &z_i|\boldsymbol{\pi} \overset{ind}{\sim} \sum_{k=0}^{+\infty} \pi_k \delta_k, \qquad \boldsymbol{\pi} \sim SB(\alpha), \qquad d_k \sim G_0. \end{aligned} \tag{5.19}$$

Again, G_0 can be chosen to be a Gamma density to exploit conjugacy or a Repulsive distribution to avoid redundancies in the estimates. Alternatively, as we already discussed, a mixture between a distribution of a bounded support (Truncated Gamma or Uniform on $(0, D)$) and a point mass in D is especially useful when D is low or a high ID is expected in the data.

5.4 Posterior Inference

The posterior distribution for model (5.19) has no closed-form expression, so we need to rely on posterior simulation techniques to perform inference. A simple solution would be employing the Blocked-Gibbs Sampling scheme illustrated in Ishwaran and James (2001). Notice that adopting the stick-breaking representation allows us to effortlessly extend the DPMM to other stick-breaking prior, like the Pitman-Yor Processes. The algorithm presented in Ishwaran and James (2001) is based on an approximation since it relies on the truncation of an infinite series. Alternatively, we can overcome this issue using a slice sampler (Kalli et al., 2011; Walker, 2007) to sample from the exact posterior. To implement an independent slice-efficient sampler, we augment the model introducing two new quantities: a latent variable u_i for each observation and a deterministic sequence $\boldsymbol{\xi} = \{\xi_1, \xi_2, \dots, \}$, where each term is defined as $\xi_j = \kappa\left(1-\kappa\right)^{j-1}$. The likelihood of model eq. (5.19) becomes:

$$f\left(\boldsymbol{\mu}, \mathcal{N}^{(q)}, \boldsymbol{z}, \boldsymbol{u}|\boldsymbol{\pi}, \mathbf{d}\right) = \prod_i^n \frac{\pi_{z_i}}{\xi_{z_i}} \mathbf{1}_{\{u_i < \xi_{z_i}\}} \mathcal{L}\left(\boldsymbol{\mu}_{i,L}, \mathcal{N}_i^{(q)}|\mathbf{z}, \mathbf{d}\right). \tag{5.20}$$

Notice that, if we integrate out $\boldsymbol{u} = (u_1, u_2, \dots, u_n)$, we get back to our original model. The introduction of the latent variables $\boldsymbol{u}$ allows for a stochastic truncation at each iteration of the sampler. This sets a finite number of mixture components needed at each MCMC sweep, making computations feasible.

5.4.1 MCMC algorithm

The full conditionals are the following:

1. The full conditional of each u_i is uniformly distributed: $u_i|\cdots \sim \mathcal{U}\left(0, \xi_{zi}\right)$
2. Sample the Stick Breaking variables $u_k|\cdots \sim Beta\left(1 + \sum_{i=1}^n \mathbf{1}_{z_i=k}, \alpha + \sum_{i=1}^n \mathbf{1}_{z_i>k}\right)$

3. Let $\boldsymbol{z}_{-i}$ denote the vector $\boldsymbol{z}$ without its i-th element. Sample the cluster indicators z_i according to:

$$\mathbb{P}\left(z_i = k|\boldsymbol{z}_{-i}, \cdots\right) \propto \frac{\pi_{z_i}}{\xi_{z_i}} \mathbb{1}_{\{u_i < \xi_{z_i}\}} f\left(\boldsymbol{\mu}_{i,L}, \mathcal{N}_i^{(q)} | z_1, \ldots, z_{i-1}, k, z_{i+1}, \ldots, z_n, \mathbf{d}\right)$$

We underline that, given the new likelihood we are considering, now the cluster labels are not independent given all the other parameters. Let us define

$$\boldsymbol{z}_i^k = (z_1, \ldots, z_{i-1}, k, z_{i-1}, \ldots, z_n)\,.$$

Then, let $N_{z_i}(\boldsymbol{z}_{-i})$ be the number of elements in the $n-1$-dimensional vector $\boldsymbol{z}_{-i}$ that are assigned to the same manifold (mixture component) as z_i. Moreover, let $m_i^{in} = \sum_l \mathcal{N}_{li}^{(q)} \mathbb{I}_{z_l = z_i}$ be the number of points sampled from the same manifold of the i-th observation that have $\boldsymbol{x}_i$ as neighbor, and let $n_i^{in}(\mathbf{z}) = \sum_l \mathcal{N}_{il}^{(q)} \mathbb{I}_{z_l = z_i} \leq q$ be the number of neighbors of $\boldsymbol{x}_i$ sampled from the same manifold. Then, we can simplify the previous formula, obtaining the following full conditional:

$$\begin{aligned}\mathbb{P}\left(z_i = k|\boldsymbol{z}_{-i}, \cdots\right) \propto & \frac{\pi_k \hat{\mathcal{P}}\left(\boldsymbol{\mu}_{i,L}|d_k\right) \mathbb{1}_{\{u_i < \xi_k\}}}{\xi_k \cdot \mathcal{Z}\left(\zeta, N_{z_i=k}(\boldsymbol{z}_{-i}) + 1\right)} \times \left(\frac{\zeta}{1-\zeta}\right)^{n_i^{in}(\boldsymbol{z}_i^k) + m_i^{in}(\boldsymbol{z}_i^k)} \\ & \left(\frac{\mathcal{Z}\left(\zeta, N_{z_i=k}(\boldsymbol{z}_{-i})\right)}{\mathcal{Z}\left(\zeta, N_{z_i=k}(\boldsymbol{z}_{-i}) + 1\right)}\right)^{N_{z_i=k}(\boldsymbol{z}_{-i})}.\end{aligned} \tag{5.21}$$

See (Facco and Laio, 2017) for a detailed derivation of this result.

4. To sample $\boldsymbol{d}|\cdots$, we use the conjugacy result that links the Gamma and the Pareto distribution. We obtain $d_k|\cdots \sim Gamma\left(a_0 + N_l, b_0 + \sum_{i:z_i=l} \log \mu_i\right)$, where N_l is the number of observations assigned to the l-th group.
If G_0 is assumed to be a truncated Gamma distribution on $(0, D)$, then

$$d_k|\cdots \sim Gamma\left(a_0 + N_l, b_0 + \sum_{i:z_i=l} \log \mu_i\right) \mathbb{1}(\cdot)_{(0,D)}.$$

5. If we assume a Gamma prior on the concentration parameter α we can sample from its full conditional following the scheme proposed in (Escobar and West, 1995).

5.4.2 Post-processing the MCMC output

We can post-process the rich MCMC output in different ways. We can still estimate the ID of each observations using the formulas in (5.5). Moreover, we may be interested in recovering a clustering among the observations and estimate the mean ID in each group along with a representative individual (RI) for that group. We will discuss more about the RI in Section 5.6.2. Let us focus on how to recover meaningful partition. A straightforward way to recover a partition is to apply the usual minimization of loss functions (Binder, Variation of Information) typical of the BNP literature (Binder, 1978; Wade and Ghahramani, 2015), computed on the posterior pairwise coclustering matrix between observations. However, the model-based clustering recovered in this way may suffer from the overlapping among the Paretos in the likelihood and consequently might be not reliable. Another simple solution is to derive a clustering structure by inspecting the MCMC posterior median estimates as in (5.5). To estimate an interesting

partition, we can apply classical clustering algorithms such as k-means, where the optimal number of groups can be fixed studying the behavior of cluster quality indexes such as *Silhouette* (Rosseeuw, 1987) or the *Calinski-Harabasz index* (Caliński T. and Harabasz J., 1974).

5.5 Applications

5.5.1 Simulation Study

To investigate how the model behaves varying the number n of observations, we simulate independent realizations from four Gaussian random variables with different, well-separated centroids (vectors centered in -5, 0, 5 and 10). The true dimensions of these random variables are 7, 3, 5 and 10, respectively. We vary the size of the samples, distributing equally the observations across the four distributions.
For each simulation, we run 100000 iterations as a burn-in period and subsequently collect 10000 sweeps as a posterior MCMC sample. We set $q = 3$, $\zeta = 0.75$ following the directions provided in Allegra et al. (2019), and consistently with the choice of q we run the model on the first 3 consecutive ratios, so $L = 3$. The hyperparameters are specified as follows: on d_k we place a repulsive $Gamma(5, 1)$ prior $\forall k$, so that $\mathbb{E}[d_k] = Var[d_k] = 5\ \forall k$. For the repulsive parameters, we set $\tau = 0.5$ and $\nu = 0.001$. For the concentration parameter of the DP we choose a Gamma(3,3). Being the data so well separated, minimizing the Variation of Information criterion (Wade and Ghahramani, 2015) appears to be sufficient to recover a meaningful partition. To compare the estimated partition $\mathcal{C}$ to the ground truth, we use the Adjusted Rand Index (ARI) (Rand, 1971; Hubert and Arabie, 1985). We collect the posterior medians for each observation and compute a mean stratified for the labels of the cluster assignments, to obtain an ID value for each group. Table 5.1 reports the results. As expected, the classification is almost perfect in all the different cases. It is interesting to note that many observations are needed to capture the true latent dimensions of the biggest manifold (true ID = 10). When the number of observations is limited, the true ID is underestimated.

	GT	$n = 40$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
d_1	7	7.5620	6.1329	6.6556	6.5499	6.9300	7.1707
d_2	3	3.8718	2.5988	3.4442	2.9069	3.1045	2.8831
d_3	5	4.5988	3.3017	4.9553	5.5028	4.9217	5.0901
d_4	10	6.3364	6.4490	7.6904	8.5021	8.5416	9.8264
ARI	—	1.0000	1.0000	1.0000	0.9426	1.0000	0.9856

TABLE 5.1: Simulation Study. The table shows the estimated ID for each of the estimated clusters, recovered minimizing the Variation of Information as in Wade and Ghahramani (2015). GT indicates the ground truth, ARI the Adjusted Rand Index. In these cases, we fix $q = 3$ and $\zeta = 0.75$

We also study how the results change varying the number of consecutive ratios employed, along with q. Table 5.2 reveals some interesting insights: on simple data, the model is fairly robust to the specifications of both L and q. In particular, the only problematic case is encountered when $q = 2$: it seems that the local information provided by the additional term in the likelihood is not sufficient to detect the real data structure. As q becomes higher, the true structure of the data is recovered. Over, in other practical applications, we found that a level of q over 4 tends to favor the creation of small, uninteresting clusters, producing "local overfitting". This behavior is not observed in this application, due to the usage of the Repulsive prior on d_k. We underline how, practically, L and q can be independent of each other as the last column Table 5.2 shows.

	q=L=2	q=L=3	q=L=4	q=L=5	q=3, L=5
d_1	4.7867	6.6490	6.4756	6.3807	6.3808
d_2	9.0233	3.0968	3.2008	2.8340	2.9158
d_3	—	5.2243	5.0723	4.6628	4.6628
d_4	—	7.6249	7.7603	8.0265	7.9882

TABLE 5.2: Simulation Study. The table shows the estimated ID for each of the estimated clusters, recovered minimizing the Variation of Information as in Wade and Ghahramani (2015). The number of observation is fixed, equal to n=500.

We finally assess the reliability of our model applying CRIME to the `5 Gaussians` dataset used also in Allegra et al. (2019). The data consists of 5000 observations in D=10, where the data are generated, in equal proportions, from Gaussians characterized by different ID (1, 2, 4, 5, and 9, respectively). Differently from our first application, here the 5 distributions overlap to a great extent, being their centroids close to each other. Figure 5.D.1 in the Appendix shows the composition of the data. We run the model with the priors previously specified and obtain the results reported in Table 5.3. The overall classification is capable to reach an ARI of 0.9306.

	d_1	d_2	d_3	d_4	d_5
GT	1	2	4	5	9
$\hat{d}$	1.0095	2.0358	4.2360	4.9681	9.0349
Clust	1020	973	989	999	1019

TABLE 5.3: Simulation Study. `5 Gaussians` dataset. The table shows the estimated ID for each of the estimated clusters.

5.5.2 Application to real data

Leukemia Dataset

We consider the famous `Golub` dataset, from the microarray study in Golub et al. (1999): $n = 72$ leukemia patients, 47 with acute lymphoblastic leukemia (`ALL`) and 25 with acute myeloid leukemia (`AML` – a worse prognosis) have each had genetic activity measured for a panel of $D = 7128$ genes. The `AML` group appears to show greater activity on average (Efron and Hastie, 2016).

We run 100000 iterations discarded as burn-in period, followed by 50000 MCMC sweeps that we retain as posterior sample. We fix $L = q = 3$, and we adopt a repulsive Gamma prior with hyperparameters $(a_0 = 5, b_0 = 1, \tau = 0.5, \nu = 0.001)$ for the ID parameters and we set $\alpha \sim Gamma(3, 3)$ for the DP concentration parameter.

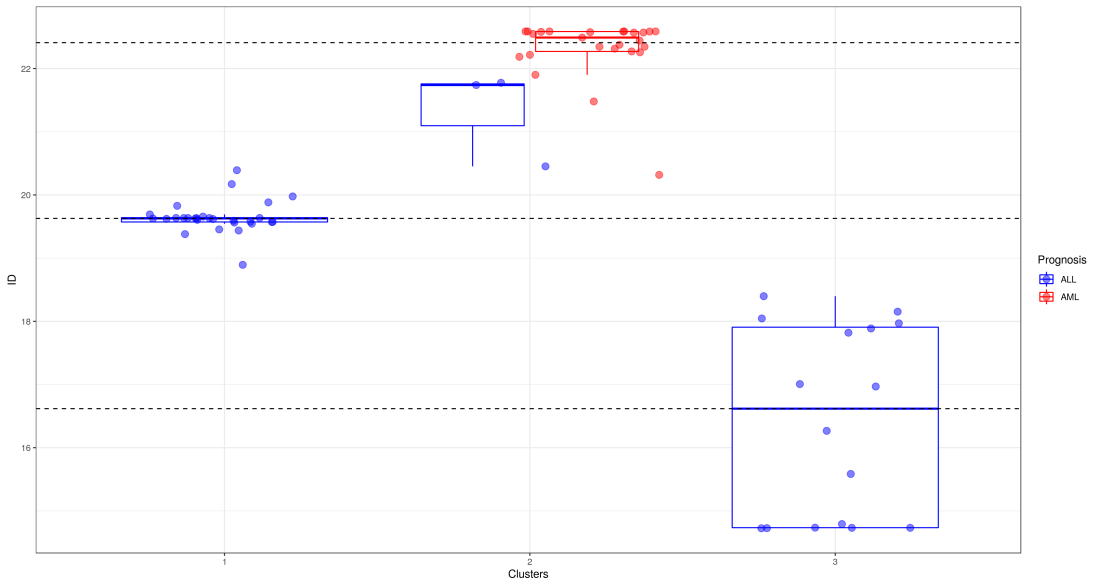


FIGURE 5.1: `Golub` dataset. The graph shows the boxplots of the median IDs estimated for the 72 patients, stratified by the recovered clusters. The different colors are associated with prognosis. The different horizontal lines highlight the median ID per cluster.

The estimated median IDs gather in three clusters, found minimizing the Variation of Information. In Table 5.4 we report the stratified mean and the standard deviations of the median estimates, along with the proportion of `AML` prognosis. Cluster 1 and 3 contain only `ALL` patients whose genes have average ID ≈ 19 and ≈ 16. The mean estimate of the second cluster's ID is less representative, being the estimates more variable. As also suggested by Figure 5.1 the second cluster, characterized by higher average ID ≈ 22, contains all the `AML` cases. This leads to the conclusions that the sets of genes of patients with `AML` lie on a more complex manifold than the `ALL` genes. This kind of information could provide valuable insights for subsequent explorative analyses, differential expression studies, and prognostic tasks.

	Cluster 1	Cluster 2	Cluster 3
n	28	28	16
Mean ID	19.65 (0.26)	22.21 (0.59)	16.41 (1.513)
`AML`%	0.00%	89.28%	0.00%

TABLE 5.4: `Golub` Dataset. Average, standard deviation of the median ID and AML proportion found in each cluster.

Schimtz dataset

The `Schmitz` dataset is a collection of more than 25000 rna-sequenced genes measured for 564 biopsy samples affected by diffuse large B-cell lymphoma, or DLBCL. DLBCL is a type of cancer that starts in white blood cells called lymphocytes. Different cases of DLBCL are phenotypically and genetically heterogeneous. Gene-expression profiling has identified subgroups of DLBCL according to the cell of origin that are associated with a differential response to chemotherapy and targeted agents: activated B-cell–like (ABC), germinal-center B-cell–like (GCB), and unclassified (Schmitz et al., 2018). It is of paramount importance trying to find biological patterns that can provide more insights regarding these subgroups.

For our analysis, we consider 481 samples for which additional demographic and medical information is available, such as gender, age, type of treatment and time until recurrence of the disease. Led by a field expert, we focus our attention on a subset of 19 genes that are known to be extremely relevant in DLBCL cases. The list of the codes of the considered genes is reported in the Appendix. For each sample, the rna-seq counts are normalized according to the relative library size. Our target is to understand if the different subgroups of DLBCL are characterized by heterogeneous IDs, similarly as in the application to the `Golub` dataset. Moreover, since extra information is available, we also investigate how the different survival times are related to the estimated IDs.
To fit the model, we employ the same hyperprior values of the previous application and we run 50000 MCMC iterations after a burn-in of double length.

	Cluster 1	Cluster 2	Cluster 3
n	267	114	100
Mean ID	10.09 (0.21)	9.69 (0 11)	11.68 (0.06)
`ABC`%	72.28%	43.85%	0.00 %
`GCB`%	5.99%	20.17%	99.00%
`Unclass`%	21.72%	35.96 %	1.00%
Mean Age	62.41	61.57	57.79
Female %	41.57%	39.47%	39.00%

TABLE 5.5: `Schmitz` Dataset. Different characteristics of the three estimated clusters.

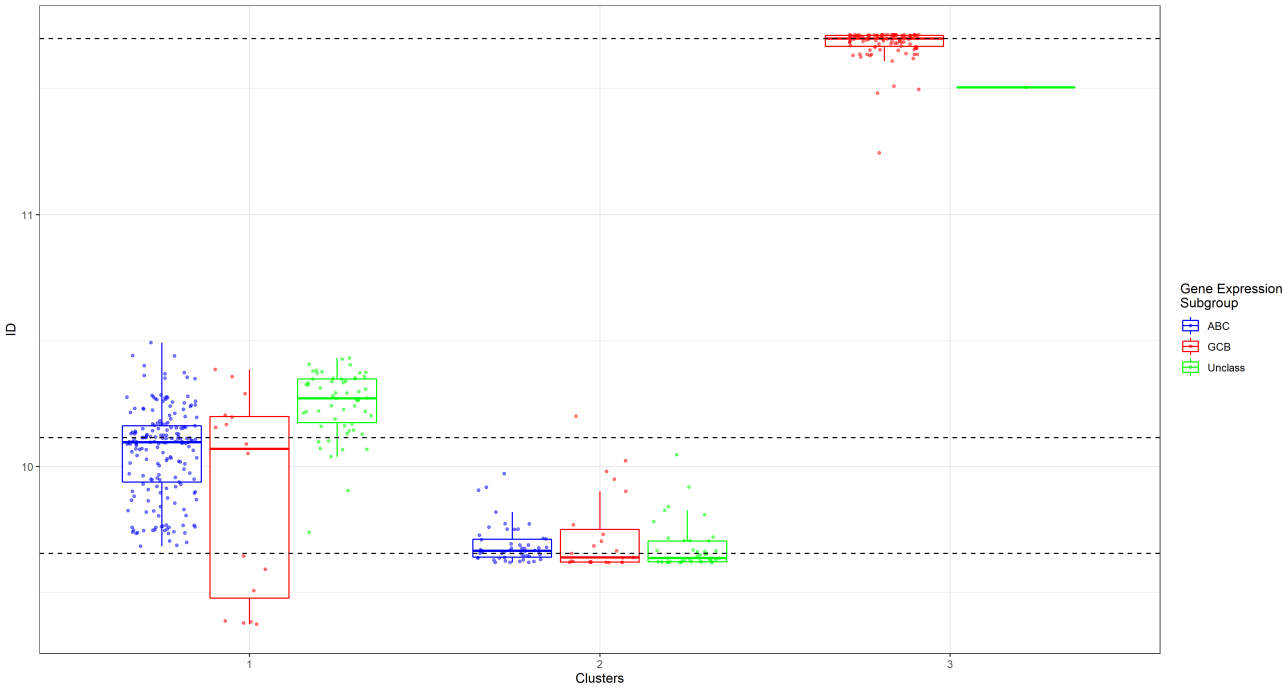


FIGURE 5.2: `Schmitz` dataset. Survivial curves based on Kaplan-Meier estimates (top panels) and table of number at risk (bottom panel) in the 3 estimated clusters.

The results provide very interesting insights. Minimizing the Variation of Information we recover 3 clusters. Descriptive statistics of the features of this partition are reported in Table 5.5. The first cluster is characterized by an higher amount of ABC samples, associated with a mean ID of $\approx$ 10 and higher average age. The second cluster, of mean ID of $\approx$ 9.5, contains similar

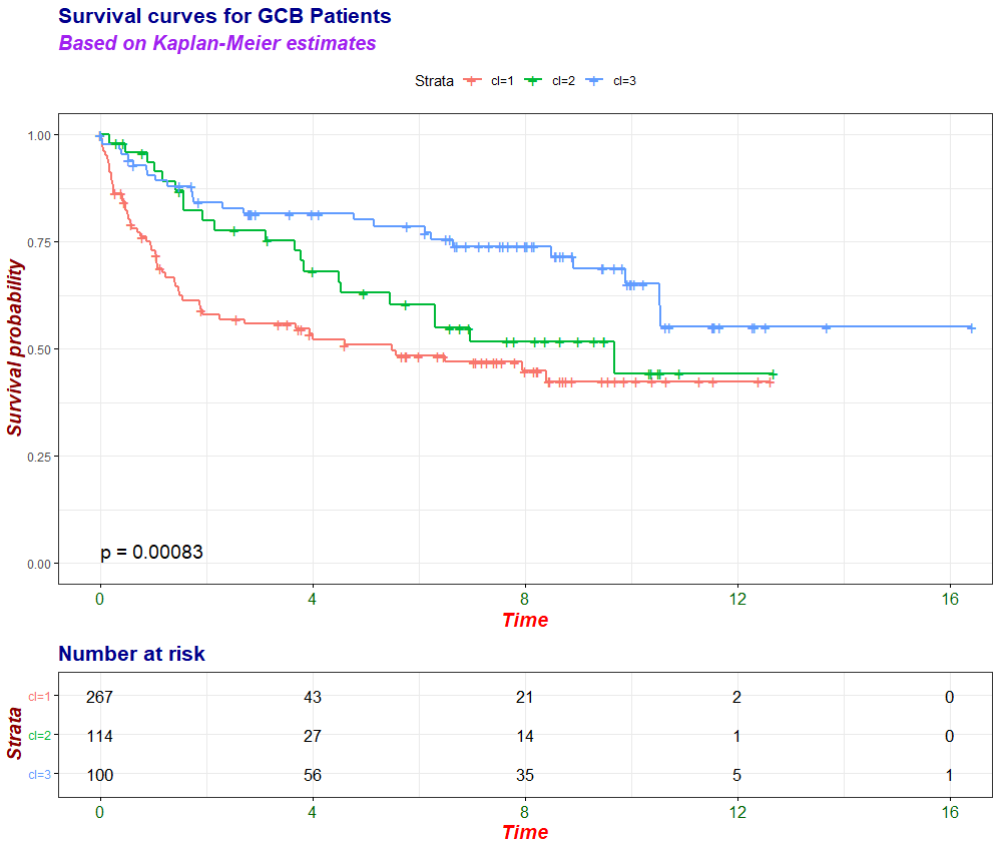


FIGURE 5.3: `Schmitz` dataset. Survivial curves based on Kaplan-Meier estimates (top panels) and table of numbers at risk (bottom panel) in the 3 retrieved clusters.

proportions of ABC, GCB and Unclassified subgroups. Interestingly, the third cluster is almost entirely composed of GCB subtypes and no ABC. Moreover, it is characterized by the highest mean ID ($\approx$ 12) and younger patients. The different proportions of subgroups in the various clusters are evident from the Figure 5.2.
We also study if the different clusters are helpful to discriminate subgroups of patients characterized by different survival times (in months, until death). The estimated Kaplan-Meier curves (Kaplan and Meier, 1958) and numbers at risk are reported in Figure 5.3. The results are remarkable: the three curves are well discriminated, with the first cluster being characterized by patients in worse conditions, being the red curve always below all the others. On the contrary, the blue curve, relative to the third cluster, shows that patients with higher ID are associate with longer survival time. This is in line with our expectations, since the first cluster contains the highest proportion of ABC subtype, the most dangerous one, while the third contains none of them. An interesting index is this context is the median follow-up time: for the first cluster it is of almost 6 months, for the second almost 10, while it is not even reached ($>$ 16 months) in the third cluster.

These two applications showed how the ID can be extremely helpful to obtain prognostic insights from biomedical dataset.

5.6 Future Directions and Conclusions

5.6.1 A Dependent Dirichlet Process Approach

The introduction of the adjacency matrix $\mathcal{N}^q$ is a clever and elegant expedient to inform the model regarding the distance structure between the data points. However, modeling the extra information in this manner requires to specify and tune hyper-parameters and may not completely solve the problem. Generally speaking, we may have access to covariates that help us to learn about the topology structure of the data, as the introduction of the adjacency matrix does in the previous case. To incorporate covariates, we can rewrite model (5.19) adopting a different approach, via Dependent Dirichlet processes (MacEachern, 1999; MacEachern, 2000; De Iorio et al., 2004; Duan et al., 2007). Suppose that we know the true manifold membership of our data. We see the data as coming from collections of distributions defined on an appropriate space S, where the index represents the actual manifold or any other available covariate. These models provide additional flexibility by allowing the mixing distribution G_{s} to change with $\mathrm{s} \in S$ while inducing dependence among the members of the collection (Rodríguez and Dunson, 2011). Specifically, many authors propose DPMM where the dependence is introduced among the atoms or among the weights of a stick-breaking representation. However, studies in the literature showed that introducing the dependence only at the level of the atoms is too limiting. For example, models with non-constant weights have richer support (MacEachern, 2000). Thus, many authors focused on modifying the classical stick breaking formulation to obtain more flexible models. In a spatial setting, Reich and Fuentes (2007) propose a stick breaking prior by assigning each location a different, unknown distribution, and smoothing the distributions in space with a series of kernel functions. Following the directions provided by Ishwaran and James (2001), Rodríguez and Dunson (2011) build a new stick breaking process characterized by variables defined as $u_l = \Phi(\alpha_l)$ with $\alpha_l \sim \mathrm{N}(\mu, \sigma^2)$, where the Beta distribution is substituted with a Normal c.d.f. This approach is easily extendable to the presence of external variables, which leads to the following Dependent Dirichlet Process Mixture Models via Probit stick breaking process for the data $(y_1, \ldots, y_n)$:

$$\begin{aligned} y_i(\mathrm{s}) \sim f_{\mathrm{s}} &= \int K(\cdot|\phi) G_{\mathrm{s}}(\mathrm{d}\phi), \qquad G_{\mathrm{s}}(\cdot) = \sum_{l=1}^{L} w_l(\mathrm{s}) \delta_{\theta_l(\mathrm{s})}(\cdot), \\ w_l(\mathrm{s}) &= \Phi(\alpha_l(\mathrm{s})) \prod_{r<l} (1 - \Phi(\alpha_r(\mathrm{s}))), \end{aligned} \tag{5.22}$$

where $K(\cdot|\phi)$ is a parametric kernel indexed by ϕ, $\{\alpha_l(\mathbf{s}) : \mathbf{s} \in S\}_{l=1}^{L-1}$ has Gaussian marginals and $\{\boldsymbol{\theta}_l(\mathbf{s}) : \mathbf{s} \in S\}_{l=1}^{L}$ are independent and identically distributed sample paths from a given stochastic process. The computational tractability can be simplified if conjugate SUN priors (Durante, 2019) are adopted for α_l. In the same spirit Rigon and Durante (2017) have proposed a tractable Logistic stick breaking prior of the form:

$$\begin{aligned} y_i|z_i = k, \boldsymbol{x}_i &\sim K_{\boldsymbol{x}_i}(y_i; \phi_k), \\ \mathbb{P}(z_i = k|\boldsymbol{x}_i) = \pi_k(\boldsymbol{x}_i) &= \nu_k(\boldsymbol{x}_i) \prod_{l=1}^{k-1} \{1 - \nu_l(\boldsymbol{x}_i)\}, \\ \nu_k(\boldsymbol{x}_i) &= \frac{\exp\left\{\psi(\mathbf{x}_i)^{\mathrm{T}} \alpha_h\right\}}{1 + \exp\left\{\psi(\mathbf{x}_i)^{\mathrm{T}} \alpha_h\right\}} = \frac{\mathrm{pr}(z_i = k|\boldsymbol{x}_i)}{\mathrm{pr}(z_i > k - 1|\boldsymbol{x}_i)} \quad \forall k \geq 1, \end{aligned} \tag{5.23}$$

where $K_{\boldsymbol{x}_i}(y_i; \phi_h)$ indicates a parametric kernel that can depend on both the covariate values specific of the i-th individual $\boldsymbol{x}_i \in \mathbb{R}^p$ and on the indexing parameter ϕ_h, ψ is a generic transformation of the covariates and each α_h is the vector of regression coefficients. For our application,

we fix ψ equal to the identity function. This formulation allows each $\nu_h(\boldsymbol{x}_i)$ to be interpreted as the probability of being allocated to component h, conditionally on the event of "surviving" to the previous $1, \dots, h-1$ components, meaning $\nu_h(\boldsymbol{x}_i) = \text{pr}(z_i = h | z_i > h-1, \boldsymbol{x}_i)$.
Given the Polya-Gamma conjugacy introduced by Polson et al. (2013), the authors propose a easy-to-implement algorithm that preserves the computational tractability assuming multivariate Normal priors for $\alpha_h \sim MVN(\mathbf{0}, cI_p)$, where I_p is a $p \times p$ unitary diagonal matrix. In this framework, we can modify (5.19) removing the homogeneity-inducing $\mathcal{N}^q$ term and incorporate the external information regarding proximity and other data features obtaining

$$\begin{aligned} \boldsymbol{\mu}_{i,L} | \boldsymbol{z}, \boldsymbol{d} &= \hat{\mathcal{P}}(\boldsymbol{\mu}_{i,L} | d_{z_i}), \\ z_i | \boldsymbol{\pi} &\overset{ind}{\sim} \sum_{k=0}^{+\infty} \pi_k(\boldsymbol{x}_i) \delta_k, \end{aligned} \tag{5.24}$$

with $\pi_k(\boldsymbol{x}_i)$ defined as in (5.23). The covariate information can encode already available measurements or, alternatively, we can extract topological information with previous exploratory analyses such as the output of a hierarchical clustering procedure.

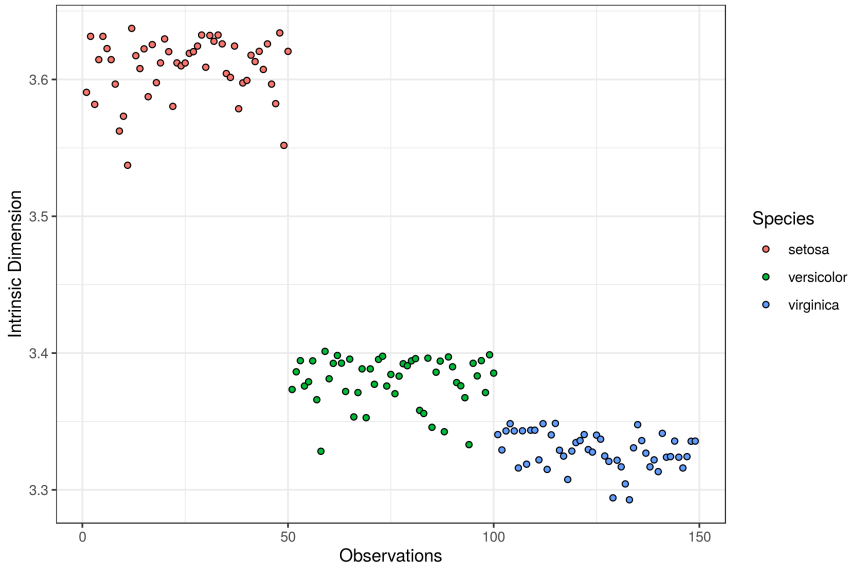


FIGURE 5.1: `Iris` dataset - Median Intrinsic Dimension a posteriori obtained fitting a Dependent Dirichlet Process approached applied to Crime model, K=10

As a preliminary application, we use again the `Iris` dataset. This time, we employ the `Species` variable as a covariate. We truncate the stick-breaking prior in (5.24) at $K = 10$ and we consider only the first ratio of distances, setting $L = 1$. We also set $c = 100$, and we run 50000 iterations after a burn-in period of the same length. Figure 5.1 shows promising results: with only the variable already present in the dataset, we can recover a result that is very similar to the one presented in Figure 5.3.

5.6.2 The representative individual.

In some applications, once the manifold assignment has been recovered, it might be interesting to summarise each group with a "representative individual" (RI). A first, simplistic solution would be to synthesize each cluster taking the considering or the coordinate-wise average or median of the observations assigned to that group. These euclidean quantities are fast and easy to compute even for high-dimensional datasets. However, the resulting RI may be different from

the observed data, substantially ignoring the topological properties of the manifold. Multivariate refinements of the median are available and we refer to Small (1990) for a review. The medoids, recovered minimizing the average dissimilarity to all the objects in the same cluster, are known to be a more informative alternative since they are always restricted to be members of the dataset (Struyf et al., 1996).

Information regarding the topology of the data is included employing the Geodesic distance: this quantity is computed taking into account the geometry of the latent manifold in which the data are embedded. Once a (possibly weighted) local graph among the data is built, the geodesic path between two nodes is defined as the path with the minimum number of edges. If the graph is weighted, it is a path with the minimum sum of edge weights. The length of a geodesic path (the sum of its weights) is called geodesic distance or shortest distance. A graph is usually represented by a distance matrix $\mathcal{D}$, whose non-zero entries indicate the presence of an edge between two nodes.
Let $\hat{\mathcal{N}}^q$ be the symmetric version of $\mathcal{N}^q$, such that $\hat{\mathcal{N}}^q_{ij} = 1$ if $\mathcal{N}^q_{ij} = 1 \vee \mathcal{N}^q_{ji} = 1$. Thus, $\hat{\mathcal{N}}^q$ describes an unweighted undirected graph. We compute $\mathcal{D}$ as

$$\mathcal{D}_{ij} = D \odot \hat{\mathcal{N}}^q, \tag{5.25}$$

where D is the usual euclidean distance matrix and $\odot$ indicates the element-wise product. We recover a weighted graph, where every point is connected with its first q-NNs, and the weight of each connection is exactly the euclidean distance between the two nodes. Once this matrix is available, we apply the Floyd—Warshall's algorithm (Floyd, 1962) to derive a geodesic distance matrix $\mathcal{D}^G_F$. Many other algorithms are available to derive $\mathcal{D}^G$: another option is to make use of the Spherelets Geodesic estimator of Li and Dunson (2019a), a state-of-the-art method that includes the value of the intrinsic dimension of the data in the computation. We leave for future research the investigation of how different distances may affect the ID estimation.

Once a geodesic distance matrix is available, we need to select the representative individual. Let $\Delta^d_G(x_1, x_2)$ be the geodesic distance between two points $x_1, x_2 \in \mathbb{R}^D$, embedded on a potentially non-linear sub-manifold of dimension $d < D$. Thus, we can estimate the representative individual $\hat{x}_d$ with

$$\hat{x}_d = \arg\min_x \sum_{i=1}^{n} \Delta^d_G(x, x_i), \tag{5.26}$$

in the spirit of the definition of Fréchet-mean. However, this proposal does not ensure that the $\hat{x}_d$ is actually a member of the dataset and, for high-dimensional samples, can be extremely computationally expensive. Instead, we obtain a representative point selecting the observation of the dataset that minimizes the average geodesic distance from all the other data points.

We apply all the aforementioned approaches to two different samples: the `Spiral` dataset, a two-dimensional spiral embedded in a $D = 3$ dimensional space, and to the `Iris` dataset. To obtain a multivariate median we employ the Liu's method (Liu et al., 1999) on the `Spiral` dataset, which is exact for two-dimensional data. Instead, for the `Iris` dataset we compute Tukey's median (Tukey, 1975). We also estimate the overall medoid using the `pam` algorithm (Reynolds et al., 2006), fixing the number of groups equal to 1. To obtain $\mathcal{D}^G_F$ we set $q = 3$. Applying the algorithm of Li and Dunson (2019a) requires the specification of an integer ID and the number $\hat{q}$ of NNs needed to compute the spherelets approximation. For the `Spiral` dataset we set $d = 2$ and we study the results for $\hat{q} = 3, 4, 5$. For the `Iris` dataset, following the insights provided by Figure 5.3, we separate three clusters according to the Species level and we roughly set the IDs equal to $d_1 = 4$, $d_2 = 3$ and $d_3 = 3$.

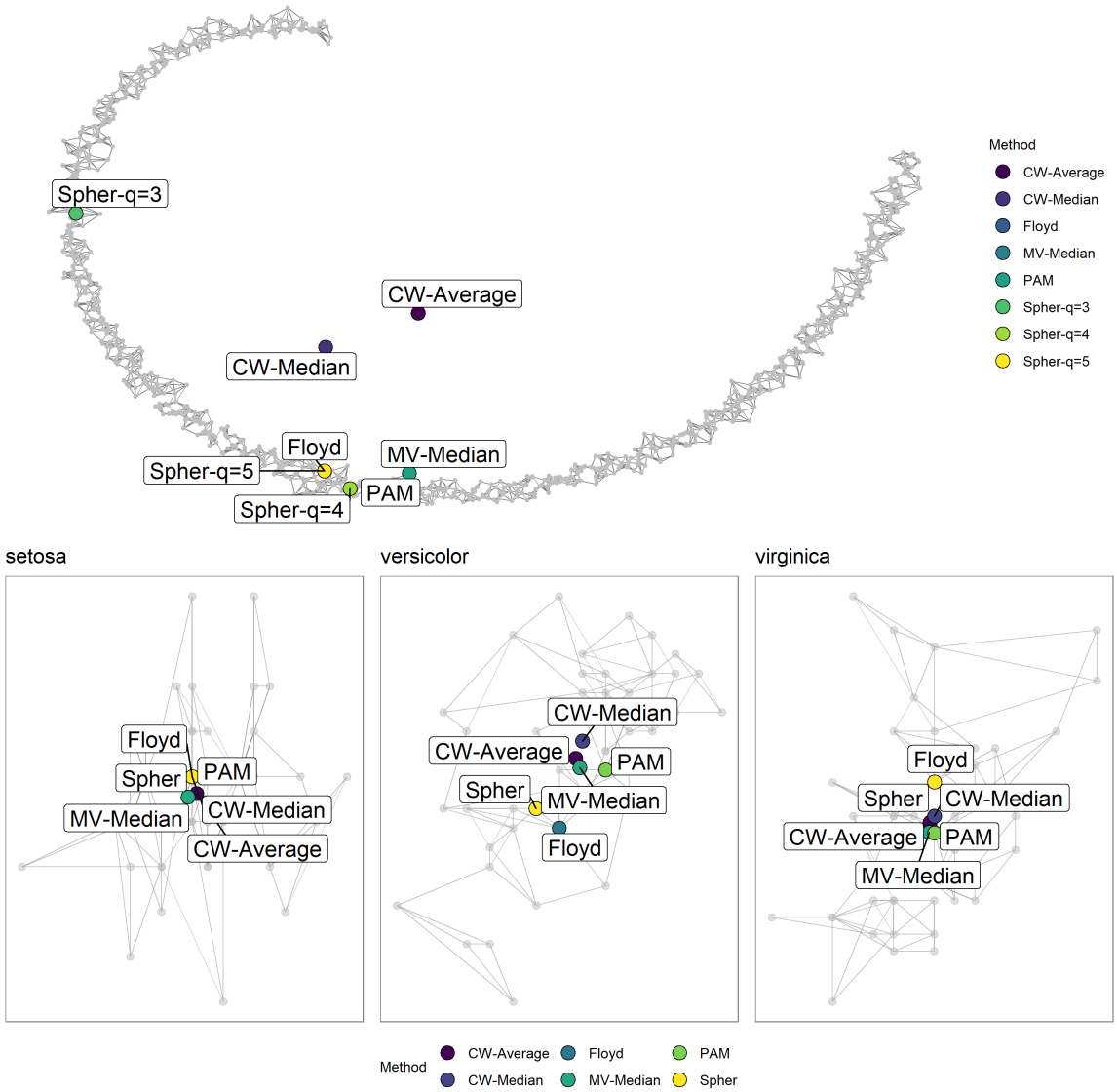


FIGURE 5.2: Top panel: `Spiral` Dataset - 1000 observations (gray dots) of intrinsic dimension $d = 2$, embedded in a $D = 3$-dimensional space. Bottom panel: `Iris` Dataset - Observations (gray dots) of dimension $d_1 = 4$, $d_2 = 3$, and $d_3 = 3$ embedded in a $D = 4$ dimensional space. The points are projected onto the second and the third dimensions. The undirected edges indicates whether or not two points are connected according to the matrix $\hat{\mathcal{N}}^q$. Representative individuals obtained with different methods are highlighted. In detail: `CW-Average`, `CW-Median` are the coordinate-wise average and median, `MV-Median` represents the multivariate median, `PAM` denotes the overall medoid, `Floyd` and `Spher` label the RIs found applying graphical-based methods

From the top panel of Figure 5.2, where the results on the `Spiral` dataset are reported, it is immediate to see why the coordinate-wise average and median (`CW-Average`, `CW-Median`) have poor representative importance: the two points are outside the manifold where the data lie, losing any interpretability. The multivariate median (`MV-Median`) and the overall medoid (`PAM`) coincide. All the graph-based methods (`Floyd`, `Spher`) provide similar results, except for the Spherelets approach when $q = 3$. The bottom panel of Figure 5.2 shows the RI for the `Iris` data, stratified by Species. The data are projected onto the second and third dimensions to allow a graphical representation. All the methods provide similar results, and we can appreciate

how the central tendency is captured. In general, methods based on graphical distances are preferable. In particular, for these simple cases, Floyd—Warshall's algorithm turns out to be computationally efficient and clear, providing a reliable RI.

5.6.3 Discussion and Conclusions

Dimensionality reduction is one of the most investigated areas of statistics. One of the key aspects of this field is the estimation of the ID of a dataset, namely the dimension of the non-linear sub-manifold where the data are postulated to lie. In this work, we have proposed a local, non-linear probabilistic mixture model that allows for the presence of heterogeneous IDs in the same dataset. We start proposing two simple modifications to Hidalgo, a finite mixture model where the IDs are regarded as mixture component parameters: instead of adopting classical conjugate independent distributions, we propose to consider truncated distributions (especially useful when the nominal dimension of the dataset D is limited) and repulsive densities. This would avoid the creation of small clusters characterized by very similar IDs, avoiding overfitting. Then, we show that the theoretical framework presented in Facco et al. (2017) and Allegra et al. (2019), where the authors provide a law for the ratio of the distances between the first and second nearest neighbor of a point, can be easily extended to a general number of consecutive ratios. With simulations, we underline that including more than one ratio helps to lower the variance of the estimates, even if the hypothesis of independence and identical distribution adopted in the MLE case can be too limiting. We exploit the derived results proposing a BNP mixture model, CRIME, for which we derive a slice sampler to perform posterior inference, applied on simulated and real datasets. We also showed some preliminary results about the derivation of a representative individual for the manifold, once a partition of the data reflecting the different intrinsic manifolds is estimated.
The developed theoretical framework is elegant and contributes to the Homogeneous Poisson Point process theory, but at the same time contains the major limitation of this model: the independence among the ratios (or the exchangeability, if we adopt the Bayesian paradigm) is difficult to justify, especially in cases where the number of data points is limited and the number of consecutive ratios in high. Models that can handle the covariance among the ratios will be the focus of future research, as well as models that take into account measurement error. A second limitation stands in the choice of q, the number of neighbors in the introduced adjacency matrix. At the moment, we justify its choice by selecting a value equal to the number of ratios considered, since both are linked to the local homogeneity on the same scale. On one hand, the model can be extended modeling the adjacency matrix $\mathcal{N}^q$ more carefully, possibly exploiting network or graph theory. On the other hand, one can also attempt a different approach, including the information regarding the local topology as a covariate with a Dependent Dirichlet Process approach. We discuss some possibilities in the previous section. That being said, many other exciting future directions are possible. We show how the model works well with functional data. This suggests that applications on more involved data structures, such as networks or tensors, as long as we define a proper metric that induces a suitable neighborhood structure among the points. It would be also interesting to extend the model from a theoretical point of view, investigating the possibility of including concepts from Extreme Value Theory. Estimating the IDs can also be useful in many statistical applications other than exploratory procedure and classification tasks: studying the variations in the IDs of the generated samples in Approximate Bayesian Computation (ABC, Beaumont et al., 2002) as a function of a distance threshold could uncover to interesting insights for better tuning that kind of models.

Appendix

5.A Proofs of theoretical results

The following useful results hold:

1. Let $X \sim Exp(\rho)$ and $Y \sim Erlang(n, \rho)$, such that $X \perp\!\!\!\perp Y$. Then, $Z = \frac{X}{Y} + 1 \sim Pareto(1, n)$.

2. If $Z \sim Pareto(x_m, \alpha)$ then $\log(Z/x_m) \sim Exp(\alpha)$.
   Proof: recall that $f_Z(z) = dx_m^d x^{-(d+1)}$ and consider $w = \log(z/x_m) \iff z = x_m \exp(w)$. The Jacobian of the (scalar) transformation is $dz = x_m \exp(w)dw$.
   Then, $f_W(w) = x_m \exp(w) dx_m^d (x_m exp(w))^{-(d+1)} = d\exp(-dw))$ which is the density of a Exponential random variable with parameter d.

3. If $X \sim Pareto(1, \alpha)$, then $Y = X^q \sim Pareto(1, \alpha/q)$.
   Proof: If $X \sim Pareto(1, \alpha)$, then $f_X(x) = \alpha x^{-(1+\alpha)}$. Then, since $X = Y^{1/q}$ and $\frac{d}{dy} y^{1/q} = \frac{1}{q} Y^{1/q-1}$ employing the Jacobian method we obtain:

$$f_Y(y) = \alpha y^{-(1/q+\alpha/q)} \frac{1}{q} y^{1/q-1} = \left(\frac{\alpha}{q}\right) y^{(-\alpha/q+1)}$$

   which is the density of a $Pareto(1, \alpha/q)$ random variable.

5.A.1 Proof of Lemma 1

Given the previous results 1. and 3., proving Lemma 1 is straightforward. Consider v_j, the volume of hyperspherical shell computed as $\omega_d \left(r_j^d - r_{j-1}^d\right)$, where r_j is the distance between an observation x and its j-th nearest neighbor (NN). We know that $v_i \overset{i.i.d.}{\sim} Exp(\rho)$ ($Erlang(1, \rho)$). Then, according to 1., $\frac{v_2}{v_1} + 1 \sim Pareto(1, 1)$. Also, $\frac{v_2}{v_1} + 1 = r_2^d / r_1^d$. We can then conclude that

$$\mu = \frac{r_2}{r_1} = \left(\frac{v_2}{v_1} + 1\right)^{1/d}$$

which means that, according to result 3., we have $\mu \sim Pareto(1, d)$.

5.A.2 Proof of Lemma 2

The marginal distributions in (5.9),(5.10) follow from elementary properties of Exponential, Gamma and Pareto random variables. We drop the observational index i for ease of exposition. Let us call $\gamma_l = \log\left(\frac{r_l}{r_{l-1}}\right)$, for $l = 2, 3, \ldots, L$. We want to derive the joint density of $(\gamma_1, \gamma_2, \ldots, \gamma_L)$. To do so, we start from the joint density of $(v_1, v_2, v_3, \ldots, v_L)$, with density $f(v_1, v_2, v_3, \ldots, v_L) = \rho^L \exp\left[-\rho \sum_{l=1}^L v_l\right]$. Consider the following one-to-one transformation:

$$\begin{cases} \gamma_1 &= v_1 \\ \gamma_2 &= \frac{1}{d}\log\left(1+\frac{v_2}{v_1}\right) \\ \gamma_3 &= \frac{1}{d}\log\left(1+\frac{v_3}{v_1+v_2}\right) \\ &\vdots \\ \gamma_L &= \frac{1}{d}\log\left(1+\frac{v_L}{\sum_{l=1}^{L-1} v_l}\right) \end{cases} \iff \begin{cases} v_1 &= \gamma_1 \\ v_2 &= \gamma_1\left(\exp\left[d\gamma_2\right]-1\right) \\ v_3 &= \gamma_1\exp\left[d\gamma_2\right]\left(\exp\left[d\gamma_3\right]-1\right) \\ \vdots &\quad \vdots \\ v_L &= \gamma_1\exp\left[d\sum_{l=2}^{L-1}\gamma_l\right]\left(\exp\left[d\gamma_L\right]-1\right) \end{cases}$$

Now, let $E_2^L = e^{d\sum_{l=2}^{L-1}\gamma_l}$. This transformation has Jacobian:

$$J = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \left(e^{d\gamma_2}-1\right) & \gamma_1 d e^{d\gamma_2} & 0 & \dots & 0 \\ e^{d\gamma_2}\left(e^{d\gamma_3}-1\right) & \gamma_1 d e^{d\gamma_2}\left(e^{d\gamma_3}-1\right) & \gamma_1 d e^{d\gamma_2}e^{d\gamma_3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d e^{d\sum_{l=2}^{L-1}\gamma_l}\left(e^{d\gamma_L}-1\right) & d\gamma_1 E_2^L\left(e^{d\gamma_L}-1\right) & d\gamma_1 E_2^L\left(e^{d\gamma_L}-1\right) & \dots & d\gamma_1 E_2^L e^{d\gamma_L} \end{bmatrix}$$

whose determinant is $|J| = \gamma_1^{L-1} d^{L-1}\prod_{l=2}^{L}\exp\left[d\cdot(L-l+1)\gamma_l\right]$.
We can write

$$f(\gamma_1,\gamma_2,\dots,\gamma_L) = \rho^L\gamma_1^{L-1}d^{L-1}\exp\left[-\rho\gamma_1 E_2^L\right]\prod_{l=2}^{L}\exp\left[d\cdot(L-l+1)\gamma_l\right]$$

We then integrate out γ_1 to find $f\left(\gamma_2,\dots,\gamma_L\right)$:

$$\begin{aligned} f\left(\gamma_2,\dots,\gamma_L\right) =& d^{L-1}\prod_{l=2}^{L}(l-1)\exp\left[-(l-1)d\gamma_l\right] \\ =& d\exp\left[-d\gamma_2\right]\cdots d(L-1)\exp\left[-d(L-1)\gamma_L\right]. \end{aligned}$$

This result reveals an interesting property. Since $f\left(\gamma_2\dots,\gamma_L\right) = \prod_{l=2}^{L} f\left(\gamma_l\right)$, we can conclude that $\gamma_2,\dots,\gamma_L$ are independent exponential random variables.

So we have obtained the following general statement:

$$\left(\gamma_2,\dots,\gamma_L\right) \sim Exp(d)\times\cdots\times Exp((L-1)\,d). \tag{5.27}$$

This holds if when the hypothesis of uniformity of the intensity of the Poisson Process holds till the L-th NN. Again, we can use this result combined with q, jointly modeling the first q ratios.

5.A.3 Proof of Lemma 3

Let $\{X_i\}_{i=1}^n$, $n\geq 2$, denote a sequence of independent exponential random variables with pairwise distinct parameters λ_i. The sum of n random variables $X_i\sim Exp(\lambda_i)$ is said to follow an Hypoexponential distribution, with density

$$f_{X_1+X_2+\cdots+X_n}(x) = \left[\prod_{i=1}^{n}\lambda_i\right]\sum_{j=1}^{n}\frac{\mathrm{e}^{-\lambda_j x}}{\prod_{\substack{l=1\\ l\neq j}}^{n}\left(\lambda_l-\lambda_j\right)}, \quad x>0$$

We want to characterize the distribution of $X = \mu_{i,n_1,n_2} = \frac{r_{n_2}}{r_{n_1}}$, with $n_2 > n_1$ integer values. Notice that

$$X = \frac{r_{n_2}}{r_{n_1}} = \frac{r_{n_2}}{r_{n_2-1}} \cdot \frac{r_{n_2-1}}{r_{n_2-2}} \cdots \frac{r_{n_1+1}}{r_{n_1}},$$

i.e. X can be rewritten as the product of $n_2 - n_1$ independent Pareto distributions. If we consider instead $Y = \log(X)$, we end up with

$$Y = \log(X) = \log\left(\frac{r_{n_2}}{r_{n_1}}\right) = \gamma_{n_2} + \ldots + \gamma_{n_1+1},$$

i.e. a sum of $L = n_2 - n_1$ independent Exponential random variables with integer parameters ranging from $n_1 d$ to $(n_2 - 1)d$.
So now we can write the distribution of Y employing an Hypoexponential density:

$$\begin{aligned} f_Y(y) =& \left[\prod_{i=1}^{L} \lambda_i\right] \sum_{j=1}^{L} \frac{\mathrm{e}^{-\lambda_j y}}{\prod_{\substack{l=1 \\ l\neq j}}^{L} (\lambda_l - \lambda_j)} \\ =& d^{(n_2-n_1)} \frac{(n_2-1)!}{(n_1-1)!} \sum_{j=1}^{n_2-n_1} \frac{\mathrm{e}^{-(n_1+j-1)dy}}{d^{n_2-n_1-1} \prod_{\substack{l=1 \\ l\neq j}}^{n_2-n_1} ((n_1+l-1)-(n_1+j-1))} \\ =& d \frac{(n_2-1)!}{(n_1-1)!} \sum_{j=1}^{n_2-n_1} \frac{\mathrm{e}^{-(n_1+j-1)dy}}{\prod_{\substack{l=1 \\ l\neq j}}^{n_2-n_1} (l-j)}, \quad y > 0. \end{aligned} \tag{5.28}$$

We notice that the following equality holds

$$\prod_{\substack{l=1 \\ l\neq j}}^{n_2-n_1} (l-j) = (j-1)!(n_2-n_1-j)!(-1)^{j+1}$$

Figure 5.A.1 shows some example of the density in (5.28) for different values of d, n_1 and n_2:
We can also derive the distribution for $X = \exp(Y)$, transforming the last density:

$$\begin{aligned} f_X(x) =& d \frac{(n_2-1)!}{(n_1-1)!} \frac{1}{x} \sum_{j=1}^{n_2-n_1} \frac{\mathrm{e}^{-(n_1+j-1)d\log x}}{\prod_{\substack{l=1 \\ l\neq j}}^{n_2-n_1} (l-j)} = d \frac{(n_2-1)!}{(n_1-1)!} \sum_{j=1}^{n_2-n_1} \frac{x^{-(n_1+j-1)d-1}}{\prod_{\substack{l=1 \\ l\neq j}}^{n_2-n_1} (l-j)}, \\ =& d \frac{(n_2-1)!}{(n_1-1)!} \sum_{j=1}^{n_2-n_1} \frac{x^{-(n_1+j-1)d-1}}{(j-1)!(n_2-n_1-j)!(-1)^{j-1}} \\ =& d \frac{(n_2-1)!}{(n_1-1)!} \sum_{k=1}^{n_2-n_1} \frac{x^{-(n_2-k)d-1}}{(k-1)!(n_2-n_1-k)!(-1)^{n_2-n_1-k}} \\ =& \frac{d}{x^{n_2 d+1}} \frac{(n_2-1)!}{(n_1-1)!} \sum_{k=1}^{n_2-n_1} \frac{x^{kd}(-1)^{n_2-n_1-k}}{(k-1)!(n_2-n_1-k)!} \\ =& \frac{d}{x^{n_2 d+1}} \frac{(n_2-1)!}{(n_1-1)!} \frac{(n_2-n_1-1)!}{(n_2-n_1-1)!} \sum_{k=1}^{n_2-n_1} \frac{x^{kd}(-1)^{n_2-n_1-k}}{(k-1)!(n_2-n_1-k)!} \\ =& \frac{d}{x^{n_2 d+1}} \frac{(n_2-1)!}{(n_1-1)!} \frac{1}{(n_2-n_1-1)!} \sum_{l=0}^{n_2-n_1-1} \binom{n_2-n_1-1}{l} (x^d)^{l+1} (-1)^{n_2-n_1-l-1} \\ =& \frac{d}{x^{(n_2-1)d+1}} \frac{(n_2-1)!}{(n_1-1)!} \frac{(x^d-1)^{n_2-n_1-1}}{(n_2-n_1-1)!} = (n_2-n_1)\binom{n_2-1}{n_1-1} \frac{d(x^d-1)^{n_2-n_1-1}}{x^{(n_2-1)d+1}}, \quad x > 1. \end{aligned} \tag{5.29}$$

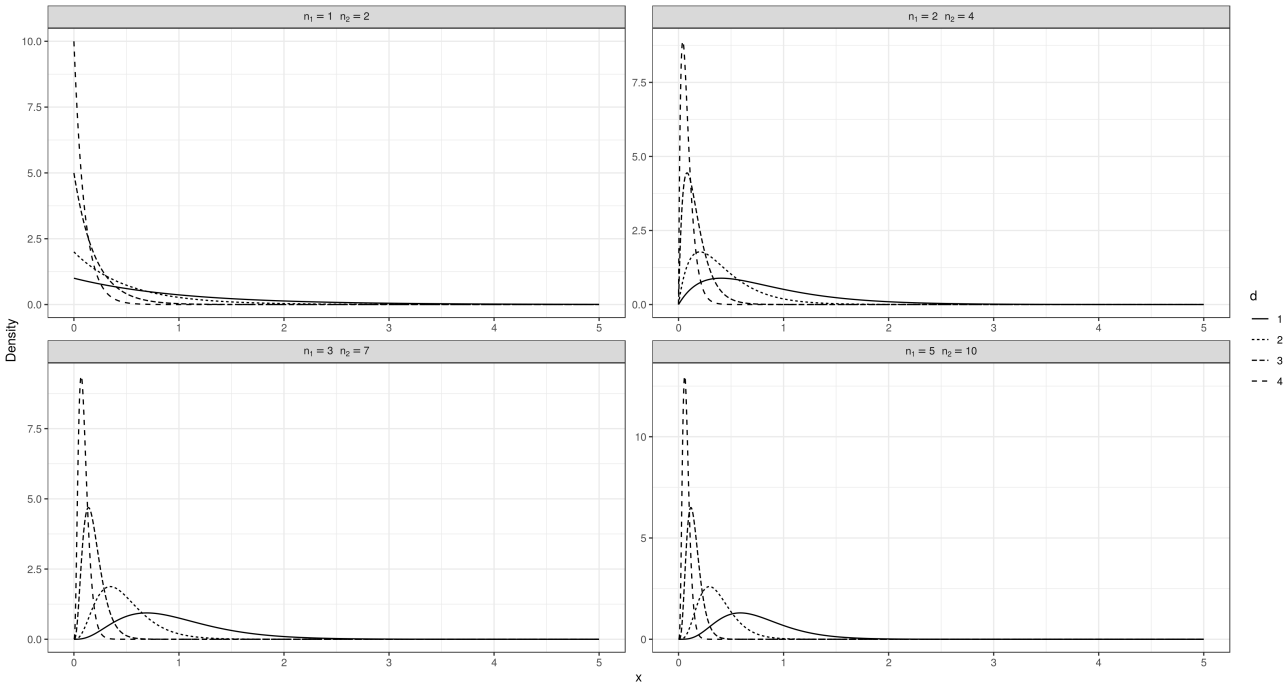


FIGURE 5.A.1: The various panels report the shapes of the density in (5.28) for different values of d, n_1 and n_2

Note that at the fourth line we applied the reflection property of the indexes of a sum: $\sum_{k=1}^{K} a_k = \sum_{k=1}^{K} a_{K-k+1}$. In the sixth line, we applied the Newton binomial formula.
Figure 5.A.2 shows some example of the density in (5.29) for different values of d, n_1 and n_2:

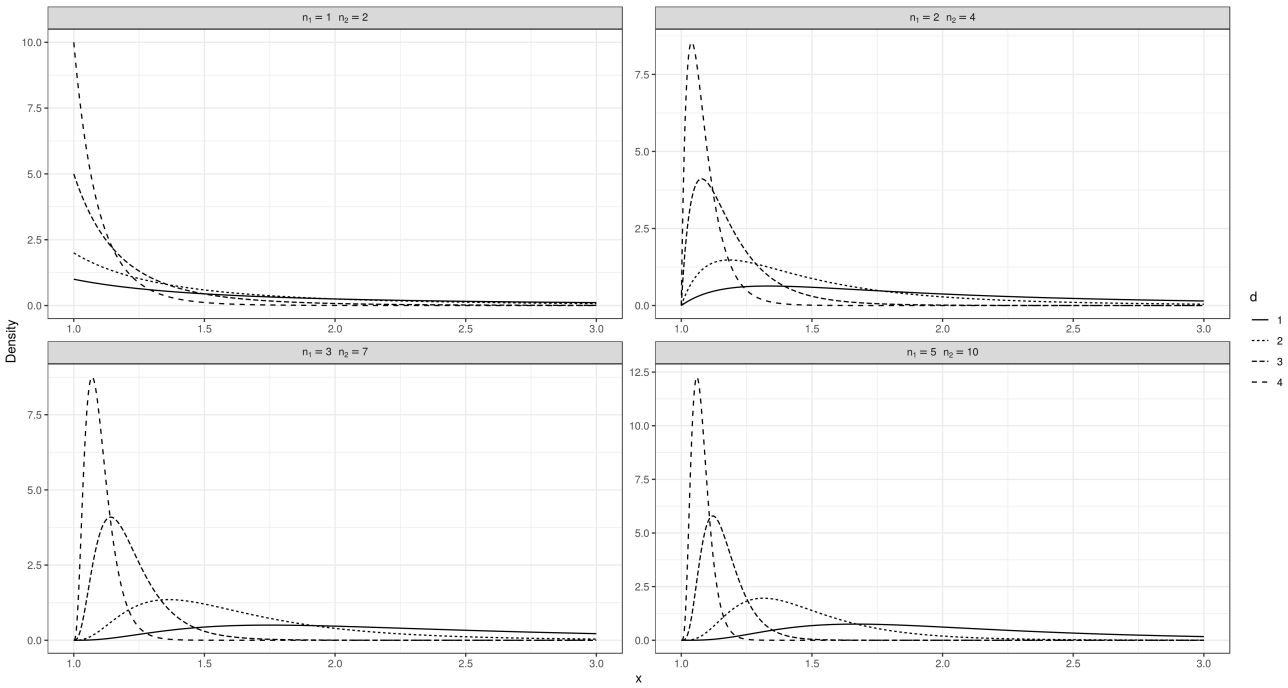


FIGURE 5.A.2: The various panels report the shapes of the density in (5.29) for different values of d, n_1 and n_2

We can specialize the last density to the case where $n_1 = n_0$ and $n_2 = 2n_0$. We recover:

$$\begin{aligned} f_X(x) =& d\frac{(2n_0-1)!}{(n_0-1)!}\sum_{j=1}^{n_0}\frac{x^{-(n_0+j-1)d-1}}{(j-1)!(n_0-j)!(-1)^{j+1}}, \\ =& \frac{(2n_0-1)!}{(n_0-1)!^2}\cdot\frac{d(x^d-1)^{n_0-1}}{x^{(2n_0-1)d+1}}, \quad x>1. \qquad\qquad x>1. \end{aligned}$$

5.A.4 Distributions of the distances r_l

We can also derive the joint distribution of the first L distances $\{r_l\}$ from a point x to its first L NNs. Recall that $f(v_1,\dots,v_L) = \prod_{l=1}^{L} f(v_l)$, where $f(v_l) = \rho\exp(-\rho v_l)$, meaning that $v_l \overset{i.i.d.}{\sim} Exp(\rho)$. We can consider the following one-to-one transformation:

$$\begin{cases} r_1 &= \left(\frac{v_1}{\omega_d}\right)^{1/d} \\ r_2 &= \left(\frac{v_1+v_2}{\omega_d}\right)^{1/d} \\ r_3 &= \left(\frac{v_1+v_2+v_3}{\omega_d}\right)^{1/d} \\ \vdots & \\ r_L &= \left(\frac{\sum_{l=1}^{L} v_l}{\omega_d}\right)^{1/d} \end{cases} \iff \begin{cases} v_1 &= \omega_d r_1^d \\ v_2 &= \omega_d\left(r_2^d - r_1^d\right) \\ v_3 &= \omega_d\left(r_3^d - r_2^d\right) \\ \vdots & \\ v_L &= \omega_d\left(r_L^d - r_{L-1}^d\right) \end{cases}$$

The determinant of the Jacobian of this transformation is $|J| = (\omega_d d)^L \prod_{l=1}^{L} r_l^{d-1}$. Thus, the distribution of the first L distances has density:

$$f(r_1,\dots,r_L) = (\rho\omega_d d)^L\left(\prod_{l=1}^{L} r_l^{d-1}\right)\exp\left[-\rho\omega_d r_L^d\right],$$

with $r_l \in \mathbb{R}^+$ with the constraint that $r_1 < r_2 < \ldots < r_L$.
It can be particularly interesting to derive the marginal distribution of the generic distance r_L. This can be easily done integrating out the previous distances r_l, $l = 1,\dots,L-1$.

$$\begin{aligned} f(r_L) &= \int_0^{r_{L-1}}\int_0^{r_{L-2}}\cdots\int_0^{r_1}(\rho\omega_d d)^L\left(\prod_{l=1}^{L} r_l^{d-1}\right)\exp\left[-\rho\omega_d r_L^d\right]dr_1\cdots dr_L \\ &= \exp\left[-\rho\omega_d r_L^d\right](\rho\omega d)^L\int_0^{r_{L-1}}\int_0^{r_{L-2}}\cdots\int_0^{r_1}\left(\prod_{l=1}^{L} r_l^{d-1}\right)dr_1\cdots dr_L \\ &= \exp\left[-\rho\omega_d r_L^d\right](\rho\omega d)^L\int_0^{r_{L-1}}\int_0^{r_{L-2}}\cdots\int_0^{r_1}\left(\prod_{l=1}^{L} r_l^{d-1}\right)dr_1\cdots dr_L \\ &= \exp\left[-\rho\omega_d r_L^d\right](\rho\omega d)^L\frac{r_L^{Ld-1}}{(L-1)!d^{L-1}}. \end{aligned}$$

This result is remarkable since we have proven that the generic distance from a point to its L-th NN follows a Generalized Gamma distribution, whose density is given by

$$f(x) = \frac{p/a^q}{\Gamma(q/p)}x^{q-1}e^{-(x/a)^p}, \qquad x,a,p,q>0,$$

Therefore, $f(r_L)$ is a Generalized Gamma density with parameters

$$p = d, \quad a = \frac{1}{\sqrt[d]{\rho\omega_d}}, \quad q = Ld.$$

There is another, much easier way to recover this result. Since $v_l \sim Exp(\rho)$ for each $l = 1, \ldots, L$, it is easy to see that the volume of the hypersphere of radius r_L, defined as $V_L = \sum_{l=1}^{L} v_l = \omega_d r_L^q$ follows an Erlang distribution: $V_L \sim Gamma(L, \rho)$. Thus,

$$V_L \sim Gamma(L,\rho) \iff r_L^d \sim Gamma(L, \omega_d\rho) \iff r_L \sim GenGamma\left(d, \frac{1}{\sqrt[d]{\rho\omega_d}}, Ld\right).$$

5.B A general formula for sampling Interval-truncated random variables

Consider the real-valued random variable X, characterized by density function $f(x)$, cumulative density function (c.d.f.) $F(x)$ and support $S_X = \mathbb{R}$. Let us consider the simplest case of a truncated random variable. Suppose we are interested in the probability density of X after restricting the support to be between two constants: $I = (a, b]$. In other words, we are interested in the distribution of X given $a < X \leq b$, given by

$$f(x|a < X \leq b) = \frac{f(x)\mathbf{1}_{(a,b]}(x)}{F(b) - F(a)}.$$

If $a = -\infty$ or $b = +\infty$, we speak about right and left truncation, respectively. In the statistical literature, numerous algorithms for sampling from truncated distributions have been proposed. One of the most used relies on the so-called *inverse c.d.f. method*. Algorithm 1 lists the steps, as they are described in Saucier (2000). Of course, the key point in order to exploit such an algorithm is the clear definition of the c.d.f. of the truncated distribution. For the univariate case, it is reasonably simple to invert the formula, analytically or numerically. Dealing with the peculiar form of the repulsive densities introduced in Petralia et al. (2012), we need a method that is capable to sample from distributions whose supports have been truncated over more than one interval.

For simplicity, consider the case in which we have two intervals, $I_1 = (a, b]$ and $I_2 = (c, d]$ where we assume $b < c$ so that $I_1 \cap I_2 = \emptyset$. Let us denote with E_1 and E_2 the events $\{a < X \leq b\}$ and $\{c < X \leq d\}$, respectively. Moreover let $P_1 = F(b) - F(a)$, $P_2 = F(d) - F(c)$ and $T_2 = P_1 + P_2$. The new density is given by

$$f(x|E_1, E_2) = \frac{f(x)\mathbf{1}_{I_1 \cup I_2}(x)}{P_1 + P_2}.$$

The new c.d.f. consequently becomes:

$$F(x|E_1, E_2) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{1}{T_2}\left(F(x) - F(a)\right) & \text{if } x \in (a, b] \\ \frac{P_1}{T_2} & \text{if } x \in (b, c] \\ \frac{P_1}{T_2} + \frac{1}{T_2}\left(F(x) - F(c)\right) & \text{if } x \leq (c, d] \\ 1 & \text{if } x \geq d \end{cases}$$

Once the mechanism is clear, it is straightforward to generalized to the case of n disjoint intervals, identified by the sequence of real numbers $a_1^1 < a_1^2 < a_2^1 < a_2^2 < \ldots < a_n^1 < a_n^2$, where a_k^1 and a_k^2 are the extremes of the k-th interval I_k, $k \in \{1, \ldots, n\}$. Let us denote the (ordered)

intervals as $I_1, I_2, \ldots, I_n$. Let $p_k = F(a_k^2) - F(a_k^1)$, $P_k = \sum_{r=1}^{k} p_r$ and $T = \sum_{r=1}^{n} p_r$. Finally, let $E_k = \{a_k^1 < X \leq a_k^2\}$, $\bar{I}_{k,k+1} = \{a_k^2 < X \leq a_{k+1}^1\}$ with $\bar{I}_{0,1} = \{-\infty < X \leq a_1^1\}$ and $\bar{I}_{n,n+1} = \{a_k^2 < X < +\infty\}$.

Algorithm 1: Basic Algorithm for sampling a truncated distribution with the inverse c.d.f. method.

Input:
NSIM: number of samples desired, a: left truncation parameter, b: right truncation parameter
Result: A sample of dimension NSIM from a Truncated Distribution

```
for i ∈ {1, ..., NSIM} do
    Generate U_i ~ U(0, 1).
    Set Y_i = F(a) + [F(b) − F(a)]U_i.
    Return X_i = F^{-1}(Y_i).
end
```

The density in this more general case is given by

$$f(x|E_1, \ldots, E_n) = \frac{f(x)\mathbf{1}_{I_1 \cup \ldots, \cup I_n}(x)}{T}.$$

The new c.d.f. consequently becomes:

$$F(x|E_1, \ldots, E_n) = \begin{cases} 0 & \text{if } x \in \bar{I}_{0,1} \\ \frac{1}{T}\left(F(x) - F(a_1^1)\right) & \text{if } x \in I_1 \\ \frac{p_1}{T} & \text{if } x \in \bar{I}_{1,2} \\ \frac{p_1}{T} + \frac{1}{T}\left(F(x) - F(a_2^1)\right) & \text{if } x \in I_2 \\ \ldots & \ldots \\ \frac{P_{k-1}}{T} & \text{if } x \in \bar{I}_{k-1,k} \\ \frac{P_{k-1}}{T} + \frac{1}{T}\left(F(x) - F(a_k^1)\right) & \text{if } x \in I_k \\ \frac{P_k}{T} & \text{if } x \in \bar{I}_{k,k+1} \\ \ldots & \ldots \\ 1 & \text{if } x \in \bar{I}_{n,n+1} \end{cases}$$

Despite its verbose formulation, this function is elementary to implement. To use the inverse c.d.f. method, we first sample a value u_i from a $\mathcal{U}(0,1)$ and we then compare the corresponding quantile x_i^* solving the equation $F(x_i^*) = u_i$. This can be easily done in Rcpp using the bisection method (Burden and Faires, 1985). As an example, we report four different scenarios: a sample of 100000 instances from a $Gamma(5,1)$ and its corresponding c.d.f. and the other three cases, where the random variable has been truncated on different intervals. In detail, the three collections of intervals are

1. $I_1 = (4, 9]$
2. $I_1 = (1, 2]$, $I_2 = (4, 7]$, $I_3 = (10, 12]$
3. $I_k = (k, k + 0.5]$, with $k \in \{0, \ldots, 15\}$

Figure 5.B.1 reports the four scenarios. We can appreciate that the algorithm, despite its simplicity, works well even in cumbersome situations like the third framework (bottom-right panels).

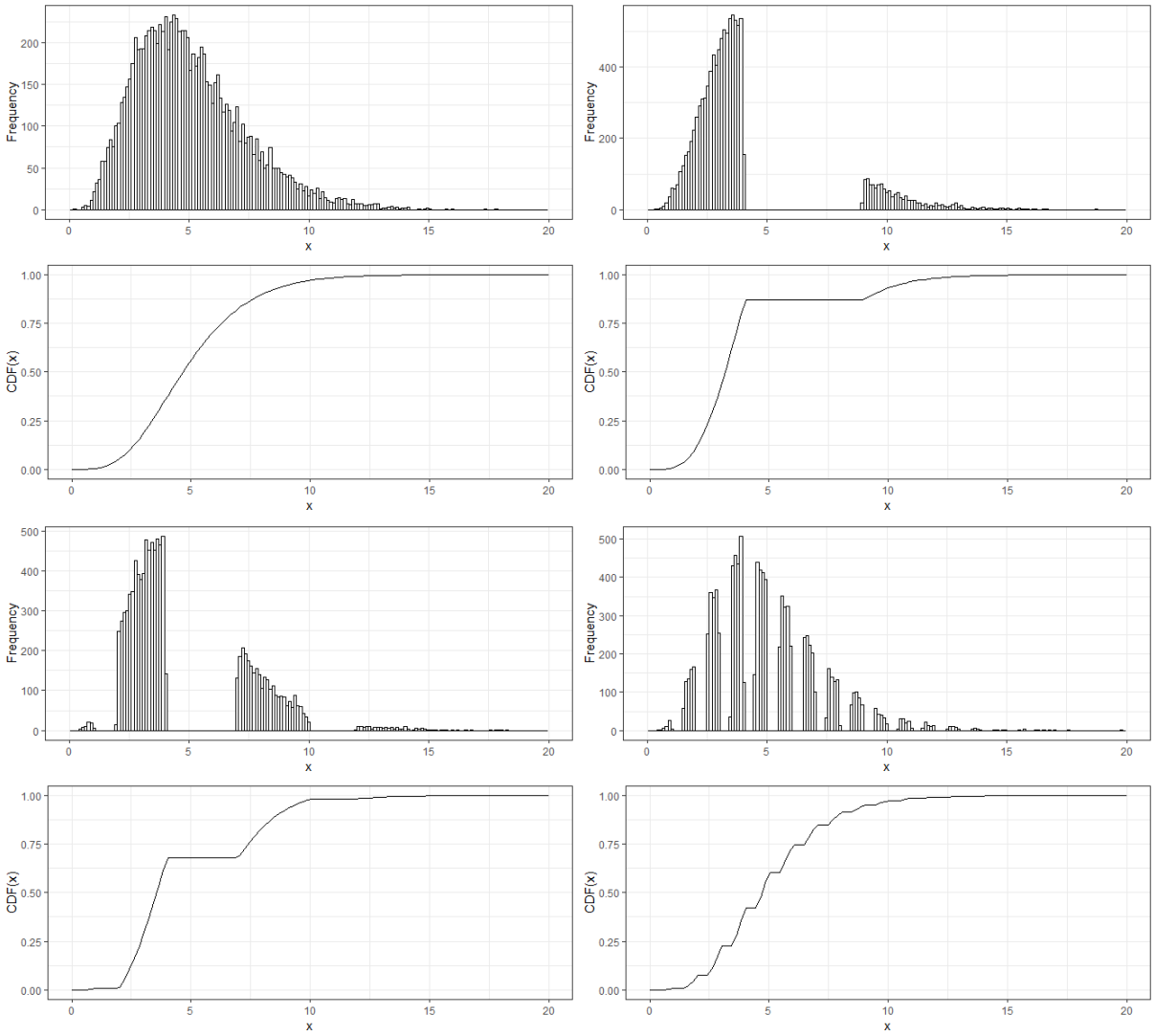


FIGURE 5.B.1: Four different applications of the the inverse c.d.f. method based on interval-truncated c.d.f.s. The plots report the histograms (top panels) sampled from specific distributions (bottom panels).

5.C Additional simulation with truncated priors

We also considered a simple simulated dataset to assess the utility of the truncated, uniform and truncated with point mass priors. We considered four uniform clouds of points, well separated in the space. The dimension of the dataset is $D = 7$, while the clouds are of dimension $d = 1, 3, 5, 7$. We see how, employing a simple truncation leads to an underestimation of the true ID of the fourth group.

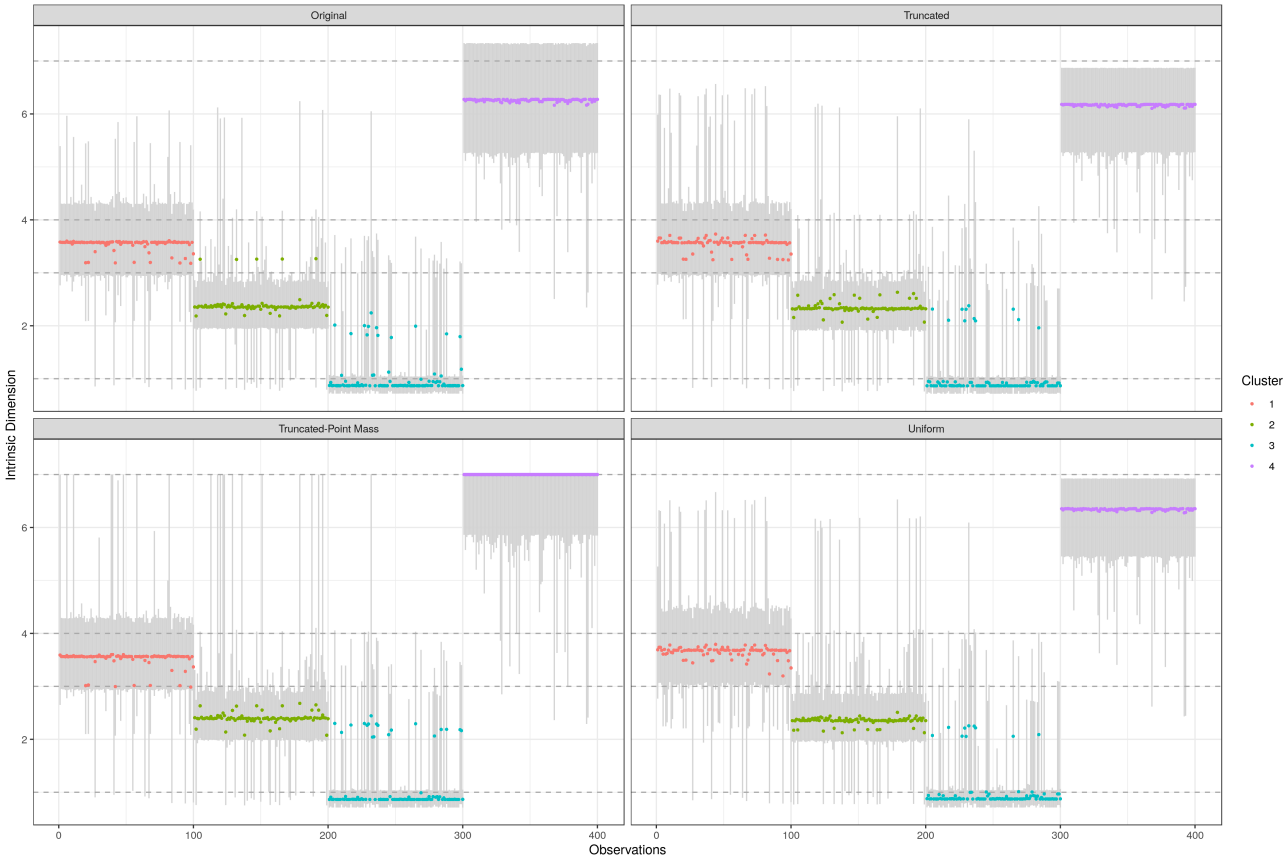


FIGURE 5.C.1: `Uniform clouds` dataset. The top-left panel reports the posterior means and the 90% credible sets for each observation, after applying Hidalgo with K=4. The panels on the right show the same output when a truncated Gamma (top) or a Uniform (bottom) prior on d is adopted. The bottom-left panel shows the results when a Truncated Gamma is mixed with a point mass in D, with prior mixture proportion equal to $\hat{\rho} = 0.9$. The dashed horizontal lines denote the true values of the different IDs.

5.D Five Gaussians Dataset

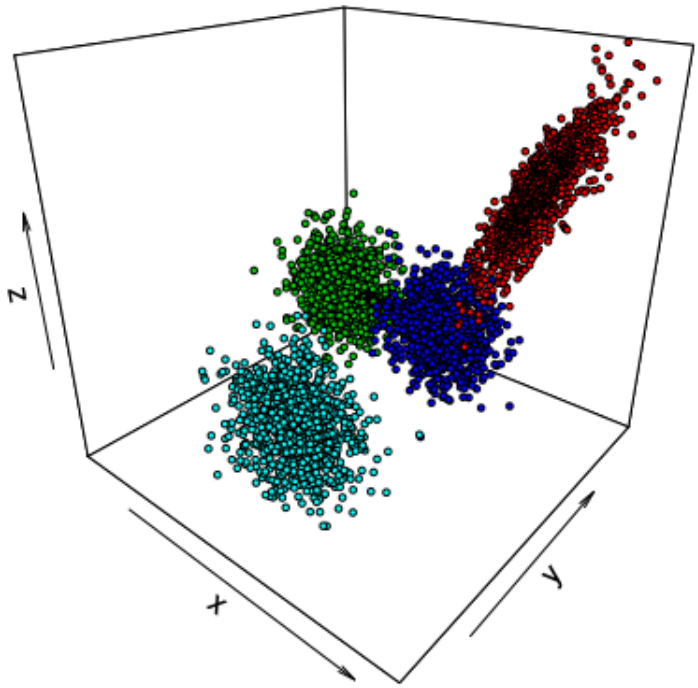


FIGURE 5.D.1: The first three coordinates (out of 10) of the `5 Gaussians` dataset used in Allegra et al. (2019). Evidently, the five distributions are overlapping. The 1-dimensional Gaussian is hidden in the main cloud of points.

5.E Schmitz dataset: list of 19 interesting genes

We report the codes of the 19 genes that have been selected by field expert for the analysis. They are: `DENND3, BCL6, LRMP, NEK6, LMO2, ITPKB, SERPINA9, MYBL1, MME, FUT8, BLNK, BMF, CCND2, SH3BP5, PIM1, IRF4, ENTPD1, ETV6, IL16`.

Chapter 6

Conclusions and Future Directions

"Sono già stato qui; e mi manca, come mi manco anch'io,
smarrito in realtà labili alla coscienza."
Sono già stato qui – Salvariello Panzatore

"The thing I like best about winter: when it's over."
A shop sign in Newport Beach

In this dissertation, we have introduced and discussed several contributions focusing specifically on Bayesian mixture models, typically in a nonparametric setting. Each method is accompanied by an algorithm for conducting posterior inference, a simulation study, and real data applications. We now summarise our main proposals and highlight future directions.

In Chapter 2, we develop a BNP methodology for partially exchangeable data, where the observations are organized into sub-populations, or units. The proposed Common Atoms Model (CAM) can be seen as a modification of the nested Dirichlet Process (nDP, Rodríguez et al., 2008), where the support of the almost surely discrete nonparametric model is constrained to be the same across all the units. The constraint does not compromise the flexibility of the nonparametric methodology and, more importantly, leads to a model that does not suffer from the degeneracy property of the nDP, whereby the posterior collapses to the fully exchangeable case whenever ties are present in the data (Camerlenghi et al., 2019b). In the nDP, the within-group clustering still contributes to a compact representation of the data, but unit-level inference across subgroups is precluded. Instead, the proposed CAM framework naturally allows unit-level inference and clustering of observations across groups, since the structure of the common atoms allows mapping group-specific distributional patterns onto shared support.
We have studied in detail the properties of the model and have developed a nested version of the slice sampler to enable fast and efficient computations when conducting posterior inference with a large number of observations. The proposed framework can also be easily extended to accommodate several types of data structures. For example, in our application to a microbiome dataset, we have embedded the CAM into a rounded mixture of Gaussian framework (Canale and Dunson, 2011), to accommodate for the presence of discrete observations. We have shown how the estimated clustering structures can be useful for understanding the sources of the heterogeneity in microbiome samples, and inform subsequent microbiome analyses.
Many future directions are possible. First, the model can be promptly robustified by employing Pitman-Yor (PY) priors instead of the more typical DP. Moreover, CAM can be improved into a model capable of taking into account the presence of covariates and/or data evolving over time, which are ubiquitous in modern biology. For example, one can construct dependent random measures with covariate-dependent weights as in MacEachern (2000). Finally, we remark that

the shared atom structure of CAM makes it particularly suited for the analysis of nested data sets where the distributions of the units are expected to differ only over a small fraction of the observations. Of course, this may not always be the case, and more general methods for handling partially exchangeable data may be needed. More precisely, one can study how to model both the shared and the idiosyncratic parts of the distributions in a more flexible way. In Chapter 1 we discussed our first attempt, the Mixed nested DP, which is based on the superposition of random measures. It would be interesting to further investigate this topic, providing more flexible models and developing tractable and fast algorithms for the estimation of more involved partially exchangeable structures.

Chapter 3 and 4 are related since they both discuss improvements of the two-groups model (Efron, 2004). According to this framework, the test statistics in a multiple testing problem are modeled as a mixture of two competing densities called null and non-null components, respectively. In both the chapters, we comply to the rationale that the null distribution f_0, and the non-null distribution f_1 should be well separated, with the first taking values around the origin and the second having longer tails.
More specifically, in Chapter 3 we propose a version of the two-groups model where both f_0 and f_1 are modeled with PY mixture models. The base measures and the hyperparameters are carefully chosen to ensure flexibility for f_1 and to constrain f_0 to be close to the theoretical null distribution, i.e. a standard Gaussian. We develop a marginal sampler and assess its performance against well-established competitors with simulation studies. Moreover, we apply this method to the test statistics from a Negative Binomial regression in a microbiome study, where the OTU frequencies are associated with case/control variables. Motivated by the application to the Prostate cancer dataset, future research for this project will be focused on how to improve the computational aspects: the marginal algorithm can be extremely costly, especially for large datasets. We have currently developed a split-merge move for this framework (Dahl, 2005), but we are planning to investigate also the use of the state-of-the-art importance conditional sampler described in Canale et al. (2019).
In Chapter 4, we focus on the requirement of separation between f_0 and f_1 and we employ non-local distributions (Johnson and Rossell, 2010) to model the non-null densities. More in detail, we first introduce a general family of weighted densities, that encompasses many known distributions. Then, we propose to model f_1 with a non-local density, i.e. a density function that vanishes in a neighborhood of the origin. The use of non-local priors to directly model the non-null likelihood ensures that the overlap among the two competing distributions is greatly reduced. We propose both a parametric and a nonparametric specification of the model. Moreover, we show that under reasonable assumptions the weighted version of the model leads to better results than its unweighted counterpart, in terms of false discovery rate, false negative rate, and statistical power. We compare the model with competitors in an extensive simulation study and then apply our methodology to publicly available datasets from genomics.
Several future directions are worth pursuing, beyond the multiple testing framework. For example, the connection between weighted distributions and penalization methods should be further investigated, as it may provide a general framework for the development of interesting priors that can induce particularly desirable results (e.g. shrinkage, sparsity, etc.). Moreover, weighted and non-local likelihoods can also be employed for classification problems in data-analysis.

In Chapter 5 we examine the notion of Intrinsic Dimension (ID) of a dataset with a local probabilistic model formulation. The ID of a dataset of dimension D, usually denoted as d, is the dimension of the hidden, possibly non-linear manifold, where the data are postulated to lie. Expecting some redundancy among the variables, we can assume $d < D$. We first proposed to improve the performance of *Hidalgo*, the heterogeneous ID estimator introduced in Allegra et al. (2019) using a more suitable prior specification for the ID parameter. Hidalgo is based on

the fact that, under mild assumptions, the ratio of the distances between the first two nearest neighbors (NN) of a point follows a Pareto distribution parameterized only by d. We extend the above methodology, by providing close form distributions for every consecutive ratio of NN distances and for ratios between the distances of NN of general order. We also develop CRIME, a Bayesian nonparametric mixture model to avoid to specify beforehand the number of hidden manifolds present in a dataset, as required by Hidalgo. We investigate the performances of the model by means of a set of simulation studies, and we also apply the model to a well-known, real dataset. We need to underline that this modeling framework is well suited for continuous data, but is not able to handle discrete realizations, especially when ties are present in the dataset. Our proposal opens up a variety of different possible directions. First, the developed theoretical contributions pave the way to the investigation of the dependence between the ID and the scale at which we observe a dataset. Hidalgo and CRIME are based on a constrained likelihood, where the additional term imposes local homogeneity in the data and at the same time informs about their topological structure. We can investigate a different modeling framework considering an unconstrained likelihood specification and introducing additional information in different ways. As an example, we showed the possibility of employing a Dependent Dirichlet Process approach. The concept of ID is extremely general, and we are currently using CRIME to study the hidden structure of complex data, (e.g. network data, functional data) and inferential methodologies (e.g. Approximate Bayesian Computation). Lastly, we are researching on how to extend the model to take into account measurement error.

Chapter 7

Acknowledgements

"Serve pane e fortuna, serve vino e coraggio,
Soprattutto ci vogliono buoni compagni di viaggio!"
(Manco a farlo apposta) Le Luci d'America – Luciano Ligabue

"It all seems so very arbitrary. I applied for a job at this company because they were hiring.
I took a desk at the back because it was empty. But no matter how you get there or where you end up,
human beings have this miraculous gift to make that place home. Let's do this."
The Office, Finale – Creed

Here I am at the only part of my thesis that I will re-read from time to time. I had so much fun writing the acknowledgments for my master's thesis and I imagined in a thousand ways of how I would structure this one. Trying to remember every meeting, every small gesture that was so important in getting me here. But, like every good academic, I am now very close to the deadline, finding myself having/wanting to write this section in a hurry. And since the long analogy with the Bayesian paradigm has already been used, I opt for the more classic and tested stream of consciousness.
It has always been hard for me to detach my personal life over the past three years from "doing my Ph.D.", as in a "work-life" trade-off. Rather, I feel that this career choice (combined perhaps with a little ambition) has influenced my mood and all my important decisions to a great extent. Wondering why, I thought about how to explain this feeling and recently I surprised myself pronouncing: "Have you seen? Look, I've done everything by myself, on my own!". The lack of truthfulness of this sentence struck me as it came out of my mouth. On my own? Please. These three years were so unique because of the large number of great people that my Ph.D. allowed me to meet. I have so much to be thankful for, so many people to acknowledge. These few lines are for you who have accompanied, inspired, raised and changed me.

"An advisor knows he has done his job well if you hate him a bit towards the end of the doctorate" (Anonymous Advisor). I will protect the identity of the author of this sentence, but perhaps it hides some truth. In any case, I must admit that I might have developed a slight Stockholm syndrome. So, first of all, a thought to my guides.
Thanks Antonietta, for being a turning point of my Ph.D.. You have not only been an advisor, but also an excellent mentor and a great manager. And looking at everything I have been able to accomplish with your boost, it is clear that you are really a bit of a magician as well as a master illusionist.
Thanks Michele, the right person at the right moment. You made me (re-)discover the beauty of doing research and (understanding right away how to effectively motivate me) you always pushed me to do better. I have still got a long way to go, but now I feel I have a solid starting point, and your example to follow.
Thank you Prof. Mecatti for all your help. If I have felt supported in all my academic choices during the past years, I also owe it to you. The same goes for the coordinator of my program,

Prof. Vittadini.

I also feel particularly lucky to have met my colleagues, who shared the Ph.D. journey with me. Thanks Luca, Riccardo, Nik, and Tia: you have taught me so much, from the powerful Greek pragmatism to knowing how to distinguish what is measurable from what is disgusting. Moreover, thanks to Riccardo, Anna, Andrea, Paolo, and Federica. With you all, I shared way more than just an office.
And thanks to you Cap, who with your enthusiasm and your friendship have literally changed my way of approaching research, news, and life in general. I can't express enough how grateful I am to have found you on my way and how much all the good that happened is owed to you.

An entire paragraph has to be dedicated to the amazing people I met in the U.S., thanks to that life-changing experience that was visiting Irvine. The UCI Stat. Department will always have a place in my heart, composed of brilliant students and faculty who are literal statistical rockstars. It has been an honor growing up there, as a researcher and as a person. In particular, thanks to Ben, Adam, Maricela, Kyle, Jaylen, Naomi, and Klaus. And to Isaac and Mary, Catherine and Peter, Bob and Connie, Lori, Carly and Casey, and Martin.
And of course to Wendy: thank you blondie, for your invaluable help with everyday life, for showing me all the cool places your country has to offer (Panera is definitely one of them!), for your precious friendship, and for all that we shared.
You all helped this inappropriate Italian to feel at home on the other side of the Ocean.

I want to also thank the countless sources of inspiration that shared a small part of my journey with me, starting from the brilliant professors and researchers with whom I had the fortune to collaborate and, more generally, to all the brilliant people that I met in USI, Sissa, Bocconi, Bicocca, Cattolica, and UCI. You all impressed and stimulated me to do my best: Stefano, Alessandro, Lucia, Daniele, Tommy, Antonio, Bernardo, Federico C., Federico A., Gaia, Dan, Hal, Alan, Veronica. Some of you have taught me a course, some others have organized workshops, others just told me few, powerful words in a particular moment. To all of you I look with admiration, in the hope of reaching your level.

Thanks to you, who live in the "real world ", outside the academy, outside statistics, who made me feel your support, especially in the moments when everything went wrong. So, thank you Ila, Mondo, Alberto, Clara, all my fellows from my theater family, my fellow swimmers, in particular Ale and Simo, and my Little Italy in California.

I am truly grateful to my family. Traveling and leaving home helped me better understand all the affection and love you feel for me (and you prove to me!) every day, so clear in every hug at the airport, whether I was leaving or coming back. All the support given to me in every moment of difficulty (from the most personal to the most stupid and bureaucratic ones) has become increasingly evident to me. I owe you a lot: everything I am doing is to make you proud.

I spun like a top across many universities for research, for teaching, for workshops and seminars. Getting out from the comfort zone of my office in Bicocca made me feel like a diver who is able to jump from the one-meter springboard but that suddenly realizes the existence of the ten-meter platform. In these last three years, I have just tried to overcome the fear of vertigo induced by the ladder. Now that I know how to climb that ladder, there is nothing left to do but learn how to jump.

They have been three intense years of climbing: I have learned a lot, and of this amount, Statistics is only the smallest part.

Eccomi arrivato all'unica parte della mia tesi che ogni tanto rileggerò. Ho provato così tanto gusto a scrivere i ringraziamenti nella mia tesi di laurea magistrale che ho fantasticato in mille modi come buttare giù quelli per il Ph.D., cercando di salvare ogni incontro, ogni piccolo gesto per me importante per ricordarli in modo efficace. Ma come ogni buon accademico, arrivo a pochissimo dalla consegna della tesi trovandomi a dover/voler buttar giù tutto di fretta. E visto che la lunga analogia con il paradigma Bayesiano me l'hanno già fregata, opto per il più classico e collaudato stream of consciuosness.

Riflettendoci, mi è sempre risultato difficilissimo distaccare la mia vita personale di questi ultimi tre anni dal "fare il dottorato", come in un binomio casa-lavoro. Credo piuttosto che questa scelta di carriera (combinata forse con un poco di ambizione) abbia permeato ogni istante e determinato più di una volta il mio mood e scelte importanti. Chiedendomi il perchè, ho pensato come spiegare questa sensazione, fino a che recentemente mi sono sorpreso a pronunciare: "Hai visto? Guarda che ho combinato da solo, con le mie forze!". La poca veridicità di quella frase mi ha colpito non appena è uscita dalla mia bocca. Da solo? Per favore. Cio che è stato fondamentale è il numero enorme di incontri e conoscenze che questo PhD mi ha permesso di fare. Ho davvero tanto per cui essere grato, a tanti. Queste righe sono per voi, che mi avete accompagnato, ispirato, cresciuto e cambiato.

"Un advisor sa di aver fatto bene il suo lavoro se verso la fine del dottorato si arriva ad odiarlo un po' " (Advisor Anonimo). Proteggerò l'indentità dell'autore di questa frase, ma forse nasconde un po' di verità. In ogni caso devo ammettere di aver comunque sviluppato una leggera sindrome di Stoccolma. Quindi, per prima cosa, un pensiero alle mie guide.
Grazie Antonietta, per aver costituito il turning point del mio dottorato. Non sei stata solo una notevole advisor, ma anche un ottimo mentore e una grande manager. E se riguardo a tutto quello che sono stato capace di fare grazie alla tua spinta, è evidente che sei davvero pure un po' maga oltre che un'abile illusionista.
Grazie Michele, persona giusta al momento giusto. Mi hai fatto riscoprire il piacere di fare ricerca e (capendo subito come pungolarmi) mi hai spinto a migliorarmi costantemente. Ho ancora un sacco di strada da fare, ma ora ho basi solide su cui poggiarmi e il tuo esempio da seguire.
Grazie Prof. ssa Mecatti per tutto il suo aiuto: se mi sono sentito sostenuto in ogni mia scelta accademica durante gli ultimi due anni lo devo anche a lei. Stesso dicasi per il coordinatore, Prof. Vittadini.

Ritengo poi di essere stato partciolarmente fortunato ad avere incontrato i colleghi che mi sono capitati. Grazie Luca, Riccardo, Nik e Tia: mi avete insegnato tanto, dal potente pragmatismo greco, al saper distinguere cio che è misurabile da cio che è disgustoso. E ancora grazie a Riccardo, Anna, Andrea, Paolo e Federica. Con voi tutti ho condiviso ben più che solo un ufficio.
E soprattutto grazie a te Cap, che con il tuo entusiasmo e la tua amicizia hai cambiato letteralmente il mio modo di approcciarmi alla ricerca, alle novità, alla vita in generale. Non posso esprimere abbastanza quanto sono grato di averti trovato sul mio cammino e quanto ti debba se questi anni sono andati per il verso giusto.

Un intero paragrafo non può che essere dedidcato alle persone incredibili che ho conosciuto negli U.S., grazie a quella esperienza di visiting a Irivine che mi ha cambiato la vita. Il dipartimento di Statistics a UCI avrà sempre un poso nel mio cuore, composto da studenti brillanti e rockstar statistiche come faculty. E' stato un onore avere la possibilità di unirmi a voi, per crescere come ricercatore e come persona. In particolare, grazie a Ben, Adam, Maricela, Kyle, Jaylen, Naomi e Klaus. E poi a Isaac e Mary, Catherine e Peter, Bob e Connie, Lori, Carly e Casey, Martin.

E ovviamente a Wendy: grazie bionda, per tutto il prezioso aiuto con la vita di ogni giorno, per avermi mostrato i posti spettacolari del tuo paese (Panera e' sicuramente nella lista!), per la tua preziosa amicizia e per tutto cio che abbiamo condiviso.
Voi tutti avete aiutato questo italiano inappropriato a sentirsi a casa anche dall'altra parte dell'Oceano.

Grazie alle innumerevoli altre fonti di ispirazione, partendo dai geniali professori e ricercatori con cui ho avuto la fortuna di collaborare e più in generale a tutti coloro che ho incontrato fra USI e Sissa, fra Bocconi e Cattolica, che mi hanno colpito e stimolato: Stefano, Alessandro, Lucia, Daniele, Tommy, Antonio, Bernardo, Federico, Federico e Gaia, Dan, Hal, Alan, Veronica. Qualcuno di voi mi ha insegnato un corso, qualcun altro organizzato workshop, altri ancora semplicemente condiviso con me poche parole, ma di enorme effetto. A voi tutti guardo con ammirazione, nella speranza di raggiungere il vostro livello.

Un grazie a voi che vivete nel "mondo reale", fuori dall'accademia, fuori dalla statistica, che mi avete fatto sentire il vostro tifo e supporto, specie nei momenti quando tutto andava storto: Ila, Mondo, Alberto, voi del teatro, Clara, Francesca, Giulia, Alex, voi del nuoto, in particolare Ale e Simo, e la mia Little Italy in California.

Sono davvero grato alla mia famiglia. Viaggiando e andandomene da casa ho capito ancora meglio tutto l'affetto che provate (e mi provate!) ogni giorno, così chiaro in ogni abbraccio in ogni aeroporto. Tutto il sostegno datomi in ogni momento di difficoltà (da quelle più personali a quelle più stupide e burocratiche) mi è diventato sempre più evidente. Vi devo molto, e ogni cosa che faccio è per rendervi fieri.

Ho girato come una trottola in molte università per ricerca, per didattica, per workshop e seminari. E questo mettere la testa fuori dalla mia confort zone dell'ufficio in Bicocca mi ha fatto sentire come un tuffatore abile dal trampolino di un metro che si accorge dell'esistenza della piattaforma dei dieci.
In questi tre anni ho solamente cercato di superare la paura delle vertigini indotte dalla scala. Ora che almeno la scala è alle spalle, non rimane che imparare a tuffarsi.

Sono stati tre anni densi, di scalata appunto, per cui devo essere davvero grato. Ho imparato davvero tanto, e di questo tanto la statistica non è che la minima parte.

"Il testo che avrei voluto scrivere
Non è di certo questo
Perciò dovrò continuare a scrivere
Perchè di certo riesco. Prima o poi."
Il testo che avrei voluto scrivere - Michele Salvemini

Bibliography

[1] Raja Hafiz Affandi, Emily Fox, and Ben Taskar. "Approximate Inference in Continuous Determinantal Processes". In: (2013), pp. 1430–1438.

[2] Raja Hafiz Affandi, Emily B. Fox, Ryan P. Adams, and Ben Taskar. "Learning the parameters of determinantal point process kernels". In: *31st International Conference on Machine Learning, ICML 2014* 4 (2014), pp. 2967–2981.

[3] Michele Allegra, Elena Facco, Alessandro Laio, and Antonietta Mira. "Data classification based on the local intrinsic dimension". In: (2019), pp. 1–27.

[4] David B. Allison, Gary L. Gadbury, Moonseong Heo, José R. Fernández, Cheol-Koo Lee, Tomas A. Prolla, and Richard Weindruch. "A mixture model approach for the analysis of microarray gene expression data". In: *Computational Statistics & Data Analysis* 39.1 (2002), pp. 1–20.

[5] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays". In: *Proceedings of the National Academy of Sciences* 96.12 (1999), pp. 6745–6750.

[6] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken Ichi Kawarabayashi, and Michael Nett. "Estimating local intrinsic dimensionality". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2015-Augus (2015), pp. 29–38.

[7] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10 (2010), R106. arXiv: 1310.0424.

[8] Charles E. Antoniak. "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems". In: *The Annals of Statistics* 2.6 (1974), pp. 1152–1174.

[9] Julyan Arbel, Pierpaolo De Blasi, and Igor Prünster. "Stochastic Approximations to the Pitman–Yor Process". In: *Bayesian Analysis* (2019).

[10] Robert Bain. "Are our brains Bayesian?" In: *Significance* 13.4 (2016), pp. 14–19.

[11] Dipankar Bandyopadhyay and Antonio Canale. "Non-parametric spatial models for clustered ordered periodontal data". In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 65.4 (2016), pp. 619–640.

[12] Anjishnu Banerjee, Jared Murray, and David B Dunson. "Bayesian Learning of Joint Distributions of Objects". In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* 31.Mdm (2013), pp. 1–9. arXiv: 1303.0449.

[13] Andrés F Barrientos, Alejandro Jara, and Fernando A Quintana. "On the Support of MacEachern's Dependent Dirichlet Processes and Extensions". In: *Bayesian Analysis* 7.2 (2012), pp. 277–310.

[14] J. Baselga, C. S. Pedamallu, A. M. Earl, F. Duke, A. D. Kostic, J. Jung, D. Gevers, A. I. Ojesina, M. Meyerson, S. Ogino, J. Tabernero, M. Michaud, C. Huttenhower, A. J. Bass, R. A. Shivdasani, B. W. Birren, W. S. Garrett, and C. Liu. "Genomic analysis identifies association of Fusobacterium with colorectal carcinoma". In: *Genome Research* 22.2 (2012), pp. 292–298.

[15] Vladimir Batagelj, Nata Sa, and Simona Korenjak Cerne. "Clustering of modal valued symbolic data". In: (2015), pp. 1–26. arXiv: arXiv:1507.06683v1.

[16] Mark A. Beaumont, Wenyang Zhang, and David J. Balding. "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4 (2002), pp. 2025–2035. arXiv: 1112.4755.

[17] Yoav Benajmini and Yosef Hochberg. "Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing". In: *J R Statist Soc B.* B 57.1 (1995), pp. 289–300. arXiv: 95/57289 [0035-9246].

[18] Sourabh Bhattacharya. "Gibbs sampling based Bayesian analysis of mixtures with unknown number of components". In: *Sankhya: The Indian Journal of Statistics* 70.1 (2008), pp. 133–155.

[19] D. A. Binder. "Bayesian cluster analysis". In: *Biometrika* 65.1 (1978), pp. 31–38.

[20] C. M. Bishop. *Neural Networks for Pattern Recognition*. 1995.

[21] David Blackwell and James B. MacQueen. "Ferguson Distributions Via Polya Urn Schemes". In: *The Annals of Statistics* 1.2 (1973), pp. 353–355.

[22] David M. Blei and Peter I. Frazier. "Distance dependent Chinese restaurant processes". In: *Journal of Machine Learning Research* 12 (2011), pp. 2461–2488.

[23] Natalia Bochkina and Judith Rousseau. "Adaptive density estimation based on a mixture of gammas". In: *Electronic Journal of Statistics* 11.1 (2017), pp. 916–962.

[24] Carlo Emilio Bonferroni. "Teoria Statistica Delle Classi e Calcolo Delle Probabilità". In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936* (1936).

[25] Z. I. Botev. "The normal law under linear restrictions: simulation and estimation via minimax tilting". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 79.1 (2017), pp. 125–148. arXiv: 1603.04166.

[26] Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W Berg, Anders Nielsen, Hans J. Skaug, Martin Mächler, and Benjamin M. Bolker. "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling". In: *The R Journal* 9.2 (2017), pp. 378–400.

[27] James H. Bullard, Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". In: *BMC Bioinformatics* 11.1 (2010), p. 94.

[28] Richard L. Burden and J. Douglas Faires. "The Bisection Algorithm". In: *Numerical Analysis* (1985).

[29] Caliński T. and Harabasz J. "A Dendrite Method For Cluster Analysis". In: *Communications in Statistics* 3.1 (1974), pp. 1–27.

[30] Francesco Camastra and Alessandro Vinciarelli. "Estimating the intrinsic dimension of data with a fractal-based method". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.10 (2002), pp. 1404–1407.

[31] Federico Camerlenghi. "Hierarchical and Nested Random Probability Measures with Statistical Applications". In: *Ph.D. Thesis* (2017).

[32] Federico Camerlenghi, Antonio Lijoi, Peter Orbanz, and Igor Prünster. "Distribution theory for hierarchical processes". In: *The Annals of Statistics* 47 (2019), pp. 67–92.

[33] Federico Camerlenghi, David B. Dunson, Antonio Lijoi, Igor Prünster, and Abel Rodríguez. "Latent nested nonparametric priors". In: *Bayesian Analysis* to appear (2019). arXiv: 1801.05048.

[34] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. "Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework". In: *Mathematical Problems in Engineering* 2015 (2015).

[35] Antonio Canale and David B. Dunson. "Bayesian kernel mixtures for counts". In: *Journal of the American Statistical Association* 106.496 (2011), pp. 1529–1539. arXiv: NIHMS150003.

[36] Antonio Canale and Igor Prünster. "Robustifying Bayesian nonparametric mixtures for count data". In: *Biometrics* 73.1 (2017), pp. 174–184.

[37] Antonio Canale, Riccardo Corradin, and Bernardo Nipoti. "Importance conditional sampling for Bayesian nonparametric mixtures". In: (2019). arXiv: 1906.08147.

[38] Antonella Capitanio. *The skew-normal and related families*. 2011, pp. 1–262.

[39] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76.1 (2017).

[40] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. "Handling Sparsity via the Horseshoe". In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics 5* (2009), pp. 73–80.

[41] George Casella and Christian P. Robert. "Rao-blackwellisation of sampling schemes". In: *Biometrika* 83.1 (1996), pp. 81–94.

[42] Gilles Celeux. "Bayesian Inference for Mixture: The Label Switching Problem". In: *Compstat* (1998), pp. 227–232.

[43] Eric Z. Chen and Hongzhe Li. "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data". In: *Bioinformatics* 32.17 (2016), pp. 2611–2617.

[44] Y.-Y. Chen, G. D. Wu, M. Bewtra, H. Li, K. Bittinger, K. Gupta, E. Gilroy, J. Chen, R. Sinha, D. Knights, L. Nessel, W. A. Walters, F. D. Bushman, R. Baldassano, R. Knight, S. A. Keilbaugh, J. D. Lewis, and C. Hoffmann. "Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes". In: *Science* 334.6052 (2011), pp. 105–108.

[45] Stuart Coles and Anthony Davison. *Statistical Modelling of Extreme Values*. 2008, pp. 1–70.

[46] Guido Consonni, Luca La Rocca, and Jim Q. Smith. "Moment Priors for Bayesian Model Choice with Applications to Directed Acyclic Graphs". In: *Bayesian Statistics 9* 9780199694 (2012).

[47] Jose A. Costa and Alfred O. Hero. "Geodesic entropic graphs for dimension and entropy estimation in Manifold learning". In: *IEEE Transactions on Signal Processing* 52.8 (2004), pp. 2210–2221.

[48] Cox and Cox. *Multi-dimensional scaling*. 2000.

[49] Jürgen Cox and Matthias Mann. "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification". In: *Nature Biotechnology* 26.12 (2008), pp. 1367–1372.

[50] David B. Dahl. "An improved merge-split sampler for conjugate Dirichlet process mixture models". In: *Technical Report* (2003), p. 32.

[51] David B. Dahl and Michael A. Newton. "Multiple hypothesis testing by clustering treatment effects". In: *Journal of the American Statistical Association* 102.478 (2007), pp. 517–526.

[52] D.B. Dahl. "Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models". In: *Journal of Computational and Graphical Statistics* 77843.2004 (2005).

[53] Paul Damien, Jon Wakefield, and Stephen G. Walker. "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.2 (1999), pp. 331–344.

[54] Maria De Iorio, Peter Müller, Gary L. Rosner, and Steven N. MacEachern. "An anova model for dependent random measures". In: *Journal of the American Statistical Association* 99.465 (2004), pp. 205–215.

[55] S. W. Dharmadhikari and K. Joag-Dev. "Mean , Median , Mode III". In: *Statistica Neerlandica* 37.4 (1983), pp. 165–168.

[56] M. Di Paola, G. Pieraccini, D. Cavalieri, P. Lionetti, S. Massart, M. Ramazzotti, J. B. Poullet, S. Collini, and C. De Filippo. "Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa". In: *Proceedings of the National Academy of Sciences* 107.33 (2010), pp. 14691–14696. arXiv: arXiv:1408.1149.

[57] T. G. Dinan and J. F. Cryan. "Melancholic microbes: A link between gut microbiota and depression?" In: *Neurogastroenterology and Motility* 25.9 (2013), pp. 713–719.

[58] Kim Anh Do, Peter Müller, and Feng Tang. "A Bayesian mixture model for differential gene expression". In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 54.3 (2005), pp. 627–644.

[59] Jason A. Duan, Michele Guindani, and Alan E. Gelfand. "Generalized spatial dirichlet process models". In: *Biometrika* 94.4 (2007), pp. 809–825.

[60] Leo L. Duan and David B. Dunson. "Bayesian Distance Clustering". In: (2018). arXiv: 1810.08537.

[61] Lutz Dümbgen. "Combinatorial stochastic processes". In: *Stochastic Processes and their Applications* 52.1 (1994), pp. 75–92.

[62] David B. Dunson and John E. Johndrow. "The Hastings algorithm at fifty". In: *Biometrika* (2019).

[63] David B. Dunson and Ju Hyun Park. "Kernel stick-breaking processes". In: *Biometrika* 95.2 (2008), pp. 307–323.

[64] Daniele Durante. "Conjugate Bayes for probit regression via unified skew-normal distributions". In: *Biometrika* (2019).

[65] Bradley Efron. "Empirical Bayes estimates for large-scale prediction problems". In: *Journal of the American Statistical Association* 104.487 (2009), pp. 1015–1028.

[66] Bradley Efron. "Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction". In: *Large-Scale Inference Empirical Bayes Methods for Estimation, Testing, and Prediction* (2012), pp. 1–263.

[67] Bradley Efron. "Large-scale simultaneous hypothesis testing: The choice of a null hypothesis". In: *Journal of the American Statistical Association* 99.465 (2004), pp. 96–104.

[68] Bradley Efron. "Rejoinder: Microarrays, Empirical Bayes and the Two-Groups Model". In: *Statistical Science* 23.1 (2008), pp. 45–47. arXiv: 0808.0603.

[69] Bradley Efron. "Size, power and false discovery rates". In: *Annals of Statistics* 35.4 (2007), pp. 1351–1377. arXiv: arXiv:0710.2245v1.

[70] Bradley Efron and Trevor Hastie. *Computer age statistical inference: Algorithms, evidence, and data science*. 2016, pp. 1–475.

[71] Bradley Efron and Robert Tibshirani. "Empirical Bayes methods and false discovery rates for microarrays". In: *Genetic Epidemiology* 23.1 (2002), pp. 70–86.

[72] Bradley Efron and Robert Tibshirani. "Using specially designed exponential families for density estimation". In: *Annals of Statistics* 24.6 (1996), pp. 2431–2461.

[73] Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher. "Empirical bayes analysis of a microarray experiment". In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1151–1160.

[74] M. D. Escobar. "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means". In: (1988).

[75] Michael D. Escobar and Mike West. "Bayesian density estimation and inference using mixtures". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 577–588. arXiv: 1212.4786.

[76] W. J. Ewens. "The sampling theory of selectively neutral alleles". In: *Theor. Popul. Biol.* 3 (1972), pp. 87–112.

[77] Elena Facco and Alessandro Laio. "The intrinsic dimension of biological data landscapes". PhD thesis. 2017.

[78] Elena Facco, Maria D'Errico, Alex Rodriguez, and Alessandro Laio. "Estimating the intrinsic dimension of datasets by a minimal neighborhood information". In: *Scientific Reports* 7.1 (2017), pp. 1–8.

[79] K. Falconer. *Fractal Geometry—Mathematical Foundations and Applications*. John Wiley & Sons, 2nd edition, 2003.

[80] Stefano Favaro and Yee Whye Teh. "MCMC for normalized random measure mixture models". In: *Statistical Science* 28.3 (2013), pp. 335–359.

[81] T. S. Ferguson. "Bayesian density estimation by mixtures of normal distributions." In: *Recent Advances in Statistics* (1983). Ed. by H Rizvi and J Rustagi, pp. 287–303.

[82] Thomas S. Ferguson. "A Bayesian Analysis of Some Nonparametric Problems". In: *The Annals of Statistics* 1.2 (1973), pp. 209–230. arXiv: arXiv:1011.1669v3.

[83] Bruno de Finetti. "Sur la condition d'equivalence partielle". In: *Actualitès Scientifiques et Industrielles* 739 (1938), pp. 5–18.

[84] Robert W. Floyd. "Algorithm 97: Shortest path". In: *Communications of the ACM* 5.6 (1962), p. 345.

[85] Keinosuke Fukanaga. *Introduction to Statistical Pattern Recognition*. Ed. by Academic press. 1972.

[86] Alan E. Gelfand. "Sampling-Based Approaches to Calculating Marginal Densities". In: *Journal of the American Statistical Association* 85.410 (1990), pp. 398–409.

[87] Alan E. Gelfand, Susan E. Hills, Amy Racine-Poon, and Adrian F.M. Smith. "Illustration of Bayesian inference in normal data models using Gibbs sampling". In: *Journal of the American Statistical Association* 85.412 (1990), pp. 972–985.

[88] Stuart Geman and Donald Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), pp. 721–741.

[89] Subhashis Ghosal and Aad van der Vaart. "Fundamentals of nonparametric Bayesian inference". In: *Fundamentals of Nonparametric Bayesian Inference* (2017), pp. 1–646.

[90] Jelle J. Goeman and Aldo Solari. "Multiple hypothesis testing in genomics". In: *Statistics in Medicine* 33.11 (2014), pp. 1946–1978. arXiv: arXiv:1306.1646v1.

[91] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring". In: *Science* 286.5439 (1999), pp. 531–527.

[92] Brian Goodman and Humphrey Gardner. "The microbiome and cancer". In: *Journal of Pathology* 244.5 (2018), pp. 667–676.

[93] Daniele Granata and Vincenzo Carnevale. "Accurate Estimation of the Intrinsic Dimension Using Graph Distances: Unraveling the Geometric Complexity of Datasets". In: *Scientific Reports* 6 (2016).

[94] Rebecca Graziani, Michele Guindani, and Peter F Thall. "Bayesian nonparametric estimation of targeted agent effects on biomarker change to predict clinical outcome". In: *Biometrics* 71.1 (2015), pp. 188–197.

[95] Peter J. Green. "Reversible jump Markov chain monte carlo computation and Bayesian model determination". In: *Biometrika* 82.4 (1995), pp. 711–732.

[96] J. E. Griffin and M. Kolossiatis. "Comparing Distributions Using Dependent Normalized Random Measure Mixtures". In: *Matrix* (2010), pp. 1–33.

[97] J. E. Griffin and M. F.J. Steel. "Order-based dependent dirichlet processes". In: *Journal of the American Statistical Association* 101.473 (2006), pp. 179–194.

[98] J. E. Griffin, M. Kolossiatis, and M. F.J. Steel. "Comparing distributions by using dependent normalized random-measure mixtures". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75.3 (2013), pp. 499–529.

[99] Michele Guindani, Peter Müller, and Song Zhang. "A Bayesian discovery procedure". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 71.5 (2009), pp. 905–925.

[100] Michele Guindani, Nuno Sepúlveda, Carlos Daniel Paulino, and Peter Müller. "A Bayesian semiparametric approach for the differential analysis of sequence counts data". In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 63.3 (2014), pp. 385–404.

[101] Thomas J. Hardcastle and Krystyna A. Kelly. "BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data". In: *BMC Bioinformatics* 11 (2010).

[102] David I. Hastie, Silvia Liverani, and Sylvia Richardson. "Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations". In: *Statistics and Computing* 25.5 (2015), pp. 1023–1037. arXiv: 1304.1778.

[103] W. K. Hastings. "Monte carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109.

[104] Spyridon J. Hatjispyros, Theodoros Nicoleris, and Stephen G. Walker. "Random density functions with common atoms and pairwise dependence". In: *Computational Statistics and Data Analysis* 101 (2016), pp. 236–249.

[105] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker. *Bayesian nonparametrics*. 2010, pp. 1–299.

[106] Matthew D. Hoffman and Andrew Gelman. "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15 (2014), pp. 1593–1623.

[107] Liang Hong and Ryan Martin. "A flexible Bayesian nonparametric model for predicting future insurance claims". In: (2016), pp. 1–21.

[108] Michael E. Houle. “Dimensionality, Discriminability, Density & Distance Distributions”. In: *ICDMW* (2013).

[109] Lawrence Hubert and Phipps Arabie. “Comparing partitions”. In: *Journal of Classification* 2.1 (1985), pp. 193–218.

[110] Antonio Irpino and Rosanna Verde. “Basic statistics for distributional symbolic variables: a new metric-based approach”. In: *Advances in Data Analysis and Classification* 9.2 (2015), pp. 143–175.

[111] Hemant Ishwaran and Lancelot F. James. “Gibbs sampling methods for stick-breaking priors”. In: *Journal of the American Statistical Association* 96.453 (2001), pp. 161–173.

[112] Hemant Ishwaran and Mahmoud Zarepour. “Exact and approximate sum representations for the Dirichlet process”. In: *Canadian Journal of Statistics* 30.2 (2002), pp. 269–283.

[113] Sonia Jain and Radford M. Neal. “A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model”. In: *Journal of Computational and Graphical Statistics* 13.1 (2004), pp. 158–182.

[114] Sonia Jain and Radford M. Neal. “Splitting and merging components of a nonconjugate Dirichlet process mixture model”. In: *Bayesian Analysis* 2.3 (2005), pp. 445–472.

[115] Sushrut Jangi and J. Thomas Lamont. “Asymptomatic colonization by clostridium difficile in infants: Implications for disease in later life”. In: *Journal of Pediatric Gastroenterology and Nutrition* 51.1 (2010), pp. 2–7.

[116] Valen E. Johnson and David Rossell. “Bayesian model selection in high-dimensional settings.” In: *Journal of the American Statistical Association* 107.498 (2012), pp. 649–660. arXiv: NIHMS150003.

[117] Valen E. Johnson and David Rossell. “On the use of non-local prior densities in Bayesian hypothesis tests”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 72.2 (2010), pp. 143–170.

[118] I. Jolliffe. “Principal component analysis”. In: (2002).

[119] Michael I. Jordan and David M. Blei. “Variational inference for Dirichlet process mixtures”. In: *Bayesian Analysis* 1.1 (2006), pp. 121–143.

[120] Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O’Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L Mason, Karen L. Madsen, and Gane K.-S. Wong. “Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics”. In: *Frontiers in Microbiology* 7 (2016), p. 459.

[121] Maria Kalli, Jim E. Griffin, and Stephen G. Walker. “Slice sampling mixture models”. In: *Statistics and Computing* 21.1 (2011), pp. 93–105.

[122] E. L. Kaplan and Paul Meier. “Nonparametric Estimation from Incomplete Observations”. In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481.

[123] Robert E. Kass and Adrian E. Raftery. “Bayes Factors”. In: *Journal of the American Statistical Association* 90.430 (1995). arXiv: 1099-1255.

[124] Andrew L. Kau, Philip P. Ahern, Nicholas W. Griffin, Andrew L. Goodman, and Jeffrey I. Gordon. “Human nutrition, the gut microbiome and the immune system”. In: *Nature* 474.7351 (2011), pp. 327–336.

[125] Leonard Kaufman and Peter J Rousseeuw. “Clustering by Means of Mediods BT - Statistical Data Analysis based on the L1 Norm”. In: *Statistical Data Analysis based on the L1 Norm* (1987), pp. 405–416.

[126] Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D. Peddada. “Zeros in microbiome data”. In: (2017).

[127] Sinae Kim, David B. Dahly, and Marina Vannucci. "Spiked dirichlet process prior for bayesian multiple hypothesis testing in random effsects models". In: *Bayesian Analysis* 4.4 (2009), pp. 707–732.

[128] Sinae Kim, Mahlet G. Tadesse, and Marina Vannucci. "Variable selection in clustering via Dirichlet process mixture models." In: *Biometrika* 93.4 (2006), pp. 877–893.

[129] J. F. C. Kingman. *Poisson Processes.* Vol. 3. 1992.

[130] M. Kolossiatis, J. E. Griffin, and M. F.J. Steel. "On Bayesian nonparametric modelling of two correlated distributions". In: *Statistics and Computing* 23.1 (2013), pp. 1–15.

[131] Athanasios Kottas and Gilbert W. Fellingham. "Bayesian semiparametric modeling and inference with mixtures of symmetric distributions". In: *Statistics and Computing* 22.1 (2012), pp. 93–106.

[132] Alex Kulesza and Ben Taskar. "Determinantal point processes for machine learning". In: (2012), pp. 1–120. arXiv: 1207.6083.

[133] Ruiting Lan and Peter R. Reeves. "Escherichia coli in disguise: Molecular origins of Shigella". In: *Microbes and Infection* 4.11 (2002), pp. 1125–1132.

[134] John W. Lau and Peter J. Green. "Bayesian model-based clustering procedures". In: *Journal of Computational and Graphical Statistics* 16.3 (2007), pp. 526–558.

[135] Frédéric Lavancier, Jesper Møller, and Ege Rubak. "Determinantal point process models and statistical inference". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 77.4 (2015), pp. 853–877. arXiv: 1205.4818.

[136] Kristin P. Lennox, David B. Dahl, Marina Vannucci, Ryan Day, and Jerry W. Tsai. "A Dirichlet process mixture of hidden Markov models for protein structure prediction". In: *Annals of Applied Statistics* 4.2 (2010), pp. 916–942. arXiv: 1011.2065.

[137] Ruth E. Ley. "Obesity and the human microbiome". In: *Current Opinion in Gastroenterology* 26.1 (2010), pp. 5–11.

[138] Didong Li and David B. Dunson. "Classification via local manifold approximation". In: (2019). arXiv: 1903.00985.

[139] Didong Li and David B. Dunson. "Geodesic Distance Estimation with Spherelets". In: (2019). arXiv: 1907.00296.

[140] Didong Li, Minerva Mukhopadhyay, and David B. Dunson. "Efficient Manifold and Subspace Approximations with Spherelets". In: (2017). arXiv: 1706.08263.

[141] J. G. Liao, Yong Lin, Zachariah E. Selvanayagam, and Weichung Joe Shih. "A mixture model for estimating the local false discovery rate in DNA microarray analysis". In: *Bioinformatics* 20.16 (2004), pp. 2694–2701.

[142] Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. "Controlling the reinforcement in Bayesian non-parametric mixture models". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 69.4 (2007), pp. 715–740.

[143] Antonio Lijoi, Bernardo Nipoti, and Igor Prünster. "Dependent mixture models: Clustering and borrowing information". In: *Computational Statistics and Data Analysis* 71 (2014), pp. 417–433.

[144] D. V. Lindley and M. J. Schervish. "Theory of Statistics." In: *The Statistician* 45.4 (2006), p. 536.

[145] J. K. Lindsey. "Comparison of Probability Distributions". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.1 (1974), pp. 38–47.

[146] J. K. Lindsey. "Construction and Comparison of Statistical Models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.3 (1974), pp. 418–425.

[147] Jun S. Liu. "The collapsed gibbs sampler in Bayesian computations with applications to a gene regulation problem". In: *Journal of the American Statistical Association* 89.427 (1994), pp. 958–966.

[148] Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. "Multivariate analysis by data depth: Descriptive statistics, graphics and inference". In: *Annals of Statistics* 27.3 (1999), pp. 783–858.

[149] Albert Y. Lo. "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates". In: *The Annals of Statistics* 12.1 (1984), pp. 351–357.

[150] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (2014), p. 550. arXiv: arXiv:1303.3997v2.

[151] Steven N. MacEachern. "Dependent dirichlet processes". In: *Manuscript* (2000).

[152] Steven N. MacEachern. "Dependent nonparametric processes". In: *ASA proceedings of the section on bayesian statistical science* (1999), pp. 50–55.

[153] Steven N. MacEachern. "Estimating normal means with a conjugate style dirichlet process prior". In: *Communications in Statistics - Simulation and Computation* 23.3 (1994), pp. 727–741.

[154] Steven N. Maceachern and Peter Müller. "Estimating mixture of dirichlet process models". In: *Journal of Computational and Graphical Statistics* 7.2 (1998), pp. 223–238.

[155] Jialiang Mao, Yuhan Chen, and Li Ma. "Bayesian graphical compositional regression for microbiome data". In: (2017). arXiv: 1712.04723.

[156] Ryan Martin and Surya T. Tokdar. "A nonparametric empirical Bayes framework for large-scale multiple testing," in: *Biostatistics* 13.3 (2012), pp. 427–439. arXiv: 1106.3885.

[157] D. McFadden. "Modeling the Choice of Residential Location". In: *Transportation Research Record* 672 (1978), pp. 75–96.

[158] M. Meilă. "Comparing clusterings — an information based distance". In: *Journal of Multivariate Analysis* 98 (2007), pp. 873–895.

[159] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. "Equation of state calculations by fast computing machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[160] D. Moltchanov. "Distance distributions in random networks". In: *Ad Hoc Networks* 10.6 (2012), pp. 1146–1166.

[161] Xochitl C. Morgan and Curtis Huttenhower. "Chapter 12: Human Microbiome Analysis". In: *PLoS Computational Biology* 8.12 (2012). arXiv: arXiv:1011.1669v3.

[162] P. Mueller and Abel Rodríguez. "Nonparametric Bayesian Inference". In: *NSF-CBMS Regional Conference Series in Probability and Statistics* (2013).

[163] Minerva Mukhopadhyay, Didong Li, and David B. Dunson. "Estimating densities with nonlinear support using Fisher-Gaussian kernels". In: (2019). arXiv: 1907.05918.

[164] Pietro Muliere and Luca Tardella. "Approximating distributions of random functionals of Ferguson-Dirichlet priors". In: *Canadian Journal of Statistics* 26.2 (1998), pp. 283–297.

[165] Patrick Muller, Giovanni Parmigiani, and Kenneth Rice. "FDR and Bayesian multiple comparisons rules". In: *Bayesian Statistics 8*. Ed. by José M Bernardo, M J Bayarri, James O Berger, A P Dawid, D Heckerman, Mike West, and Adrian F M Smith. Vol. 0. 1995. Oxford University Press, 2006, pp. 349–370.

[166] Peter Müller, Fernando Quintana, and Gary Rosner. "A method for combining inference across related nonparametric Bayesian models". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 66.3 (2004), pp. 735–749.

[167] Peter Müller, Alaattin Erkanli, and Mike West. "Bayesian curve fitting using multivariate normal mixtures". In: *Biometrika* 83.1 (1996), pp. 67–79.

[168] Peter Müller, Fernando Andrés Quintana, Alejandro Jara, and Tim Hanson. *Bayesian Nonparametric Data Analysis*. 2015.

[169] Omkar Muralidharan. "An empirical Bayes mixture method for effect size and false discovery rate estimation". In: *Annals of Applied Statistics* 6.1 (2012), pp. 422–438. arXiv: arXiv:1010.1425v1.

[170] Radford M. Neal. "Dirichlet Process Mixture Models". In: *Journal of Computational and Graphical Statistics* 9.2 (2000), pp. 249–265.

[171] Radford M. Neal. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models". In: *Journal of Computational and Graphical Statistics* 9.2 (2000), pp. 249–265. arXiv: arXiv:1011.1669v3.

[172] Radford M. Neal. "Slice sampling (with discussion)". In: *The Annals of Statistics* 31.3 (2003), pp. 705–767. arXiv: 1003.3201v1.

[173] Michael A. Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. "Detecting differential gene expression with a semiparametric hierarchical mixture method". In: *Biostatistics* 5.2 (2004), pp. 155–176.

[174] A. O'Hagan and Tom Leonard. "Bayes estimation subject to uncertainty about parameter constraints". In: *Biometrika* 63.1 (1976), pp. 201–203.

[175] John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. "Nested hierarchical dirichlet processes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (2015), pp. 256–270.

[176] Wei Pan, Jizhen Lin, and Chap T. Le. "A mixture model approach to detecting differentially expressed genes with microarray data". In: *Functional and Integrative Genomics* 3.3 (2003), pp. 117–124.

[177] Omiros Papaspiliopoulos. "A note on posterior sampling from Dirichlet mixture models". In: 8 (2008), pp. 1–8.

[178] Omiros Papaspiliopoulos and Gareth O. Roberts. "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models". In: *Biometrika* 95.1 (2008), pp. 169–186. arXiv: 0710.4228.

[179] Francesca Petralia, Vinayak Rao, and David B. Dunson. "Repulsive Mixtures". In: (2012), pp. 1–9. arXiv: 1204.5243.

[180] Karl W. Pettis, Thomas A. Bailey, Anil K. Jain, and Richard C. Dubes. "An Intrinsic Dimensionality Estimator from Near-Neighbor Information". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.1 (1979), pp. 25–37.

[181] Jim Pitman. "Combinatorial stochastic processes". In: *Lecture Notes for St. Flour Summer School. Springer-Verlag* (2002).

[182] Jim Pitman. "Exchangeable and partially exchangeable random partitions". In: *Probability Theory and Related Fields* 102.2 (1995), pp. 145–158.

[183] Jim Pitman. "Some developments of the Blackwell-MacQueen urn scheme". In: *Statistics, Probability, and Game Theory: Papers in honor of David Blackwell*. Ed. by Thomas S Ferguson, Lloyd S Shapley, and James B MacQueen. IMS Lecture Notes-Monograph Series, 1996, pp. 245–267.

[184] Jim Pitman and Marc Yor. "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator". In: *Annals of Probability* 25.2 (1997), pp. 855–900.

[185] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. "CODA: convergence diagnosis and output analysis for MCMC". In: *R News* 6.1 (2006), pp. 7–11.

[186] Nicholas G. Polson, James G. Scott, and Jesse Windle. "Bayesian inference for logistic models using Pólya-Gamma latent variables". In: *Journal of the American Statistical Association* 108.504 (2013), pp. 1339–1349.

[187] Mihai Pop, Alan W. Walker, Joseph Paulson, Brianna Lindsay, Martin Antonio, M. Anowar Hossain, Joseph Oundo, Boubou Tamboura, Volker Mai, Irina Astrovskaya, Hector Corrada Bravo, Richard Rance, Mark Stares, Myron M. Levine, Sandra Panchalingam, Karen Kotloff, Usman N. Ikumapayi, Chinelo Ebruke, Mitchell Adeyemi, Dilruba Ahmed, Firoz Ahmed, Meer Taifur Alam, Ruhul Amin, Sabbir Siddiqui, John B. Ochieng, Emmanuel Ouma, Jane Juma, Euince Mailu, Richard Omore, J. Glenn Morris, Robert F. Breiman, Debasish Saha, Julian Parkhill, James P. Nataro, and O. Colin Stine. "Diarrhea in young children from low-income countries leads to large-scale alterations in wintestinal microbiota composition". In: *Genome Biology* 15.6 (2014), R76.

[188] Ian Porteous, Alex Ihler, Padhraic Smyth, and Max Welling. "Gibbs sampling for (coupled) infinite mixture models in the stick-breaking representation". In: *Proceedings of UAI* 22.4 (2006), pp. 385–392. arXiv: 1206.6845.

[189] Stan Pounds and Stephan W. Morris. "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values". In: *Bioinformatics* 19.10 (2003), pp. 1236–1242.

[190] Adrian E. Raftery, Michael A. Newton, Jaya M. Satagopan, and Pavel N. Krivitsky. "Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity". In: *Bayesian Statistics* 8 (2007), pp. 1–45.

[191] J. Ramsay and B.W. Silverman. *Functional Data Analysis*. 2005.

[192] William M. Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical Association* 66.336 (1971), pp. 846–850.

[193] Brian J. Reich and Montserrat Fuentes. "A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields". In: *The Annals of Applied Statistics* 1.1 (2007), pp. 249–264.

[194] A. P. Reynolds, G. Richards, B. De La Iglesia, and V. J. Rayward-Smith. "Clustering rules: A comparison of partitioning and hierarchical clustering algorithms". In: *Journal of Mathematical Modelling and Algorithms* 5.4 (2006), pp. 475–504.

[195] Tommaso Rigon and Daniele Durante. "Tractable Bayesian Density Regression via Logit Stick-Breaking Priors". In: (2017). arXiv: 1701.02969.

[196] J. Riihimäki and A. Vehtari. "Gaussian processes with monotonicity information". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 9 (2010), pp. 645–652.

[197] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. "{limma} powers differential expression analyses for {RNA}-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47.

[198] Christian Robert, Ronald Christensen, Wesley Johnson, Adam Branscum, and Timothy Hanson. "Bayesian Ideas and Data Analysis". In: *Chance* 25.2 (2015), pp. 58–60.

[199] Christian P. Robert. "Multimodality and label switching : a discussion". In: *Workshop on mixtures, ICMS* (2010).

[200] Christian P. Roberts. *The Bayesian Choice*. 2017, pp. 33–65. arXiv: arXiv:1011.1669v3.

[201] Gareth O. Roberts and Jeffrey S. Rosenthal. "Coupling and Ergodicity of Adaptive MCMC". In: *Journal of Applied Probability* 44.2 (2007), pp. 458–475.

[202] Gareth O. Roberts and Jeffrey S. Rosenthal. "Examples of adaptive MCMC". In: *Journal of Computational and Graphical Statistics* 18.2 (2009), pp. 349–367.

[203] Mark D. Robinson and Gordon K. Smyth. "Moderated statistical tests for assessing differences in tag abundance". In: *Bioinformatics* 23.21 (2007), pp. 2881–2887.

[204] Abel Rodriguez and David B. Dunson. "Functional clustering in nested designs: Modeling variability in reproductive epidemiology studies". In: *Annals of Applied Statistics* 8.3 (2014), pp. 1416–1442.

[205] Abel Rodríguez and David B. Dunson. "Nonparametric Bayesian models through probit stick-breaking processes". In: *Bayesian Analysis* 6.1 (2011), pp. 145–178.

[206] Abel Rodríguez, David B. Dunson, and Alan E. Gelfand. "The nested dirichlet process". In: *Journal of the American Statistical Association* 103.483 (2008), pp. 1131–1144.

[207] Carlos E. Rodríguez and Stephen G. Walker. "Label switching in Bayesian mixture models: Deterministic relabeling strategies". In: *Journal of Computational and Graphical Statistics* 23.1 (2014), pp. 25–45.

[208] Peter J. Rosseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[209] David Rossell and Donatello Telesca. "Nonlocal Priors for High-Dimensional Estimation". In: *Journal of the American Statistical Association* 112.517 (2017), pp. 254–265. arXiv: 1402.5107.

[210] David Rossell, Donatello Telesca, and Valen E. Johnson. "High-dimensional bayesian classifiers using non-local priors". In: *Studies in Classification, Data Analysis, and Knowledge Organization* (2013), pp. 305–313.

[211] Richard Saucier. *Computer Generation of Statistical Distributions*. March. 2000.

[212] R. Schmitz, G. W. Wright, D. W. Huang, C. A. Johnson, J. D. Phelan, J. Q. Wang, S. Roulland, M. Kasbekar, R. M. Young, A. L. Shaffer, D. J. Hodson, W. Xiao, X. Yu, Y. Yang, H. Zhao, W. Xu, X. Liu, B. Zhou, W. Du, W. C. Chan, E. S. Jaffe, R. D. Gascoyne, J. M. Connors, E. Campo, A. Lopez-Guillermo, A. Rosenwald, G. Ott, J. Delabie, L. M. Rimsza, K. Tay Kuang Wei, A. D. Zelenetz, J. P. Leonard, N. L. Bartlett, B. Tran, J. Shetty, Y. Zhao, D. R. Soppet, S. Pittaluga, W. H. Wilson, and L. M. Staudt. "Genetics and pathogenesis of diffuse large B-Cell lymphoma". In: *New England Journal of Medicine* 378.15 (2018), pp. 1396–1407.

[213] A. J. Sethuraman. "A Constructive Definition of Dirichlet Priors". In: *Statistica Sinica* 4 (1994), pp. 639–650.

[214] Babak Shahbaba and Wesley O. Johnson. "Bayesian nonparametric variable selection as an exploratory tool for discovering differentially expressed genes". In: *Statistics in Medicine* 32.12 (2013), pp. 2114–2126.

[215] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[216] Thomas J. Sharpton. "An introduction to the analysis of shotgun metagenomic data". In: *Frontiers in Plant Science* 5.June (2014), p. 209. arXiv: 15334406.

[217] Guiling Shi, Chae Young Lim, and Tapabrata Maiti. "Model selection using mass-nonlocal prior". In: *Statistics and Probability Letters* 147 (2019), pp. 36–44.

[218] Pixu Shi, Anru Zhang, and Hongzhe Li. "Regression analysis for microbiome compositional data". In: *Annals of Applied Statistics* 10.2 (2016), pp. 1019–1040. arXiv: 1603.00974.

[219] Minsuk Shin, Anirban Bhattacharya, and Valen E. Johnson. "Scalable Bayesian Variable Selection Using Nonlocal Prior Densities in Ultrahigh-Dimensional Settings". In: *Statistica Sinica* (2015). arXiv: 1507.07106.

[220] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, and William R. Sellers. "Gene expression correlates of clinical prostate cancer behavior". In: *Cancer Cell* 1.2 (2002), pp. 203–209.

[221] Christopher G. Small. "A Survey of Multidimensional Medians". In: *International Statistical Review / Revue Internationale de Statistique* 58.3 (1990), p. 263.

[222] Leonard A. Smith. "Intrinsic limits on dimension calculations". In: *Physics Letters A* 133.6 (1988), pp. 283–288.

[223] Arne Smits and Wolfgang Huber. *DEP: Differential Enrichment analysis of Proteomics data*. 2017.

[224] Gordon K. Smyth. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004).

[225] M. Sperrin, T. Jaki, and E. Wit. "Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models". In: *Statistics and Computing* 20.3 (2010), pp. 357–366.

[226] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 64.4 (2002), pp. 583–616.

[227] Stan Development Team. *{RStan}: the {R} interface to {Stan}*. 2019.

[228] John D. Storey. "The optimal discovery procedure: A new approach to simultaneous significance testing". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 69.3 (2007), pp. 347–368.

[229] Anja Struyf, Mia Hubert, and Peter J. Rousseeuw. "Clustering in an object-oriented environment". In: *Journal of Statistical Software* 1 (1996), pp. 1–30.

[230] Erik B. Sudderth and Michael I. Jordan. "Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes". In: *Advances in Neural Information Processing Systems 21 (NIPS 2008)*. 2009, pp. 1585–1592.

[231] Wenguang Sun and T. Tony Cai. "Oracle and adaptive compound decision rules for false discovery rate control". In: *Journal of the American Statistical Association* 102.479 (2007), pp. 901–912.

[232] Alexander Swidsinski, Vera Loening-Baucke, Hans Verstraelen, Sylwia Osowska, and Yvonne Doerffel. "Biostructure of Fecal Microbiota in Healthy Subjects and Patients With Chronic Idiopathic Diarrhea". In: *Gastroenterology* 135.2 (2008).

[233] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Hierarchical Dirichlet processes". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581. arXiv: arXiv:1210.6738v2.

[234] Lavanya Sita Tekumalla, Priyanka Agrawal, and Indrajit Bhattacharya. "Nested Hierarchical Dirichlet Processes for Multi-Level Non-Parametric Admixture Modeling". In: (2015). arXiv: 1508.06446.

[235] Ville Tolvanen. *Gaussian Processes with Monotonicity Constraint for Big Data*. Aalto University - School of Electrical Engineering, 2014.

[236] R. D. Tuddenham and M. M. Snyder. "Physical growth of California boys and girls from birth to eighteen years." In: *Publications in child development. University of California, Berkeley* 1.2 (1954), pp. 183–364.

[237] J. W. Tukey. "Mathematics and the picturing of data". In: *Proceedings of the international congress of mathematicians* (1975), pp. 523–532.

[238] Sara Wade and Zoubin Ghahramani. "Bayesian cluster analysis: Point estimation and credible balls". In: *arXiv:stat.MEsta* (2015), pp. 1–33. arXiv: 1505.03339.

[239] Stephen G. Walker. "Sampling the Dirichlet mixture model with slices". In: *Communications in Statistics: Simulation and Computation* 36.1 (2007), pp. 45–54.

[240] Sumio Watanabe. "A Widely Applicable Bayesian Information Criterion". In: *Journal of Machine Learning Research* 14 (2013), pp. 867–897.

[241] R. H. Whittaker. "Vegetation of the Siskiyou Mountains, Oregon and California". In: *Ecological Monographs* 30.3 (2006), pp. 279–338. arXiv: /www.jstor.org/stable/1943563 [https:].

[242] Daniela M. Witten. "Classification and clustering of sequencing data using a poisson model". In: *Annals of Applied Statistics* 5.4 (2011), pp. 2493–2518. arXiv: 1202.6201.

[243] Fangzheng Xie and Yanxun Xu. "Bayesian Repulsive Gaussian Mixture Model". In: (2017). arXiv: 1703.09061.

[244] Yanxun Xu, Peter Müller, and Donatello Telesca. "Bayesian inference for latent biologic structure with determinantal point processes (DPP)". In: *Biometrics* 72.3 (2016), pp. 955–964. arXiv: 1506.08253.

[245] Natalya Yutin and Michael Y. Galperin. "A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia." In: *Environmental microbiology* 15.10 (2013), pp. 2631–41. arXiv: NIHMS150003.

[246] Xiaofei Zhang, Arne H. Smits, Gabrielle B.A. van Tilburg, Pascal W.T.C. Jansen, Matthew M. Makowski, Huib Ovaa, and Michiel Vermeulen. "An Interaction Landscape of Ubiquitin Signaling". In: *Molecular Cell* 65.5 (2017), 941–955.e8.

[247] Bin Zhu, Allison E. Ashley-Koch, and David B. Dunson. "Generalized Admixture Mapping for Complex Traits". In: *Genes|Genomes|Genetics* 3.7 (2013), pp. 1165–1175. arXiv: arXiv:1111.5551v1.

[248] Daiane Aparecida Zuanetti, Peter Müller, Yitan Zhu, Shengjie Yang, and Yuan Ji. "Clustering distributions with the marginalized nested Dirichlet process". In: *Biometrics* 74.2 (2018), pp. 584–594.