Scuola di Dottorato
Università degli Studi di Milano-Bicocca

Department of

Economics, Management, and Statistics

Ph.D. program: **Economics and Statistics** Cycle: **XXXV°**
Curriculum: **Statistics**

Innovative approaches to Bayesian Clustering Methods: parametric and nonparametric perspectives

Surname: **Giampino**
Name: **Alice**
Registration number: **790347**

Supervisor: Prof. **Bernardo Nipoti**
Co-Supervisor: Prof. **Michele Guindani**
Tutor: Prof. **Sonia Migliorati**

Academic Year: 2022-2023

Abstract

Throughout this thesis, we embrace the Bayesian mixture models setting, harnessing their flexibility and adaptability to address a range of challenging research questions concerning data clustering. This manuscript is a collection of three projects. The first two projects are dedicated to the use of Bayesian Nonparametric (BNP) methods. In the concluding project, we focus on a parametric setting, and provide a novel methodological framework to investigate specific research inquiries arising in topic modeling.

The manuscript addresses distinct research questions, each approached through the lens of Bayesian methodology. In Chapter 2, we tackle the challenge of simultaneous clustering of users and items within datasets riddled with missing information, a common occurrence in data from social platforms. We propose an innovative co-clustering method that accommodates informative censoring, providing a robust solution for handling missing data and extracting valuable insights. Chapter 3 shifts the focus towards modeling the evolution of data partitions over time, developing a model for changepoint detection using time-varying random partition models. The proposed approach builds upon the principles of dynamic linear models in time series, extending them by incorporating latent state equations that model the evolution of partitions of units over time. In the final chapter, we introduce a novel model based on distributions defined on the simplex to address the intriguing question of whether such distributions can capture various forms of dependence among topics in a corpus of textual documents. Our investigation leads us to the definition of a model characterized by positive correlation across topics, highlighting the versatility and applicability of simplex-based distributions in modeling complicated relationships within textual datasets.

In summary, this thesis aims at providing a thoughtful perspective on Bayesian mixture models and their applications, while also presenting innovative solutions to various research questions, demonstrating the breadth and depth of Bayesian methodology in tackling complex data analysis problems.

To find truth,
one must travel through a dense fog.

David Dweck

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the last decade, the amount of data that scientists and economic activities have at their disposal has increased dramatically. Such data is often unstructured and *heterogeneous*. For this reason, the necessity to develop new, efficient, and flexible models has emerged. In this context, mixture models have gained prominence as a powerful and versatile framework in statistical modeling, widely employed across various fields to address the challenges including cluster analysis, discriminant analysis, survival analysis (Peel and MacLahlan, 2000). Mixture models find extensive applications in different domains, such as image segmentation, clustering, anomaly detection, biological data modeling and finance for modeling asset returns. The ability to capture complex data structures and the adaptability of mixture models make them very successful and an indispensable tool in data analysis and modeling. These models are based on the intuitive notion that observed data often originates from a mixture of underlying probability distributions, each reflecting distinct data sources, latent classes, or components. The direct benefit of their flexibility allows to capture a wide range of data characteristics, from multimodality to heterogeneity (McLachlan and Basford, 1988).

1.1 Mixture models

Mixture models stand as a cornerstone in the realm of statistical modeling since 19^{th} century. Newcomb (1886) showed that a two-component mixture can be used to combine observations to detect outliers, especially when there is heterogeneity in the data.

From a mathematical point of view, suppose that $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space, where Ω is a sample space, $\mathcal{A}$ is a σ-algebra on Ω, and $\mathbb{P}$ is a probability measure defined on $\mathcal{A}$. Let $\mathbf{X} : \Omega \to \mathcal{X}$ be a random vector defined on Ω and taking values in some space $\mathcal{X} \subseteq \mathbb{R}^d$.

Then, for any $\omega \in \Omega$, $\mathbf{x} = \mathbf{X}(\omega)$ is a realization of $\mathbf{X}$. We will refer to a collection of realizations of $\mathbf{X}$ as observations or data.

Definition 1.1.1 (Finite Mixture distribution)**.** $\mathbf{X}$ has a finite mixture distribution if its probability density function (pdf) has the form

$$p(\mathbf{x}) = \zeta_1 f_1(\mathbf{x}) + \cdots + \zeta_K f_K(\mathbf{x}), \text{ for } \mathbf{x} \in \mathcal{X} \tag{1.1}$$

where

$$\zeta_k > 0, \text{ for } k = 1, \ldots, K, \text{ and } \zeta_1 + \cdots + \zeta_K = 1$$

and

$$f_k(\cdot) \geq 0, \quad \int_{\mathcal{X}} f_k(\mathbf{x}) d\mathbf{x} = 1, \text{ for } k = 1, \ldots, K.$$

The parameters $\zeta_1, \ldots, \zeta_K$ are the mixing weights and $f_1(\cdot), \ldots, f_K(\cdot)$ are the component densities of the mixture (see, e.g., Gelman et al., 2013).

A more explicit representation $p(\cdot)$ can be obtained if $f_1(\cdot), \ldots, f_K(\cdot)$ have specified parametric forms. In this case, we can rewrite (1.1) as

$$p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\zeta}) = \zeta_1 f_1(\mathbf{x}|\boldsymbol{\theta}_1) + \cdots + \zeta_K f_K(\mathbf{x}|\boldsymbol{\theta}_K), \text{ for } \mathbf{x} \in \mathcal{X}, \tag{1.2}$$

where $\boldsymbol{\theta}_k$ denotes the vector of the parameters of $f_k(\cdot)$, ζ_k the k-th mixing weight, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_K)$ denote the vectors of all the parameters of the mixture model. Mixture models are flexible as it is not required that all components belong to the same parametric family. However, the majority of applications in the literature focus, for simplicity, on components from the same parametric family of distributions. In the latter situation, the finite mixture density can be defined as follows

$$p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\zeta}) = \zeta_1 f(\mathbf{x}|\boldsymbol{\theta}_1) + \cdots + \zeta_K f(\mathbf{x}|\boldsymbol{\theta}_K) = \sum_{k=1}^{K} \zeta_k f(\mathbf{x}|\boldsymbol{\theta}_k), \quad \mathbf{x} \in \mathcal{X} \tag{1.3}$$

where $f(\cdot|\boldsymbol{\theta}_k)$ denotes a generic member of the parametric family and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ belong to the same parameter space Θ.

One of the the most notable and popular examples of mixture distributions is defined as mixture of (d-dimensional) Gaussian distributions, which is usually referred to as Gaussian mixture or normal mixture, indicated as Gaussian mixture model (GMM, see, e.g., Peel

and MacLahlan, 2000). In this case, the k-th component density has the form

$$\phi_d\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \frac{1}{|2\pi\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)' \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)\right\}, \text{ for } \mathbf{x} \in \mathbb{R}^d.$$

Hence, in general the density function of a Gaussian mixture has the form

$$p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\zeta}) = \zeta_1 \phi_d\left(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) + \cdots + \zeta_K \phi_d\left(\mathbf{x}|\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K\right), \text{ for } \mathbf{x} \in \mathbb{R}^d$$

with $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_K)$ and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$ represent, respectively, the vectors of the means and the variance-covariance matrices of the Gaussian components of the mixture. Figure 1.1 illustrates a simple example of a univariate Gaussian mixture model obtained by assigning equal weight to three components. This visualization helps to grasp how the mixture components explain the underlying data distribution.

Mixture models offer a robust framework for modeling data distributions characterized by multiple components. These models, based on the principle of combining simpler probability distributions, enable us to represent and understand complex data sources, and have become a valuable tool in various fields. The mathematical formulation, as exemplified by the GMM in Figure 1.1, provides a clear structure for modeling the data, and visualizations aid in comprehending how these components contribute to the overall distribution.

Once the components have been defined, it is possible to incorporate mixtures within a Bayesian framework. This involves specifying probability distributions for the component parameters and the component weights. The mixture distribution, characterized by probabilities $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_K)$ is a representation of the variability in $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ across the population of interest. At this stage, we are assuming that the number of mixture components, denoted as K, is known and fixed. If we have observations $\mathbf{x}$ for which their mixture components are known, the mixture model in (1.2) can be readily adapted by incorporating the prior distributions for $\boldsymbol{\zeta}$ and $\boldsymbol{\theta}$, that is

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\zeta}) &= \zeta_1 f_1\left(\mathbf{x}|\boldsymbol{\theta}_1\right) + \cdots + \zeta_K f_K\left(\mathbf{x}|\boldsymbol{\theta}_K\right), \text{ for } \mathbf{x} \in \mathcal{X}, \\ \boldsymbol{\theta}|\boldsymbol{\zeta} &\sim G, \\ (\zeta_1, \ldots, \zeta_K) &\sim \mathcal{D}_K, \end{aligned} \tag{1.4}$$

where G is a discrete distribution with support on the K locations of the components, and $\mathcal{D}_K$ is a discrete prior distribution, defined on the K-dimensional simplex, for the K weights. At the heart of this framework lies the concept of exchangeability, a fundamental

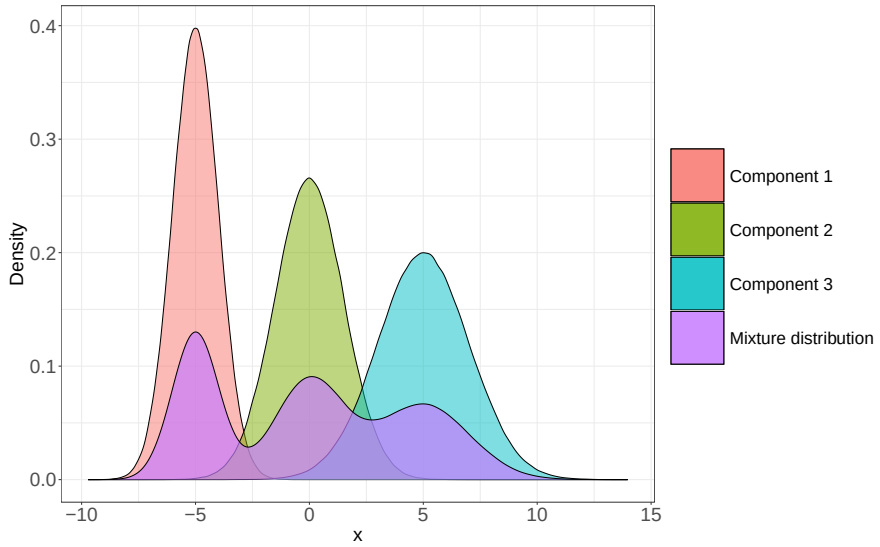


Figure 1.1: *A visual representation of a univariate Gaussian mixture model with three components, here arbitrarily denoted as components 1, 2 and 3.*

principle in Bayesian statistics.

1.1.1 Exchangeability

Exchangeability, thoroughly studied by de Finetti (1937), represents a foundational notion in probability theory. Exchangeability is a way to formalize an assumption of *homogeneity* of the data. An ideally infinite sequence of observations $X_1, X_2, \ldots$, defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in $(\mathbb{X}_0, \mathcal{X}_0)$, is said to be exchangeable if the order in which the observations are recorded is irrelevant as far as their joint distribution is concerned. More formally,

Definition 1.1.2. A sequence of observations $\{X_n\}_{n\geq 1}$ is exchangeable if and only if $(X_1, \ldots, X_n)$ converges in distribution to $\left(X_{\sigma(1)}, \ldots, X_{\sigma(n)}\right)$, that is

$$(X_1, \ldots, X_n) \stackrel{d}{=} \left(X_{\sigma(1)}, \ldots, X_{\sigma(n)}\right) \tag{1.5}$$

where σ is any permutation of the set $\{1, \ldots, n\}$, for any $n \geq 1$.

That is, the exchangeability of an infinite sequence of observations is defined by means of the invariance of its distribution under finite permutations of the observations themselves.

We can rewrite (1.5) as

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \mathbb{P}\left(X_{\sigma(1)} \in A_1, \ldots, X_{\sigma(n)} \in A_n\right) \tag{1.6}$$

for any $A_1, \ldots, A_n \in \mathcal{X}_0$, for any $n \geq 1$, and for any permutation σ of $\{1, \ldots, n\}$. Let $\tilde{p} : (\Omega, \mathcal{A}) \to (\mathrm{P}, \mathcal{P})$ be a random probability measure on $(\mathbb{X}_0, \mathcal{X}_0)$.

Before stating the de Finetti's representation theorem, it is convenient to recall the definition of a Polish space.

Definition 1.1.3 (Polish space)**.** A space $\mathbb{X}_0$ is a Polish space if it is a separable topological space whose topology is metrisable by a complete metric.

Theorem 1.1.1. *(de Finetti, 1937) If $\mathbb{X}_0$ is a Polish space, the following conditions are equivalent:*

i) $\{X_n\}_{n\geq 1}$ *is a sequence of exchangeable observations;*

ii) there exists a random probability measure $\tilde{p}$ on $(\mathbb{X}_0, \mathcal{X}_0)$, such that $\{X_n\}_{n\geq 1}$ are conditionally i.i.d. given $\tilde{p}$;

iii) there exists a probability measure G on $(\mathrm{P}, \mathcal{P})$ such that

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \int_{\mathbf{P}} \prod_{i=1}^{n} p\left(A_i\right) G(\mathrm{d}p)$$

for any $A_1, \ldots, A_n \in \mathcal{X}_0$ and $n \geq 1$.

Bruno de Finetti, in his celebrated representation theorem (Theorem 1.1.1) states that a sequence of observations is exchangeable if and only if its distribution is a mixture of laws of a sequence of independent and identically distributed (i.i.d.) random variables. If $\tilde{p} \sim G$ is a random probability measure with distribution G, the measure G is uniquely defined on $(\mathrm{P}, \mathcal{P})$ by the distribution $\rho = \mathbb{P} \circ X^{-1}$ of $X : G$ is called the de Finetti measure of the sequence $\{X_n\}_{n\geq 1}$, and in the context of Bayesian statistics plays the role of prior distribution. Indeed, when observations are exchangeable, their distribution can be represented in a hierarchical form as

$$\begin{aligned} X_n \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \quad n \geq 1 \\ \tilde{p} &\sim G, \end{aligned} \tag{1.7}$$

where G is the prior distribution for $\tilde{p}$. Thus, by virtue of Theorem 1.1.1, the use of priors in Bayesian methods is justified from a theoretical standpoint. Furthermore, if G has support on some finite-dimensional subspace of P, as for example the space of distribution on $\mathbb{X}_0$ with a specific parametric form, then the model is called parametric. On the other hands, if G has an infinite-dimensional support, then model in (1.7) is termed nonparametric.

1.2 Dirichlet Process

One of the cornerstones of Bayesian nonparametric is the Dirichlet process (DP). The framework introduced by de Finetti in the 1930s encompassed the nonparametric scenario. However, it lacked a tractable nonparametric distribution G. The Dirichlet process emerged as the pioneering tractable nonparametric distribution in the academic literature (Ferguson, 1973). Hence, the connection with mixture models is straightforward. The Dirichlet process offers a more flexible and powerful framework for modeling data when the number of components is uncertain or potentially infinite, making it a valuable tool in Bayesian nonparametric modeling.

In the framework of finite mixture models, one assumes that there are $K \leq n$ subpopulations that compose the total population, with K known and fixed, corresponding to the number of components in the mixture as in Equation 1.4. The components can have a common parametric family with the k-th component depending on a specific parameter vector $\boldsymbol{\theta}_k$. In this setting, exchangeability implies that the data points can be assigned to the components in any order without affecting the underlying model's structure. In the context of DP-based mixture models, this means that we can keep adding new components as more data points arrive, and the model can adapt to the data distribution. Thanks to the flexibility of the DP, the number of components K is not fixed a priori and becomes subject of inference. In order to introduce this process, it is useful to start from the Dirichlet distribution and its properties.

1.2.1 Dirichlet distribution

The Dirichlet distribution, named after the mathematician Peter Gustav Lejeune Dirichlet, is the multivariate generalization of the Beta distribution. It is usually denoted as $\text{Dir}(\alpha)$ where α is the concentration parameter with components $\alpha_i > 0$ for $i = 1, \ldots, n$. Thanks to its analytical properties, the Dirichlet is the most commonly adopted prior distribution for parameters defined on the simplex, used in Bayesian statistics. Before defining the Dirichlet

distribution, it is convenient to recall the Beta and the Gamma distributions.

Definition 1.2.1 (Beta distribution)**.** The random variable $Y \sim \text{Beta}(a, b)$ is said to have a Beta distribution if its probability density function on the unit interval $(0, 1)$ is given by

$$f_Y(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \quad 0 < y < 1 \text{ and } a, b > 0,$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ denotes the gamma function.

Definition 1.2.2 (Gamma distribution)**.** The random variable $Y \sim \text{Gam}(a, b)$ is said to have a Gamma distribution if its probability density function is given by

$$f_Y(y) = \frac{b^a}{\Gamma(a)} y^{a-1}e^{-by}, \quad y > 0.$$

The Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_n)$, for $\alpha_i > 0$, as $i = 1, \ldots, n$, can be described as the law of the random vector

$$(Y_1/Y, \ldots, Y_n/Y), \quad Y = \sum_{i=1}^n Y_i,$$

where $Y_i \overset{\text{ind}}{\sim} \text{Gam}(\alpha_i, 1)$ for any $i = 1, \ldots, n$. It is defined on a bounded support, i.e. the n-part simplex, defined as

$$S^n = \left\{ (x_1, \ldots, x_n) : x_i > 0, \quad i = 1, \ldots, n, \quad \sum_{i=1}^n x_i = 1 \right\}.$$

Hence, the pdf of the Dirichlet distribution is defined as follows

$$f_Y(\boldsymbol{y} = (y_1, \ldots, y_n)) = \frac{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n y_i^{\alpha_i - 1}, \quad \alpha_i \geq 0. \tag{1.8}$$

We can show that, by construction, the Dirichlet pdf with two components $f(y_1, y_2)$, coincides with the pdf of a Beta distribution with parameters α_1 and α_2. In the following we briefly recall some of the well-known properties of Dirichlet distribution (for proofs and details see Ferguson, 1973; Minka, 2000). The most relevant, when adopted as prior distribution, is that the Dirichlet distribution is conjugated to the multinomial likelihood, property that greatly facilitates posterior updates in Bayesian analysis. The Dirichlet distribution is commonly used as distribution for the weights of the components in the Bayesian mixture

model in (1.4). Furthermore, it is easy to recover marginal and conditional distributions as well as to maintain identifiability in the parameter estimation. The moments of the distribution are available in closed form. Finally, we remark that the infinite-dimensional generalization of the Dirichlet distribution is the Dirichlet process. In other terms, the Dirichlet distribution coincides with the finite dimensional distributions of the nonparametric prior defined in Ferguson (1973).

1.2.2 Ferguson-Dirichlet process

The Dirichlet process, commonly referred to as DP and also known as the Ferguson-Dirichlet process, is arguably the most famous example of Bayesian Nonparametric (BNP) model. The DP is a stochastic process whose realizations are discrete distributions with probability 1. Consequently, they are valuable for providing versatile mixing components in discrete mixture models.

Theorem 1.2.1. *(Ferguson, 1973). Let $(\mathbb{X}_0, \mathcal{X}_0)$ be a Polish space, α a finite measure on $(\mathbb{X}_0, \mathcal{X}_0)$ such that $\alpha(\mathbb{X}_0) = a > 0$. Then there exists a random probability measure $\tilde{p}$ with finite dimensional Dirichlet distribution. Its law is uniquely determined on the space $(\mathrm{P}, \mathcal{P})$ and $\tilde{p}$ is termed the Ferguson-Dirichlet process.*

In other terms, following Theorem 1.2.1, we can state the following definition

Definition 1.2.3 (Dirichlet process)**.** A random measure $\tilde{p}$ on $(\mathbb{X}_0, \mathcal{X}_0)$ is said to possess a Dirichlet process distribution, $\mathrm{DP}(\alpha)$, with base measure α, if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathbb{X}$,

$$(\tilde{p}(A_1), \ldots, \tilde{p}(A_k)) \sim \mathrm{Dir}(\alpha(A_1), \ldots, \alpha(A_k)).$$

where α is a given finite positive Borel measure on $(\mathbb{X}_0, \mathcal{X}_0)$ (Ghosal and Van der Vaart, 2017).

1.2.3 Posterior inference

Let $X_1, X_2, \ldots$ be a sequence of exchangeable random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in $(\mathbb{X}_0, \mathcal{X}_0)$, with de Finetti measure G, that is

$$X_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p} \quad i \in \mathbb{N}$$
$$\tilde{p} \sim G.$$

The focal point of Bayesian inference is on recovering the posterior distribution of $\tilde{p}$ and determining the predictive distribution. The former requires to obtain the distribution of $\tilde{p}$ conditionally given $X_1, \ldots, X_n$, which is a distribution on the space $(\mathrm{P}, \mathcal{P})$. The latter, instead, allows to predict the outcome of the next observation X_{n+1} given the previously observed values $X_1, \ldots, X_n$, i.e. to determine $\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n)$, known as 1-step ahead prediction.

In general, obtaining these distributions might require mathematical or, alternatively, computational effort. However, if G coincides with a $\mathrm{DP}(\alpha)$, the convenient properties of the Dirichlet process come to help. In particular, thanks to its conjugacy property, the posterior distribution of a Dirichlet process is again a Dirichlet process. As a result, closed-form Bayesian estimators for parameters of interest can be readily obtained. In the following theorem, more explicit forms to derive the posterior and the predictive distributions are provided.

Theorem 1.2.2. *(Ferguson, 1973) Let $(X_n)_{n\geq 1}$ be an exchangeable sequence of observations on $(\Omega, \mathcal{A}, \mathbb{P})$, with $\mathrm{DP}(\boldsymbol{\alpha})$ as the de Finetti measure of this sequence, namely*

$$X_i \mid \tilde{p} \overset{iid}{\sim} \tilde{p}, \quad i \in \mathbb{N}$$
$$\tilde{p} \sim \mathrm{DP}(\alpha),$$

where α is a finite measure on $(\mathbb{X}_0, \mathcal{X}_0)$.

Then, the posterior distribution of $\tilde{p}$ equals $\mathrm{DP}(\alpha_n)$, being $\alpha_n = \alpha + \sum_{i=1}^n \delta_{X_i}$.

Moreover, the predictive distribution is

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) = \frac{a}{a+n}\frac{\alpha(A)}{a} + \frac{n}{a+n}\left(\frac{1}{n}\sum_{j=1}^{n}\delta_{X_j}(A)\right),$$

for any $A \in \mathcal{X}_0$ and with $a = \alpha(\mathbb{X}_0)$.

Last but not least, the marginal distribution of X_i can be calculated similarly to the posterior distribution formula. Positive probability of the ties among X_i is implied by the discrete nature of the DP prior. Let $X_i \mid \tilde{p} \overset{iid}{\sim} \tilde{p}$, with $\tilde{p} \sim G$, be a random sample for $i = 1, \ldots, n$. In particular, $\tilde{p} \sim \mathrm{DP}(\alpha, G_0)$ with G_0 being the normalization of α, i.e. $G_0 = \alpha/\alpha(\mathbb{X}_0)$. The marginal distribution $p(X_1, \ldots, X_n) = \int \prod_{i=1}^n G(X_i)\, d\pi(G)$ is recovered by exploiting the Pólya urn representation of Blackwell and MacQueen (1973). It can be specified, by resorting to the chain rule, as $p(X_1, \ldots, X_n) = p(X_1)\prod_{i=2}^n p(X_i \mid X_1, \ldots, X_{i-1})$,

where

$$\mathbb{P}\left(X_i \in A \mid X_1 \ldots, X_{i-1}\right) = \frac{1}{a+i-1}\sum_{h=1}^{i-1}\delta_{X_h}\left(A\right) + \frac{a}{a+i-1}G_0\left(A\right), \tag{1.9}$$

for $i \geq 2$ (see, e.g., Müller et al., 2015). The marginal joint distribution of $(X_1, \ldots, X_n)$ is exchangeable since the X_i are i.i.d. given $\tilde{p}$. For all the other properties of the Dirichlet process and it equivalent definitions, one can refer to Ghosal and Van der Vaart (2017) and Müller et al. (2015).

1.2.4 Stick-breaking representation

A very useful approach to define random probability measures is the stick-breaking construction. This strategy is general, and, as a particular case, the Dirichlet process can be recovered. The stick-breaking representation of the Dirichlet process has been first proved by Sethuraman (1994).

Exploiting the stick-breaking method, one can define almost surely discrete random probability measures on $(\mathbb{X}_0, \mathcal{X}_0)$ of the following type

$$\tilde{p} = \sum_{j\geq 1}\tilde{p}_j\delta_{\tilde{\theta}_j},$$

where the $\tilde{\theta}_j$s are i.i.d. random atoms distributed as P_0, which is a probability on $(\mathbb{X}_0, \mathcal{X}_0)$. The $\tilde{p}_j$s are positive stick-breaking weights with the property $\sum_{j\geq 1}\tilde{p}_j = 1$ almost surely, which ensure that $\tilde{p}$ is a probability. To define the random masses $\tilde{p}_1, \tilde{p}_2, \ldots$, let us consider a sequence of random variables $Y_1, Y_2, \ldots$ in $[0, 1]$, for which we provide a constructive definition nicely described by the following metaphor. Consider a stick of length 1, break it into two pieces of length Y_1 and $1 - Y_1$, and set $\tilde{p}_1 = Y_1$. The remaining stick of length $1 - Y_1$ is again broken into two pieces of relative lengths Y_2 and $(1 - Y_2)$. Hence, we set $\tilde{p}_2 = (1 - Y_1)\, Y_2$, the remaining stick has length $(1 - Y_1)\,(1 - Y_2)$. Iterating such a procedure we define the following infinite sequence of weights

$$\tilde{p}_1 = Y_1, \quad \tilde{p}_2 = (1 - Y_1)\, Y_2, \quad \ldots \quad \tilde{p}_j = Y_j\prod_{i=1}^{j-1}(1 - Y_i), \quad \ldots$$

A graphical representation of the procedure is shown in Figure 1.2. The following theorem for i.i.d. stick-breaking weights guarantees that the weights sum up to 1, as long as the random variables $Y_1, Y_2, \ldots$ are i.i.d. and not degenerate at 0.

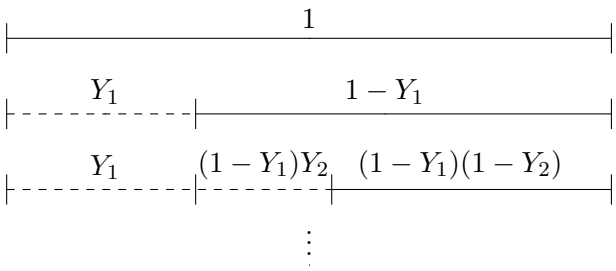


Figure 1.2: *Graphical representation of the stick-breaking procedure.*

Theorem 1.2.3. *Assume that $Y_1, Y_2, \ldots$ are i.i.d. random variables in $[0, 1]$ then*

$$\sum_{j \geq 1} \tilde{p}_j = 1 \iff \mathbb{P}(Y_1 > 0) > 0$$

To show that the Dirichlet process is a stick-breaking prior having i.i.d. stick-breaking weights, we can state the following

Lemma 1.2.1 (Dirichlet equation)**.** *Let α be a measure on $(\mathbb{X}_0, \mathcal{X}_0)$ with $\alpha(\mathbb{X}_0) = a > 0$, consider $Y \sim \text{Beta}(1, a)$ and $\tilde{\theta} \sim G_0$, with $G_0 = \alpha/a$. Consider the following distributional equation*

$$\tilde{p} = Y \delta_{\tilde{\theta}} + (1 - Y)\tilde{p} \tag{1.10}$$

where the variable is a random probability measure $\tilde{p}$. If $\tilde{\theta}$ and Y are independent, the Ferguson-Dirichlet process is the only solution of Equation 1.10.

The representation of Sethuraman is provided in the following theorem.

Theorem 1.2.4. *(Sethuraman, 1994) Let $\tilde{\theta}_1, \tilde{\theta}_2, \ldots \overset{iid}{\sim} G_0$, where $G_0 = \alpha/a$, assume that $Y_1, Y_2, \ldots \overset{iid}{\sim} \text{Beta}(1, a)$, where the two sequences are independent. The random probability measure*

$$\tilde{p} := \sum_{j \geq 1} Y_j \prod_{i=1}^{j-1} (1 - Y_i)\, \delta_{\tilde{\theta}_j}$$

has distribution $\text{Dir}(\alpha)$.

Using this representation, it becomes evident that, as the Dirichlet distribution serves as the component weights distribution for finite mixtures, the Dirichlet process might conveniently be chosen to play the corresponding role in infinite mixture models.

1.2.5 Dirichlet Process Mixture models

When it comes to modeling absolutely continuous distributions, such as in density estimation problems, the Dirichlet process by itself is not suitable due to its discreteness. However, an effective nonparametric model for continuous distributions emerges when we convolve the Dirichlet process with an appropriate continuous kernel function, giving rise to what is known as the Dirichlet process mixture (DPM) model (Antoniak, 1974).

Definition 1.2.4 (Kernel function)**.** A kernel function on $\mathbb{X}_0 \times \Theta$ is any function $\psi : \mathbb{X}_0 \times \Theta \to \mathbb{R}^+$such that

1. $\psi(x;\theta)$ is bi-measurable (i.e., that is measurable in both the arguments);
2. for any $\theta \in \Theta, x \mapsto \psi(x;\theta)$ is a probability density function.

Definition 1.2.5 (Dirichlet process mixture)**.** A Dirichlet process mixture (DPM) on $\mathbb{X}_0$ is a random density $\tilde{f}$ defined as

$$\tilde{f} = \int_{\Theta} \psi(x;\theta) \mathrm{d}\tilde{p}(\theta) \tag{1.11}$$

with $\tilde{p} \sim \mathrm{DP}(\alpha)$.

In the special case of a Gaussian kernel, we can derive one of the most popular DPM models, namely the location-scale DP mixture of univariate Gaussian kernels. Let $\mathbb{X}_0 = \mathbb{R}$, $\psi(x;\boldsymbol{\theta})$ with $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$, and $\Theta = \mathbb{R} \times \mathbb{R}^+$. The random density can be defined as

$$\tilde{f}(x) = \int_{\mathbb{R}\times\mathbb{R}^+} \psi\left(x; \boldsymbol{\theta} = \left(\mu, \sigma^2\right)\right) \mathrm{d}\tilde{p}(\boldsymbol{\theta}),$$

with the discrete random probability measure, $\tilde{p}$, used to model the joint distribution of mean and variance of the Gaussian kernel. A simplification of this model is called the location DP mixture of univariate Gaussian kernels. In this case, the Dirichlet process is used to model only the distribution of the mean of the Gaussian kernel. The variance is an additional parameter of the kernel. Consider $\theta = \mu$ and $\phi = \sigma^2$, $\Theta = \mathbb{R} \times \mathbb{R}^+$. The random density $\tilde{f}$ is specified as follows

$$\tilde{f}_{\sigma^2}(x) = \int_{\mathbb{R}} \psi\left(x; \mu, \sigma^2\right) \mathrm{d}\tilde{p}(\mu),$$

when σ^2 is fixed. If σ^2 has assigned a prior p_0, $\tilde{f}$ becomes

$$\tilde{f}(x) = \int_{\mathbb{R}_{\mathbb{R}+}} \psi\left(x; \mu, \sigma^2\right) \mathrm{d}\tilde{p}(\mu) \mathrm{d}p_0\left(\sigma^2\right).$$

The mixture model in (1.4) with the mixing measure G being a Dirichlet process prior can be written as a hierarchical model. Specifically, we introduce a vector of latent variables $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, and consider the following augmented version of the DPM:

$$\begin{aligned} X_i \mid \theta_i &\overset{\text{ind}}{\sim} \psi\left(x_i; \theta_i\right), \quad i = 1, \ldots, n \\ \theta_i \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \quad i = 1, \ldots, n \\ \tilde{p} &\sim \mathrm{DP}(\alpha). \end{aligned} \tag{1.12}$$

1.2.5.1 Partitions and Clustering

The hierarchical representation of Equation 1.12 allows us to recover, in a straightforward way, the partition induced by $\boldsymbol{\theta}$. For $i, j = 1, \ldots, n$ and $i \neq j$, the discreteness of the Dirichlet implies that $P(\theta_i = \theta_j)$ is greater than 0, i.e. $\boldsymbol{\theta}$ displays ties with positive probability. Consequently, the vector $\boldsymbol{\theta}$ will exhibit $k < n$ distinct values $(\theta_1^*, \ldots, \theta_k^*)$ with corresponding frequencies given by $(n_1, \ldots, n_k)$, where $\sum_{j=1}^k n_j = n$. In essence, the latent variables $\boldsymbol{\theta}$ yield a partition of the n elements into k blocks, where each block corresponds to a specific value θ_j^*. As a result, we can identify two observations X_i and X_j as belonging to the same cluster if the corresponding latent parameters θ_i and θ_j belong to the same block, i.e. $\theta_i = \theta_j$.

Notably, in the context of the Dirichlet process, the probability of observing any particular partition of n elements into k blocks with frequencies $(n_1, \ldots, n_k)$ is given by the exchangeable partition probability function (EPPF, see Pitman, 1995) $\Pi_k^{(n)}(n_1, \ldots, n_k)$, expressed as:

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{a^k}{(a)_n} \prod_{j=1}^k (n_j - 1)!, \tag{1.13}$$

with $(a)_n$ denoting the ascending factorial, i.e. $(a)_n = a(a+1)(a+2)\ldots(a+n-1) = \Gamma(a+n)/\Gamma(a)$. It is important to note that the EPPF is symmetric in its arguments and serves as a probability distribution over the space $\mathcal{S}_n$ encompassing all possible partitions of the n elements. Let $\mathcal{S}_{n,k}$ be the set encompassing all conceivable partitions of n elements

into k blocks. Then, it can be verified that

$$\sum_{k=1}^{n} \sum_{\mathcal{S}_{n,k}} \Pi_k^{(n)}(n_1, \ldots, n_k) = 1.$$

We are now in the position to derive the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X} = (X_1, \ldots, X_n)$. This distribution is proportional to the joint distribution of $(\boldsymbol{X}, \boldsymbol{\theta})$,

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{X}) &= \Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} g_0\left(\theta_j^*\right) \prod_{i=1}^{n} \psi\left(X_i \mid \theta_i\right) \\ &= \Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} g_0\left(\theta_j^*\right) \prod_{i \in \mathcal{C}_j} \psi\left(X_i \mid \theta_j^*\right) \end{aligned}$$

obtained after marginalizing with respect to $\tilde{p}$. Here, g_0 is the probability density function associated with the base measure G_0, and $\mathcal{C}_j = \left\{i \in \{1, \ldots, n\} : \theta_i = \theta_j^*\right\}$ identifies the set of indexes allocated to the j-th cluster, accordingly $n_j = |\mathcal{C}_j|$. Consequently, the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X}$ can be expressed as

$$p(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{\Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} g_0\left(\theta_j^*\right) \prod_{i \in \mathcal{C}_j} \psi\left(X_i \mid \theta_j^*\right)}{\sum_{k=1}^{n} \sum_{\mathcal{S}_{n,k}} \Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} \int_{\Theta} g_0(t_j) \prod_{i \in \mathcal{C}_j} \psi(X_i \mid r_j) \, \mathrm{d}r_j}.$$

As already discussed, the partition of $\boldsymbol{\theta}$ translates into a partition of the observations $\boldsymbol{X}$, clustered into k groups. The partition of the set of experimental units $\{1, \ldots, n\}$ is referred as $\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_j\}$. The sets $\mathcal{C}_j$ are random due to the fact that the θ_is are random. As highlighted in Müller et al. (2015), the Dirichlet process mixture implies a model on a random partition π of the experimental units. The posterior inference on clustering of the data is described by the posterior model $p(\pi \mid \boldsymbol{X})$.

Due to the large dimensionality of $\mathcal{S}_{n,k}$, the posterior distribution of π is, in general, hard to study. To this end, various computational approaches have been proposed in the literature (see, e.g., Neal, 2000). In contrast to parametric Bayesian analysis, one of the significant advantages of employing BNP methods lies in their flexibility in modeling distribution functions without restrictive assumptions. However, this flexibility comes with its own set of computational challenges. The remarkable progresses of BNP models in recent decades is largely attributable to advancements in simulation-based computational techniques, with Markov chain Monte Carlo methods (MCMC) playing a prominent role (see, e.g., Müller et al., 2015). MCMC methods for mixture models study the posterior

distribution of π by sampling from the posterior distribution of $\boldsymbol{\theta}$. The partitions visited by the chain provide us with valuable insights on how data can be clustered, including both the number of groups and the allocation of observations to distinct groups. Collecting this information through the sampling algorithm allows us to investigate, for example, the variable K, which represents the number of clusters in the dataset. Notably, the DPM model does not require the number K to be fixed a priori. Instead, it treats K as a random variable, and its posterior distribution is influenced by both prior beliefs and the observed data. Exploiting the EPPF in (1.13) of the DP, we can derive the prior induced by the Dirichlet process on K as follows

$$\begin{aligned}
\mathbb{P}(K=k) &= \sum_{\mathcal{S}_{n,k}} \Pi_k^{(n)}(n_1,\ldots,n_k) \\
&= \sum_{\mathcal{S}_{n,k}} \frac{a^k}{(a)_n} \prod_{j=1}^{k} (n_j-1)! \\
&= \frac{a^k}{(a)_n} \sum_{\mathcal{S}_{n,k}} \prod_{j=1}^{k} (n_j-1)! \\
&= \frac{a^k}{(a)_n} |s(n,k)|,
\end{aligned}$$

where $s(n,k)$ represents a Stirling number of the first kind. MCMC algorithms play a pivotal role in the estimation and fitting of models incorporating DP priors. These algorithms are indispensable tools in the Bayesian settings, enabling us to draw inferences and explore the posterior distributions of parameters and latent variables in a wide range of models. Among the various MCMC techniques employed in the context of nonparametric mixture models, four of the most notable ones are: the marginal algorithm (Escobar, 1988; Neal, 2000), the slice sampler (Walker, 2007; Kalli et al., 2011), the retrospective sampler (Papaspiliopoulos and Roberts, 2008), and the importance conditional sampler (Canale et al., 2022).

1.3 Main Contributions of the Thesis

In this chapter, we proposed an essential review of Bayesian mixture models, with a specific emphasis on mixtures based on the Dirichlet process. We have focused on the significance of this process, revisited its fundamental assumptions, provided a concise summary of its properties and notation, and discussed the methods for deriving posterior and predictive distributions. The remainder of this manuscript is dedicated to the investigation of specific

research questions, examined within the framework of Bayesian methodology. It is worth noting a clear distinction: while the next two chapters harness the flexibility and adaptability of nonparametric methods, the final chapter takes a slightly different approach, by embracing a parametric approach.

More specifically, in Chapter 2, we consider a significant problem in the field of data analysis: simultaneously clustering users and items, task that is particularly challenging when dealing with datasets plagued by missing information. This issue becomes especially relevant when considering the vast and diverse datasets generated by social platforms, where missing data is a common occurrence. In this chapter, we propose a novel approach to address this issue, a comprehensive model for co-clustering data that accounts for informative censoring. The proposed model not only facilitates efficient clustering of users and items but also handles the complexities introduced by missing data, enabling a more robust and accurate analysis of underlying patterns in the dataset. This approach represents an innovative contribution to the field of BNP methods, particularly in data-rich environments, with missing values being considered as a source of information.

In Chapter 3, our focus shifts to a compelling question: how to model the evolution of partitions over time. To tackle this issue, we explore how to identify changepoints in multivariate time series based on time-varying random partitions. The temporal dimension adds a layer of complexity to the already challenging task of modeling the clustering of units with a random partition model. We present a compelling framework that not only provides valuable insights into the ever-evolving nature of partitions over time but also effectively allows capturing changepoints in their evolution. Our method contributes substantially to the existing literature on temporal modeling with random partition models, which to date is limited to few contributions.

The final chapter of this manuscript, Chapter 4, is dedicated to defining a new model for textual data using distributions defined on the simplex. This chapter centers around the intriguing question of whether such distributions can accommodate various forms of dependence among topics. Our examination leads us into the realm of topic modeling with positive correlation, where we explore the potential of these distributions in uncovering underlying patterns and dependencies within complex datasets. Throughout this chapter, we push the boundaries of what simplex-based distributions can achieve in terms of modeling complicated relationships, shedding light on their versatility and applicability across diverse contexts.

Bibliography

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 1(2):353–355.

Canale, A., Corradin, R., and Nipoti, B. (2022). Importance conditional sampling for Pitman–Yor mixtures. *Statistics and Computing*, 32(3):40.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

Escobar, M. D. (1988). *Estimating the means of several normal populations by nonparametric estimation of the distribution of the means*. Yale University.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2013). *Bayesian data analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and computing*, 21:93–105.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York.

Minka, T. P. (2000). Estimating a Dirichlet distribution. Technical report, MIT Media Laboratory.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*, volume 1. Springer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American journal of Mathematics*, pages 343–366.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.

Peel, D. and MacLahlan, G. (2000). Finite mixture models. *John & Sons*.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54.

Chapter 2

A Bayesian Model for Co-clustering Ordinal Data with Informative Censoring

2.1 Introduction

Multivariate ordinal data, such as movie ratings and politicians' votes, have been extensively analyzed with both parametric and nonparametric approaches (e.g., Bennett et al., 2007; Zhou et al., 2008; van Dijk et al., 2009). This type of data plays a central role in various domains, with recommendation systems being a notable example. In recommendation systems, individual preferences are formalized as ordinal ratings provided for various items. Examining such applications helps us in shedding light on the specific challenges the analysis of these data might present and in outlining the main objectives of our contribution. First of all, the ease of accessibility of diverse data sources, driven by technological advancements, has significantly amplified the challenge of extracting meaningful information from the large volume of available data (Abello et al., 2013). Moreover, recommendation systems inputs are often characterized by sparsity, with individual preferences expressed for only a subset of the available items, as it happens for instance in e-commerce platforms: any statistical analysis must address the issue of handling missing observations. A common strategy is to consider them as absent information and thus discard them. On the contrary, missing data itself can convey useful information: for example, censoring in movie ratings, or politicians' votes, might indicate lack of interest for a specific movie type, or objection to a specific political position. Finally, when the goal is to provide similarity-based recommendations, it is interesting to explore both the clustering of individual preferences and the clustering of items. In this chapter, we present a modelling strategy that, while not designed exclusively to deal with the problem of recommending items that might be pertinent to specific individuals, is able to address the aforementioned challenges. More specifically,

we propose a method to do simultaneous clustering, i.e. co-clustering, of individuals and items, while accounting for the large dimension and the sparse nature of the data.

Co-clustering methods, also known under the name of biclustering or two-mode clustering, were first introduced by Hartigan (1972): they are meant to simultaneously cluster rows and columns of a matrix. Over the past three decades, these approaches have found widespread application, particularly in the realm of biological data analysis, where they have been used to jointly cluster genes and conditions (Madeira and Oliveira, 2004; Cheng and Church, 2000). Other applications of these methods can be found in data mining (Busygin et al., 2008) and recommendation system (Choi et al., 2018). A non-exhaustive list of references that propose methods for co-clustering includes Meeds and Roweis (2007), where, in the context of recommendation systems, a solution based on the Pitman–Yor process is proposed; Wang et al. (2011), where a Mondrian process is used to model dependence of row and column clusters and a reversible jump MCMC for posterior sampling; Wang et al. (2012), an extension of the model of Meeds and Roweis (2007) where a new approach, called Infinite Hidden Relational Model, is proposed to predict interaction values for new objects. Within this line of research, some contributions have considered the problem of missing observations: Shan and Banerjee (2008) address it by considering only a complete subset of data; Reisner et al. (2019) present an `R` package to co-cluster data in the presence of missing values, which implements a geometric approach that exploits an optimal rearrangement of rows and column of the data matrix. Our work contributes to this literature and is innovative as it regards missing data as an informative component, a concept we refer to as "informative censoring". This also allows us to avoid the common practice of discarding or imputing missing data. Additionally, by adopting a nonparametric approach, we circumvent the challenge of setting the number of clusters for rows and columns. Instead, we treat these quantities as model parameters for which we can make inference based on the evidence provided by the data.

The problem of modeling multivariate ordinal data has attracted a great deal of attention in the recent Bayesian and Bayesian nonparametric literature (see, e.g., DeYoreo and Kottas, 2018, 2020; Webb and Forster, 2008). A cleaver idea to deal with these data, discrete by their nature, consists in introducing continuous unobservable latent variables and a set of cutoffs that allow to relate observations and latent variables, in the spirit of (Albert and Chib, 1993). Kottas et al. (2005) combine this idea with the definition of a Dirichlet process mixture of multivariate Normal kernel to model the continuous latent variables. The flexibility of their proposal conveniently allows for arbitrary cutoffs to be fixed without affecting posterior inference. The model we define combines this framework with a matrix

factorization approach to handle the large dimensional nature of the data. The latter idea is in line with the literature on recommendation systems through a collaborative filtering approach (Koren, 2008). The principle at the core of Bayesian matrix factorization models (e.g. Salakhutdinov and Mnih, 2008) is that the preferences of a user are determined by a small number of unobserved factors. For this reason, the $n \times p$ preference matrix, reporting the preferences that n individuals assign to p items, is modeled as the product of a $d \times n$ user coefficient matrix and a $d \times p$ factor matrix, where d represents the latent dimension. The aim is to find the best rank-d approximation to the observed $n \times p$ target matrix under a given loss function. An extension of this model can include extra information on either individuals or items, element that can improve the model predictive ability, especially when data are sparse (Porteous et al., 2010).

The rest of the chapter is organized as follows. Section 2.2 is dedicated to the specification of the model. The strategy for posterior distributions, and posterior inference with details on its implementation can be found in Section 2.3 and Section 2.4. The performance of the model is investigated by means of the analysis of synthetic and real data, as presented in Section 2.5 and 2.6, respectively. Finally, the Appendix includes additional results for the full conditional distributions and the data analysis.

2.2 Model

Let $\boldsymbol{X}$ be a tensor of continuous observations, with dimensions $n \times p \times q$, where n is the number of individuals, p is the number of items and q is the size of each observation $\boldsymbol{X}_{ij}$ (individual i, item j). We assume that d is the size of latent factor (with $d \ll n, p$), and we let $\boldsymbol{\theta}_i$ be the factor matrix for the i-th individual, with size $d \times 1 \times q$, $\boldsymbol{\psi}_j$ be the factor matrix for the j-th item, with size $d \times 1 \times q$, and $\boldsymbol{\Xi}$ be a variance-covariance matrix, with size $q \times q$. We propose to model the factor matrices $\boldsymbol{\theta}_i$ and $\boldsymbol{\psi}_j$ with independent Dirichlet processes, which leads to the following nonparametric Bayesian factor model:

$$\begin{aligned} \boldsymbol{X}_{ij} | \boldsymbol{\theta}_i, \boldsymbol{\psi}_j &\overset{\text{ind}}{\sim} N_q(\boldsymbol{\theta}_i^\top \boldsymbol{\psi}_j, \boldsymbol{\Xi}); \\ \boldsymbol{\theta}_i | H &\overset{\text{iid}}{\sim} H; \qquad \boldsymbol{\psi}_j | F \overset{\text{iid}}{\sim} F; \\ (H, F) &\sim \text{DP}(M_H, H_0)\text{DP}(M_F, F_0). \end{aligned} \tag{2.1}$$

where the base measures (H_0, F_0) are specified as matrix Normal distributions (see, e.g., Viroli, 2011). The model in (2.1) can be seen as a multivariate extension of the setting presented in Porteous et al. (2010).

Our objective is to explore the flexibility of (2.1) to model ordinal observations with censored components. As a running example, one can think of a movie platform, with the ordinal observations being the rating assigned by users to movies, and the censored components corresponding to movies not rated by specific individuals. With this purpose, we resort to the idea presented in Kottas et al. (2005), where a Bayesian nonparametric model for continuous distributions is used, at a latent level, to model multivariate ordinal data. Although the same strategy can be applied to a larger dimensional framework, we confine ourselves to the case $q = 2$, as two dimensions suffice in handling movie ratings and censored observations. More specifically, we let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ be the vector of ratings assigned by individual i to the p movies on the database, where we assume that each Y_{ij} takes values in $\{1, 2, \ldots, C_j\}$. In realistic scenarios, it seems reasonable to expect that some of the components of each observation $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ might be missing, that is there might be movies that were not rated by i-th individual. We formalize this by endowing each observation $\boldsymbol{Y}_i$ with a vector $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{ip})$, with δ_{ij}, for $j = 1, \ldots, p$, indicating whether the j-th component Y_{ij} of $\boldsymbol{Y}_i$ was observed ($\delta_{ij} = 1$) or not ($\delta_{ij} = 0$). We formalise the fact that the j-th element of $\boldsymbol{Y}_i$ is not observed by writing $Y_{ij} \in [C_j]$, where $[n] := \{1, 2, \ldots, n\}$ denotes the set of the first n positive integers. The strategy of Kottas et al. (2005) can then be applied by introducing the latent variables $\boldsymbol{Z}_i$, for $i = 1, \ldots, n$, such that, for any pair of positive integers $\ell_1 \leq \ell_2$,

$$Y_{ij} \in \{\ell_1, \ldots, \ell_2\} \text{ if } \gamma_{j,\ell_1 - 1} < Z_{ij} \leq \gamma_{j,\ell_2}$$

where $-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,C_{j-1}} < \gamma_{j,C_j} = \infty$ are the cutoffs for each $j = 1, \ldots, p$. Only two special cases must be considered here, namely $\ell_1 = \ell_2 = \ell$ (observed component), and $\ell_1 = 1$ and $\ell_2 = C_j$ (unobserved component). That is

$$\begin{aligned} Y_{ij} &= \ell && \text{if } \gamma_{j,\ell-1} < Z_{ij} \leq \gamma_{j,\ell}; \\ Y_{ij} &\in [C_j] && \text{if } -\infty = \gamma_{j,0} < Z_{ij} \leq \gamma_{j,C_j} = \infty. \end{aligned}$$

Similarly, for the censored variables δ_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, we introduce a continuous latent variable D_{ij}, in the spirit of Albert and Chib (1993), such that $\delta_{ij} = 1$ if and only if $D_{ij} \geq 0$. If we let $\boldsymbol{\Xi} = \text{diag}(\sigma^2, \tau^2)$, then we can exploit (2.1) to model the latent vectors $\boldsymbol{X}_{ij} = (Z_{ij}, D_{ij})$, and thus obtain the following hierarchical model for the

observations (Y_{ij}, δ_{ij}),:

$$
\begin{aligned}
Pr(Y_{ij} = \ell, \mid Z_{ij}, \delta_{ij}) \;&=\; \frac{1}{C_j^{1-\delta_{ij}}} \mathbb{1}_{(\gamma_{j,\ell-1}, \gamma_{j,\ell}]}(Z_{ij})^{\delta_{ij}} \\
\delta_{ij} &= \begin{cases} 1, & \text{if } D_{ij} \geq 0 \\ 0, & \text{if } D_{ij} < 0 \end{cases} \\
Z_{ij} \mid \boldsymbol{U}_i, \boldsymbol{V}_j &\overset{\text{ind}}{\sim} N(\boldsymbol{U}_i^\top \boldsymbol{V}_j, \sigma^2); \\
D_{ij} \mid \boldsymbol{R}_i, \boldsymbol{W}_j &\overset{\text{ind}}{\sim} N(\boldsymbol{R}_i^\top \boldsymbol{W}_j, \tau^2); \\
\begin{pmatrix} \boldsymbol{U}_i \\ \boldsymbol{R}_i \end{pmatrix} = \begin{bmatrix} U_{i1} & U_{i2} & \dots & U_{id} \\ R_{i1} & R_{i2} & \dots & R_{id} \end{bmatrix} &\mid H \overset{\text{iid}}{\sim} H \\
H &\sim \text{DP}(M_H, H_0) \\
\begin{pmatrix} \boldsymbol{V}_j \\ \boldsymbol{W}_j \end{pmatrix} = \begin{bmatrix} V_{1j} & V_{2j} & \dots & V_{dj} \\ W_{1j} & W_{2j} & \dots & W_{dj} \end{bmatrix} &\mid F \overset{\text{iid}}{\sim} F \\
F &\sim \text{DP}(M_F, F_0)
\end{aligned}
\tag{2.2}
$$

where, as in (2.1) H and F are independent Dirichlet processes and $d \ll n, p$. The first equation defining the model is a way to formalize the idea that, if $\delta_{ij} = 1$, i.e. $\mathbb{1}(D_{ij} > 0) = 1$, then, conditionally on Z_{ij}, the distribution of Y_{ij} is degenerate at the value ℓ such that Z_{ij} belongs to $(\gamma_{j,\ell-1}, \gamma_{j,\ell}]$; if $\delta_{ij} = 0$, i.e. $\mathbb{1}(D_{ij} > 0) = 0$, then the distribution of Y_{ij} does not depend on the value of the latent variable Z_{ij} and is uniform on $\{1, \dots, C_j\}$.
The model is completed by specifying the base measures H_0 and F_0, which we define as independent matrix normal distributions, that is

$$
\begin{aligned}
H_0 &= N_{2,d}\left(\begin{pmatrix} m_{U,i1} & m_{U,i2} & \dots & m_{U,id} \\ m_{R,i1} & m_{R,i2} & \dots & m_{R,id} \end{pmatrix}, \boldsymbol{\Phi}, \boldsymbol{\Sigma} \right); \\
F_0 &= N_{2,d}\left(\begin{pmatrix} m_{V,1j} & m_{V,2j} & \dots & m_{V,dj} \\ m_{W,1j} & m_{W,2j} & \dots & m_{W,dj} \end{pmatrix}, \boldsymbol{\Omega}, \boldsymbol{\Upsilon} \right),
\end{aligned}
\tag{2.3}
$$

where $\boldsymbol{\Phi}$ and $\boldsymbol{\Omega}$ are (2×2) variance-covariance matrices of the rows of the matrix Normals,

$$
\begin{aligned}
\boldsymbol{\Phi} &= \begin{pmatrix} \phi_1 & \phi_2 \\ \phi_2 & \phi_3 \end{pmatrix}; \\
\boldsymbol{\Omega} &= \begin{pmatrix} \omega_1 & \omega_2 \\ \omega_2 & \omega_3 \end{pmatrix},
\end{aligned}
$$

and, similarly, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Upsilon}$ are $(d \times d)$ variance-covariance matrices of the columns. All the matrices we defined are to be considered positive definite. Moreover, the parametrization we consider is such that it is worth stressing that while observations $\boldsymbol{Y}_i$ and $\boldsymbol{\delta}_i$, and latent variables $\boldsymbol{Z}_i$ and $\boldsymbol{D}_i$ have dimension p, the underlying factor model works with a smaller number d of factors. Furthermore, we observe that the introduction of a latent layer of continuous random variables considerably simplifies the task of writing the joint distribution of all the random elements in the model, which is the starting point of next section.

2.3 Posterior Distribution

In view of the definition of a MCMC algorithm for posterior inference, described in Section 2.4, we study the joint conditional distribution of the random elements that constitute the model in (2.2), given the data. More specifically, we show the derivation of the conditional distribution that is obtained after marginalizing with respect to the random probability measures (H, F). This step conveniently simplifies the task of posterior sampling by analytically integrating out the infinite-dimensional parameters of the model.

For the sake of compactness, we introduce the following notation, we let $\boldsymbol{\theta} = (\boldsymbol{U}, \boldsymbol{R}) = ((\boldsymbol{U}_1, \boldsymbol{R}_1), \ldots, (\boldsymbol{U}_n, \boldsymbol{R}_n))$, $\boldsymbol{\psi} = (\boldsymbol{V}, \boldsymbol{W}) = ((\boldsymbol{V}_1, \boldsymbol{W}_1), \ldots, (\boldsymbol{V}_p, \boldsymbol{W}_p))$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, $\boldsymbol{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n)$, and denote the data as $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ and $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n)$. Then, we are interested in studying the conditional distribution of $(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{D})$, given the data $(\boldsymbol{Y}, \boldsymbol{\delta})$, that is $p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{D} \mid \boldsymbol{Y}, \boldsymbol{\delta}) \propto p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{Y}, \boldsymbol{\delta})$. Conditionally on (H, F), we can write

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{D} \mid \boldsymbol{Y}, \boldsymbol{\delta}, (H, F)) &\propto p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\delta}) p(\boldsymbol{\delta} \mid \boldsymbol{D}) p(\boldsymbol{Z}, \boldsymbol{D} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid (H, F)) \\
&= \Big[\prod_{i=1}^{n} \prod_{j=1}^{p} p(Y_{ij} \mid Z_{ij}, \delta_{ij}) p(\delta_{ij} \mid D_{ij}) p(Z_{ij}, D_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\psi}_j) \Big] \\
&\qquad \times p(\boldsymbol{\theta} \mid H) p(\boldsymbol{\psi} \mid F).
\end{aligned}
$$

given that $(H, F) \sim \mathrm{DP}(M_H, H_0)\mathrm{DP}(M_F, F_0)$, it is possible to analytically marginalize (H, F) out of $p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{D} \mid \boldsymbol{Y}, \boldsymbol{\delta}, (H, F))$. Thus, we get

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{D} \mid \boldsymbol{Y}, \boldsymbol{\delta}) &\propto \mathbb{E}_{(H,F)} \big[p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\delta}) p(\boldsymbol{\delta} \mid \boldsymbol{D}) p((\boldsymbol{Z}, \boldsymbol{D}) \mid \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid (H, F)) \big] \\
&= p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\delta}) p(\boldsymbol{\delta} \mid \boldsymbol{D}) p((\boldsymbol{Z}, \boldsymbol{D}) \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \mathbb{E}_{(H,F)} \big[p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid (H, F)) \big] \\
&= p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\delta}) p(\boldsymbol{\delta} \mid \boldsymbol{D}) p((\boldsymbol{Z}, \boldsymbol{D}) \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \mathbb{E}_{H} \big[p(\boldsymbol{\theta} \mid H) \big] \mathbb{E}_{F} \big[p(\boldsymbol{\psi} \mid F) \big].
\end{aligned}
$$

The distribution of the latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, i.e. $\mathbb{E}_H[p(\boldsymbol{\theta} \mid H)]$ and $\mathbb{E}_F[p(\boldsymbol{\psi} \mid F)]$, are characterized by the so-called Pólya urn scheme (Blackwell et al., 1973), which is based on the predictive distribution of $\boldsymbol{\theta}_{i+1}$ and $\boldsymbol{\psi}_{j+1}$ given the observation of $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_i\}$ and $\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_j\}$. Moreover, given the independence of H and F, the two predictive distributions are independent. Specifically, for the case $H \sim \mathrm{DP}(M_H, H_0)$, we have

$$p(\boldsymbol{\theta}_{i+1} \in \cdot \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_i) = \frac{M_H}{M_H + i} H_0(\cdot) + \frac{1}{M_H + i} \sum_{r=1}^{i} \delta_{\boldsymbol{\theta}_r}(\cdot),$$

analogously for $F \sim DP(M_F, F_0)$ and the predictive distribution for $\boldsymbol{\psi}_{j+1}$. Thus we can write

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{Y} \mid \boldsymbol{V}, \boldsymbol{\delta}) \propto & \prod_{i=1}^{n} \Big[\prod_{j=1}^{p} p(Y_{ij} \mid Z_{ij}, \delta_{ij}) p(\delta_{ij} \mid D_{ij}) p(Z_{ij}, D_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\psi}_j) \Big] \\ & \times \prod_{i=1}^{n} p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{i-1}) \prod_{j=1}^{p} p(\boldsymbol{\psi}_j \mid \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_{j-1}). \end{aligned}$$

The distribution of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ obtained by marginalizing with respect to (H, F) is more conveniently written by means of the exchangeable partition probability function (EPPF) of the two Dirichlet processes. Given the almost sure discreteness of H and F, with positive probability there will be ties between the components of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, leading to a total of k_n and k_p distinct values. It is convenient to denote these distinct values as $(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{k_n}^*)$ and $(\boldsymbol{\psi}_1^*, \ldots, \boldsymbol{\psi}_{k_p}^*)$, whose frequencies in $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are summarized in the vectors $(n_1, \ldots, n_{k_n})$ and $(p_1, \ldots, p_{k_p})$. The EPPFs of the Dirichlet process tell us that the probability of observing any specific realization of $\boldsymbol{\theta}$ (or $\boldsymbol{\psi}$) displaying k_n (or k_p) distinct values with frequencies $(n_1, \ldots, n_{k_n})$ (or $(p_1, \ldots, p_{k_p})$). That is,

$$\begin{aligned} \Pi_{k_n}^{(n)}(n_1, \ldots, n_{k_n}) &= \frac{M_H^{k_n}}{(M_H)_n} \prod_{r=1}^{k_n} (n_r - 1)! \\ \Pi_{k_p}^{(p)}(p_1, \ldots, p_{k_p}) &= \frac{M_F^{k_p}}{(M_F)_p} \prod_{t=1}^{k_p} (p_t - 1)!. \end{aligned}$$

As a result we can write

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{D} \mid \boldsymbol{Y}, \boldsymbol{\delta}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{p} p(Y_{ij} \mid Z_{ij}, \delta_{ij}) p(\delta_{ij} \mid D_{ij}) p(Z_{ij}, D_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\psi}_j) \\ &\quad \times \Pi_{k_n}^{(n)}(n_1, \ldots, n_{k_n}) \prod_{r=1}^{k_n} h_0(\boldsymbol{\theta}_r^*) \times \Pi_{k_p}^{(p)}(p_1, \ldots, p_{k_p}) \prod_{t=1}^{k_p} f_0(\boldsymbol{\psi}_t^*) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{p} p(Y_{ij} \mid Z_{ij}, \delta_{ij}) p(\delta_{ij} \mid D_{ij}) \\ &\quad \times \prod_{r=1}^{k_n}\prod_{t=1}^{k_p}\prod_{i \in \mathcal{C}_r}\prod_{j \in \mathcal{C}_t} p(Z_{ij}, D_{ij} \mid \boldsymbol{\theta}_r^*, \boldsymbol{\psi}_t^*) \\ &\quad \times \prod_{r=1}^{k_n} h_0(\boldsymbol{\theta}_r^*) \Pi_{k_n}^{(n)}(n_1, \ldots, n_{k_n}) \prod_{t=1}^{k_p} f_0(\boldsymbol{\psi}_t^*) \Pi_{k_p}^{(p)}(p_1, \ldots, p_{k_p}), \end{aligned} \tag{2.4}$$

or, more compactly,

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{D} \mid \boldsymbol{Y}, \boldsymbol{\delta}) &\propto \Pi_{k_n}^{(n)}(n_1, \ldots, n_{k_n}) \Pi_{k_p}^{(p)}(p_1, \ldots, p_{k_p}) \prod_{r=1}^{k_n} h_0(\boldsymbol{\theta}_r^*) \prod_{t=1}^{k_p} f_0(\boldsymbol{\psi}_t^*) \\ &\quad \times \prod_{i \in \mathcal{C}_r}\prod_{j \in \mathcal{C}_t} p(Y_{ij} \mid Z_{ij}, \delta_{ij}) p(\delta_{ij} \mid D_{ij}) p(Z_{ij}, D_{ij} \mid \boldsymbol{\theta}_r^*, \boldsymbol{\psi}_t^*), \end{aligned} \tag{2.5}$$

where h_0 and f_0 denote the probability density functions corresponding to the base measures H_0 and F_0, and $\mathcal{C}_r = \{i \in \{1, \ldots, n\} : \boldsymbol{\theta}_i = \boldsymbol{\theta}_r^*\}$ and $\mathcal{C}_t = \{j \in \{1, \ldots, p\} : \boldsymbol{\psi}_j = \boldsymbol{\psi}_t^*\}$. From (2.5) one can obtain the full conditional distributions of $\boldsymbol{\theta}_i, \boldsymbol{\psi}_j$, $\boldsymbol{Z}_i$ and $\boldsymbol{D}_i$, for any $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

2.4 Posterior Inference

Given the convoluted form of the posterior distribution in (2.5), we devise a MCMC sampling strategy to investigate its properties. Specifically, we build a Gibbs sampling algorithm with parameter updates made efficient by the closed-form of the full conditional distributions. Given the base measures defined in Equation 2.3, we let $\boldsymbol{m}_U = (m_{U,i1}, m_{U,i2}, \ldots, m_{U,id}), \boldsymbol{m}_R = (m_{R,i1}, m_{R,i2}, \ldots, m_{R,id}), \boldsymbol{m}_V = (m_{V,1j}, m_{V,2j}, \ldots, m_{V,dj})$

and $\boldsymbol{m}_W = (m_{W,1j}, m_{W,2j}, \ldots, m_{W,dj})$. We start with the simple observation that, if

$$\begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{R} \end{pmatrix} \sim N_{2,d}\left(\begin{pmatrix} \boldsymbol{m}_U \\ \boldsymbol{m}_R \end{pmatrix}, \boldsymbol{\Phi}, \boldsymbol{\Omega} \right)$$

with parameters $\boldsymbol{m}_U$ and $\boldsymbol{m}_V$ as described above, and variance-covariance matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$, then, we can rewrite the joint distribution of $\boldsymbol{U}$ and $\boldsymbol{R}$, by using the chain rule, as

$$\boldsymbol{U}|\boldsymbol{R} \overset{\text{ind}}{\sim} N_d\left(\widetilde{\boldsymbol{m}}_U, \widetilde{\boldsymbol{\Sigma}}\right); \quad \boldsymbol{R} \overset{\text{ind}}{\sim} N_d\left(\boldsymbol{m}_R, \phi_3 \boldsymbol{\Sigma}\right)$$

where $\widetilde{\boldsymbol{m}}_U = \boldsymbol{m}_U + \frac{\phi_2}{\phi_3}(\boldsymbol{R} - \boldsymbol{m}_R)$ and $\widetilde{\boldsymbol{\Sigma}} = \left(\phi_1 - \frac{\phi_2^2}{\phi_3}\right)\boldsymbol{\Sigma}$. Similarly, if

$$\begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{W} \end{pmatrix} \sim N_{2,d}\left(\begin{pmatrix} \boldsymbol{m}_V \\ \boldsymbol{m}_W \end{pmatrix}, \boldsymbol{\Omega}, \boldsymbol{\Psi} \right)$$

with parameters $\boldsymbol{m}_V$ and $\boldsymbol{m}_W$ as described above, and variance-covariance matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}$, then, we can rewrite the joint distribution of $\boldsymbol{V}$ and $\boldsymbol{W}$, by using the chain rule, as

$$\boldsymbol{V}|\boldsymbol{W} \overset{\text{ind}}{\sim} N_d\left(\widetilde{\boldsymbol{m}}_V, \widetilde{\boldsymbol{\Upsilon}}\right); \quad \boldsymbol{W} \overset{\text{ind}}{\sim} N_d\left(\boldsymbol{m}_W, \omega_3 \boldsymbol{\Upsilon}\right),$$

where $\widetilde{\boldsymbol{m}}_V = \boldsymbol{m}_V + \frac{\omega_2}{\omega_3}(\boldsymbol{W} - \boldsymbol{m}_W)$ and $\widetilde{\boldsymbol{\Upsilon}} = \left(\omega_1 - \frac{\omega_2^2}{\omega_3}\right)\boldsymbol{\Upsilon}$. This observation will be useful in dealing with the distributions H_0 and F_0 appearing in (2.5).
The full conditional distributions of $\boldsymbol{Z}_i$ and $\boldsymbol{D}_i$ are proportional to

$$\begin{aligned} p(\boldsymbol{Z}_i|\ldots) \propto \exp\Big\{ -\frac{1}{2} tr[\boldsymbol{\Delta}^{-1}(\boldsymbol{Z}_i - \boldsymbol{V}^\top \boldsymbol{U}_i)(\boldsymbol{Z}_i - \boldsymbol{V}^\top \boldsymbol{U}_i)^\top] \Big\} \\ \times \prod_{j=i}^{p} \mathbb{1}_{(\gamma_{j,V_{ij}-1}, \gamma_{j,V_{ij}}]}(Z_{ij})^{\mathbb{1}_{\{D_{ij}>0\}}}; \\ p(\boldsymbol{D}_i|\ldots) \propto \exp\Big\{ -\frac{1}{2} tr\big[\boldsymbol{\Gamma}^{-1}(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{R}_i)(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{R}_i)^\top\big] \Big\} \\ \times \prod_{j=i}^{p} \Big(\delta_{ij} \mathbb{1}_{\{D_{ij}\geq 0\}} + (1-\delta_{ij})\mathbb{1}_{\{D_{ij}<0\}}\Big), \end{aligned}$$

For the full conditional distributions of $(\boldsymbol{U}_i, \boldsymbol{R}_i)$ and, similarly for $(\boldsymbol{V}_j, \boldsymbol{W}_j)$, we can write

$$\begin{aligned} p\left((\boldsymbol{u}_i, \boldsymbol{r}_i)|(\boldsymbol{U}_{-i}, \boldsymbol{R}_{-i}), \ldots\right) \propto \frac{M_H}{M_H + n - 1} \int \exp\left\{ -\frac{1}{2} tr\left[\boldsymbol{\Delta}^{-1}(\boldsymbol{Z}_i - \boldsymbol{V}^\top \boldsymbol{u})(\boldsymbol{Z}_i - \boldsymbol{V}^\top \boldsymbol{u})^\top\right] \right\} \\ \times \exp\left\{ -\frac{1}{2} tr\left[\boldsymbol{\Gamma}^{-1}(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{r})(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{r})^\top\right] \right\} \end{aligned}$$

$$
\begin{aligned}
&\times |\widetilde{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} tr\left[\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{u}-\widetilde{\boldsymbol{m}}_U)(\boldsymbol{u}-\widetilde{\boldsymbol{m}}_U)^\top\right]\right\} \\
&\times |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} tr\left[\phi_3^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{r}-\boldsymbol{m}_R)(\boldsymbol{r}-\boldsymbol{m}_R)^\top\right]\right\} \mathrm{d}(\boldsymbol{u},\boldsymbol{r}) \\
&\times f_d\left(\boldsymbol{u}_i;\widetilde{\boldsymbol{m}}_U,\widetilde{\boldsymbol{\Sigma}}\right) f_d\left(\boldsymbol{r}_i;\boldsymbol{m}_R,\phi_3\boldsymbol{\Sigma}\right) \\
+\sum_{k=1}^{k_n} \frac{n_k}{M_H+n-1} &\exp\left\{-\frac{1}{2} tr\left[\boldsymbol{\Delta}^{-1}(\boldsymbol{Z}_i-\boldsymbol{V}^\top\boldsymbol{u}_k^*)(\boldsymbol{Z}_i-\boldsymbol{V}^\top\boldsymbol{u}_k^*)^\top\right]\right\} \\
&\times \exp\left\{-\frac{1}{2} tr\left[\boldsymbol{\Gamma}^{-1}(\boldsymbol{D}_i-\boldsymbol{W}^\top\boldsymbol{r}_k^*)(\boldsymbol{D}_i-\boldsymbol{W}^\top\boldsymbol{r}_k^*)^\top\right]\right\} \delta_{(\boldsymbol{u}_k^*,\boldsymbol{r}_k^*)}(\boldsymbol{u}_i,\boldsymbol{r}_i),
\end{aligned}
$$

where $f_d\left(\boldsymbol{x};\boldsymbol{m},\boldsymbol{S}\right)$ is used to denote the density of a d-dimensional multivariate Normal random vector, with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{S}$. We observe that a convenient simplification is obtained by assuming independence between columns, case that allows us to split a matrix variate normal into the product of two independent multivariate normals. As a result, when $\phi_2 = 0$, $\boldsymbol{U}_i$ is independent of $\boldsymbol{R}_i$, and we can write

$$
\begin{aligned}
Pr\big((\boldsymbol{U}_i,\boldsymbol{R}_i) \text{ is new } |\ldots\big) \propto &\frac{M_H}{M_H+n-1}(2\pi)^{-p}|\boldsymbol{\Delta}\boldsymbol{\Gamma}|^{-1/2}|\boldsymbol{\Sigma}\boldsymbol{\Sigma}|^{-1/2}\phi_3^{-1/2}\phi_1^{-1/2} \\
&\times |(\phi_1^{-1}\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)\left(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}+\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top\right)|^{-1/2} \\
&\times \exp\Big\{-\frac{1}{2} tr\big(\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top+\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i\boldsymbol{D}_i^\top \\
&+\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R\boldsymbol{m}_R^\top+\phi_1^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_U\boldsymbol{m}_U^\top\big)\Big\} \\
&-\frac{1}{2} tr[(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i+\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R) \\
&\quad\times(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i+\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R)^\top \\
&\quad\times(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top+\phi_3^{-1}\boldsymbol{\Sigma}^{-1})^{-1}] \\
&-\frac{1}{2} tr[(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i+\phi_1^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_U) \\
&\quad\times(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i+\phi_1^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_U)^\top \\
&\quad\times(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top+\phi_1^{-1}\boldsymbol{\Sigma}^{-1})^{-1}]\}
\end{aligned} \tag{2.6}
$$

Moreover, new values for $(\boldsymbol{U}_i,\boldsymbol{R}_i)$ can be sampled from the following independent distributions

$$
\begin{aligned}
\boldsymbol{U}_i|rest \overset{\text{ind}}{\sim} N_d\Big((\phi_1^{-1}\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}&(\phi_1^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_U+\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i); \\
&(\phi_1^{-1}\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}\Big),
\end{aligned}
$$

$$\boldsymbol{R}_i|rest \overset{\text{ind}}{\sim} N_d\Big((\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i); \\ (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}\Big).$$

Furthermore, the probability that $(\boldsymbol{U}_i, \boldsymbol{R}_i)$ coincides with an already observed value, that is a value that appears in $(\boldsymbol{U}_{-i}, \boldsymbol{R}_{-i})$, is

$$\begin{aligned} Pr\big((\boldsymbol{U}_i, \boldsymbol{R}_i) = (\boldsymbol{u}_r^*, \boldsymbol{r}_k^*)|\dots\big) \propto & \frac{n_k}{M_H + n - 1}(2\pi)^{-p}|\boldsymbol{\Delta}|^{-1/2}|\boldsymbol{\Sigma}|^{-1/2} \\ & \times \exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Delta}^{-1}(\boldsymbol{Z}_i - \boldsymbol{V}^\top\boldsymbol{u}_r^*)(\boldsymbol{Z}_i - \boldsymbol{V}^\top\boldsymbol{u}_r^*)^\top\right]\right\} \\ & \times \exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Gamma}^{-1}(\boldsymbol{D}_i - \boldsymbol{W}^\top\boldsymbol{r}_k^*)(\boldsymbol{D}_i - \boldsymbol{W}^\top\boldsymbol{r}_k^*)^\top\right]\right\}, \end{aligned} \tag{2.7}$$

where $(\boldsymbol{u}_r^*, \boldsymbol{r}_k^*)$ are the elements of $(\boldsymbol{U}, \boldsymbol{R})$ in the k-th cluster and n_k is the cardinality of the same cluster. Once the probabilities in (2.6) and (2.7) have been computed up to a proportionality constant, their exact value can be recovered by exploiting the fact that they sum up to 1.

More details are provided in Appendix 2.A. We can derive in a similar way the full conditional distributions for $(\boldsymbol{V}_j, \boldsymbol{W}_j)$. Finally, for the hyperparameters σ^2 and τ^2, we specify the following a priori distributions

$$\sigma^2 \sim IG(\alpha_\sigma, \beta_\sigma),$$
$$\tau^2 \sim IG(\alpha_\tau, \beta_\tau),$$

where IG stands for the *Inverse-Gamma* distribution. The corresponding full conditionals are given by

$$\sigma^2|rest \sim IG\left(\alpha_\sigma + \frac{np}{2}, \beta_\sigma + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}(Z_{ij} - \boldsymbol{U}_i^\top\boldsymbol{V}_j)^2\right),$$
$$\tau^2|rest \sim IG\left(\alpha_\tau + \frac{np}{2}, \beta_\tau + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}(D_{ij} - \boldsymbol{R}_i^\top\boldsymbol{W}_j)^2\right).$$

It is well known that algorithms based on Pólya urn schemes can suffer of slow mixing (see, e.g., the discussion in Ishwaran and James, 2001). A solution to deal with this problem is the introduction of an acceleration step that consists in updating the distinct values of the latent variables from their full conditional distributions.

We describe the acceleration step we introduce in the algorithm by focusing on the update of $(\boldsymbol{U}_k^*, \boldsymbol{R}_k^*)$, the update of $(\boldsymbol{V}_l^*, \boldsymbol{W}_l^*)$ being analogous. We let $C_k = \{i \in \{1, \dots, n\} : (\boldsymbol{U}_i, \boldsymbol{R}_i) = (\boldsymbol{U}_k^*, \boldsymbol{R}_k^*)\}$ and obtain

$$\begin{aligned} Pr((\boldsymbol{U}_k^*, \boldsymbol{R}_k^*) \in (\mathrm{d}\boldsymbol{u}, \mathrm{d}\boldsymbol{r}) | \boldsymbol{Z}, \boldsymbol{D}, (\boldsymbol{V}, \boldsymbol{W})) \propto H_0(\mathrm{d}\boldsymbol{u}, \mathrm{d}\boldsymbol{r}) \\ \times \prod_{i \in C_k} \exp\left\{ -\frac{1}{2} tr\left[\boldsymbol{\Delta}^{-1} (\boldsymbol{Z}_i - \boldsymbol{V}^\top \boldsymbol{u})(\boldsymbol{Z}_i - \boldsymbol{V}^\top \boldsymbol{u})^\top \right] \right\} \\ \times \exp\left\{ -\frac{1}{2} tr\left[\boldsymbol{\Gamma}^{-1} (\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{r})(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{r})^\top \right] \right\}. \end{aligned}$$

Sampling from this distribution is straightforward because it can be decomposed into the product of two d-dimensional Normal distributions. We let $\bar{\boldsymbol{Z}}_k$ and $\bar{\boldsymbol{D}}_k$ be respectively the mean of the observations of the variable $\boldsymbol{Z}$ and $\boldsymbol{D}$ in the k-th cluster and, by considering again the simplifying assumption $\phi_2 = 0$, we obtain

$$\begin{aligned} \boldsymbol{U}_k^* | rest \overset{\text{ind}}{\sim} N_d\big((\phi_1^{-1} \boldsymbol{\Sigma}^{-1} + n_k \boldsymbol{V} \boldsymbol{\Delta}^{-1} \boldsymbol{V}^\top)^{-1} (\phi_1^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{m}_U + n_k \boldsymbol{V} \boldsymbol{\Delta}^{-1} \bar{\boldsymbol{Z}}_k); \\ (\phi_1^{-1} \boldsymbol{\Sigma}^{-1} + n_k \boldsymbol{V} \boldsymbol{\Delta}^{-1} \boldsymbol{V}^\top)^{-1} \big) \\ \boldsymbol{R}_k^* | rest \sim N_d\big((\phi_3^{-1} \boldsymbol{\Sigma}^{-1} + n_k \boldsymbol{W} \boldsymbol{\Gamma}^{-1} \boldsymbol{W}^\top)^{-1} (\phi_3^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{m}_R + n_k \boldsymbol{W} \boldsymbol{\Gamma}^{-1} \bar{\boldsymbol{D}}_k); \\ (\phi_3^{-1} \boldsymbol{\Sigma}^{-1} + n_k \boldsymbol{W} \boldsymbol{\Gamma}^{-1} \boldsymbol{W}^\top)^{-1} \big) \end{aligned}$$

for more details see Appendix A.1.

Algorithm 1 summarizes the steps of the gibbs sampler we are now in the position to define, by combining the sequential updates of the model parameters according to the full conditionals we derived, with the acceleration step.

Algorithm 1 Gibbs sampling with the acceleration step

set admissible initial values for the latent vectors $\boldsymbol{U}^{(0)}, \boldsymbol{V}^{(0)}, \boldsymbol{R}^{(0)}$ and $\boldsymbol{W}^{(0)}$
for each iteration $r = 1 \ldots, R$ **do**:
  for each $i = 1, \ldots, n$ and $j = 1, \ldots, p$ **do**:
    sample $(\boldsymbol{U}_i, \boldsymbol{R}_i)^{(r)}$ from

$$\frac{M_H}{M_H + n - 1} L(\boldsymbol{U}_i^{(r)}, \boldsymbol{R}_i^{(r)}, \boldsymbol{W}, \boldsymbol{V}; \boldsymbol{Z}_i, \boldsymbol{D}_i) H_0(\cdot) + \sum_{k=1}^{k_n^{(r)}} \frac{n_k^{(r)}}{M_H + n - 1} L(\boldsymbol{U}_i^{(r)}, \boldsymbol{R}_i^{(r)}, \boldsymbol{W}, \boldsymbol{V}; \boldsymbol{Z}_i, \boldsymbol{D}_i) \delta_{(\boldsymbol{U}_i^*, \boldsymbol{R}_i^*)^{(r)}}(\cdot)$$

    sample $(\boldsymbol{V}_j, \boldsymbol{W}_j)^{(r)}$ from

$$\frac{M_H}{M_H + p - 1} L(\boldsymbol{V}_j^{(r)}, \boldsymbol{W}_j^{(r)}, \boldsymbol{U}, \boldsymbol{R}; \boldsymbol{Z}_j, \boldsymbol{D}_j) F_0(\cdot) + \sum_{k=1}^{\kappa_D^{(r)}} \frac{n_k^{(r)}}{M_H + p - 1} L(\boldsymbol{V}_j^{(r)}, \boldsymbol{W}_j^{(r)}, \boldsymbol{U}, \boldsymbol{R}; \boldsymbol{Z}_j, \boldsymbol{D}_j) \delta_{(\boldsymbol{V}_j^*, \boldsymbol{W}_j^*)^{(r)}}(\cdot)$$

  end for
  set $(\boldsymbol{U}^*, \boldsymbol{R}^*)^{(r)} = \left((\boldsymbol{U}_1^*, \boldsymbol{R}_1^*)^{(r)}, \ldots, (\boldsymbol{U}_{k_n^{(r)}}^*, \boldsymbol{R}_{k_n^{(r)}}^*)^{(r)}\right)$ be the vector of distinct parameters in $(\boldsymbol{U}, \boldsymbol{R})^{(r)}$
  set $(\boldsymbol{V}^*, \boldsymbol{W}^*)^{(r)} = \left((\boldsymbol{V}_1^*, \boldsymbol{W}_1^*)^{(r)}, \ldots, (\boldsymbol{V}_{\kappa_p^{(r)}}^*, \boldsymbol{W}_{\kappa_p^{(r)}}^*)^{(r)}\right)$ be the vector of distinct parameters in $(\boldsymbol{V}, \boldsymbol{W})^{(r)}$
  for each $k = 1, \ldots, k_n^{(r)}$ **do**:
    let $C_k^{(r)}$ be the set of indexes i such that $(\boldsymbol{U}_i, \boldsymbol{R}_i)^{(r)} = (\boldsymbol{U}_k^*, \boldsymbol{R}_k^*)^{(r)}$;
    update $(\boldsymbol{U}_k^*, \boldsymbol{R}_k^*)^{(r)}$ from

$$Pr\left((\boldsymbol{U}_k^*, \boldsymbol{R}_k^*) \in (\mathrm{d}\boldsymbol{U}, \mathrm{d}\boldsymbol{R}) | \ldots\right) \propto H_0(\mathrm{d}\boldsymbol{U}, \mathrm{d}\boldsymbol{R}) \prod_{i \in C_k^{(r)}} L(\boldsymbol{U}, \boldsymbol{R}, \boldsymbol{W}, \boldsymbol{V}; \boldsymbol{Z}_i, \boldsymbol{D}_i)$$

  end for
  for each $l = 1, \ldots, \kappa_p^{(r)}$ **do**:
    let $\zeta_l^{(r)}$ be the set of indexes j such that $(\boldsymbol{V}_j, \boldsymbol{W}_j)^{(r)} = (\boldsymbol{V}_l^*, \boldsymbol{W}_l^*)^{(r)}$;
    update $(\boldsymbol{V}_l^*, \boldsymbol{W}_l^*)^{(r)}$ from

$$Pr\left((\boldsymbol{V}_l^*, \boldsymbol{W}_l^*) \in (\mathrm{d}\boldsymbol{V}, \mathrm{d}\boldsymbol{W}) | \ldots\right) \propto F_0(\mathrm{d}\boldsymbol{V}, \mathrm{d}\boldsymbol{W}) \prod_{j \in \zeta_l^{(r)}} L(\boldsymbol{U}, \boldsymbol{R}, \boldsymbol{W}, \boldsymbol{V}; \boldsymbol{Z}_j, \boldsymbol{D}_j)$$

  end for
  for each $i = 1, \ldots, n$ and $j = 1, \ldots, p$ **do**:
    sample $Z_{ij}^{(r)}$ and $D_{ij}^{(r)}$ respectively from

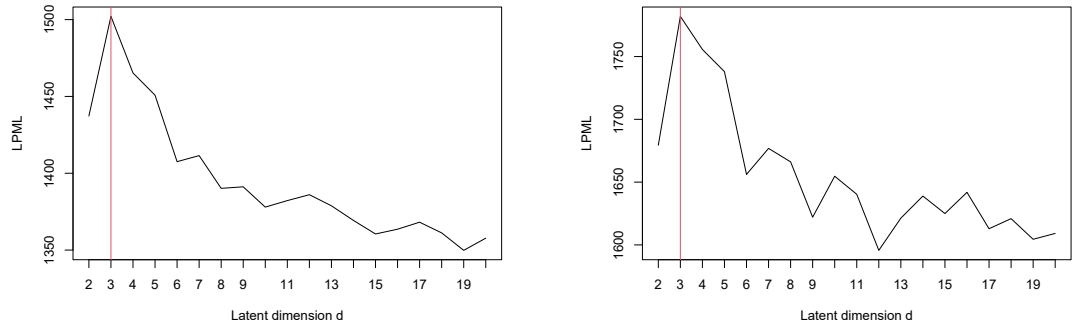
$$N(\boldsymbol{V}_j^{(r)\top} \boldsymbol{U}_i^{(r)}; \sigma^2) \mathbb{1}_{(\gamma_{j,Y_{ij}-1}, \gamma_{j,Y_{ij}}]}(Z_{ij}^{(r)})^{\mathbb{1}_{\{D_{ij}>0\}}};$$
$$N(\boldsymbol{W}_j^{(r)\top} \boldsymbol{R}_i^{(r)}; \tau^2) \left(\delta_{ij} \mathbb{1}_{\{D_{ij} \geq 0\}} + (1 - \delta_{ij}) \mathbb{1}_{\{D_{ij}<0\}}\right).$$

    where $N(\cdot, \cdot)$ is a Normal distribution.
  end for
end for
end

2.5 Simulation studies

We investigate the performance of model (2.2) in co-clustering individuals and items in synthetic data simulated from different scenarios. Our study consist in two parts: the first one aims to study the sensitivity of the model to the size of the data, the second one is designed to study how the ability of the model to detect the correct cluster is affected by the amount and the nature of censored observations. In both cases data are generated by specifying finite mixture models for the matrices $(\boldsymbol{U}_i\,;\boldsymbol{R}_i)$ and $(\boldsymbol{V}_j\,;\boldsymbol{W}_j)$, and by using the parametric part of the model in (2.2) as the conditional distribution of the observations (Y_{ij},δ_{ij}) given $(\boldsymbol{U}_i\,;\boldsymbol{R}_i)$ and $(\boldsymbol{V}_j\,;\boldsymbol{W}_j)$. In the first part of the study, we simulate ordinal observations so to mimic the movie ratings data that will be considered in the illustration of Section 2.6.2. Specifically, we assume that the ratings $\boldsymbol{Y}_i$ can take values in $\{1,2,3\}$ and consider a scenario characterized by three types of users and three types of movies. Thus inducing nine bivariate clusters, or co-clusters. Finally, we censor 5% of the observations, picked at random among the records corresponding to the lowest ratings, thus imitating a mechanism of informative censoring. In the second part of the study, we simulate binary observations thus reproducing a type of data similar to politicians votes data of the illustration that will be presented in Section 2.6.1. Specifically, we assume observations $\boldsymbol{Y}_i$ take values in $\{0,1\}$, or equivalently that votes can be either "No" or "Yes", and we consider a scenario characterized be three main parties and three types of votes, thus inducing nine bivariate clusters. In the first part we consider different values for n and p; in the second one instead we let the amount of censored observations and the type of censoring to vary. For each considered scenario, we independently generate and analyse 100 datasets. Example of generated datasets for the two types of data are represented in Figures 2.13, 2.14 and 2.15.

The first step we need to take in order to implement model 2.2 consists in selecting the cardinality of the latent dimension d. We propose to choose d on a case-by-case basis by comparing the predictive ability of models with different values of d. For this reason we decide to implement and evaluate the *LogPseudo Marginal Likelihood* (LPML, Geisser and Eddy, 1979) for a set of values of d. Rather than repeating this step for all the generated datasets, we consider a simple scenario for both ordinal and binary data, with $n = p = 50$ and 5% of censored observations, and we evaluate the LPML for a range of models corresponding to values of d in $\{2,\dots,20\}$. The results (see Figures 2.1a and 2.1b) suggest that the best predictive ability is achieved with $d = 3$, for both ordinal and binary data. Hence, the value 3 for the latent dimension d is used throughout this section, for all the scenarios.



(a) *Ordinal data.* ***(b)*** *Binary data.*

Figure 2.1: *LPML for different values of the latent dimension d.*

2.5.1 Part one

We investigate the performance of our model when the dimension of the data changes. We simulate ordinal data with dimension $(n, p) \in \{(50, 50), (100, 100), (200, 200)\}$. As expected for any Gibbs sampling algorithm, the computational time increases more than linearly with the data dimension np. Table 2.1 shows the impact of data size on the efficiency of the computational process.

Computational Cost						
Size	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
50x50	682	715	725	718	732	735
100x100	3501	3541	3555	3554	3569	3595
200x200	27588	28302	28698	28874	29674	30525

Table 2.1: *Computational cost of the MCMC algorithm, in seconds per 1000 iterations, for datasets of varying size.*

The results of this simulation are compared with the use of the Adjusted Rand Index (ARI, Rand, 1971; Hubert and Arabie, 1985; Vinh et al., 2009). Given the specific nature of the clustering problem we are considering, we exploit the ARI to compare true and estimated bivariate partitions, which are induced, respectively, by true and estimated row and column partitions. To stress this fact, we refer to this index as to bivariate ARI. Figure 2.2 shows the bivariate ARI when the data dimension changes. The performance of the model appears rather stable across different dimensions of the dataset. Nonetheless, it can be observed that the median value of the bivariate ARI is slightly larger for larger

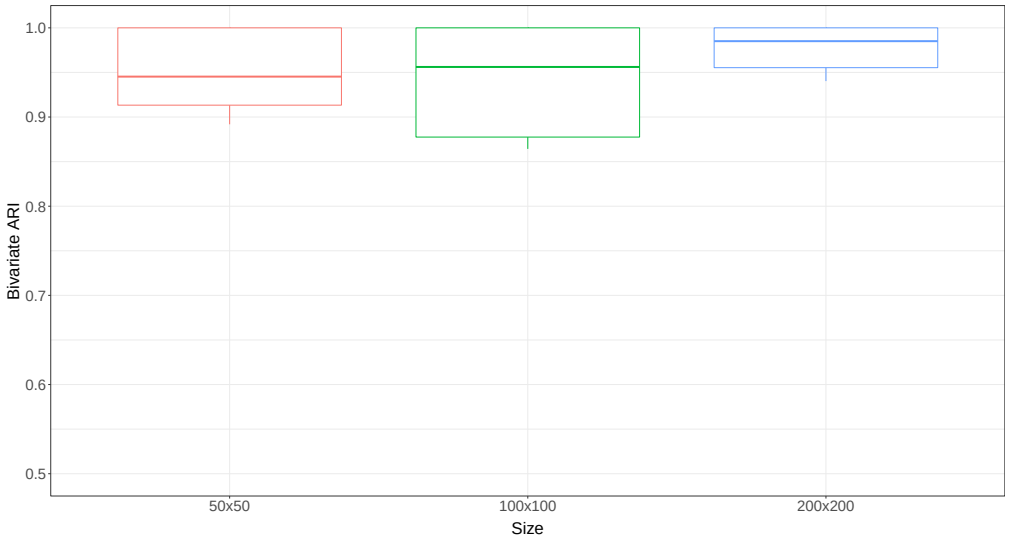


Figure 2.2: *Bivariate ARI boxplot comparison over different sizes, for ordinal data. Results are based on 100 replicates.*

dimensional datasets, thus indicating that more data helps in recovering the true latent clusters.

2.5.2 Part two

Next, we examine the performance of our model when the percentage and type of censoring vary. Binary data are generated by setting $n = p = 50$. Once a dataset is formed, a portion consisting of 5%, or 15%, of the observations is censored. Censoring are introduced by following two different principles: when the observations to be censored are picked uniformly at random from the np data entries, we say the censoring are non-informative; when the observations to be censored are picked uniformly at random from the entries that are equal to 0, then we say the censorings are informative. Datasets are analysed by means of model (2.2). As for Part 1, we assess the performance of our model by comparing the true and the estimated bivariate partitions by means of the bivariate ARI. The results of our study are presented in Figure 2.3, which shows that the latent clusters are better identified in the settings with 5% of missing values rather than when the percentage of censored entries is increased to 15%. Moreover, the scenarios with informative censoring tend to lead to larger bivariate ARI values, thus showing that our model is able to exploit this additional source of information.

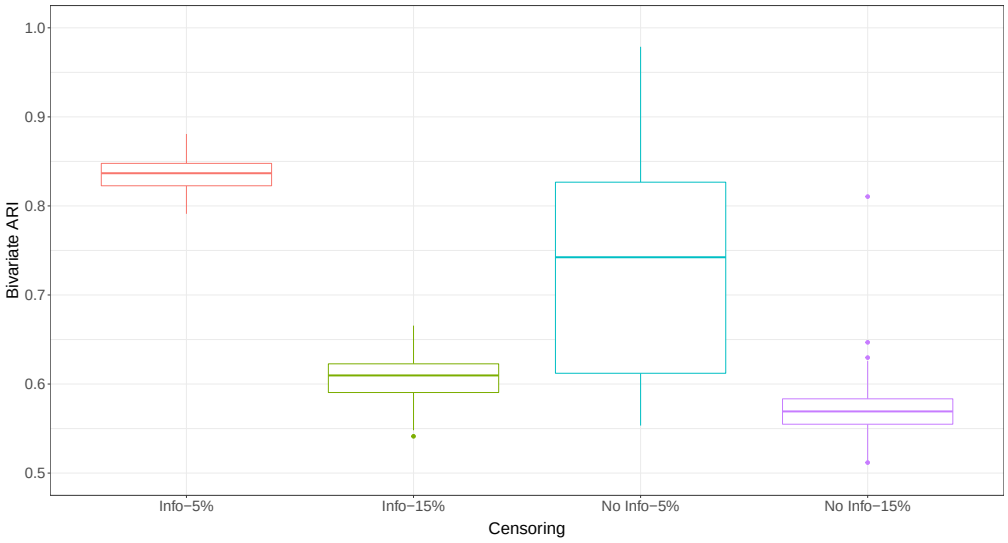


Figure 2.3: *Bivariate ARI boxplot comparison over different censoring types, for binary data. Results are based on 100 replicates.*

2.6 Real data application

2.6.1 U.S. Senators Data

Political data is of significant interest for examining how politicians vote within their respective parties, whether they align with party lines or exhibit distinct patterns. Such analysis can reveal, for example, contrasting behaviors within a single political party. The dataset we consider was retrieved from Voteview[1] and comprises records of 100 U.S. senators and their voting decisions in 35 voting sessions held between May 2, 2022, and May 16, 2022. The data is visually represented in Figure 2.4. It is worth noting that 3.37% of the entries are missing, thus referring to sessions in which the politician did not vote. The specific nature of this dataset lets us anticipate the presence of polarized clusters corresponding to party affiliations. Furthermore, it is reasonable to assume that censored observations hold valuable information, given that the choice to abstain from voting in a particular session is frequently a political statement in its own right. As for the simulated data considered in Section 2.5, we resort to the LPML to set the value of the latent dimension d. Our analysis suggests that the best predictive performance is achieved when setting $d = 3$, as demonstrated in Figure 2.5.

[1] https://voteview.com/

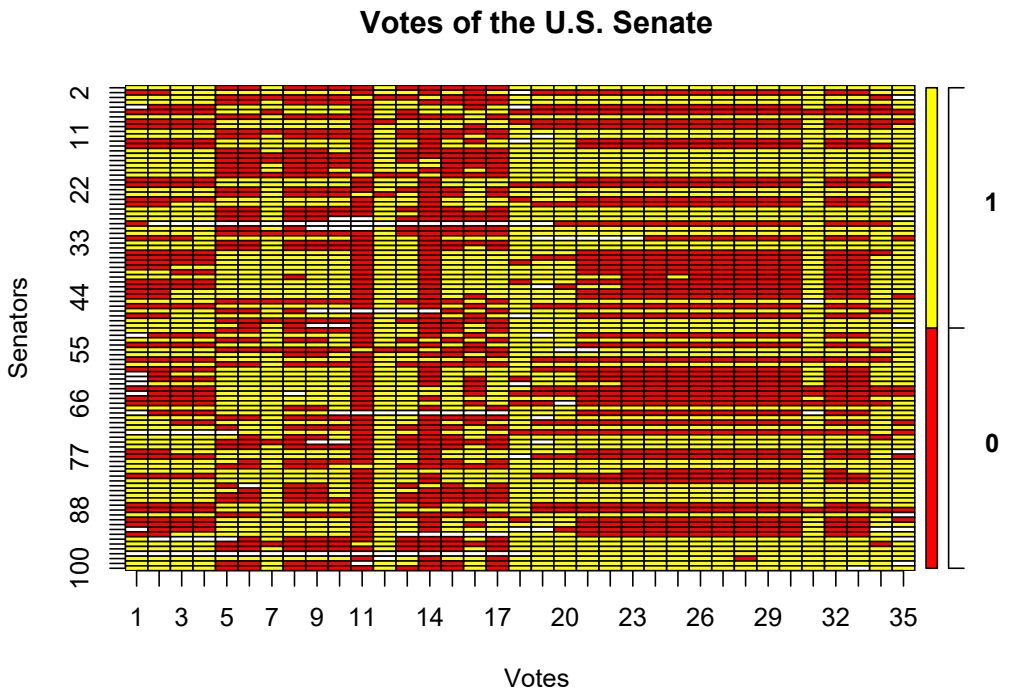


Figure 2.4: *Representation of the votes of the U.S. Senators, the right legend represents the modalities (i.e., "No"=0, "Yes"=1). White cells indicate missing votes.*

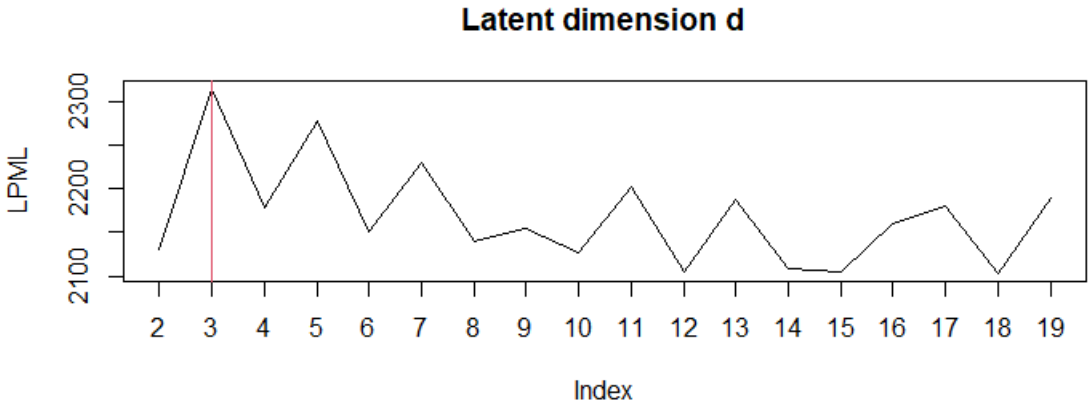


Figure 2.5: *LPML to evaluate the latent dimension d for the dataset U.S. Senate.*

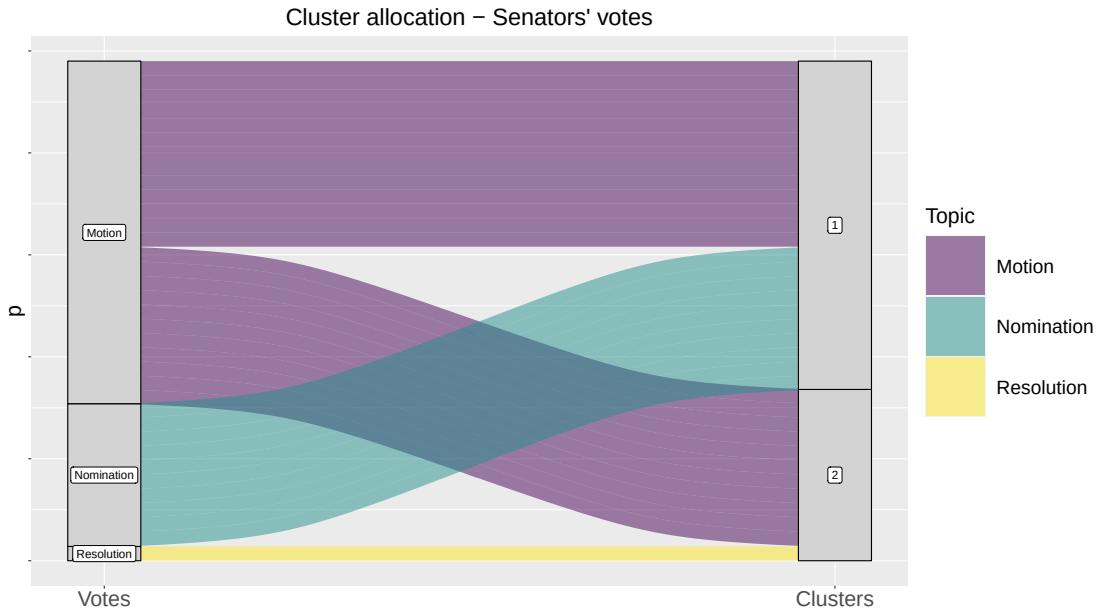


Figure 2.6: *Alluvial diagram comparing topic of voting sessions and identified clusters for the U.S. Senators' votes.*

The results of our cluster analysis for votes and Senators are displayed in Figures 2.6 and 2.7, respectively. These alluvial diagrams provide interesting insight by comparing the identified clusters with available information on type of voting sessions and party affiliations. Votes are clustered into two groups while the topics covered during the voting sessions can be classified into: nominations to promote a new person for a specific position, motions and resolutions. Our analysis result identifies one large cluster with all the nominations and just over half of the motions, and a smaller one containing the prevalence of motions and all the resolutions. This outcome may suggest that similarities exist between the way senators voted on certain motions and nominations. Furthermore, as expected, the clusters for Senators appear to be highly polarized towards their respective party affiliations, it is interesting to observe that Independent senators are clustered together with senators from the Democratic party. We also note that there exists a third smaller cluster suggesting that the votes of nine Republican senators are aligned with the preferences of five Democratic politicians.

We compared the results our analysis with those obtained by using the `R` package *biclustermd* (Reisner et al., 2019). It is important to note that this alternative approach to co-clustering problems requires the number of clusters to be set, and that it involves to impute the missing values with the mean of the data, thus ignoring the fact that, as we

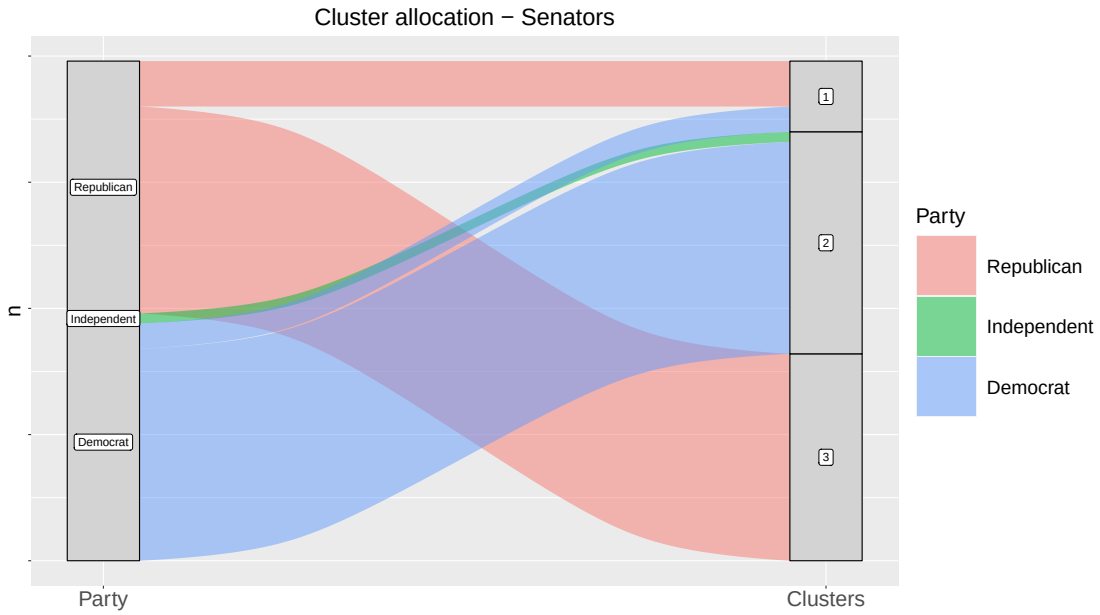


Figure 2.7: *Alluvial diagram comparing party affiliation and identified clusters for the U.S. Senators.*

discussed, the censoring in the U.S. senators data might actually contain useful information.

The clusters we found with *biclustermd* are rather coherent with our results. For the votes, only one record was clustered in a different group, as shown in Figure 2.17. As for the senators' clusters, some differences were apparent, especially referring to the smallest group. Our method captured the similarities between some Democrats' and Republicans' preferences, while the *biclustermd* identified a small cluster with only Republicans, as shown in Figure 2.16. More information about the results can be found in Appendix 2.B.

2.6.2 Movielens Data

In the last decades, the widespread use of movie platforms has led to the collection of a large amount of data pertaining to both users and movies. Movielens[2] is a dataset available in the `R` package *dslabs* (Irizarry and Gill, 2021). To contain the amount of missing information, we selected a subset of the original dataset that includes 60 users and 28 movies, resulting in 17.98% of missing data (Figure 2.9). The ratings take values in $\{0.5, 1, \ldots, 5\}$. Figure 2.8 shows the counts per ratings in this dataset. The distribution of the ratings appears rather concentrated around values in the range $[4, 5]$, possibly because the dataset under

[2] http://files.grouplens.org/datasets/movielens/ml-latest-small.zip

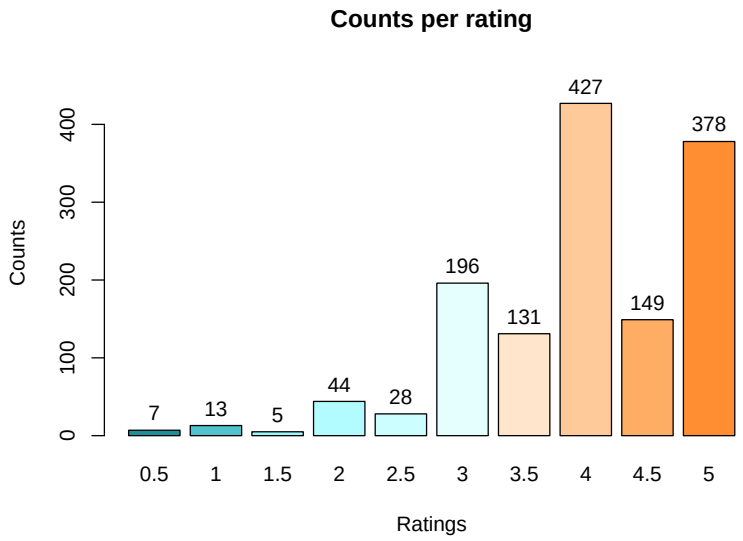


Figure 2.8: *Counts per rating for the Movielens dataset.*

analysis consists of very well-known and acclaimed movies. The frequencies corresponding to half-points, i.e., $\{0.5, 1.5, 2.5, 3.5, 4.5\}$, are lower than those for full rating points, i.e., $\{1, 2, 3, 4, 5\}$. However, this aspect is not critical to our model, as the cutoffs involved in the analysis help representing all the modalities of the ordinal data despite of the frequencies. The flexibility of our fully Bayesian nonparametric specification allows us to effectively handle this specific feature of the data.

By resorting to the LPML, we determined that the optimal predictive performance is achieved with a latent dimension of $d = 2$, as illustrated in Figure 2.10. Notably, the movie cluster consists of four distinct groups, which are arbitrarily denoted as clusters 1, 2, 3 and 4. Cluster 1 represents the genre "Drama/Thriller", the only movie that, based on its tags, appears surprisingly grouped with this cluster is "Toy Story". Cluster 2 groups "Adventure / Action" movies and its composition appears rather homogeneous. The movies composing cluster 3 appear to be characterized by a plot involving a journey to be completed to resolve the misadventures of their characters. The last cluster consists of more satirical movies and its composition seems rather homogeneous. Table 2.2 reports the titles of the 28 movies in the dataset, their genres, and their cluster allocation.

Our analysis has identified six distinct user clusters with frequencies $\{8, 31, 4, 7, 9, 1\}$, arbitrarily denoted as clusters 1, 2, 3, 4, 5, and 6. To protect user privacy, no individual information was made available. For this reason, in order to gain some insight on the

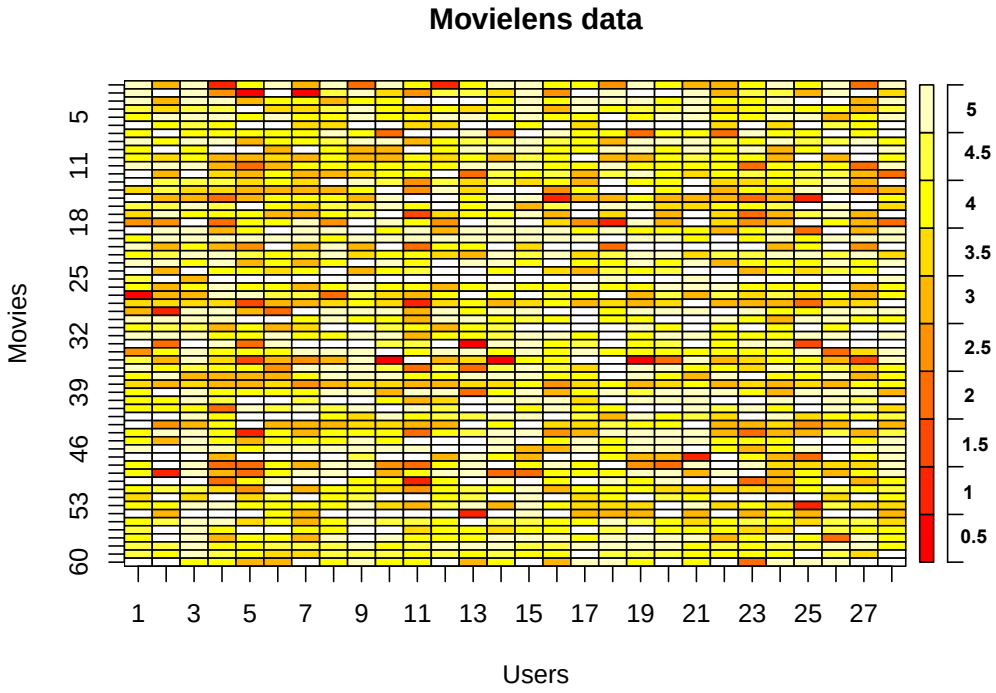


Figure 2.9: *Visual representation of Movielens data. White cells indicate not available ratings.*

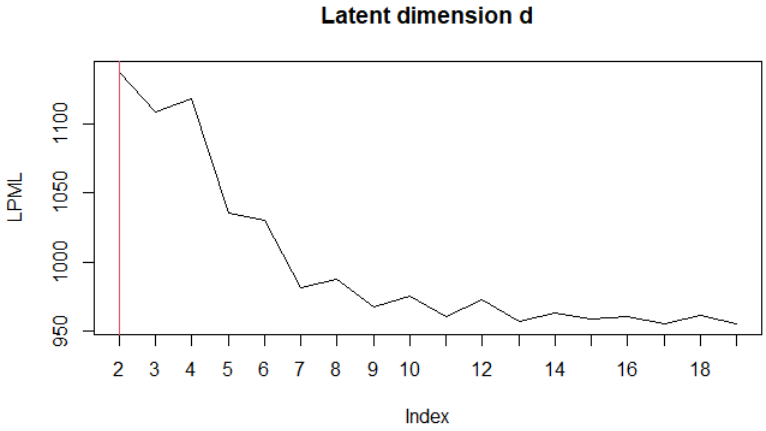


Figure 2.10: *LPML to evaluate the latent dimension d for the dataset Movielens.*

Title	Genre	Cluster
"Seven"	Mystery|Thriller	1
"Braveheart"	Action|Drama|War	2
"Pulp Fiction"	Comedy|Crime|Drama|Thriller	1
"Forrest Gump"	Comedy|Drama|Romance|War	2
"Speed"	Action|Romance|Thriller	2
"The Fugitive"	Thriller	3
"Jurassic Park"	Action|Adventure|Sci-Fi|Thriller	2
"The Silence of the Lambs"	Crime|Horror|Thriller	1
"The Shawshank Redemption"	Crime|Drama	1
"Star Wars: Ep. VI"	Action|Adventure|Sci-Fi	2
"Men in Black"	Action|Comedy|Sci-Fi	2
"The Sixth Sense"	Drama|Horror|Mystery	1
"American Beauty"	Drama|Romance	1
"Star Wars: Ep. IV"	Action|Adventure|Sci-Fi	2
"The Godfather"	Crime|Drama	1
"Die Hard"	Action|Crime|Thriller	2
"E.T. the Extra-Terrestrial"	Children|Drama|Sci-Fi	3
"Monty Python and the Holy Grail"	Adventure|Comedy|Fantasy	4
"Star Wars: Ep. V"	Action|Adventure|Sci-Fi	2
"Raiders of the Lost Ark"	Action|Adventure	2
"Goodfellas"	Crime|Drama	4
"The Terminator"	Action|Sci-Fi|Thriller	2
"Groundhog Day"	Comedy|Fantasy|Romance	3
"Back to the Future"	Adventure|Comedy|Sci-Fi	2
"Ferris Bueller's Day Off"	Comedy	3
"The Matrix"	Action|Sci-Fi|Thriller	1
"Toy Story"	Adventure|Animation|Children|Comedy|Fantasy	1
"Fargo"	Comedy|Crime|Drama|Thriller	4

Table 2.2: *Title, genre and cluster allocation for the movies in Movielens dataset.*

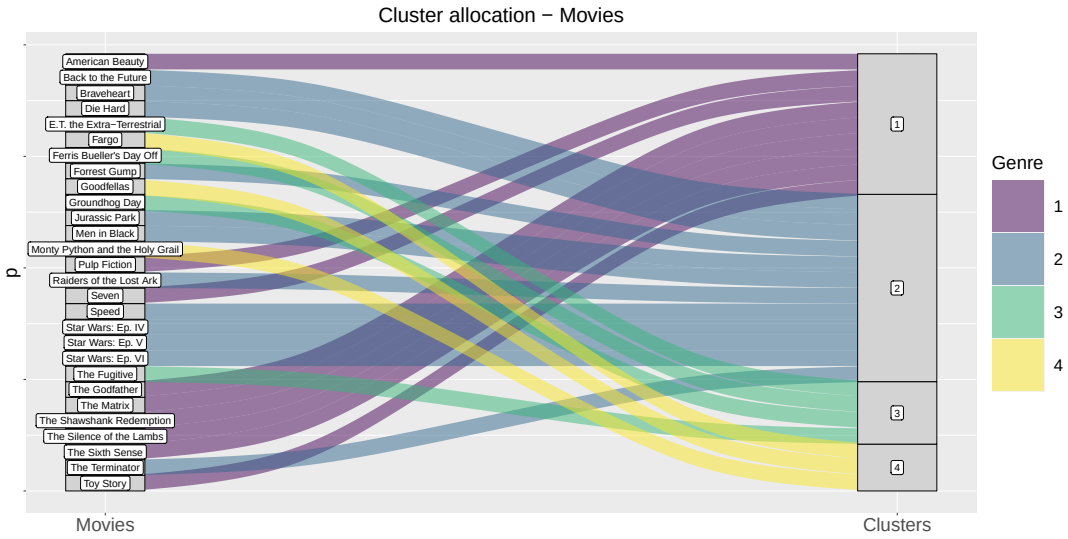


Figure 2.11: *Cluster results for movies for Movielens data.*

composition of each users cluster, we study how users in each group have rated movies in the four movie clusters, namely "Drama/Thriller", "Adventure/Action", 'Journey" and "Satire". By inspecting Figure 2.12, it is apparent that different users clusters are characterized by different degrees of appreciations for the four movie genres we identified. It is worth stressing that while a high rating is certainly away to express appreciation for a movie, a missing rating might indicate lack of interest for a movie. In the left column on Figure 2.12, we can observe the voting patterns of users across different clusters. We notice that cluster 2 and cluster 5 share some similarities in their voting patterns, as do cluster 3 and cluster 6. However, upon examining the right column, which presents the percentage of missing values for each user cluster, we can clearly discern differences in the missing data. This visualization emphasizes the importance of accounting for missing values as valuable information.

For a comparison, we analysed the same data by means of the `R` package *biclustermd*, which is specifically designed to handle informative missing values in the data matrix. The results we obtained for the users clusters are quite similar to ours, despite some discrepancies in the number of users assigned to each cluster, that is $\{8, 5, 10, 15, 9, 13\}$. On the other hand, the movie clusters exhibit a slightly different grouping, as illustrated in Figure 2.18. The first cluster represents "Drama/Thriller/Adventure" movies, the second cluster comprises "Drama/Thriller" movies, and the movie "Ferris Bueller's Day Off" appears to have been misclassified. The third cluster consists of "Comedy" movies, and the final cluster

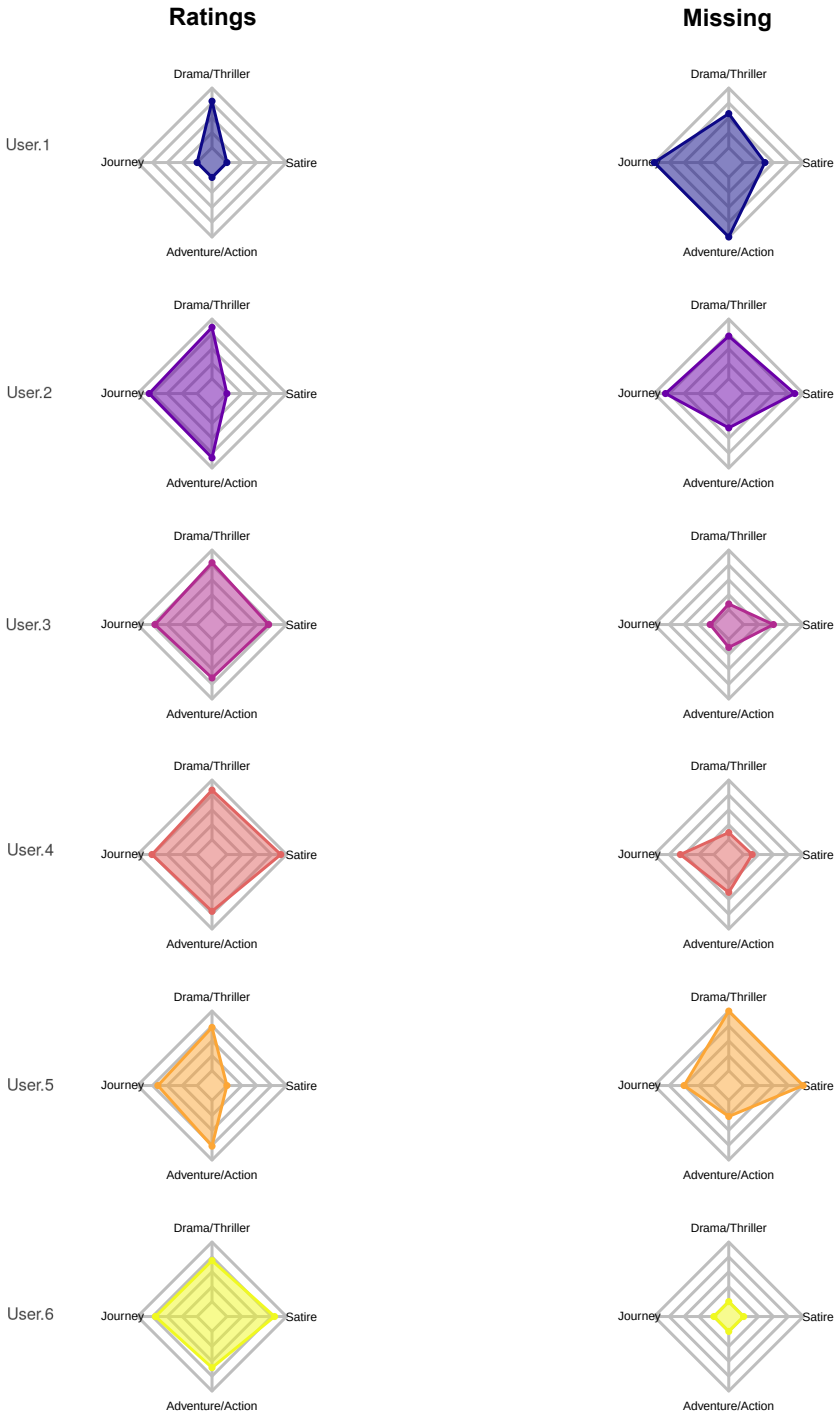


Figure 2.12: *Radar charts showing the characterization of users' clusters given the main genre of clusters' movies. Ratings are on the left and percentage of missing values on the right.*

features "Thriller/Drama/Action" movies, with "Monty Python and the Holy Grail" being an exception.

2.7 Discussion and future direction

Our proposed nonparametric Bayesian method provides a novel and efficient approach for modeling multivariate ordinal data with informative censoring. This method employs a matrix factorization model specification that can handle high dimensionality problems, while also allowing for a multivariate framework. Additionally, the use of continuous latent variables specification makes the model easy to implement and capable of handling practically any type of data.

To perform co-clustering analysis, we introduced two independent Dirichlet processes that provide a flexible approach to the problem. Our model displayed good results in both simulation studies and real data applications. When compared to an alternative approach, namely the `R` package *biclustermd*, we could observe rather similar results, with the notable difference that *biclustermd* requires the number of clusters to be specified a priori, as seen in Reisner et al. (2019).

One of the advantages of our model is its ability to exploit information from both the observations and the missing values, thus helping to find the latent cluster structure, even when the observations do not display substantial differences between clusters. Another advantage of our approach is that the number of clusters does not need to be defined a priori. Moreover, the ability to combine all the information given by the data, e.g. rates and missing data from Movielens, helps to profile the users based on their preferences while guaranteeing the protection of sensible data. This makes our method results a powerful tool in recommendation system settings.

However, there are some limitations to our model, such as its sensitivity to the specification of the variances σ^2 and τ^2, which was left fixed in our analysis. Alternatively, a prior distribution can be assigned to σ^2 and τ^2, as specified in Section 2.4. It is also possible to set σ^2 and τ^2 as the values that maximize the predictive ability of the resulting model, by resorting to a suitable criterion, such as, eg., the LPML.

It is important to note that our model does not scale well when n and p become large, so addressing these limitations is important for the model to be applied to data with large n and p, like Netflix Prize data[3]. Additionally, high dimensional data with large portion

[3] https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data

of missing values (e.g., $> 80\%$) can be challenging, which is an important consideration for those interested in using our method. To improve the method's scalability and accuracy, we are exploring how to combine our method with the scalable multistep Monte Carlo algorithm of Ni et al. (2020) to simultaneously cluster rows and columns in large datasets. Furthermore, it is essential to note that the percentage and type of missing values play a cardinal role in the performance of the model.

Overall, our nonparametric Bayesian method provides an efficient way to model multivariate ordinal data with informative censoring. The method is flexible, and can handle practically any type of data. It also provides a useful alternative to existing methods with the goal of performing co-clustering analysis.

Bibliography

Abello, J., Pardalos, P. M., and Resende, M. G. C. (2013). *Handbook of massive data sets*, volume 4. Springer.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Bennett, J., Lanning, S., et al. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York.

Blackwell, D., MacQueen, J. B., et al. (1973). Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 1(2):353–355.

Busygin, S., Prokopyev, O., and Pardalos, P. M. (2008). Biclustering in data mining. *Computers & Operations Research*, 35(9):2964–2987.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Ismb*, volume 8, pages 93–103.

Choi, S., Ha, H., Hwang, U., Kim, C., Ha, J.-W., and Yoon, S. (2018). Reinforcement learning based recommender system using biclustering technique. *arXiv preprint arXiv:1801.05532*.

DeYoreo, M. and Kottas, A. (2018). Bayesian nonparametric modeling for multivariate ordinal regression. *Journal of Computational and Graphical Statistics*, 27(1):71–84.

DeYoreo, M. and Kottas, A. (2020). Bayesian nonparametric density regression for ordinal responses. In *Flexible Bayesian Regression Modelling*, pages 65–89. Elsevier.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Irizarry, R. A. and Gill, A. (2021). *dslabs: Data Science Labs*. R package version 0.7.4.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association*, 96(453):161–173.

Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434.

Kottas, A., Müller, P., and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14(3):610–625.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM transactions on computational biology and bioinformatics*, 1(1):24–45.

Meeds, E. and Roweis, S. (2007). Nonparametric Bayesian biclustering. Technical report, Citeseer.

Ni, Y., Müller, P., Diesendruck, M., Williamson, S., Zhu, Y., and Ji, Y. (2020). Scalable Bayesian nonparametric clustering and classification. *Journal of Computational and Graphical Statistics*, 29(1):53–65.

Porteous, I., Asuncion, A., and Welling, M. (2010). Bayesian matrix factorization with side information and Dirichlet process mixtures. In *AAAI*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Reisner, J., Pham, H., Olafsson, S., Vardeman, S. B., and Li, J. (2019). biclustermd: An R Package for Biclustering with Missing values. *R J.*, 11(2):69.

Salakhutdinov, R. and Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887.

Shan, H. and Banerjee, A. (2008). Bayesian co-clustering. In *2008 Eighth IEEE International Conference on Data Mining*, pages 530–539. IEEE.

van Dijk, B., van Rosmalen, J., and Paap, R. (2009). A Bayesian approach to two-mode clustering. Technical report.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21:511–522.

Wang, P., Domeniconi, C., Rangwala, H., and Laskey, K. B. (2012). Feature enriched non-parametric Bayesian co-clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 517–529. Springer.

Wang, P., Laskey, K. B., Domeniconi, C., and Jordan, M. I. (2011). Nonparametric Bayesian co-clustering ensembles. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 331–342. SIAM.

Webb, E. L. and Forster, J. J. (2008). Bayesian model determination for multivariate ordinal and binary data. *Computational statistics & data analysis*, 52(5):2632–2649.

Zhou, Y., Wilkinson, D., Schreiber, R., and Pan, R. (2008). Large-scale parallel collaborative filtering for the netflix prize. In *International conference on algorithmic applications in management*, pages 337–348. Springer.

Appendix

In this Appendix we report the main passages to recover the full conditional distributions specified in Section 2.4 and details on the acceleration step. We also provide additional plots concerning the computational analysis.

2.A Full conditionals

The full conditional distribution of $(\boldsymbol{U}_i, \boldsymbol{R}_i)$ is as follows

$$
\begin{aligned}
Pr((\boldsymbol{U}_i, \boldsymbol{R}_i) \text{ is new } | \ldots) \propto & \frac{M_H}{M_H + n - 1} \int \exp\Big\{ -\frac{1}{2} tr\big[\boldsymbol{\Delta}^{-1}(\boldsymbol{Z}_i - \boldsymbol{V}^\top \boldsymbol{u})(\boldsymbol{Z}_i - \boldsymbol{V}^\top \boldsymbol{u})^\top\big]\Big\} \\
& \times \exp\Big\{ -\frac{1}{2} tr\big[\boldsymbol{\Gamma}^{-1}(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{r})(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{r})^\top\big]\Big\} \\
& \times |\widetilde{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2} tr\big[\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{u} - \widetilde{\boldsymbol{m}}_U)(\boldsymbol{u} - \widetilde{\boldsymbol{m}}_U)^\top\big]\Big\} \\
& \times |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2} tr\big[\phi_3^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{r} - \boldsymbol{m}_R)(\boldsymbol{r} - \boldsymbol{m}_R)^\top\big]\Big\} \mathrm{d}(\boldsymbol{u}, \boldsymbol{r}) \\
\propto & \frac{M_H}{M_H + n - 1} \int \exp\Big\{ -\frac{1}{2} tr\big[\boldsymbol{\Delta}^{-1}(\boldsymbol{V}^\top \boldsymbol{u} - \boldsymbol{Z}_i)(\boldsymbol{V}^\top \boldsymbol{u} - \boldsymbol{Z}_i)^\top + \widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{u} - \widetilde{\boldsymbol{m}}_U)(\boldsymbol{u} - \widetilde{\boldsymbol{m}}_U)^\top\big]\Big\} \\
& \times \exp\Big\{ -\frac{1}{2} tr\big[\boldsymbol{\Gamma}^{-1}(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{r})(\boldsymbol{D}_i - \boldsymbol{W}^\top \boldsymbol{r})^\top + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{r} - \boldsymbol{m}_R)(\boldsymbol{r} - \boldsymbol{m}_R)^\top\big]\Big\} \mathrm{d}(\boldsymbol{u}, \boldsymbol{r}) \\
& \propto \frac{M_H}{M_H + n - 1} \int \exp\Big\{ -\frac{1}{2} tr\big[\boldsymbol{\Delta}^{-1}(\boldsymbol{V}^\top \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{V} - 2\boldsymbol{Z}_i \boldsymbol{u}^\top \boldsymbol{V}) + \widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{u}\boldsymbol{u}^\top - 2\widetilde{\boldsymbol{m}}_U \boldsymbol{u}^\top)\big]\Big\} \\
& \times \exp\Big\{ -\frac{1}{2} tr\big[\boldsymbol{\Gamma}^{-1}(\boldsymbol{W}^\top \boldsymbol{r}\boldsymbol{r}^\top \boldsymbol{W} - 2\boldsymbol{D}_i \boldsymbol{r}^\top \boldsymbol{W}) + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{r}\boldsymbol{r}^\top - 2\boldsymbol{m}_R \boldsymbol{r}^\top)\big]\Big\} d(\boldsymbol{u}, \boldsymbol{r}) \\
\propto & \frac{M_H}{M_H + n - 1} (2\pi)^{-p} |\boldsymbol{\Delta}\boldsymbol{\Gamma}|^{-1/2} |\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}|^{-1/2} \phi_3^{-1/2} |(\widetilde{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top) \\
& \qquad \big(\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top\big)|^{-1/2} \\
& \times \exp\Big\{ -\frac{1}{2} tr\big(\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i \boldsymbol{Z}_i^\top + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R \boldsymbol{m}_R^\top + \boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i \boldsymbol{D}_i^\top\big)\Big\} \\
& \times \exp\Big\{ -\frac{1}{2} tr[(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R)(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R)^\top \\
& \qquad (\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top + \phi_3^{-1}\boldsymbol{\Sigma}^{-1})^{-1}]\Big\} \\
& \times \int \exp\Big\{ -\frac{1}{2} tr\big[(\widetilde{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)(\boldsymbol{u} - (\widetilde{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i)) \\
& \qquad \times (\boldsymbol{u} - (\widetilde{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i))^\top\big]\Big\} \\
& \times \exp\Big\{ -\frac{1}{2} tr\big[(\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)(\boldsymbol{r} - (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i))
\end{aligned}
$$

$$\times (\boldsymbol{r} - (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i))^\top]\Big\}\mathrm{d}(\boldsymbol{u},\boldsymbol{r})$$

$$\propto \frac{M_H}{M_H+n-1}(2\pi)^{-p}|\boldsymbol{\Delta}\boldsymbol{\Gamma}|^{-1/2}|\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}|^{-1/2}\phi_3^{-1/2}|(\widetilde{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)\big(\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top\big)|^{-1/2}$$
$$\times \exp\Big\{-\frac{1}{2}tr\big(\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R\boldsymbol{m}_R^\top + \boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i\boldsymbol{D}_i^\top\big)\Big\}$$
$$\times \exp\Big\{-\frac{1}{2}tr[(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R)(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R)^\top$$
$$\times (\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top + \phi_3^{-1}\boldsymbol{\Sigma}^{-1})^{-1}]\Big\}$$
$$\times \int N_d\big(\boldsymbol{r}; (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + \boldsymbol{\Gamma}^{-1}\boldsymbol{W}\boldsymbol{D}_i); (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}\big)$$
$$\times \exp\Big\{-\frac{1}{2}tr[(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U)(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U)^\top$$
$$\times (\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top + \widetilde{\boldsymbol{\Sigma}}^{-1})^{-1} + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U\widetilde{\boldsymbol{m}}_U^\top]\Big\}\mathrm{d}\boldsymbol{r}$$

where in the integrals two multivariate Normal distributions are identified and specified as follows:

$$\boldsymbol{U}_i|\boldsymbol{R}_i,\ldots \overset{\text{ind}}{\sim} N_d\big((\widetilde{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i); (\widetilde{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}\big) = N_d^{(U|R)};$$
$$\boldsymbol{R}_i|\ldots \overset{\text{ind}}{\sim} \int N_d\big(\boldsymbol{r}; (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + \boldsymbol{\Gamma}^{-1}\boldsymbol{W}\boldsymbol{D}_i);$$
$$m(\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}\big)\exp\Big\{-\frac{1}{2}tr[(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U)$$
$$\times (\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U)^\top(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top + \widetilde{\boldsymbol{\Sigma}}^{-1})^{-1} + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U\widetilde{\boldsymbol{m}}_U^\top]\Big\}\mathrm{d}\boldsymbol{r} = N_d^{(R)}.$$

If the parameter ϕ_2 is considered equal to 0, we obtain:

$$Pr\big((\boldsymbol{U}_i,\boldsymbol{R}_i) \text{ is new } |\ldots)\big) \propto \frac{M_H}{M_H+n-1}(2\pi)^{-p}|\boldsymbol{\Delta}\boldsymbol{\Gamma}|^{-1/2}|\boldsymbol{\Sigma}\boldsymbol{\Sigma}|^{-1/2}\phi_3^{-1/2}\phi_1^{-1/2}$$
$$\times |(\phi_1^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)\big(\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top\big)|^{-1/2}$$
$$\times \exp\Big\{-\frac{1}{2}tr\big(\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top + \boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i\boldsymbol{D}_i^\top + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R\boldsymbol{m}_R^\top + \phi_1^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_U\boldsymbol{m}_U^\top\big)$$
$$-\frac{1}{2}tr[(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R)(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R)^\top(\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top + \phi_3^{-1}\boldsymbol{\Sigma}^{-1})^{-1}]$$
$$-\frac{1}{2}tr[(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i + \phi_1^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_U)(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i + \phi_1^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_U)^\top(\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top + \phi_1^{-1}\boldsymbol{\Sigma}^{-1})^{-1}]\Big\}$$

since $\phi_2 = 0$, $\boldsymbol{U}_i$ is independent from $\boldsymbol{R}_i$:

$$\boldsymbol{U}_i|rest \overset{\text{ind}}{\sim} N_d\big((\phi_1^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}(\phi_1^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_U + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}_i);$$

$$(\phi_1^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}) = N_d^{(U)};$$
$$\boldsymbol{R}_i|rest \overset{\text{ind}}{\sim} N_d\big((\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{D}_i);$$
$$(\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}) = N_d^{(R)}.$$

A.1 Acceleration step

The main passages to recover the distribution for the acceleration step for $(\boldsymbol{U}_k^*, \boldsymbol{R}_k^*)$ are the following

$$\begin{aligned}
&Pr((\boldsymbol{U}_k^*, \boldsymbol{R}_k^*) \in (\mathrm{d}\boldsymbol{u}, \mathrm{d}\boldsymbol{r})|\boldsymbol{Z}, \boldsymbol{D}, (\boldsymbol{V}, \boldsymbol{W})) \propto H_0(\mathrm{d}\boldsymbol{u}, \mathrm{d}\boldsymbol{r}) \\
&\quad\times \prod_{i\in C_k} \exp\Big\{-\frac{1}{2}tr\big[\boldsymbol{\Delta}^{-1}(\boldsymbol{Z}_i - \boldsymbol{V}^\top\boldsymbol{u})(\boldsymbol{Z}_i - \boldsymbol{V}^\top\boldsymbol{u})^\top\big]\Big\} \\
&\quad\times \exp\Big\{-\frac{1}{2}tr\big[\boldsymbol{\Gamma}^{-1}(\boldsymbol{D}_i - \boldsymbol{W}^\top\boldsymbol{r})(\boldsymbol{D}_i - \boldsymbol{W}^\top\boldsymbol{r})^\top\big]\Big\} \\
&\quad\propto |\widetilde{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}tr\big[\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{u} - \widetilde{\boldsymbol{m}}_U)(\boldsymbol{u} - \widetilde{\boldsymbol{m}}_U)^\top\big]\Big\} \\
&\quad\times |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}tr\big[\phi_3^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{r} - \boldsymbol{m}_R)(\boldsymbol{r} - \boldsymbol{m}_R)^\top\big]\Big\} \\
&\quad\times \prod_{i\in C_k} \exp\Big\{-\frac{1}{2}tr\big[\boldsymbol{\Delta}^{-1}(\boldsymbol{Z}_i - \boldsymbol{V}^\top\boldsymbol{u})(\boldsymbol{Z}_i - \boldsymbol{V}^\top\boldsymbol{u})^\top\big]\Big\} \\
&\quad\times \exp\Big\{-\frac{1}{2}tr\big[\boldsymbol{\Gamma}^{-1}(\boldsymbol{D}_i - \boldsymbol{W}^\top\boldsymbol{r})(\boldsymbol{D}_i - \boldsymbol{W}^\top\boldsymbol{r})^\top\big]\Big\} \\
&\quad\propto \exp\Big\{-\frac{1}{2}tr\big[\boldsymbol{\Delta}^{-1}\sum_{i\in C_k}(\boldsymbol{Z}_i - \boldsymbol{V}^\top\boldsymbol{u})(\boldsymbol{Z}_i - \boldsymbol{V}^\top\boldsymbol{u})^\top + \widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{u} - \widetilde{\boldsymbol{m}}_U)(\boldsymbol{u} - \widetilde{\boldsymbol{m}}_U)^\top\big]\Big\} \\
&\quad\times \exp\Big\{-\frac{1}{2}tr\big[\boldsymbol{\Gamma}^{-1}\sum_{i\in C_k}(\boldsymbol{D}_i - \boldsymbol{W}^\top\boldsymbol{r})(\boldsymbol{D}_i - \boldsymbol{W}^\top\boldsymbol{r})^\top + \phi_3^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{r} - \boldsymbol{m}_R)(\boldsymbol{r} - \boldsymbol{m}_R)^\top\big]\Big\} \\
&\quad\propto \exp\Big\{-\frac{1}{2}tr\big[(\widetilde{\boldsymbol{\Sigma}}^{-1} + n_k\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)(\boldsymbol{u} - (\widetilde{\boldsymbol{\Sigma}}^{-1} + n_k\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U + n_k\boldsymbol{V}\boldsymbol{\Delta}^{-1}\bar{\boldsymbol{Z}}_k)) \\
&\quad\times (\boldsymbol{u} - (\widetilde{\boldsymbol{\Sigma}}^{-1} + n_k\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{m}}_U + n_k\boldsymbol{V}\boldsymbol{\Delta}^{-1}\bar{\boldsymbol{Z}}_k))^\top\big]\Big\} \\
&\quad\times \exp\Big\{-\frac{1}{2}tr\big[(\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + n_k\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top) \\
&\qquad\qquad\times (\boldsymbol{r} - (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + n_k\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + n_k\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\bar{\boldsymbol{D}}_k)) \\
&\qquad\qquad\times (\boldsymbol{r} - (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + n_k\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + n_k\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\bar{\boldsymbol{D}}_k))^\top\big]\Big\}
\end{aligned}$$

where n_k is the number of observations in the k-th cluster. Let $\bar{\boldsymbol{Z}}_k$ and $\bar{\boldsymbol{D}}_k$ be respectively the mean of the observations of the variable $\boldsymbol{Z}$ and $\boldsymbol{D}$ in the k-th cluster and, ϕ_2 is set

equal to 0, we obtain:

$$\begin{aligned}\boldsymbol{U}_k^*|rest &\overset{\text{ind}}{\sim} N_d\big((\phi_1^{-1}\boldsymbol{\Sigma}^{-1} + n_k\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}(\phi_1^{-1}\boldsymbol{\Sigma}^{-1}m_U + n_k\boldsymbol{V}\boldsymbol{\Delta}^{-1}\bar{\boldsymbol{Z}}_k);\\ &\qquad (\phi_1^{-1}\boldsymbol{\Sigma}^{-1} + n_k\boldsymbol{V}\boldsymbol{\Delta}^{-1}\boldsymbol{V}^\top)^{-1}\big)\big) = N_d^{(U^*)};\\ \boldsymbol{R}_k^*|rest &\sim N_d\big((\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + n_k\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}(\phi_3^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{m}_R + n_k\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\bar{\boldsymbol{D}}_k);\\ &\qquad (\phi_3^{-1}\boldsymbol{\Sigma}^{-1} + n_k\boldsymbol{W}\boldsymbol{\Gamma}^{-1}\boldsymbol{W}^\top)^{-1}\big) = N_d^{(R^*)}.\end{aligned}$$

2.B Additional Plots

In this section, we provide additional plots for the simulation analysis and for the real data applications.

B.1 Simulated data

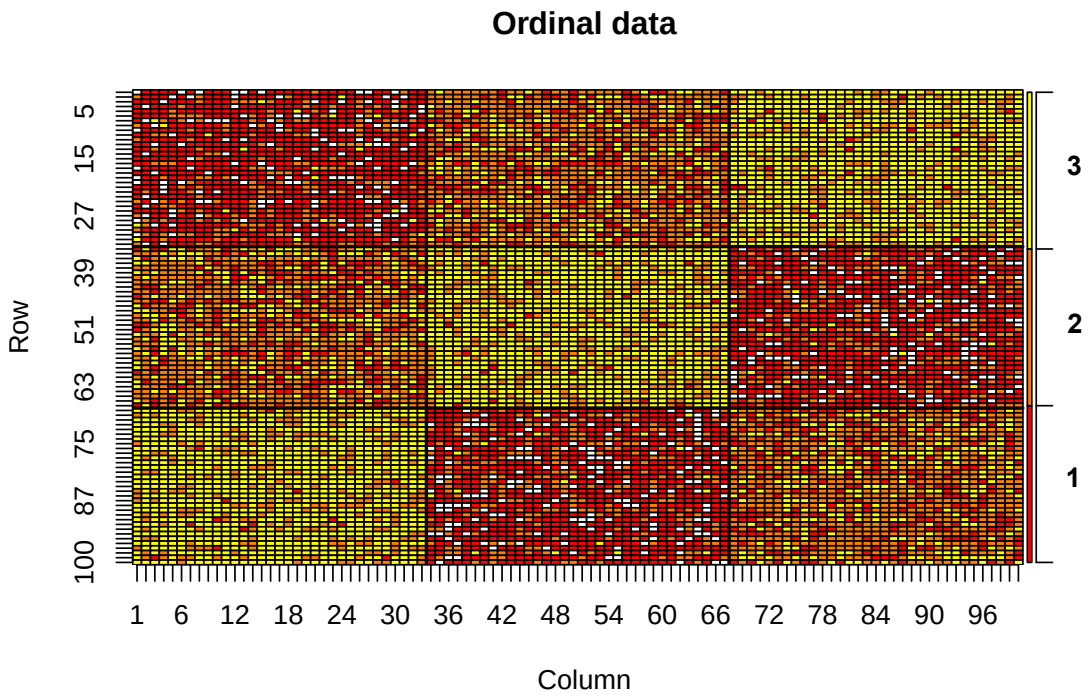


Figure 2.13: *Graphical representation of a simulated dataset, with ordinal data.*

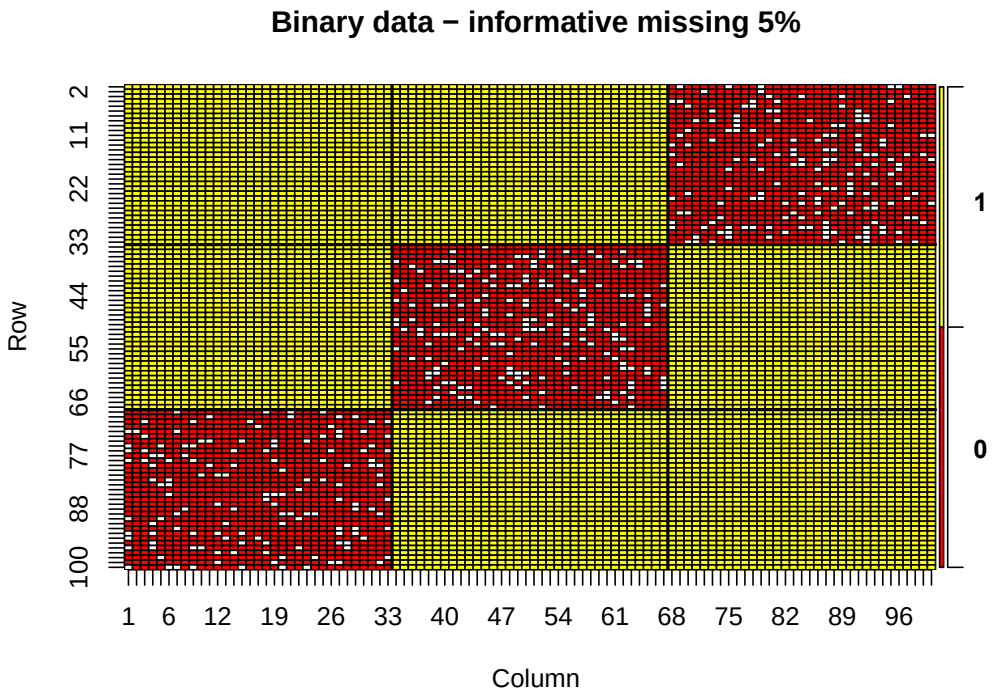


Figure 2.14: *Graphical representation of a simulated dataset, with binary data and 5% of informative missing values.*

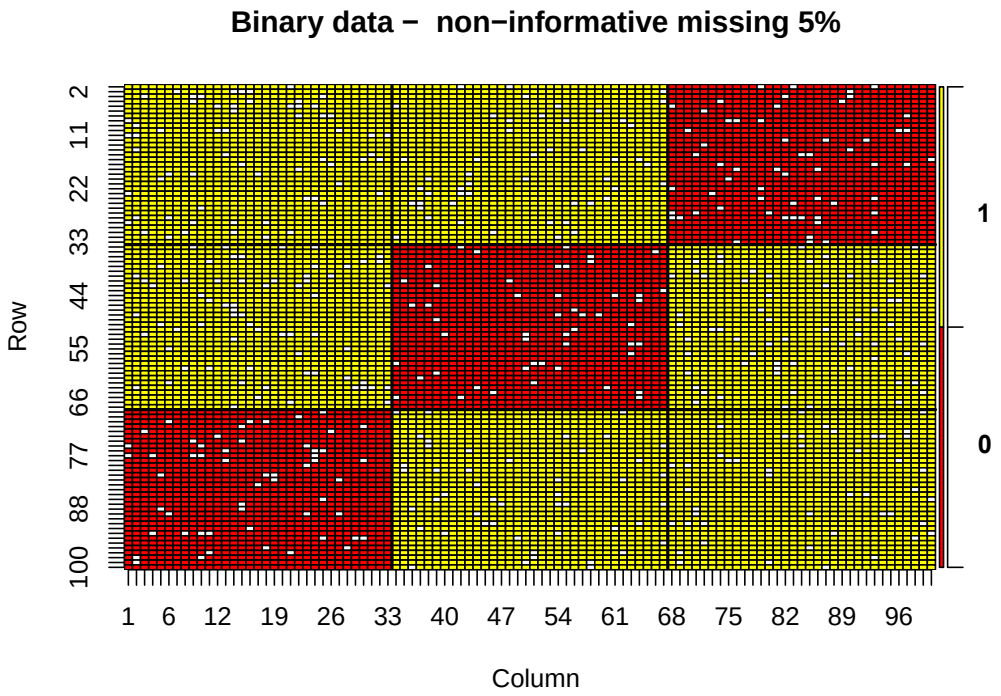


Figure 2.15: *Graphical representation of a simulated dataset, with binary data and 5% of non-informative missing values.*

B.2 U.S. Senate data

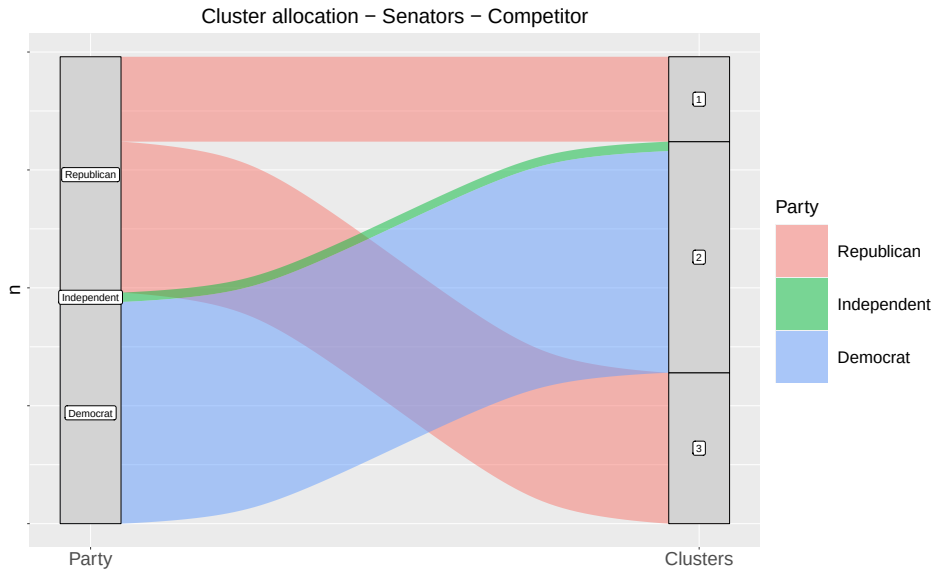


Figure 2.16: *Alluvial diagram comparing party affiliations and clusters identified by biclustermd for the U.S. Senators.*

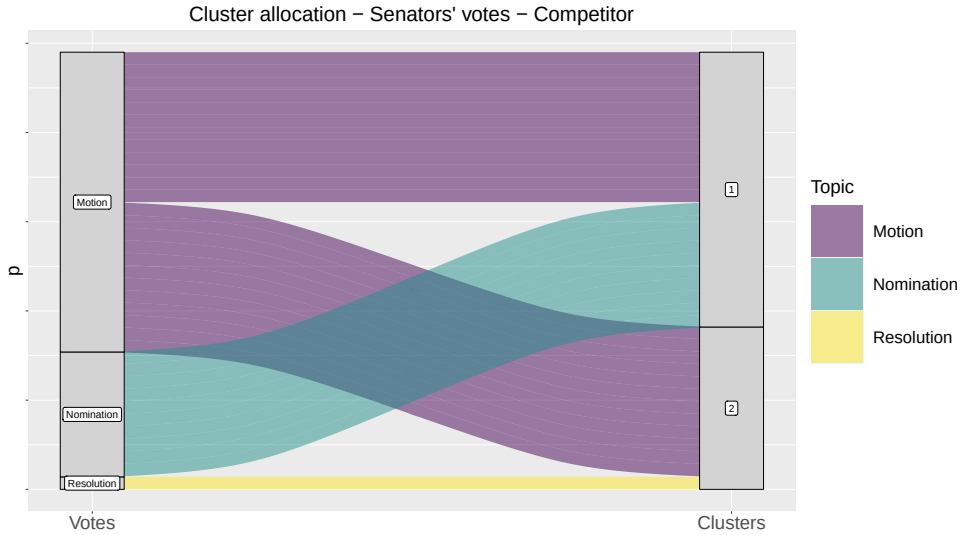


Figure 2.17: *Alluvial diagram comparing topic of voting sessions and clusters identified by biclustermd for the U.S. Senators' votes.*

B.3 Movielens data

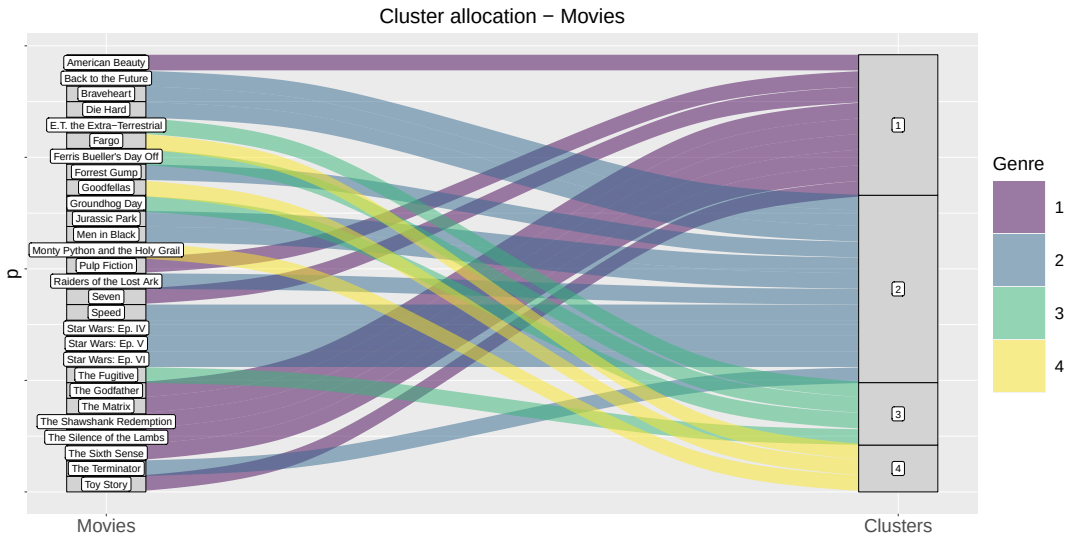


Figure 2.18: *Alluvial diagram for movie clusters identified with biclustermd.*

Chapter 3

Changepoint Detection with Local Level Dynamic Random Partition Models

3.1 Introduction

Recently, there has been an increased demand for models that can effectively describe features of complex multivariate time series data. This surge in interest is particularly prominent in fields such as biomechanics, motion analysis, human-computer interaction, and sports science. These models are used to break down continuous human motion or activity into distinct phases and states. The insights derived from such analyses can then be used to improve sports performances and delve deeper into the complexity of human biomechanics, for example, to understand gait cycles, gesture phases, athletic movements, and even cognitive states. In the analysis of human gesture data, information is typically gathered through various sensing technologies, such as motion capture systems, accelerometers, or videos. These technologies yield rich datasets that capture the temporal patterns of human movements. A key concept is to leverage the information accumulated over time while considering the time series jointly.

Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ denote a multivariate time series, where each $\boldsymbol{Y}_t$ is a n-dimensional vector, i.e., $\boldsymbol{Y}_t = \{Y_{1,t}, Y_{2,t}, \ldots, Y_{n,t}\}$, observed on n units over T time points. To illustrate, consider our application in Section 3.5, where we examine scalar velocity data obtained from $n = 4$ acceleration sensor units placed on the hands and wrists of a subject. Dynamic linear models (DLMs) are commonly employed for the analysis of time-series data due to their flexibility and adaptability in handling diverse situations (Petris et al., 2009). They define a class of state-space models and are characterized by a system of two equations: an observation equation, which describes the observed data as a linear combination of latent

state variables with noise, and a state equation that describes how latent states evolve over time, thereby tracking the underlying dynamics of the system. We introduce our contribution by referring to a simple yet fundamental DLM, the *local level model* (LLM), which describes the observed data as composed of a level component plus random noise,

$$Y_{i,t} = \beta_{i,t} + \varepsilon_i, \tag{3.1}$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$, where $N(\mu, \sigma^2)$ denotes the Normal distribution with mean μ and variance σ^2. The vector $\boldsymbol{\beta}_t = \{\beta_{1,t}, \ldots, \beta_{n,t}\}$ represents the underlying level or trend of the time series at time t. In a typical local level model, the evolution of $\boldsymbol{\beta}_t$ over time is modeled as a random walk, i.e., the level at time t is predicted to be the same as the level at time $t-1$, plus some random noise. If the variance of the random noise is small, this assumption implies smooth processes over time. Despite its simplicity, the LLM illustrates the fundamental characteristics of many time-series models, and will serve as a basic example throughout.

Studying how measurements from different units cluster or group over time, as well as how these clusters evolve during an activity, can provide valuable insights in many applications. For instance, when analyzing biometric data, it is important to understand how different body parts cooperate during different stages of a gesture or movement. Furthermore, detecting changepoints in the clustering of body parts between gesture stages can reveal important insights into the motion and inform gesture detection algorithms.

In a Bayesian setting, BNP models are a popular choice for clustering the dynamic behavior of latent variables like $\beta_{i,t}$ over time (as discussed in Quintana et al., 2022). BNP models do not require the upfront specification of the number of clusters; instead, they allow for posterior inference on cluster allocations directly from the data. Existing BNP approaches for clustering time series data vary in terms of motivation, application, and how time-dependence is introduced. For example, some approaches build on the stick-breaking representation of the Dirichlet Process (Ferguson, 1973; Sethuraman, 1994). In this context, Antoniano-Villalobos and Walker (2016) have developed a stationary Markov model where both the transition and stationary densities are nonparametric infinite mixture models. Nieto-Barajas and Contreras-Cristán (2014) have clustered temporal data while considering several features typical of time series data (e.g., trends, seasonality). BNP autoregressive model are discussed in Kalli and Griffin (2018), De Iorio et al. (2019), DeYoreo and Kottas (2018) and Beraha et al. (2022), among others. Alternatively, other authors have explored generalizations of the Pólya urn scheme of Blackwell et al. (1973), see, e.g. Caron et al.

(2007), Caron et al. (2017) and Cassese et al. (2019).

All the methods mentioned above identify clusters based on the values of associated parameters. In the context of the illustrative local level model (3.1), these models can potentially assign observations to different clusters over time if, for instance, $\beta_{i,t} = \beta_{j,t}$ for some $j \neq i, j = 1, \ldots, n$ and $\beta_{i,t+1} = \beta_{j,t+1}$ are significantly different. This occurs because model-based clustering with Bayesian nonparametric models essentially relies on specifying a mixture model that depends on a discrete random mixing measure. The probabilistic distribution over random partitions is essentially a by-product of this mixture setup; in essence, clusters in a mixture model are identified based on levels of activity. Furthermore, as highlighted by Page et al. (2022), even when a sequence of random probability measures is highly correlated, induced random partitions from previous dependent Bayesian nonparametric priors tend to exhibit weak dependencies. This can result in estimated partitions that fail to capture the evident dependencies present in the sequence of random probability measures.

A solution to address this challenge involves directly modeling the sequence of random partitions. In this context, Page et al. (2022) introduced a dependent random partition model (Hartigan, 1990; Barry and Hartigan, 1992) that incorporates an auxiliary variable denoted as $\gamma_{i,t}$ $(i = 1, \ldots, n; t = 1, \ldots, T)$, which helps determining whether a unit at time $t-1$ should be considered for possible cluster reallocation at time t. More specifically, when $\gamma_{i,t} = 0$, unit i must be re-assigned at time t, potentially to a new cluster or randomly assigned to the same cluster as the previous time. Instead, if $\gamma_{i,t} = 1$, unit i is almost surely assigned to the previous cluster. A recently published paper by Quinlan et al. (2022) presents a method that aims to correlate data partitions with the detection of changepoints in multivariate time series. However, they specified only a single unique partition of contiguous clusters for each time series.

In this chapter, we introduce a more straightforward Markov structure for modeling partitions that evolve over time. Our approach develops a random partition model capable of connecting the partition of data points to the previous partitions over time. It builds upon the principles of DLMs but extends them to incorporate latent state equations now operating within the partition context and time evolution defined by the partitions themselves. Therefore, while in the analysis of time series data it is usual to identify sudden changes in the observed values of a stochastic process as a changepoint, in this paper, we refer to a changepoint as a change in a latent partition of units. More specifically, we employ a Markov dependent structure, where the partition at time t, denoted as π_t, is modeled conditionally on the partition at time $t-1$ to account for temporal persistence and facilitate

changepoint detection. Furthermore, the selection of changepoints at each time point t is driven by a mixture that chooses between one of two partitions at each time point. At each time, the chosen partition can either coincide with the one at the previous time or follow a flexible and general random partition model. In the presence of a changepoint, the partition at time t becomes independent of the partition at time $t-1$. Unlike the existing Bayesian nonparametric models cited above, our clustering approach does not rely on the values of parameters, such as the $\beta_{i,t}$'s in equation (3.1). Instead, we directly treat dependent random partition allocations as latent structures that drive the dynamics of the observations. In contrast to the dependent random partition model proposed by Page et al. (2022), we jointly consider all the units when identifying changepoints, taking into account the multivariable nature of the dependent partitions. In Page et al. (2022), the temporal allocation is guided by subject-specific $\gamma_{i,t}$'s. As a result, a change of partitions over time (dynamic distribution of random partitions) is obtained as a by-product of individual clustering allocations. This approach may lead to a higher number of cluster configurations, potentially more false positive changes, and may hinder the identification of straightforward, clear, partition changepoints through time, which is crucial, for instance, in identifying gesture phases. Our modeling approach is relatively straightforward to implement in comparison to existing methods for dependent random partition models, as it leverages the efficiency of a Gibbs sampling method. Furthermore, while constructing the prior distribution for the partition bears some resemblance to the use of spike-and-slab priors for variable selection (Tadesse and Vannucci, 2021), dealing with partitions introduces additional complexity in the motivation, modeling and computation.

The remaining of the chapter is organized as follows: in Section 3.2, we present our proposed model and discuss its main properties. Section 3.3 covers posterior inference and computational methods. Section 3.4 describes simulation studies highlighting key aspects of our model. In Section 3.5, we present an application to the analysis of human gesture data, and finally, Section 3.6 provides concluding remarks and outlines future directions.

3.2 Local level dynamic partition model

We describe the key features of the proposed dynamic partition model, a local level dynamic partition model (LLDPM), taking model (3.1) as reference. More specifically, we assume that for each unit i, $i \in \{1, \ldots, n\}$, the observations are generated from some general likelihood (observation equation) $Y_{i,t} \mid \beta_{i,t} \overset{\text{ind}}{\sim} p\left(y_{i,t} \mid \beta_{i,t}\right)$, $t = 1, \ldots T$. In the following, we will consider a Gaussian kernel. The dynamics of the latent state equation are characterized

in terms of time-varying partitions of the n units over time. In order to describe such dynamics, we introduce a dependent RPM that models temporal dependence in terms of sequences of partition by considering an auxiliary variable, γ_t, which determines whether a partition at time $t-1$ will be considered for possible cluster reallocation at time t, $t = 1, \ldots, T$. More specifically, together with each $\beta_{i,t}$, we introduce a binary *changepoint* auxiliary variable $\gamma_t \in \{0, 1\}$, to detect changes in the partitions of the n units from time $t-1$ to time t. When considering a partition π_t at time t, we denote the number of clusters/blocks identified in the partition at time t among the n units as $|\pi_t|$. Further, let $C_{1,t}, \ldots, C_{|\pi_t|,t}$ represent the clusters of the n units as implied by the partition π_t at time t. Thus, given a partition π_{t-1} at time $t-1$, we assume that the *partition-based state equation* is characterized as a mixture over two partition models, corresponding to the case of exchangeable (independent) and fully dependent partitions across the two time points,

$$\pi_t \mid \boldsymbol{\pi}_{1:(t-1)}, \boldsymbol{\gamma}_{2:(t-1)} \sim (1-\gamma_t)\, \delta_{\pi_{t-1}}(\pi_t) + \gamma_t\, p^*(\pi_t), \quad t = 2, \ldots, T \tag{3.2}$$

where $\boldsymbol{\pi}_{1:t} = (\pi_1, \ldots, \pi_{t-1})$ is the vector of previously recorded partitions, $\boldsymbol{\gamma}_{2:(t-1)} = (\gamma_2, \ldots, \gamma_{t-1})$ indicates the vector of previous changepoints, and $p^*(\pi_t)$ indicates the distribution of a *base* random partition model. In the following, for simplicity, we assume that the base distribution over the partitions is given by the Chinese restaurant process (CRP, Pitman, 2002) with mass parameter α. We indicate it as $p^*(\pi) = p_{\text{CRP}}(\pi; \alpha)$. In the state equation (3.2), we allow γ_t to potentially depend on past times. Nevertheless, in the following we assume that $\gamma_t \overset{ind}{\sim} \text{Bern}(\eta_t)$, where the probability of a changepoint $\eta_t \sim \text{Beta}(a, b)$, that is we assume that the changepoints are independent of the vector $\boldsymbol{\gamma}_{2:(t-1)}$ and partitions $\boldsymbol{\pi}_{1:(t-2)}$. Thus, for any $t = 2, \ldots, T$, conditionally on π_{t-1} and γ_t, the partition π_t is independent of $\pi_{1:(t-2)}$ and of $\gamma_{2:(t-1)}$. The model is completed by an initialization condition on the initial partition, e.g. $\pi_1 \sim p_1^*(\cdot)$ and by priors on the values of the local mean parameters $\beta_{i,t}$. Here, we assume $p_1^*(\cdot) = p^*(\cdot) = p_{\text{CRP}}(\cdot; \alpha)$. We then consider auxiliary variables $s_{i,t} \in \{1, \ldots |\pi_t|\}$ indicating the cluster memberships/assignments, i.e. if units $i, j \in C_{k,t}$, for some $k = 1, \ldots, |\pi_t|$, then $s_{i,t} = s_{j,t} = k$. We follow the typical assumption of Bayesian nonparametric models and assume that at each time point the values of the parameters $\beta_{i,t}$ coincide within a cluster, although the specific values could differ at different times. That is, we assume

$$\beta_t \mid \pi_t \sim \prod_{j=1}^{|\pi_t|} P_0\left(\beta_{j,t}^*\right), \tag{3.3}$$

for some base distribution $P_0(\cdot)$ where $P_0(\cdot)$ is the base measure for the parameters $\boldsymbol{\beta}_t$ $(t = 1, \ldots, T)$. We can rewrite the joint distribution of the data and the latent partition and selection parameters, after marginalizing with respect to the $\boldsymbol{\beta_t}$'s, $t = 1, \ldots, T$, as follows

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{\pi}, \boldsymbol{\gamma}) &= p(\boldsymbol{\gamma}) p(\boldsymbol{\pi} \mid \boldsymbol{\gamma}) \int p(\boldsymbol{\beta} \mid \boldsymbol{\pi}) p(\boldsymbol{Y} \mid \boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \\ &= p^*(\pi_1) \prod_{t=1}^{T} \eta_t^{\gamma_t} (1-\eta_t)^{1-\gamma_t} \prod_{t=1}^{T} \left[(1-\gamma_t) \delta_{\pi_{t-1}}(\pi_t) + \gamma_t \, p^*(\pi_t) \right] \\ &\times \prod_{t=1}^{T} \prod_{j=1}^{|\pi_t|} \int \prod_{i \in C_j(\pi_t)} p(y_{i,t} \mid \beta^*_{j,t}) P_0(\mathrm{d}\beta^*_{j,t}) \end{aligned} \tag{3.4}$$

where we have indicated $\boldsymbol{\beta}$ as the collection of the local level parameters. Let

$$G(\boldsymbol{Y}_t \mid \pi_t) = \prod_{j=1}^{|\pi_t|} \int \prod_{i \in C_j(\pi_t)} p(y_{i,t} \mid \beta^*_{j,t}) P_0(\mathrm{d}\beta^*_{j,t}).$$

Then, we can rewrite (3.4) in a compact form as

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{\pi}, \boldsymbol{\gamma}) &= p^*(\pi_1) \prod_{t=1}^{T} \eta_t^{\gamma_t} (1-\eta_t)^{1-\gamma_t} \prod_{t=1}^{T} \left[(1-\gamma_t), \delta_{\pi_{t-1}}(\pi_t) + \gamma_t \, p^*(\pi_t) \right] G(\boldsymbol{Y}_t \mid \pi_t) \\ &= p(\pi_{1:T}, \gamma_{2:T}) \; G(\boldsymbol{Y}_t \mid \pi_t). \end{aligned} \tag{3.5}$$

We can further integrate with respect to $\gamma_{2:T}$ and consider $p(\boldsymbol{Y}, \boldsymbol{\pi}) = p(\pi_{1:T}) \; G(\boldsymbol{Y}_t \mid \pi_t)$, where the distribution of $\pi_{1:T}$ is

$$p(\pi_{1:T}) = p^*(\pi_1) \prod_{t=2}^{T} \left[(1-\eta_t) \delta_{\pi_{t-1}}(\pi_t) + \eta_t \, p^*(\pi_t) \right]. \tag{3.6}$$

3.2.1 Marginal properties

In this section, we explore the prior properties of model (3.5) and the proposed prior. We first formally prove that at each time t, the marginal distribution of the random partition π_t is the same as that of the base process $p^*(\cdot)$, for $t = 2, \ldots, T$.

Proposition 1. *Let $\pi_1 \sim p^*(\pi_1)$ be a random probability random model. Let $\pi_{2:T}$ and*

$\gamma_{2:T}$ be characterized by the joint distributions defined in model (3.5)*. Then, for every $t = 2, \dots, T$, the marginal distribution of the random partition π_t is the base random probability measure $p^*(\cdot)$.*

In particular, if $p^*(\cdot) = p_{\text{CRP}}(\cdot; \alpha)$, then such distribution is the random partition model implied by a Dirichlet process with total mass parameter α. The previous preposition extends also to the whole class of Gibbs-type priors. Gibbs–type priors is a more general set of prior where some of the well-known special cases are the Dirichlet and the Pitman-Yor processes (Ferguson, 1973; Pitman and Yor, 1997) as well as mixtures of symmetric Dirichlet distributions (Gnedin and Pitman, 2005), the normalized inverse Gaussian processes (Lijoi et al., 2005) and their generalization given by normalized generalized gamma processes (Lijoi et al., 2007). For an insightful investigation of the large n behavior of Gibbs-type priors, and the role of the Pitman-Yor process within this class of random probability measures, one can refer to Arbel and Favaro (2021). De Blasi et al. (2013) shows that exchangeable product partition models with probability of each partition depending only on the cardinality of each cluster coincide with the family of Gibbs-type priors. More in detail, we can consider different exchangeable partition probability functions (EPPFs). The EPPF provides a simple way to define probabilistic partition models based on the number and sizes of blocks, independently of the object labels. The clusters are exchangeable since the probability distribution over partitions is invariant under permutations of the object labels. A Gibbs-type EPPF (see, e.g., De Blasi et al., 2013) has the following form,

$$\mathbb{P}_\alpha \left[\pi_t = \{C_{1,t}, \dots, C_{|\pi_t|,t}\}\right] = V_{n,|\pi_t|} \prod_{j=1}^{|\pi_t|} \frac{\Gamma\left(|C_{j,t}| - \alpha\right)}{\Gamma(1-\alpha)},$$

where $|\cdot|$ indicates the cardinality of each block/cluster, $-\infty \leq \alpha < 1$, $V_{n,|\pi_t|}, k_t = 1, \dots, |\pi_t|$ is a weight that determines how the probability mass is allocated over partitions with different numbers of blocks. A special case is the Chinese restaurant process (Pitman, 2006),

$$p(\pi_t \mid \alpha) = \frac{\alpha^{|\pi_t|}}{\prod_{i=1}^{n}(\alpha + i - 1)} \prod_{i=1}^{|\pi_t|} (|C_{i,t}| - 1)!$$

by setting $V_{n,|\pi_t|} = \alpha^{|\pi_t|}/(\alpha)_n$, where $\alpha > 0$, and $(a)_q = \Gamma(a+q)/\Gamma(a)$ for any $a > 0$ and integer $q \geq 0$.

3.2.2 Hierarchical representation of the LLDPM

In this section, we show that our model allows for an equivalent representation that does not rely on assuming a sequence of temporally dependent partitions; instead, this representation establishes a hierarchical structure in the dependence of the partitions, which can be used to describe changes of the partition between different groups or experimental conditions. We start by focusing on the joint distribution of two subsequent random partitions in model (3.5). Without loss of generality, we consider $T = 2$ and consider the joint distribution of π_1 and π_2,

$$p(\pi_1, \pi_2) = p^*(\pi_1; \alpha) \left[(1-\eta)\delta_{\pi_1}(\pi_2) + \eta\, p^*(\pi_2; \alpha)\right].$$

We propose an alternative construction that defines the same distribution for a vector of two random partitions. More specifically, let $\tilde{\pi}$ be another random partition, and assume the following joint model for $\tilde{\pi}$, π_1 and π_2,

$$\begin{aligned} \tilde{\pi} \;&\sim\; p^*(\pi) \\ \pi_t \mid \tilde{\pi}, \gamma_1, \gamma_2 \;&\overset{\text{ind}}{\sim}\; (1-\gamma_t)\,\delta_{\tilde{\pi}}(\pi_t) + \gamma_t\, p^*(\pi_t), \qquad t = 1, 2, \end{aligned} \tag{3.7}$$

where we still assume $\gamma_t \overset{\text{iid}}{\sim} \text{Bern}(\eta_t)$, $t = 1, 2$. Note that in model (3.7), the distribution of π_2 is independent of that of π_1 given partition $\tilde{\pi}$. That is, model (3.7) effectively defines a hierarchical partition model. We can write the joint distribution of $(\tilde{\pi}, \pi_1, \pi_2)$ and (γ_1, γ_2) as

$$\begin{aligned} p(\tilde{\pi}, \pi_1, \pi_2, \gamma_1, \gamma_2) &= p(\tilde{\pi}, \pi_1, \pi_2 \mid \gamma_1, \gamma_2)\, p(\gamma_1, \gamma_2) \\ &= p(\tilde{\pi})\, p(\pi_1 \mid \tilde{\pi}, \gamma_1, \gamma_2)\, p(\pi_2 \mid \tilde{\pi}, \gamma_1, \gamma_2)\, p(\gamma_1)\, p(\gamma_2) \\ &= p^*(\tilde{\pi}) \left\{ \prod_{t=1}^{2} [(1-\gamma_t)\delta_{\tilde{\pi}}(\pi_t) + \gamma_t\, p^*(\pi_t)] \times \eta_t^{\gamma_t} (1-\eta_t)^{1-\gamma_t} \right\}. \end{aligned}$$

The distribution of (π_1, π_2) is obtained by marginalizing the last expression with respect to (γ_1, γ_2) and $\tilde{\pi}$,

$$\begin{aligned} p(\tilde{\pi}, \pi_1) &= \sum_{\tilde{\pi} \in \mathscr{P}} p^*(\tilde{\pi}) \prod_{t=1}^{2} [(1-\eta_t)\delta_{\tilde{\pi}}(\pi_t) + \eta_t\, p^*(\pi_t)] \\ &= p^*(\pi_1) \left[(1-\tilde{\eta})\, \delta_{\pi_1}(\pi_2) + \tilde{\eta}\, p^*(\pi_1)\, p^*(\pi_2)\right], \end{aligned} \tag{3.8}$$

where $\mathscr{P}$ indicates the power set of the partitions, with 2^n elements, and $\tilde{\eta} = 1 - (1 - \eta_1)(1 - \eta_2)$. By comparing (3.5) and (3.8), we appreciate that the two distributions coincide provided that $\tilde{\eta} = \eta$, which is achieved if $(1 - \eta_1)(1 - \eta_2) = (1 - \eta)$. If we further make the assumption that $\eta_1 = \eta_2$, then $\eta_1 = 1 - \sqrt{1 - \eta}$, such that $\gamma_t \overset{\text{iid}}{\sim} \text{Bern}(1 - \sqrt{1 - \eta})$ in (3.7), $t = 1, 2$. Thus, marginally, the joint distribution on the partitions posited by the local-level partition model can be seen as a case of a hierarchical partition model. This representation can be extended to $T \geq 2$ groups and it underscores how the time-varying partition model (3.8) can be seen as a special case of a partially exchangeable model on the partitions, we assume the same marginal distributions of the partitions, $p^*(\pi_t)$, and equality of the probabilities η_t's.

3.3 Posterior Inference

Posterior inference for the parameters in the model (3.5) is carried out using a MCMC algorithm. More specifically, we follow a Gibbs sampling scheme (Neal, 2000). Given the representation in (3.6), we evaluate the full conditional distributions of the model by jointly updating the pair (γ_t, π_t), for $t = 1, \ldots, T$ in each Gibbs iteration, $b = 1, \ldots, B$. This strategy consists of five main steps. We briefly describe the updates of the model parameters at a generic iteration b. Full details of the posterior distributions and of our implementation are in the Supplementary material (Appendix 3.B).

1. *update* π_t: we update the partition from its full conditional distribution $p(\pi_t^{(b)} \mid \gamma_t^{(b-1)}, \ldots)$ where the dots indicate all remaining parameters. For notation simplicity, the subscript $^{(b)}$ is deleted in formulas whenever it is clear from the context:

$$p(\pi_t^{(b)} \mid \gamma_t^{(b-1)}, \gamma_{t+1}^{(b-1)}, \ldots) \propto \Big[(1 - \eta_t)\, \delta_{\pi_{t-1}}(\pi_t) + \eta_t\, p_{\text{CRP}_\alpha}(\pi_t)\Big] \times \Big[(1 - \gamma_{t+1})\, \delta_{\pi_{t+1}}(\pi_t) + \gamma_{t+1}\, p_{\text{CRP}_\alpha}(\pi_{t+1})\Big]\, G(\boldsymbol{Y}_t \mid \pi_t).$$

   The expression of this full conditional highlights that the update depends also on $\gamma_{t+1}^{(b-1)}$.Thus, we need to distinguish between the following two cases:

   a) If $\gamma_{t+1}^{(b-1)} = 1$, then the conditioning implies a changepoint at time $t + 1$, and we do not borrow any information about the partition at time t by looking one-step

ahead. Thus, the partition at time t arises from a mixture,

$$p(\pi_t^{(b)} \mid \gamma_t^{(b-1)}, \gamma_{t+1}^{(b-1)} = 1, \ldots) \propto (1-\eta_t)\, G(\boldsymbol{Y}_t \mid \pi_{t-1})\, \delta_{\pi_{t-1}}(\pi_t) + \eta_t\, p_{\mathrm{CRP}_\alpha}(\pi_t)\, G(\boldsymbol{Y}_t \mid \pi_t),$$

i.e., π_t coincides with π_{t+1} with probability that is proportional to $(1-\eta_t)\, G(\boldsymbol{Y}_t \mid \pi_{t+1})$. Alternatively, it is generated as a random draw of a new partition from a random partition model with distribution proportional to $p_{\mathrm{CRP}_\alpha}(\pi_t) \times G(\boldsymbol{Y}_t \mid \pi_t)$. The probability of choosing this mixture component is proportional to $\eta_t \sum_{\pi_t \in \mathscr{P}_t} p_{\mathrm{CRP}_\alpha}(\pi_t)\, G(\boldsymbol{Y}_t \mid \pi_t) = \eta_t\, g_t$. We discuss how we evaluate these quantities further below.

b) If $\gamma_{t+1}^{(b-1)} = 0$:

$$p(\pi_t^{(b)} \mid \gamma_t^{(b-1)}, \gamma_{t+1}^{(b-1)} = 0, \ldots) \propto (1-\eta_t)\, \delta_{\pi_{t-1}}(\pi_t)\, G(\boldsymbol{Y}_t \mid \pi_{t-1}) + \eta_t\, p_{\mathrm{CRP}_\alpha}(\pi_{t+1})\, G(\boldsymbol{Y}_t \mid \pi_t);$$

that is, given $\gamma_{t+1} = 0$, π_t coincides either with π_{t-1} with probability proportional to $(1-\eta_t)\, \delta_{\pi_{t-1}}(\pi_t) G(\boldsymbol{Y}_t \mid \pi_{t-1})$ or with π_{t+1} with probability proportional to $\eta_t\, p_{\mathrm{CRP}_\alpha}(\pi_{t+1}) G(\boldsymbol{Y}_t \mid \pi_t)$. Hence, since $\gamma_{t+1} = 0$, we have that $\mathrm{p}(\pi_t \equiv \pi_{t+1}) = 1$.

2. *update* γ_t: the update of the auxiliary variable γ_t depends on the partitions π_t and π_{t-1} as follows,

   a) $p(\gamma_t^{(b)} = 1 \mid \pi_t^{(b)}, \ldots) \propto \eta_t^{(b-1)}\, p_{\mathrm{CRP}_\alpha}(\pi_t)$;
   b) $p(\gamma_t^{(b)} = 0 \mid \pi_t^{(b)}, \ldots) \propto (1-\eta_t^{(b-1)},) \delta_{\pi_{t-1}^{(b)}}(\pi_t^{(b)})$.

   If $\pi_t^{(b)} = \pi_{t-1}^{(b)}$, then

   – $p(\gamma_t^{(b)} = 1 \mid \pi_t^{(b)}, \ldots) \propto \eta_t^{(b-1)}\, p_{\mathrm{CRP}_\alpha}(\pi_{t-1})$
   – $p(\gamma_t^{(b)} = 0 \mid \pi_t^{(b)}, \ldots) \propto (1-\eta_t^{(b-1)})$

   However, if $\pi_t^{(b)} \neq \pi_{t-1}^{(b)}$, then $p(\gamma_t^{(b)} = 1) = 1$.

3. *update* $\boldsymbol{\beta}_t$: update $\boldsymbol{\beta}_t$ from $p(\boldsymbol{\beta}_t^{(b)} \mid \ldots)$, such that

$$p(\boldsymbol{\beta}_t^{(b)} \mid \pi_t^{(b)}, \ldots) \propto \prod_{j=1}^{|\pi_t|} \prod_{i \in C_j(\pi_t)} P(y_{i,t} \mid \beta_{j,t}^*)\, P_0(\beta_{j,t}^*).$$

4. *update* the remaining parameters are update with straightforward Gibbs sampling updates.

5. *reshuffling step* for $\boldsymbol{\pi}$: we update the visited partitions using a sampling importance resampling step considering different groups of sequential partitions.

Evaluating quantities over the power set $\mathcal{P}$. Step 1(a) highlights the computational challenges that arise when employing the changepoint detection approach outlined in equation (3.2), e.g., with respect to commonly used spike-and-slab variable selection priors. In order to explore the partition space and calculate the appropriate probabilities for selecting a specific partition, at each iteration it is necessary to compute quantities like $g_t = \sum_{\pi_t \in \mathscr{P}_t} p_{\text{CRP}_\alpha}(\pi_t)\, G(\boldsymbol{Y}_t \mid \pi_t)$, for each $t = 1, \ldots, T$. These computations can become computationally daunting as n increases. For example, in our application of Section 3.5, we consider $n = 8$ sensors; hence, we would need to compute these quantities across 256 partitions. It is important to ensure the scalability of the method. Importantly, it is worth noting that these quantities only rely on the partition π_t. Therefore, as long as we can efficiently approximate the distribution of such partitions under the base partition model, we should be able to obtain a reliable approximation for the required quantities. We propose implementing an auxiliary MCMC run prior to initializing the primary MCMC algorithm for model fitting. In this preliminary run, we fit an independent CRP_α model at each time point, saving all the partitions generated throughout the MCMC iterations. Then, we can approximate g_t by considering the set $\mathscr{S}_t$ comprising all the posterior partitions obtained through the CRP_α in the auxiliary MCMC run, as follows

$$\hat{g}_t = \frac{1}{\mid \mathscr{S}_t \mid} \sum_{\pi_t \in \mathscr{S}_t} p_{\text{CRP}_\alpha}(\pi_t) G(\boldsymbol{Y}_t \mid \pi_t). \tag{3.9}$$

Similarly, to sample a new distribution from $p_{\text{CRP}_\alpha}(\pi_t)\, G(\boldsymbol{Y}_t \mid \pi_t)$, we leverage the realized partitions from the auxiliary MCMC run and randomly select a new partition π_t from $\mathscr{S}_t$, randomly selecting a new partition π_t from $\mathscr{S}_t$, with $t = 1, \ldots, T$.

3.3.1 Changepoint detection

Changepoint detection inherently involves making multiple comparisons, since the decisions are temporally dependent. To address this multi-comparison problem, we use a decision theoretic approach to detect the presence of a changepoint, relying on the posterior probability of the changepoint (PPC_t), $p(\gamma_t = 1 \mid \text{data})$, at each time point. More specifically, we consider a compound decision-theoretic approach, which is based on a loss function that takes simultaneously into account the sequence of decisions and it is defined as a linear combination of measures of the false positive and true positive (or false negative) deci-

sions (Sun and Cai, 2007). In a Bayesian context, Müller et al. (2004) and Muller et al. (2006) have demonstrated that when assuming both independent hypotheses and independent (marginal) loss functions, the optimal approach for minimizing the resulting posterior expected loss involves thresholding the PPC_t values ($t = 1, \ldots, T$) as estimated from the output of the MCMC. However, such a procedure does not inherently control for the false discovery rate (FDR), unless such a control is explicitly accounted for. Therefore, it becomes necessary to calculate the optimal threshold in order to control the FDR at a specific desired level, ζ. This can be done by considering the Bayesian FDR (Newton et al., 2004),

$$FDR_m(h) = \frac{\sum_t^T (1 - PPC_t)\, \mathbb{1}_{(PPC_t > h)}}{\sum_t^T \mathbb{1}_{(PPC_t > h)} \vee 1} \tag{3.10}$$

where h is the chosen threshold and $\mathbb{1}_{(PPC_t>h)}$ indicates the indicator function such that if $PPC_t > h$ then $\mathbb{1}_{(PPC_t>h)} = 1$ and 0 otherwise. The optimal threshold corresponds to the minimum value of the PPC_t which still ensures that the FDR_m is less than ζ; in formulas, $h^* = \min\{h : \mathrm{FDR_m}(h) \leq \zeta\}$. We note that Müller et al. (2004) show how such a decision rule can be obtained by considering loss functions defined as a linear combination of the false discovery rate and the true positive (or false negative) rate.

The previous testing procedure can be classified as a *marginal* approach, since it fails to consider existing dependencies either among hypotheses or in the decisions themselves. Sun et al. (2015) extended this framework to the spatial setting, explicitly taking into account dependencies among the hypotheses, as induced by a spatial model. More recently, Chandra and Bhattacharya (2019) introduced non-marginal loss functions and non-marginal decision rules, which take dependencies into account during the decision-making process by considering dependent decisions directly within the loss functions. More in detail, their procedure incorporates additional information about dependencies among the tests in the definition of the error and non-error terms associated with subgroups of hypotheses. The approach penalizes the decision for each hypothesis based on the incorrect decisions regarding other dependent tests, thus defining a compound loss in which decisions regarding dependent tests rely on each other. We adapt their framework to our case. More specifically, we indicate with G_t the set of hypotheses related to having a changepoint at time t, for $t = 1, \ldots, T$. These hypotheses correspond to the null hypothesis $H_{0,t}$ (no changepoint) and the alternative hypotheses $H_{1,t-1}, H_{1,t}, H_{1,t+1}$ of a changepoint at times $t-1$, t, $t+1$. Considering this set of alternative hypotheses together is crucial, since - for example - a false changepoint detection at time $t-1$ may induce a false changepoint detection at time t, even if the null hypothesis is true at both times. Let d_t represent the decision at time t, i.e., $d_t = 1$ if the

t-th hypothesis is rejected and $d_t = 0$ if it is not. Let r_t denote the truth at time t, i.e., $r_t = 1$ if $H_{0,t}$ is true, $r_t = 0$ otherwise. Let $\boldsymbol{d} = (d_1, \dots, d_T)$, $\boldsymbol{r} = (r_1, \dots, r_T)$. We consider the compound loss function:

$$L(\boldsymbol{d}, \boldsymbol{r}) = -\text{TPR}(\boldsymbol{d}, \boldsymbol{r}) + \lambda \, \text{ER}(\boldsymbol{d}, \boldsymbol{r}),$$

where λ is a positive constant, and

$$TPR = TPR(\boldsymbol{d}, \boldsymbol{r}) = \sum_{t=1}^{T} d_t \, r_t / D$$

is the true positive rate, defined as the ratio between the number of cases where the t-th decision correctly identifies a changepoint and the number of positive decisions $D = \sum_{t=1}^{T} d_t$. In order to penalize false detections at each time t, we define a measure of error as the ratio between the total number of false detections in the set $G_t = \{t-1, t, t+1\}$ and D,

$$\begin{aligned} ER &= \left\{ \sum_{t=2}^{T} d_{t-1}(1 - r_{t-1}) + \sum_{t=1}^{T} d_t(1 - r_t) + \sum_{t=1}^{T+1} d_{t+1}(1 - r_{t+1}) \right\} / D = \\ &= \left\{ 2\, d_1\, (1 - r_1) + 3 \sum_{t=1}^{T-1} d_t\, (1 - r_t) + 2\, d_T\, (1 - r_T) \right\} / D. \end{aligned}$$

We aim to minimize the posterior expected loss with respect to $\boldsymbol{d}$. Then, following steps similar to those in Theorem 1 of Müller et al. (2004), it is possible to show that the optimal decision rule is a threshold on the posterior probabilities $PPC_t = E_{\boldsymbol{\theta}|\boldsymbol{X}_n}(r_t)$. In addition, since the expression of the error rate (3.3.1) considers the term $3 \sum_{t=1}^{m-1} d_t(1 - r_t)$, the resulting non-marginal FDR, say FDR_{nm}, is such that controlled $FDR_{nm}(h) = 3FDR_m(h)$, where FDR_m is defined in (3.10). Then, the FDR_{nm} is controlled at level $\zeta/3$, leading to a more stringent procedure than the marginal one. This becomes especially relevant when dealing with autoregressive data, which are characterized by higher structural dependence.

3.4 Simulation studies

We present two simulation studies to illustrate the performances of our LLDPM model under different data-generating mechanisms. More specifically, we investigate the model's

ability to accurately detect changepoints in independent and autocorrelated data scenarios, while also recovering the latent cluster structure. We compare with four alternative models:

1. the Dependent Random Partition Model (DRPM) introduced by Page et al. (2022) stands out - to our knowledge - as the only model-based approach that has introduced time dependency directly through partitions;

2. the Linear Dependent Dirichlet process (LDDP, Quintana et al., 2022) incorporates time explicitly within the atoms of the dependent process. In the implementation of this model, as well as all the subsequent models, the data is consolidated into a vector, with time being treated as a covariate. Consequently, we will refer to these models as alternative approaches in our analysis, rather than direct competitors.

3. the Weighted Dependent Dirichlet process (WDDP, Quintana et al., 2022) incorporates the time in the weights of the Dirichlet process;

4. the Griffiths-Milne Dependent Dirichlet approach (GMDDP Lijoi et al., 2014) to build bivariate vectors of dependent and identically distributed random measures, as implemented in the `R` package *BNPmix* (Corradin et al., 2020).

A more general class of alternative models for dependent data partitions could be obtained by adapting to the partition framework the flexible class of Compound random measures of Griffin and Leisen (2017). It is worth noting that our approach is model-based and does not require data concatenation to introduce time dependency. Additionally, we have adopted a more lenient approach when calculating the similarity between partitions at two consecutive time steps in the competitor models. These alternative approaches are not specifically designed to capture changepoints in partitions over time. Specifically, we utilized a similarity threshold of 90% based on the Adjusted Rand Index (Rand, 1971) to identify changepoints and enhance the performance of all competitor models.

3.4.1 Simulations with independent data

We start by generating data according to model (3.5). We evaluate the performance over 50 replicated datasets and different number of observations, namely $n = \{20, 50, 100\}$, over $T = 100$ time points. In each scenario, we generate independent normally distributed data including 8 changepoints, and we set the number of clusters to range from 2 to 3. An illustrative example of the data can be found in Appendix 3.C, Figure 3.6. The likelihood function is defined as a location Normal kernel with a priori variance of $\sigma^2 = 0.01$. Our

choice of hyperparameters for the model is as follows: we set the base random partition model $p^*(\cdot)$ to the CRP and set the concentration parameter $\alpha = 1$. The distribution P_0 is assumed to follow a Gaussian Normal distribution with a mean value of $\mu_t = 0$ and a variance of $\tau^2 = 0.5^2$. We further specify the hyperprior distributions as follows: $\sigma^2 \sim$ Inverse Gamma$(a_0 = 15, b_0 = 3)$ and $\tau^2 \sim$ Inverse Gamma$(a_0 = 15, b_0 = 3)$, where "IG" indicates the Inverse Gamma distribution. Additionally, we set $\eta_t \sim$ Beta$(a_t = 0.1, b_t = 0.9)$ in such a way that it assigns a higher probability to the absence of a changepoint. When fitting the DRPM model, we adopted parameters as suggested in Page et al. (2022), except that we fixed the variances as we did in our model. The temporal dependence parameter in the DRPM model follows a distribution of $Beta(b_t = 0.9, a_t = 0.1)$. Finally, for all other alternative models, we used the same variances as in our model and hyperprior parameter values as suggested for each method.

We implemented the MCMC algorithm using a Gibbs sampling scheme as detailed in Section 3.3. The performances of the algorithm are shown in Table 3.1, where we compare the computational cost of the analysis of datasets of different sizes. The computational time increases with the number of time series involved in the analysis, while the number of time points is kept fixed and equal to 100. We ran the model for 10,000 iterations, with the first 50% of iterations as burn-in. Subsequently, we calculated the optimal partition for each time point based on the posterior similarity and the estimated posterior probability η_t, for $t = 1, \ldots, T$. To determine whether a specific time point t should be considered a changepoint, we employed the False Discovery Rate method (Muller et al., 2006). As explained in Section 3.3, we implemented a penalized version of FDR with a control threshold set at 0.01/3

Computational Cost						
Units	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20	485.8	487.8	487.9	488.4	490.0	490.7
50	781.6	782.2	783.6	783.8	785.1	786.5
100	949.4	950.1	951.1	952.0	954.3	955.1

Table 3.1: *Computational cost, in seconds per 1000 iterations of the MCMC algorithm, with varying number of time series.*

Table 3.2 displays the performance of the five models as the data dimension varies. Notably, the LLDPM model outperforms all other models across all the metrics. However, as the sample size increases, all models exhibit a decrease in performance due to the greater number of subjects that have the potential to switch clusters.

The results highlight the strong performance of GMDDP when considering the metrics

Data dimension	Measure	LLDPM	DRPM	LDDP	WDDP	GMDDP
unit=20	specificity	0.6104 (0.413)	0.6106 (0.413)	0.566 (0.288)	0.3796 (0.021)	1.000 (0.000)
	accuracy	0.9902 (0.010)	0.6288 (0.375)	0.6008 (0.265)	0.4292 (0.019)	1.000 (0.000)
	recall	1.000 (0.000)	0.8375 (0.196)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	precision	0.9014 (0.094)	0.4209 (0.367)	0.2714 (0.238)	0.123 (0.004)	1.000 (0.000)
	F1	0.9455 (0.054)	0.4724 (0.337)	0.3836 (0.243)	0.2191 (0.006)	1.000 (0.00)
	AUC	0.9947 (0.006)	0.7263 (0.194)	0.6472 (0.176)	0.6898 (0.010)	1.000 (0.00)
unit=50	specificity	0.6309 (0.396)	0.6317 (0.397)	0.1961 (0.131)	0.0233 (0.021)	0.9989 (0.008)
	accuracy	0.9664 (0.023)	0.6558 (0.363)	0.2604 (0.121)	0.1014 (0.020)	0.9990 (0.007)
	recall	1.000 (0.000)	0.9325 (0.098)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	precision	0.7331 (0.149)	0.4292 (0.343)	0.0995 (0.013)	0.0818 (0.002)	0.9924 (0.054)
	F1	0.8378 (0.097)	0.5123 (0.330)	0.1808 (0.022)	0.1512 (0.002)	0.9952 (0.034)
	AUC	0.9817 (0.012)	0.7821 (0.193)	0.5961 (0.066)	0.5116 (0.011)	0.9995 (0.004)
unit=100	specificity	0.6228 (0.355)	0.6146 (0.352)	0.1397 (0.126)	0.02717 (0.022)	0.9839 (0.080)
	accuracy	0.9514 (0.031)	0.6390 (0.321)	0.2086 (0.116)	0.1050 (0.020)	0.9852 (0.073)
	recall	1.0000 (0.000)	0.9200 (0.123)	1.0000 (0.000)	1.0000 (0.000)	1.0000 (0.000)
	precision	0.6611 (0.167)	0.3444 (0.281)	0.0934 (0.013)	0.0821 (0.002)	0.9671 (0.163)
	F1	0.7843 (0.118)	0.4408 (0.281)	0.1707 (0.021)	0.1517 (0.003)	0.9721 (0.138)
	AUC	0.9740 (0.017)	0.7673 (0.168)	0.5685 (0.063)	0.5134 (0.011)	0.9920 (0.040)

Table 3.2: *Section 3.4.1: Summary statistics for changepoint detection with independent data and four competing models. The values correspond to the average (standard errors) over 50 simulations.*

related to changepoint detection. However, it is important to recall that GMDDP requires data vectorization and includes time as a covariate. Additionally, it must be considered that for all models considered as competitors of our proposal, the similarity threshold was suitably chosen to optimize their overall performance in detecting changepoints. Moreover, when evaluating the accuracy of recovering the true latent cluster structure, as depicted in Figure 3.1, GMDDP does not fare well and struggles to correctly identify clusters. On the other hand, LLDPM demonstrates very good performance in both changepoint detection and recovering the latent cluster structure. Hence, GMDDP may not be a reliable choice when the goal is to simultaneously perform changepoint detection and cluster analysis. DRPM, on the other hand, exhibits strong performance in changepoint detection, with an average specificity similar to that of our model in each scenario. However, as the number of subjects increases, its ability to recover the correct cluster structure is adversely affected.

3.4.2 Simulations with autoregressive data

In a second simulation study, we aim to assess the performance of the model in scenarios where the data are generated from an AR(1) process. Specifically, we consider datasets with 30 time points and 20 units. The changepoints are generated changing the magnitude

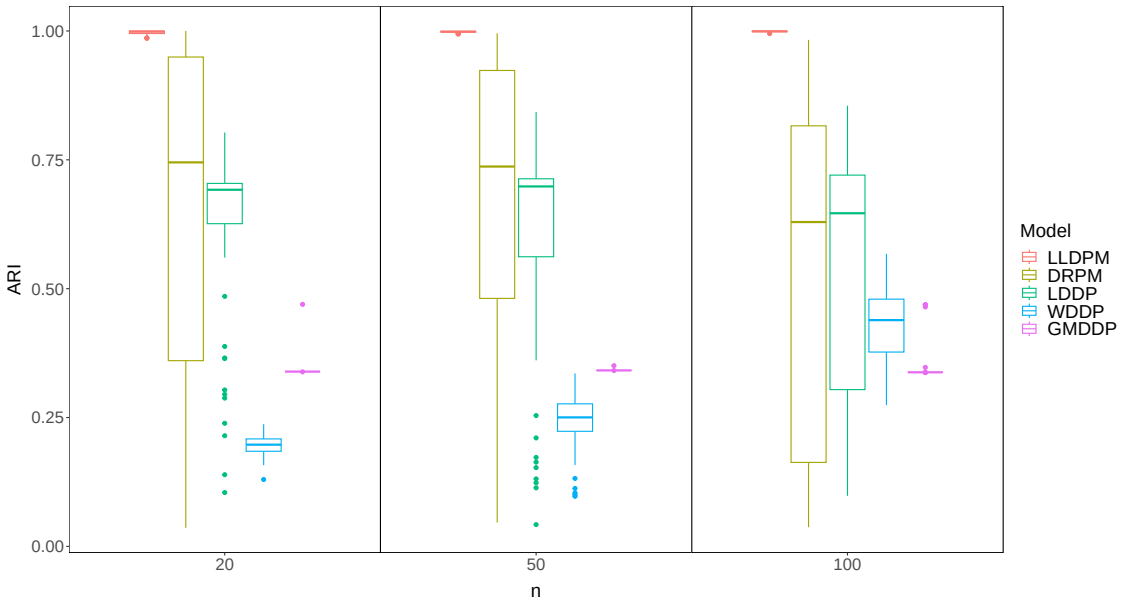


Figure 3.1: *Section 3.4.1: Boxplots of Adjusted Rand Index values evaluating the clustering performance with independent data, for four competing models,* $n = \{20, 50, 100\}$*, over 50 replicated datasets and 100 time points.*

of the time series at time points divisible by 5 and 9. The number of clusters is setting to be 1 if no changes occur, 3 and 2 if the magnitudes are changed accordingly. An illustrative example of such data can be found in Appendix 3.C, Figure 3.7. More specifically, we generate the data according to the following model:

$$Y_{i,t} = \lambda Y_{i,t-1} + \beta_{i,t} + \varepsilon_{i,t}.$$

We consider four different scenarios, each with a distinct value for the autoregressive parameter λ, chosen from the set $\{0.25, 0.5, 0.75, 0.9\}$. The error term $\varepsilon_{i,t}$ follows an independent and identically distributed normal distribution with mean 0 and variance σ^2. The results for 50 replicates are presented in Table 3.3.

In model fitting we decide to set α accordingly to the a priori number of cluster $\mathbb{E}(\mathrm{K})$. The expected exact a priori number of cluster is $\mathbb{E}(K) = \sum_{i=1}^{n} \alpha/(\alpha + i - 1)$ (Pitman, 2002), and asymptotically $\alpha \log[(\alpha + n)/\alpha]$ (Antoniak, 1974). We decided to set $\mathbb{E}(\mathrm{K})$=2, consequently the mass parameter of the Dirichlet process is $\alpha = 0.32$. To aid the competitor models in the changepoint identification procedure, we applied a 90% threshold for the posterior similarity between partitions.

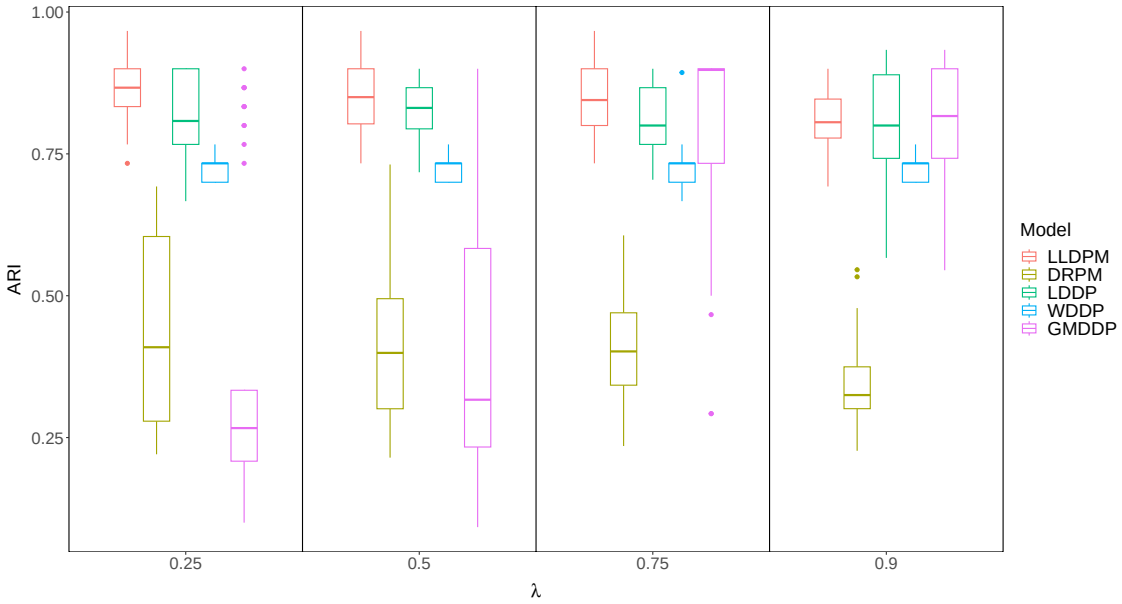


Figure 3.2: *Section 3.4.2: Boxplots of Adjusted Rand Index values evaluating the clustering performance with AR(1) data, for four competing models, and different values of the autoregressive coefficient* $\lambda = \{0.25, 0.5, 0.75, 0.9\}$*, over 50 replicated datasets and 100 time points*

LLDPM consistently demonstrates strong performance in recovering changepoints across all the metrics considered. In contrast, the GMDDP demonstrates lower performance in this scenario compared to the previous one. Surprisingly, the LDDP of Quintana et al. (2022) exhibits the best performance in terms of changepoint recovery. However, similar to the GMDDP in the previous scenario, the LDDP is less reliable in terms of correctly identifying the cluster structure. Considering both scenarios together, LDDP performs poorly in the first simulation study, leading to the suggestion that users may benefit from a more stable model that is not dependent on specific scenarios, such as LLDPM. Furthermore, DRPM performs well with autoregressive data, but its performance diminishes with increased dependence. It is worth highlighting that LLDPM and DRPM are the only models specifically designed for studying partition dynamics. In contrast, LDDP, WDDP, and GMDDP were not constructed for this specific purpose. Consequently, our simulations suggest that partition-based approaches are generally more reliable and stable across various scenarios.

AR(1) coefficient	Measure	LLDPM	DRPM	LDDP	WDDP	GMDDP
λ=0.25	specificity	0.8683 (0.151)	0.8728 (0.152)	0.9971 (0.014)	0.8121 (0.066)	0.7700 (0.125)
	accuracy	0.9147 (0.059)	0.5156 (0.044)	0.9987 (0.006)	0.5029 (0.032)	0.6173 (0.068)
	recall	0.9175 (0.073)	0.2031 (0.097)	1.000 (0.000)	0.2323 (0.063)	0.4838 (0.065)
	precision	0.9252 (0.065)	0.7473 (0.211)	0.9976 (0.012)	0.6043 (0.132)	0.7219 (0.115)
	F1	0.9195 (0.055)	0.2972 (0.095)	0.9988 (0.006)	0.3280 (0.078)	0.5743 (0.064)
	AUC	0.9145 (0.059)	0.5284 (0.040)	0.9986 (0.007)	0.5204 (0.031)	0.6308 (0.063)
λ=0.5	specificity	0.9082 (0.177)	0.9059 (0.178)	0.9943 (0.020)	0.8111 (0.061)	0.5514 (0.198)
	accuracy	0.8913 (0.066)	0.5276 (0.039)	0.9973 (0.009)	0.5015 (0.036)	0.5800 (0.100)
	recall	0.9038 (0.082)	0.1966 (0.134)	1.000 (0.000)	0.2306 (0.070)	0.6050 (0.104)
	precision	0.8992 (0.074)	0.8368 (0.208)	0.9953 (0.016)	0.5772 (0.085)	0.6219 (0.111)
	F1	0.8984 (0.061)	0.2886 (0.109)	0.9976 (0.008)	0.3249 (0.088)	0.6059 (0.082)
	AUC	0.8904 (0.066)	0.5421 (0.044)	0.9971 (0.010)	0.5188 (0.034)	0.5866 (0.097)
λ=0.75	specificity	0.9226 (0.081)	0.9286 (0.066)	0.9671 (0.065)	0.7726 (0.071)	0.3514 (0.299)
	accuracy	0.8707 (0.060)	0.5352 (0.036)	0.9847 (0.030)	0.4993 (0.061)	0.6533 (0.156)
	recall	0.8838 (0.082)	0.1910 (0.077)	1.000 (0.000)	0.2602 (0.122)	0.9175 (0.092)
	precision	0.8862 (0.089)	0.7940 (0.168)	0.9747 (0.049)	0.5618 (0.089)	0.6383 (0.139)
	F1	0.8797 (0.062)	0.2972 (0.096)	0.9866 (0.026)	0.3474 (0.102)	0.7458 (0.107)
	AUC	0.8697 (0.062)	0.5430 (0.040)	0.9836 (0.032)	0.5161 (0.058)	0.6345 (0.164)
λ=0.9	specificity	0.9127 (0.095)	0.9143 (0.096)	0.4586 (0.296)	0.7500 (0.094)	0.2811 (0.269)
	accuracy	0.8533 (0.056)	0.5181 (0.054)	0.7407 (0.141)	0.4680 (0.052)	0.6290 (0.133)
	recall	0.8600 (0.078)	0.1714 (0.091)	0.9875 (0.028)	0.2213 (0.066)	0.9334 (0.157)
	precision	0.8701 (0.072)	0.7380 (0.223)	0.6977 (0.130)	0.5090 (0.131)	0.6020 (0.149)
	F1	0.8618 (0.053)	0.2648 (0.123)	0.8112 (0.089)	0.3041 (0.081)	0.7420 (0.083)
	AUC	0.8529 (0.057)	0.5300 (0.050)	0.7230 (0.150)	0.4856 (0.053)	0.5987 (0.136)

Table 3.3: *Section 3.4.2: Summary statistics for changepoint detection with AR(1) data and four competing models. The values correspond to the average (standard errors) over 50 simulations.*

3.5 Application to Gesture Phase Segmentation

In this section, we present an analysis of video-recorded data for human gesture segmentation. The goal is to segment videos into distinct phases exhibiting different motion patterns, e.g., to identify time lapses of the video that need to be removed from a clip (Parvathy et al., 2021). More specifically, we employ the Gesture Phase Segmentation dataset, originally described by Madeo et al. (2013), which is publicly accessible for download at the following URL: https://archive.ics.uci.edu/ml/datasets/gesture+phase+segmentation. The dataset contains sensor data recordings of users recounting comic book stories facing an Xbox Microsoft Kinect^{TM} sensor. The dataset provides scalar velocity and acceleration values for the left hand, right hand, left wrist, and right wrist at regular time intervals (frames). These values were obtained by normalizing hand and wrist positions relative to the head and spine position using a fixed displacement offset of 3 to measure velocity. Our analysis focuses on the processed version of this data, which includes information about the video phases: D (rest position, from the portuguese "descanso"), P (preparation), S (stroke), H (hold), and R (retraction). To prepare the data for analysis, we follow Hadj-Amar et al. (2023) and implement a preprocessing involving several stages. Initially, we apply a 2-point moving average filter to smoothen the time series. Subsequently, we perform downsampling by selecting every 5 points uniformly, as recommended by Romanuke (2021). This approach enables us to consider longer time lags while minimizing parameters and computational complexity. The data was then transformed using the square root function, which has been shown by Hadj-Amar et al. (2023) to be suitable for accommodating Gaussian distribution. Finally, we standardize the time series to ensure uniformity and comparability. The processed data is visualized in Figure 3.3.

Gesture phase segmentation poses several challenges. Firstly, it is a subjective task with no definitive starting point for each phase, resulting in varying segmentations by different experts for the same video. Moreover, some phases exhibit similar patterns, like stationary hand gestures such as the rest position and hold. Additionally, this data type can include nuisance movements, such as touching glasses while speaking, causing fluctuations in sensor-recorded scalar velocity (Madeo et al., 2013). Thus, we simplify our analysis by focusing our interest on just two phases: R (rest) and A (activities, specifically reading).

While Madeo et al. (2013) analyzed this dataset using Support Vector Machines (SVM) to segment gesture data streams, they didn't consider the temporal dependence between partitions. Our goal, in contrast, is to detect the latent cluster structure at each frame while capturing changepoints, using our proposed LLDPM random partition model with

time dependence.

To implement model (3.5), we consider an a priori expected number of clusters set to 2, corresponding to $\alpha = 0.5$. The sampling variance is assumed as $\sigma^2 \sim IG(a_0 = 15, b_0 = 3)$ and we set $\tau^2 \sim IG(a_0 = 15, b_0 = 3)$. The prior probability of a changepoint is assumed $\eta_t \sim \text{Beta}(a_t = 0.1, b_t = 0.9)$, suggesting that the model assumes a relatively low prior probability of a changepoint at each time point. For the distribution of the coefficient $\beta_{i,t}$, P_0, we assume a Gaussian Normal distribution with a mean of $\mu_t = 0$ and variance $\tau^2 = 0.5^2$. Our MCMC chains run for 10,000 iterations, with the first half discarded as a burn-in period.

Although we conducted comparisons with DRPM, WDDP, and GMDDP, we will focus here on presenting the results solely for LLDPM. Figures and results for the other models can be found in Appendix D.1. The LLDPM model primarily identifies changepoints during the activity phase, which corresponds to the preparation, retraction, and stroke phases in the video. Most of these changepoints occur during transitions between different phases, with a higher frequency observed during the reading phases. These changepoints could represent movements associated with the subject's gesticulation and body language while narrating or changes in hand and wrist positions while engaging in the assigned activity.

In the comparison with the other methods, we note that partition-based models tend to identify fewer changepoints. More specifically, the LLDPM identified a number of changepoints falling between those of DRPM and the other alternative competitors, instilling greater confidence in the model's performance, especially in real data applications where the ground truth is unknown. DRPM identified only 15 changepoints (4.3%), primarily at the beginning of the time series. On the other hand, LDDP, WDDP, and GMDDP detected 14% of changepoints, classifying 50 out of 349 observations as changepoints (see Appendix D.1).

Figure 3.5 illustrates the clustering of the eight sensors within a specific time window (for representation purposes), chosen by minimizing the lower bound of the posterior expected Variation of Information (VI). The numbers on the plot correspond to the sensors. Notably, a distinct pattern allocation emerges, where all the sensors are grouped into a single primary cluster. Additionally, a secondary pattern takes shape, forming two clusters: one containing all the sensors from the left arm (odd numbers) and the other with sensors from the right arm (even numbers) grouped together.

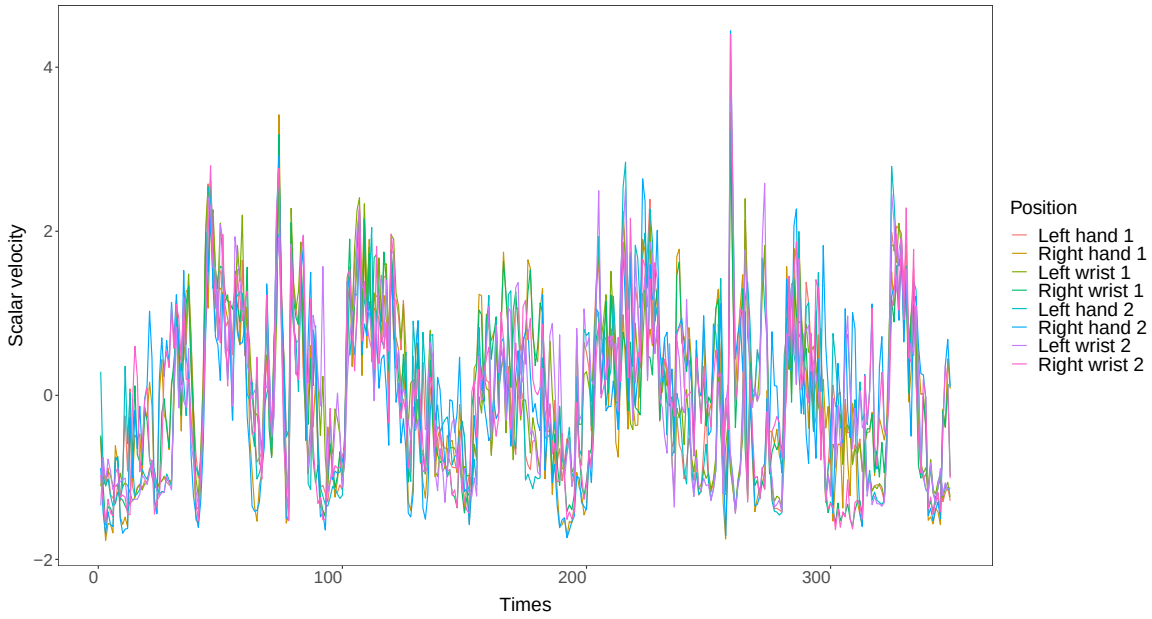


Figure 3.3: *Section 3.5: Human Gesture data. Scalar velocity of the left and right hand and the wrists after preprocessing ($T = 349$).*

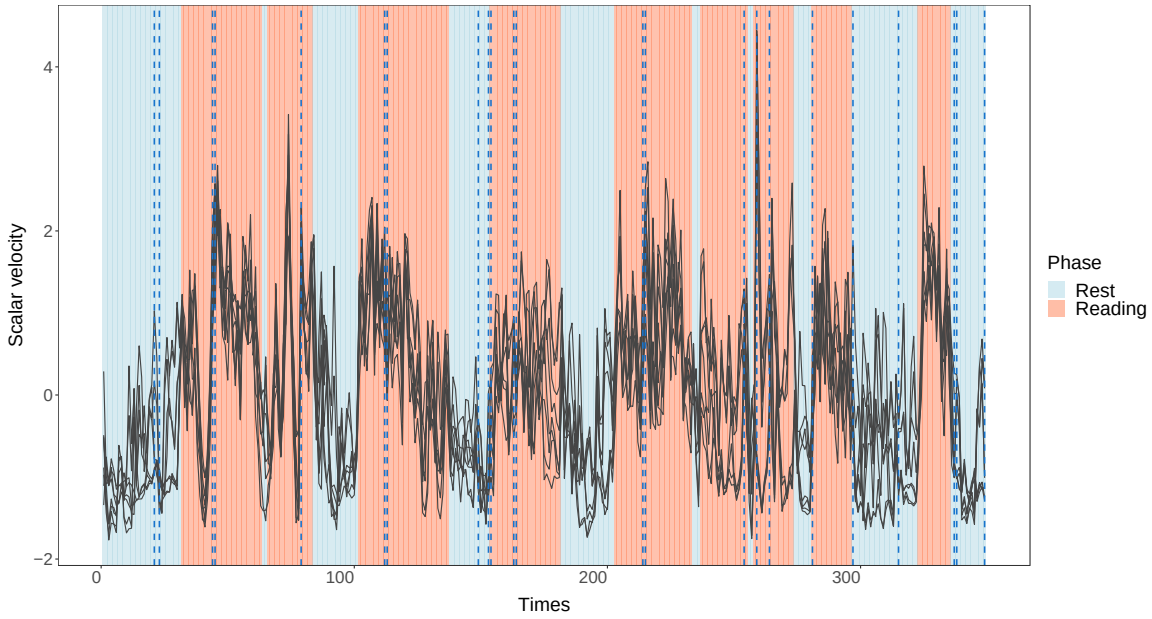


Figure 3.4: *Section 3.5: Human Gesture data. Estimated Changepoints for the LLDPM with a priori expected number of cluster 2. The two phases are visible in the background, while the vertical lines represent the detected changepoints.*

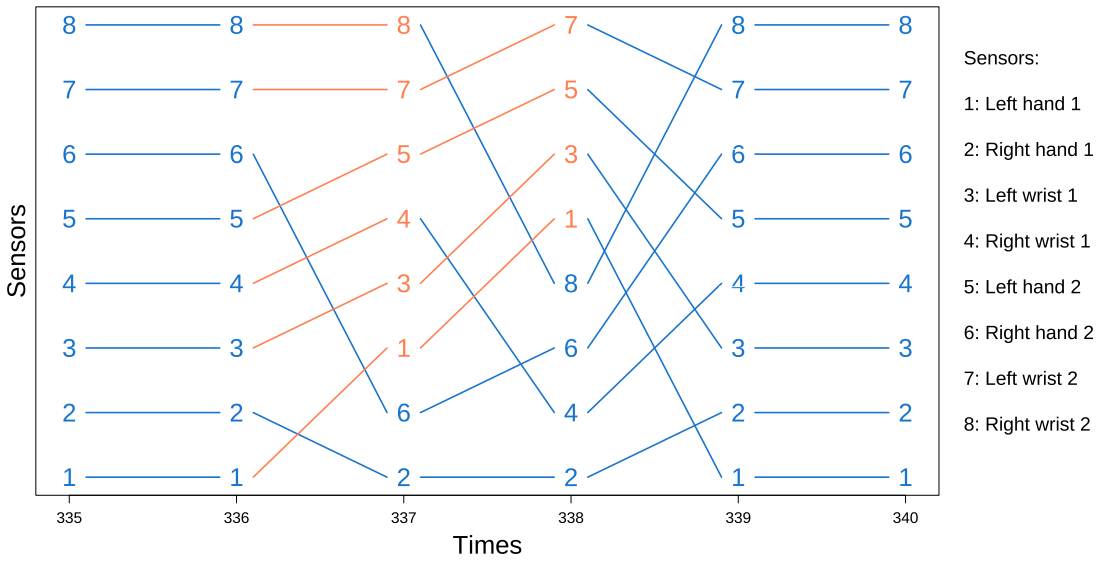


Figure 3.5: *Section Clusters for Gesture Phase Segmentation data calculated by minimum VI. The colours represent different clusters. On the x-axis are reported a time window from time 335 to time 340. The numbers correspond to the sensors.*

3.6 Discussion

In this chapter, we introduced an approach for modeling partitions with temporal dependence, with an application to human gesture sensor data analysis. Our novel random partition model links data partitions across different time points, extending the Dynamic Linear Model for multivariate time series analysis. By characterizing the evolution of the local level equation across partitions and detecting changepoints, our model effectively captures changes in cluster structures over time. Our approach maximizes the use of temporal information while introducing dependence only when supported by the data, by incorporating spike-and-slab priors to model independence between partitions, thus increasing model flexibility. We have illustrated the model's performance with synthetic data, highlighting its accuracy in recovering cluster structures and correctly identifying changepoints. In the application to gesture data, our model appears to provide reasonable inference on changepoint detection, particularly during activity phases. Notably, the model identifies a primary pattern allocation grouping all sensors and a secondary pattern with distinct left-arm and right-arm sensor clusters.

Leveraging the representation in Section 3.2.2, our model can be adapted to the analysis of multivariate data over multiple samples, where the samples are characterized by dependencies originating from sources other than temporal factors, such as different experimental conditions. Thus, the proposed model can consider diverse data structures and complex inter-condition dependencies. Importantly, even in this extended scenario, the partitions continue to follow a random partition model marginally, such as the Chinese restaurant process, highlighting the model's adaptability and versatility in capturing complex dependencies in various datasets.

The study of how gesture phases cluster offers valuable insights into the mechanics of human movement and holds the potential to significantly enhance gesture recognition systems. We can integrate the cluster inference derived from our random partition model as a predictor into a more general outcome model, e.g., to determine how various gesture intervals might be associated with a propensity for risk-taking or other behavioral outcomes. Thus, understanding these gesture phase clusters can lead to a deeper insight into human behavior, taking into account all probabilistic uncertainties.

Finally, it is worth noting that while we have focused on a relatively straightforward local-level model, an exciting avenue for future research is the development of dependent random partition models capable of modeling latent state equations that emulate the behavior of more complex dynamic linear models. This extension could encompass higher-order

time dependencies, allowing for the exploration of more complex temporal relationships and patterns in data. Similarly, we have assumed that the changepoint selection indicator $\gamma_{i,t}$ is independent across times. More in general, it could depend on time-varying covariates, such as respiratory data or measurements of expended effort in the analysis of human gesture data. Such developments have the potential to significantly enhance our ability to capture and understand time-varying clustering structures in complex time series data across a wide range of domains and applications.

Bibliography

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Antoniano-Villalobos, I. and Walker, S. G. (2016). A nonparametric model for stationary time series. *Journal of Time Series Analysis*, 37(1):126–142.

Arbel, J. and Favaro, S. (2021). Approximating predictive probabilities of gibbs-type priors. *Sankhya A*, 83(1):496–519.

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *The annals of Statistics*, pages 260–279.

Beraha, M., Guglielmi, A., Quintana, F. A., de Iorio, M., Eriksson, J. G., and Yap, F. (2022). Bayesian nonparametric vector autoregressive models via a logit stick-breaking prior: an application to child obesity. *arXiv preprint arXiv:2203.12280*.

Blackwell, D., MacQueen, J. B., et al. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355.

Caron, F., Davy, M., and Doucet, A. (2007). Generalized Pólya urn for time-varying Dirichlet process mixtures. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, page 33–40, Arlington, Virginia, USA. AUAI Press.

Caron, F., Neiswanger, W., Wood, F., Doucet, A., and Davy, M. (2017). Generalized pólya urn for time-varying pitman-yor processes. *Journal of Machine Learning Research*, 18(27).

Cassese, A., Zhu, W., Guindani, M., and Vannucci, M. (2019). A Bayesian Nonparametric Spiked Process Prior for Dynamic Model Selection. *Bayesian Analysis*, 14(2):553 – 572.

Chandra, N. K. and Bhattacharya, S. (2019). Non-marginal decisions: A novel bayesian multiple testing procedure. *Electronic Journal of Statistics*, 13(1):1535–1570.

Corradin, R., Canale, A., and Nipoti, B. (2020). Bnpmix: Bayesian nonparametric mixture models. *R package version 0.2*, 7:431.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013). Are gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE transactions on pattern analysis and machine intelligence*, 37(2):212–229.

De Iorio, M., Favaro, S., Guglielmi, A., and Ye, L. (2019). Bayesian nonparametric temporal dynamic clustering via autoregressive Dirichlet priors. *arXiv preprint arXiv:1910.10443*.

DeYoreo, M. and Kottas, A. (2018). Modeling for dynamic ordinal regression relationships: An application to estimating maturity of rockfish in california. *Journal of the American Statistical Association*, 113(521):68–80.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Gnedin, A. V. and Pitman, J. (2005). Exchangeable gibbs partitions and stirling triangles. *Записки научных семинаров ПОМИ*, 325(0):83–102.

Griffin, J. E. and Leisen, F. (2017). Compound random measures and their use in bayesian non-parametrics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):525–545.

Hadj-Amar, B., Jewson, J., and Vannucci, M. (2023). Bayesian sparse vector autoregressive switching models with application to human gesture phase segmentation. *arXiv preprint arXiv:2302.05347*.

Hartigan, J. A. (1990). Partition models. *Communications in statistics-Theory and methods*, 19(8):2745–2756.

Kalli, M. and Griffin, J. E. (2018). Bayesian nonparametric vector autoregressive models. *Journal of econometrics*, 203(2):267–282.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):715–740.

Lijoi, A., Nipoti, B., and Prünster, I. (2014). Bayesian inference with dependent normalized completely random measures.

Madeo, R. C., Lima, C. A., and Peres, S. M. (2013). Gesture unit segmentation using support vector machines: segmenting gestures from rest positions. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 46–52.

Muller, P., Parmigiani, G., and Rice, K. (2006). Fdr and bayesian multiple comparisons rules.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 4(1):155–176.

Nieto-Barajas, L. E. and Contreras-Cristán, A. (2014). A bayesian nonparametric approach for time series clustering. *Bayesian Analysis*, 9(1):147–170.

Page, G. L., Quintana, F. A., and Dahl, D. B. (2022). Dependent modeling of temporal sequences of random partitions. *Journal of Computational and Graphical Statistics*, 31(2):614–627.

Parvathy, P., Subramaniam, K., Prasanna Venkatesan, G., Karthikaikumar, P., Varghese, J., and Jayasankar, T. (2021). Development of hand gesture recognition system using machine learning. *Journal of Ambient Intelligence and Humanized Computing*, 12:6793–6800.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic linear models with R*. Springer Science & Business Media.

Pitman, J. (2002). Combinatorial stochastic processes lecture notes for st. flour summer school.

Pitman, J. (2006). *Combinatorial stochastic processes: Ecole d'eté de probabilités de saint-flour xxxii-2002*. Springer.

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

Quinlan, J. J., Page, G. L., and Castro, L. M. (2022). Joint random partition models for multivariate change point analysis. *Bayesian Analysis*, 1(1):1–28.

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). The dependent Dirichlet process and related models. *Statistical Science*, 37(1):24–41.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Romanuke, V. (2021). Time series smoothing improving forecasting. *Applied Computer Systems*, 26(1):60–70.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102:901–912.

Sun, W., Reich, B. J., Cai, T. T., Guindani, M., and Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 77:59–80.

Tadesse, M. G. and Vannucci, M. (2021). Handbook of bayesian variable selection.

Appendix

3.A Proof of Proposition

In the following, we provide the proof of the proposition stated in the article.

A.1 Proposition 1

Proof. (*the proof can be written in a more compact form*)

We consider two separate cases, namely the distribution of π_1 and the distribution of any π_t, with $t = 2, \dots, T$. We call $\mathscr{P}$ the space of all partitions of size n, and we have

$$
\begin{aligned}
p(\pi_1) &= \sum_{t=2}^{T} \sum_{\pi_t \in \mathscr{P}} p(\pi_{1:T}) \\
&= p^*(\pi_1) \sum_{t=2}^{T} \sum_{\pi_t \in \mathscr{P}} \prod_{t=2}^{T} \left[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t \, p^*(\pi_t) \right] \\
&= p^*(\pi_1) \sum_{t=3}^{T} \sum_{\pi_t \in \mathscr{P}} \prod_{t=4}^{T} \left[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t \, p^*(\pi_t) \right] \\
&\times \sum_{\pi_2 \in \mathscr{P}} \left[(1-\eta_2)\delta_{\pi_1}(\pi_2) + \eta_2 \, p^*(\pi_2) \right] \left[(1-\eta_3)\delta_{\pi_2}(\pi_3) + \eta_3 \, p^*(\pi_3) \right] \\
&= p^*(\pi_1) \sum_{t=3}^{T} \sum_{\pi_t \in \mathscr{P}} \prod_{t=4}^{T} \left[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t \, p^*(\pi_t) \right] \\
&\times \sum_{\pi_2 \in \mathscr{P}} \left[(1-\eta_2)(1-\eta_3)\delta_{\pi_1}(\pi_2)\delta_{\pi_3}(\pi_2) + (1-\eta_2)\eta_3 \delta_{\pi_1}(\pi_2) p^*(\pi_3) \right. \\
&+ \left. \eta_2(1-\eta_3)\delta_{\pi_3}(\pi_2) p^*(\pi_2) + \eta_2\eta_3 p^*(\pi_2) p^*(\pi_3) \right] \\
&= p^*(\pi_1) \sum_{t=3}^{T} \sum_{\pi_t \in \mathscr{P}} \prod_{t=4}^{T} \left[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t \, p^*(\pi_t) \right] \\
&\times \left[(1-\eta_2)(1-\eta_3)\delta_{\pi_1}(\pi_3) + (1-\eta_2)\eta_3 p^*(\pi_3) + \eta_2(1-\eta_3) p^*(\pi_3) + \eta_2\eta_3 p^*(\pi_3) \right] \\
&= p^*(\pi_1) \sum_{t=3}^{T} \sum_{\pi_t \in \mathscr{P}} \prod_{t=4}^{T} \left[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t \, p^*(\pi_t) \right] \\
&\times \left[(1-\eta_2)(1-\eta_3)\delta_{\pi_1}(\pi_3) + (\eta_2 + \eta_3 - \eta_2\eta_3) p^*(\pi_3) \right]
\end{aligned}
$$

$$= p^*(\pi_1) \sum_{t=3}^{T} \sum_{\pi_t \in \mathscr{P}} \left[(1-\tilde{\eta}_3)\delta_{\pi_1}(\pi_3) + \tilde{\eta}_3 p^*(\pi_3)\right] \prod_{t=4}^{T} \left[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t\, p^*(\pi_t)\right],$$

where, for any $t = 2, \ldots, T$, $\tilde{\eta}_t$ is defined by

$$\begin{aligned} &\tilde{\eta}_2 = \eta_2 \\ &\tilde{\eta}_t = 1 - (1-\eta_t) \prod_{r=2}^{t-1} (1-\tilde{\eta}_r), \qquad\qquad \text{for } t = 3, \ldots, T. \end{aligned}$$

By iterating the same procedure, we get

$$\begin{aligned} p(\pi_1) &= p^*(\pi_1) \sum_{\pi_T \in \mathscr{P}} \left[(1-\tilde{\eta}_T)\delta_{\pi_1}(\pi_T) + \tilde{\eta}_T\, p_{\mathrm{CRP}}(\pi_T; \alpha)\right] \\ &= p^*(\pi), \end{aligned}$$

that is, π_1 is marginally distributed as $p^*(\cdot)$. As for a generic π_t, with $t = 2, \ldots, T$, we have:

$$\begin{aligned} p(\pi_t) &= \sum_{\substack{r=1 \\ r \neq t}}^{T} \sum_{\pi_r \in \mathscr{P}} p(\pi_{1:T}) \\ &= \sum_{\substack{r=2 \\ r \neq t}}^{T} \sum_{\pi_r \in \mathscr{P}} \prod_{r=3}^{T} \left[(1-\eta_r)\delta_{\pi_{r-1}}(\pi_r) + \eta_r\, p^*(\pi_r)\right] \\ &\times \sum_{\pi_1 \in \mathscr{P}} p^*(\pi_1) \left[(1-\eta_2)\delta_{\pi_2}(\pi_1) + \eta_2\, p^*(\pi_2)\right] \\ &= \sum_{\substack{r=2 \\ r \neq t}}^{T} \sum_{\pi_r \in \mathscr{P}} \prod_{r=3}^{T} \left[(1-\eta_r)\delta_{\pi_{r-1}}(\pi_r) + \eta_r\, p^*(\pi_r; \alpha)\right] \left[(1-\eta_2)p^*(\pi_2) + \eta_2\, p^*(\pi_2)\right] \\ &= \sum_{\substack{r=2 \\ r \neq t}}^{T} \sum_{\pi_r \in \mathscr{P}} p^*(\pi_2) \prod_{r=3}^{T} \left[(1-\eta_r)\delta_{\pi_{r-1}}(\pi_r) + \eta_r\, p^*(\pi_r)\right]. \end{aligned}$$

By iterating the same procedure for the first $t-1$ terms of the first sum, we get

$$p(\pi_t) = p^*(\pi_t) \sum_{r=t+1}^{T} \sum_{\pi_r \in \mathscr{P}} \prod_{r=t+1}^{T} \left[(1-\eta_r)\delta_{\pi_{r-1}}(\pi_r) + \eta_r\, p^*(\pi_r)\right].$$

From this point the proof proceeds exactly as for the case of π_1, which we already studied.

That is, π_t is marginally distributed as $p^*(\cdot)$. □

3.B Posterior MCMC details

B.1 General case

Starting from the distribution of $p(\pi_t^{(b)} \mid \gamma_t^{(b-1)}, \ldots)$, at each iteration $b = 1, \ldots, B$, we can write the full conditional distributions (forgetting the subscription $^{(b)}$ in the conditional part of the formula to have less notation involved), as follows

$$\begin{aligned} p(\pi_t^{(b)} \mid \gamma_t^{(b-1)}, \ldots) &\propto \sum\nolimits_{\gamma_t^{(b-1)} \in \{0,1\}} \eta_t^{\gamma_t} (1-\eta_t)^{1-\gamma_t} \Big[(1-\gamma_t)\delta_{\pi_{t-1}}(\pi_t) + \gamma_t p_{\mathrm{CRP}_\alpha}(\pi_t)\Big] \times \\ &\quad \Big[(1-\gamma_{t+1})\delta_{\pi_{t+1}}(\pi_t) + \gamma_{t+1} p_{\mathrm{CRP}_\alpha}(\pi_{t+1})\Big] G(\boldsymbol{Y}_t \mid \pi_t) \\ &\propto \Big[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t p_{\mathrm{CRP}_\alpha}(\pi_t)\Big] \times \\ &\quad \Big[(1-\gamma_{t+1})\delta_{\pi_{t+1}}(\pi_t) + \gamma_{t+1} p_{\mathrm{CRP}_\alpha}(\pi_{t+1})\Big] G(\boldsymbol{Y}_t \mid \pi_t). \end{aligned}$$

Given this conditional distribution we have to consider two different cases, when $\gamma_{t+1} = 1$ and $\gamma_{t+1} = 0$,

1. If $\gamma_{t+1}^{(b-1)} = 1$:

$$\begin{aligned} p(\pi_t^{(b)} \mid \gamma_t^{(b-1)}, \ldots) &\propto \Big[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t p_{\mathrm{CRP}_\alpha}(\pi_t)\Big] p_{\mathrm{CRP}_\alpha}(\pi_{t+1}) G(\boldsymbol{Y}_t \mid \pi_t) \\ &\propto \Big[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t p_{\mathrm{CRP}_\alpha}(\pi_t)\Big] G(\boldsymbol{Y}_t \mid \pi_t) \\ &\propto (1-\eta_t) G(\boldsymbol{Y}_t \mid \pi_{t-1})\delta_{\pi_{t-1}}(\pi_t) + \eta_t p_{\mathrm{CRP}_\alpha}(\pi_t) G(\boldsymbol{Y}_t \mid \pi_t); \end{aligned}$$

   that is, given that $\gamma_{t+1} = 1$, π_t coincides with π_{t+1} with probability proportional to $(1-\eta_t)G(\boldsymbol{Y}_t \mid \pi_{t+1})$. Besides, it is generated from a distribution proportional to $p_{\mathrm{CRP}_\alpha}(\pi_t)G(\boldsymbol{Y}_t \mid \pi_t)$ (second issue) with probability proportional to $\sum_{\pi_t \in \mathscr{P}_t} p_{\mathrm{CRP}_\alpha}(\pi_t)G(\boldsymbol{Y}_t \mid \pi_t)$ (first issue). However, to generate this new partition we encounter two main issues: how to evaluate the sum over all the possible partition and how to generate a new partition weighted by $G(\boldsymbol{Y}_t \mid \pi_t)$. Solutions to these problems are presented in Section 3.3.

2. If $\gamma_{t+1}^{(b-1)} = 0$:

$$p(\pi_t^{(b)} \mid \gamma_t^{(b-1)}, \ldots) \propto \Big[(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) + \eta_t p_{\mathrm{CRP}_\alpha}(\pi_t)\Big]\delta_{\pi_{t+1}}(\pi_t) G(\boldsymbol{Y}_t \mid \pi_t)$$
$$\propto (1-\eta_t)\delta_{\pi_{t-1}}(\pi_t) G(\boldsymbol{Y}_t \mid \pi_{t-1}) + \eta_t p_{\mathrm{CRP}_\alpha}(\pi_{t+1}) G(\boldsymbol{Y}_t \mid \pi_t);$$

that is, given $\gamma_{t+1} = 0$, π_t coincides with π_{t-1} with probability proportional to $(1-\eta_t)\delta_{\pi_{t-1}}(\pi_t)G(\boldsymbol{Y}_t \mid \pi_{t-1})$ and it coincides with π_{t+1} with probability proportional to $\eta_t p_{\mathrm{CRP}_\alpha}(\pi_{t+1})G(\boldsymbol{Y}_t \mid \pi_t)$. Hence, since $\gamma_{t+1} = 0$, we have that $P(\pi_t \equiv \pi_{t+1}) = 1$.

To update γ_t, in each iteration of the gibbs sampling $b = 1, \ldots, B$, we have to evaluate:

- $p(\gamma_t^{(b)} = 1 \mid \pi_t^{(b)}, \ldots) \propto \eta_t^{(b-1)} p_{\mathrm{CRP}_\alpha}(\pi_t)$;
- $p(\gamma_t^{(b)} = 0 \mid \pi_t^{(b)}, \ldots) \propto (1-\eta_t^{(b-1)})\delta_{\pi_{t-1}^{(b)}}(\pi_t^{(b)})$;

In particular,

1. If $\pi_t^{(b)} = \pi_{t-1}^{(b)}$, then
   - $p(\gamma_t^{(b)} = 1 \mid \pi_t^{(b)}, \ldots) \propto \eta_t^{(b-1)} p_{\mathrm{CRP}_\alpha}(\pi_{t-1})$
   - $p(\gamma_t^{(b)} = 0 \mid \pi_t^{(b)}, \ldots) \propto (1-\eta_t^{(b-1)})$
2. If $\pi_t^{(b)} \neq \pi_{t-1}^{(b)}$, then $P(\gamma_t^{(b)} = 1) = 1$.

Then, we can update the parameters $\boldsymbol{\beta}$ such that

$$P(\boldsymbol{\beta}_t^{(b)} \mid \pi_t^{(b)}, \ldots) \propto \prod_{j=1}^{|\pi_t|} \prod_{i \in C_j(\pi_t)} P(y_{i,t} \mid \beta_{j,t}^*) P_0(\beta_{j,t}^*).$$

Moreover, the posterior distribution of $\eta_t^{(b)}$ with prior distribution $\eta_t \sim Beta(a, b)$ is again a $Beta$ distribution but with updated parameters: $p(\eta_t^{(b)} \mid \ldots) \sim Beta(a + \gamma_t^{(b)}; b + 1 - \gamma_t^{(b)})$.

B.2 The location Normal kernel scenario

In this Section we provide more details related to the specification of the distribution for the MCMC algorithm described in Section 3.3. In particular, in the following there are the distribution involved in the gibbs sampling scheme when we specified a location Normal kernel for $Y_{i,t}$ ($i = 1, \ldots, n$ and $t = 1, \ldots, T$).

Starting from the full model (3.5). Let

$$
\begin{aligned}
Y_{i,t}|\beta_{i,t} &\overset{ind}{\sim} N(\beta_{i,t};\sigma^2);\\
\beta_t \mid \pi_t &\sim \prod_{j=1}^{|\pi_t|} P_0\left(\beta^*_{j,t}\right);\\
P_0 &= N(\mu;\tau^2).
\end{aligned}
$$

We can express the joint distribution as

$$
p(\boldsymbol{Y},\boldsymbol{\pi},\boldsymbol{\gamma}) = \prod_{t=1}^{T} \eta_t^{\gamma_t}(1-\eta_t)^{1-\gamma_t} \prod_{t=1}^{T} \left[(1-\gamma_t)\delta_{\pi_{t-1}}(\pi_t) + \gamma_t p_{\mathrm{CRP}_\alpha}(\pi_t)\right] G(\boldsymbol{Y}_t \mid \pi_t).
$$

where in this specific setting, we have that

$$
\begin{aligned}
\log G(\mathbf{Y_t}|\pi_\mathbf{t}) = \sum_{\mathbf{j=1}}^{|\pi_\mathbf{t}|} &-\frac{n_j}{2}\log(2\pi) - n_j\log\sigma - \log\tau + \frac{1}{2}\log\frac{\sigma^2\tau^2}{n_j\tau^2+\sigma^2}\\
&-\frac{1}{2}\sum_{i=1}^{n_j} y_{i,t}^2 - \frac{1}{2\tau^2}\mu^2 + \frac{1}{2}\frac{\sigma^2\tau^2}{n_j\tau^2+\sigma^2}\cdot\left(\frac{\mu}{tau^2} + \frac{\sum_{i=1}^{n_j} y_{i,t}}{\sigma^2}\right).
\end{aligned}
$$

Before proceed with the MCMC scheme, we implement a warmup step that includes sampling from a CRP_α partitions for each time t. To speed up this step, it is convenient to run it in parallel. Then, in each gibbs sampling iteration $b = 1,\ldots,B$. Our strategy iterates four main steps:

1. update $\pi_t^{(b)}$ from $p(\pi_t^{(b)} \mid \gamma_t^{(b-1)},\ldots)$;
2. update $\gamma_t^{(b)}$ from $p(\gamma_t \mid \pi_t^{(b)},\ldots)$;
3. update $\boldsymbol{\beta}_t^{(b)}$ from $p(\boldsymbol{\beta}_t^{(b)} \mid \ldots)$;
4. update all the hyperparameters;
5. reshuffling step.

Following this procedure for each iteration b allows to obtain the MCMC chains for the posterior inference. The full conditional distributions are provided in the previous section.

B.3 Reshuffling step

Sampling from π_t given $\boldsymbol{Y}_t$ it is achieved by means of the warmup step where we use a Pólya urn to generate several realizations of $\pi_t|\boldsymbol{Y}_t$ and create a list $\mathscr{S}_t$ of partitions, from which we can draw whenever we need to sample from $\pi_t|\boldsymbol{Y}_t$. This is done for $t = 1, \ldots, T$ leading to T lists of partitions: $\{\mathscr{S}_1, \ldots, \mathscr{S}_t, \ldots, \mathscr{S}_T\}$.

An acceleration step or reshuffling step would require an update of the visited partitions at the end of each iteration. Say that after a given iteration the current configuration is, for example,

$$\underbrace{\pi_1 \equiv \pi_2 \equiv \pi_3}_{\pi_1^*} \neq \underbrace{\pi_4.}_{\pi_2^*}$$

In this case we would update:

$$\pi_1^* \mid \boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3; \tag{3.11}$$

$$\pi_2^* \mid \boldsymbol{Y}_4 \tag{3.12}$$

where 3.12 is straightforward and 3.11 requires more care.
Our model is such that, after marginalizing $\boldsymbol{\beta}$,

$$\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3 \sim \prod_{j=1}^{3} G(\boldsymbol{Y}_j \mid \pi_1^*)$$

and we need to sample from

$$P(\pi_1^* \mid \boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3) \propto P(\pi_1^*) \prod_{j=1}^{3} P(\boldsymbol{Y}_j \mid \pi_1^*) \propto p_{\mathrm{CRP}_\alpha}(\pi_1^*) \prod_{j=1}^{3} G(\boldsymbol{Y}_j \mid \pi_1^*). \tag{3.13}$$

An option is to use a Pólya urn (details not obviuos) but this would be a problem cause it involves a Gibbs sampler within a Gibbs sampler. To avoid this mechanism we exploit the lists $\{\mathscr{S}_1, \ldots, \mathscr{S}_t, \ldots, \mathscr{S}_T\}$ with a sampling importance resampling step:

i. Say we sample:

- m_1 draws from $\mathscr{S}_1$ (i.e. from $\pi_1 \mid \boldsymbol{Y}_1$);
- m_2 draws from $\mathscr{S}_2$ (i.e. from $\pi_2 \mid \boldsymbol{Y}_2$);
- m_3 draws from $\mathscr{S}_3$ (i.e. from $\pi_3 \mid \boldsymbol{Y}_3$).

We get a $m = m_1 + m_2 + m_3$ candidate partitions from $\mathscr{S}_1^* \subset \mathscr{S}_1 \cup \mathscr{S}_2 \cup \mathscr{S}_3$.

ii. We assign a weight w_j for each partition π_j in $\mathscr{S}_1^*$, with $j = 1, \ldots, m$. Each weight is given, up to a proportional constant, by evaluating 3.13 at $\pi_j^* = \pi_j$.

iii. We sample one new value for π_1^* from $\mathscr{S}_1^*$, with probabilities $w_1, \ldots, w_m$.

3.C Simulated Data

In this Appendix we present additional plot referring to examples of synthetic data presented in Section 3.4.

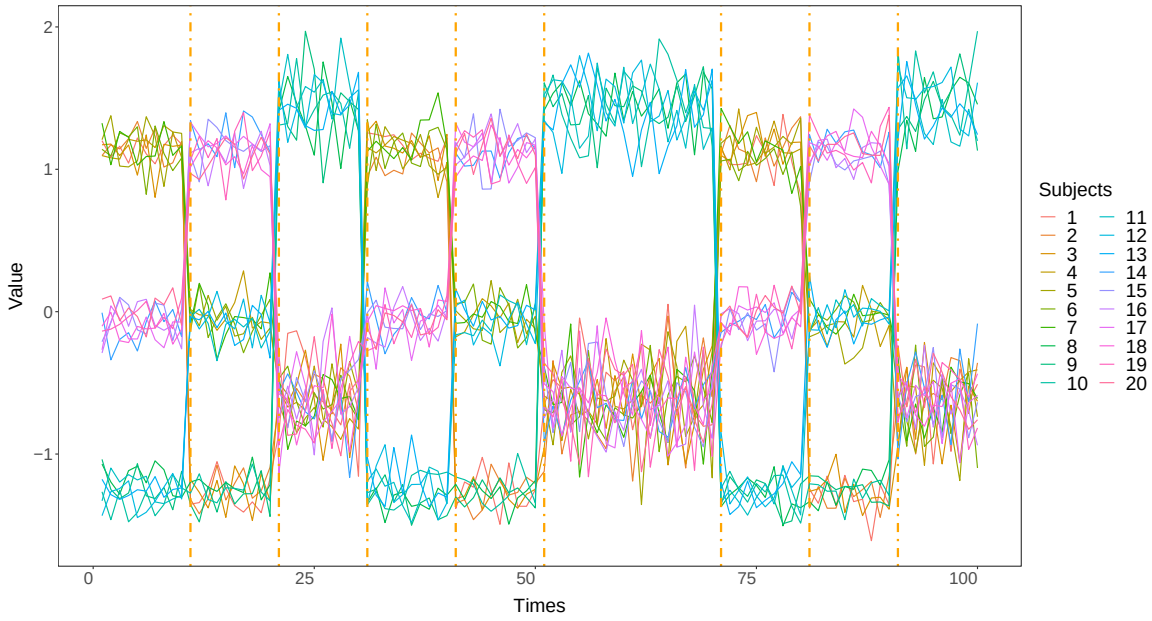


Figure 3.6: *Independent data with $n = 20$ subjects and 100 time points. The orange vertical lines correspond to changepoints.*

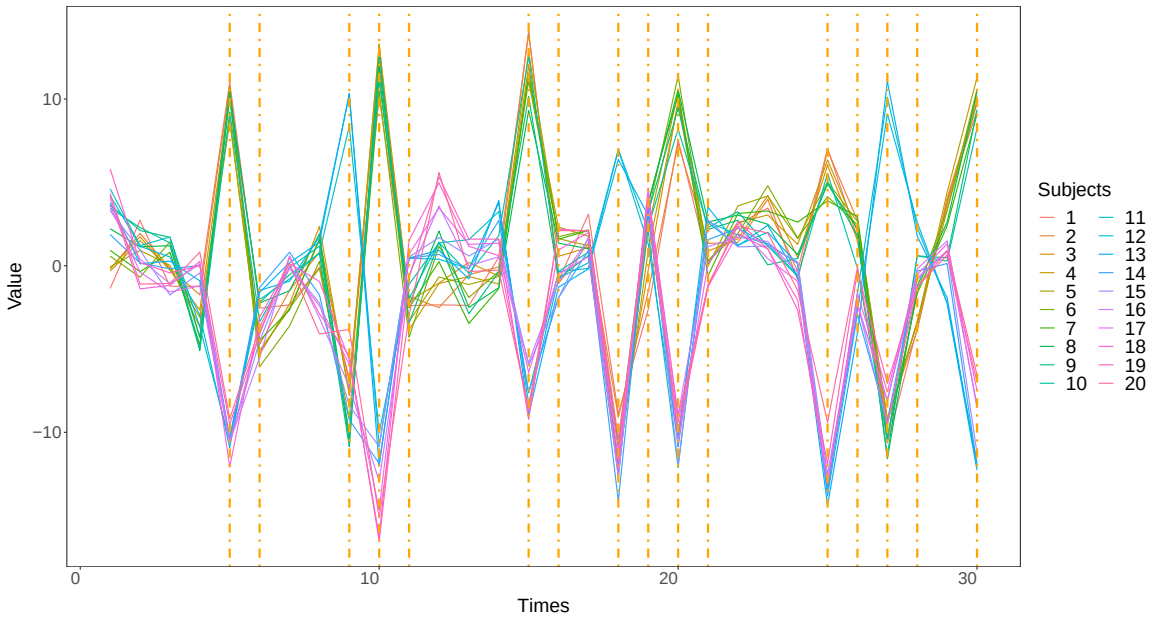


Figure 3.7: *Autoregressive data with $n = 20$ subjects and 30 time points. The orange vertical lines correspond to changepoints. The autoregressive coefficient is 0.9.*

3.D Real data

In this Appendix we present additional plot referring to data and performances of the models presented in Section 3.5.

D.1 Gesture Phase Segmentation

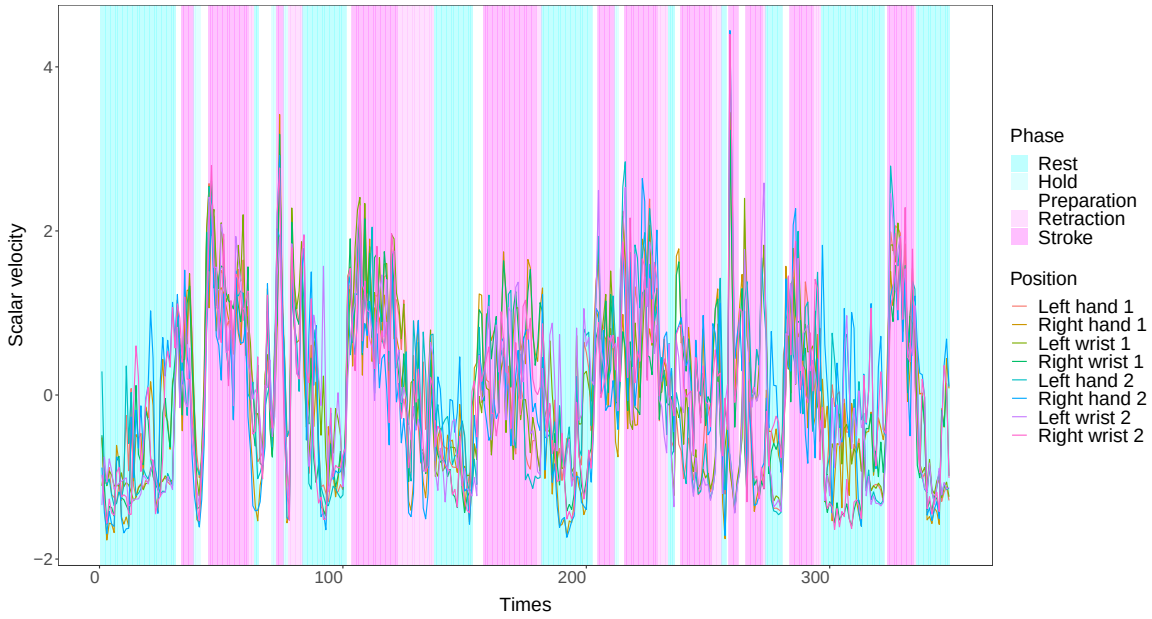


Figure 3.8: *Gesture data after preprocessing with 349 time points and the phases of the video on the background.*

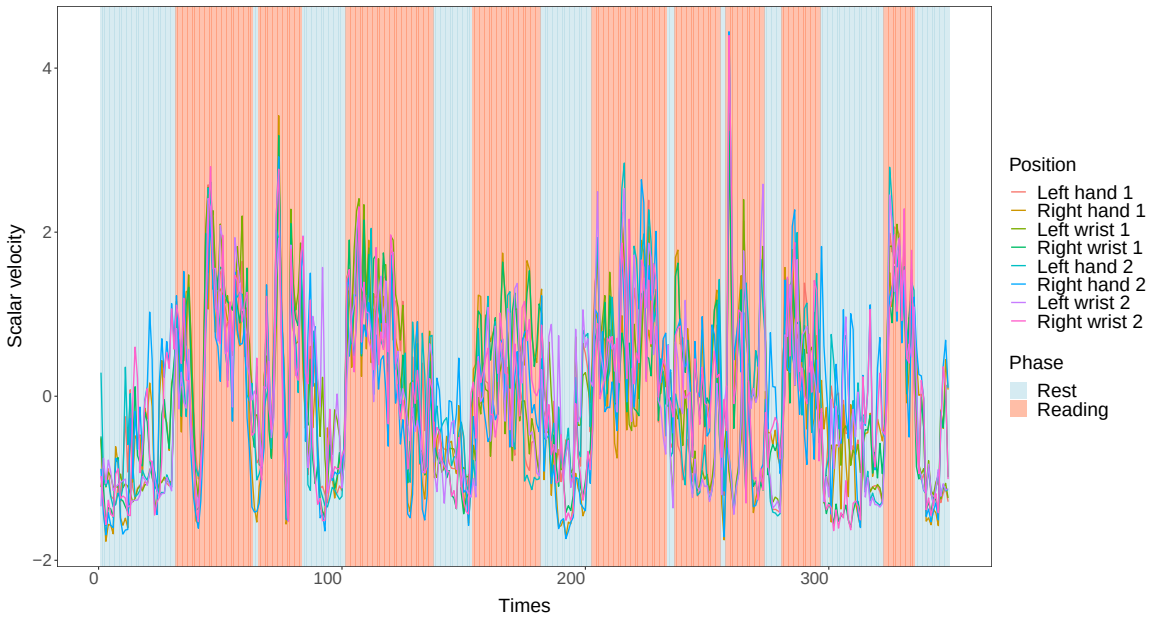


Figure 3.9: *Gesture data after preprocessing with 349 time points and the grouped phases of the video on the background.*

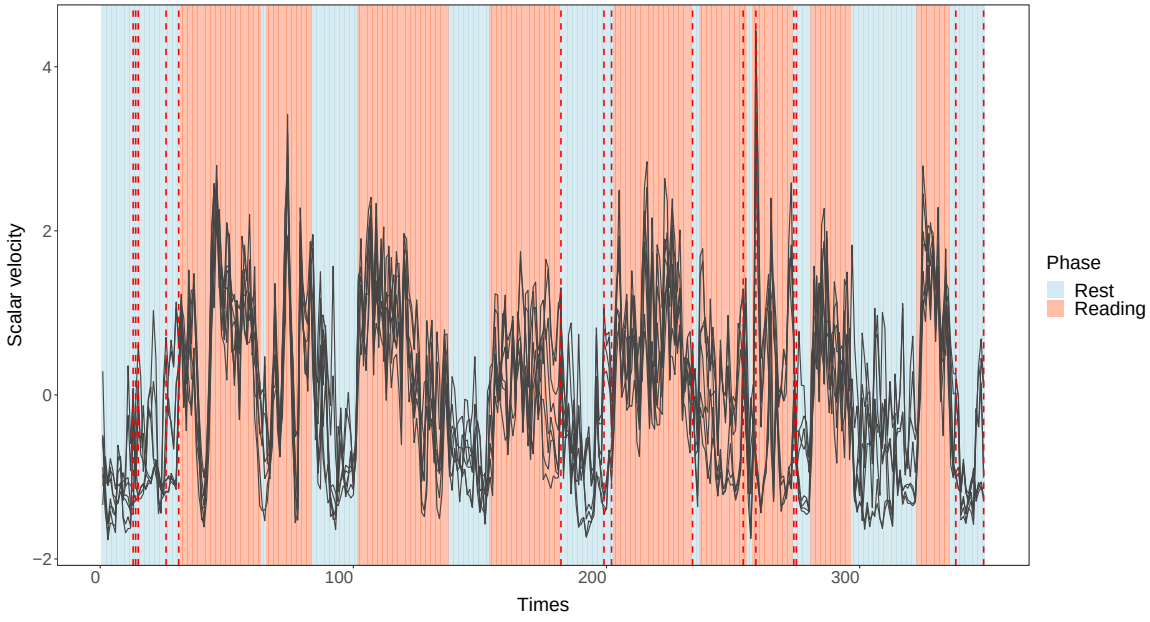


Figure 3.10: *Changepoints detection with a priori expected number of cluster 2 for Gesture Phase data with DRPM model. In the background is possible to see the two phases, while the vertical lines are the changepoints.*

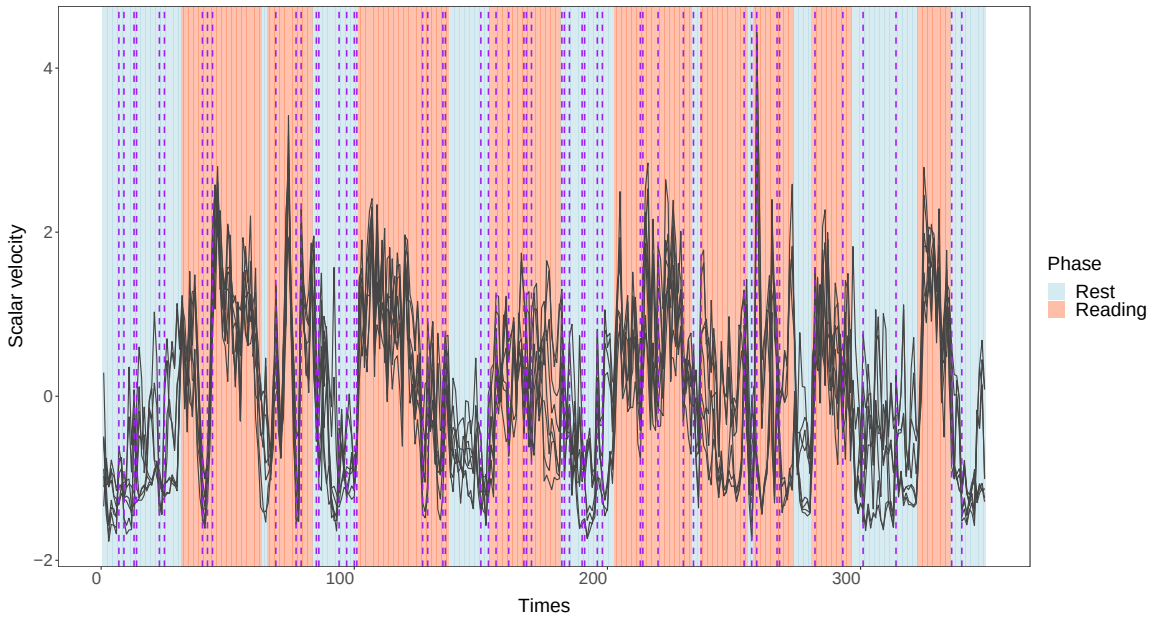


Figure 3.11: *Changepoints detection with a priori expected number of cluster 2 for Gesture Phase data with LDDP model. In the background is possible to see the two phases, while the vertical lines are the changepoints.*

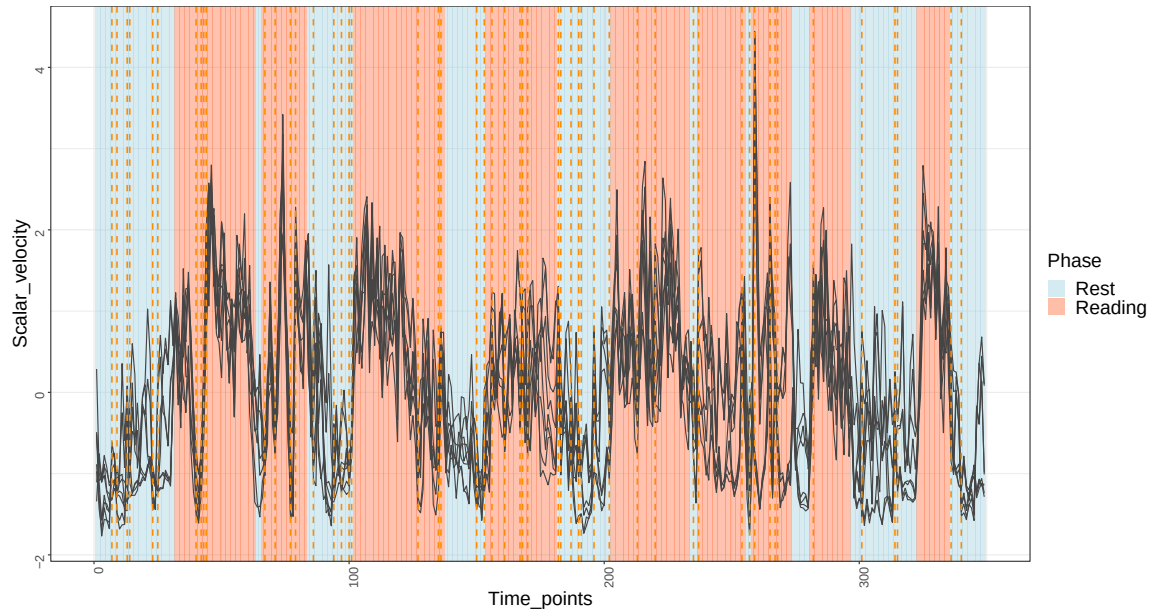


Figure 3.12: *Changepoints detection with a priori expected number of cluster 2 for Gesture Phase data with WDDP model. In the background is possible to see the two phases, while the vertical lines are the changepoints.*

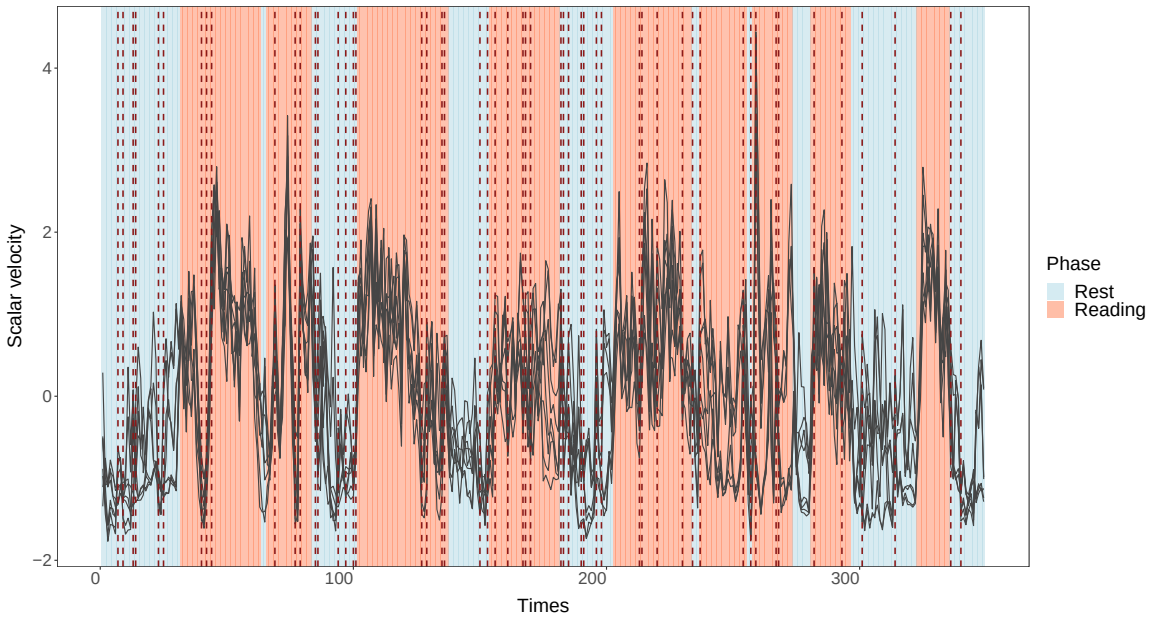


Figure 3.13: *Changepoints detection with a priori expected number of cluster 2 for Gesture Phase data with GMDDP model. In the background is possible to see the two phases, while the vertical lines are the changepoints.*

Chapter 4

A generalization of the latent Dirichlet allocation

4.1 Introduction

Information retrieval, social media analysis, semantic mining, spam filters, and genomics are only a few of the main fields where topic modeling is of great interest. These models can be used for searching results by understanding the topics in documents and matching them to user queries or by identifying the topics of interest to users for recommending relevant articles, products, or contents. Furthermore, combining topic modeling with sentiment analysis helps to determine the sentiment expressed within specific topics. The analysis of topics in text documents has been extensively studied to enable computers to obtain meaning from language processing with the purpose of content summarization, document clustering and content recommendation. Over recent years, several techniques of text analysis have been explored, especially challenging for computer science and statistics researchers. The goal is to provide a document representation in terms of latent topics distribution. An example with three documents represented as a mixture of three topics is provided in Figure 4.1. Using these techniques, it is possible to identify the text's primary subjects and determine their importance within the text using topic modeling approaches.

Topic modeling is a statistical technique used in natural language processing (NLP) and text mining to uncover the underlying themes or topics within a collection of documents. It is particularly useful for organizing, summarizing, and understanding large text corpora. The fundamental idea behind topic modeling is that each document is a mixture of several topics, and each observed word in a document is attributed to one of these topics. By discovering these topics and their proportions in each document, we gain insight into the main themes and structure of the corpus. Moreover, topics are considered latent variables, i.e. they are not directly observable but inferred from the distribution of words in docu-

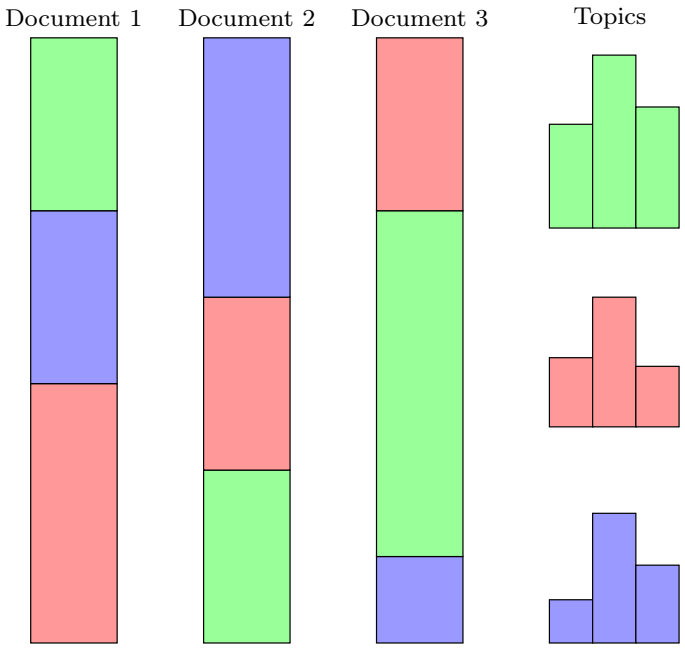


Figure 4.1: *On the left, representation of three documents as mixtures of three topics (red, blue and green). On the right, histograms of words attributed to the different topics.*

ments. For these techniques the "bag-of-words" assumption is implied, i.e. the arrangement of words in a document holds no significance; what truly matters is the frequency of each word's occurrence. This is an assumption of exchangeability for the words in a document (Aldous et al., 1985).

Let us consider a set of D text documents, labelled as $\mathcal{C}$, also called "corpus". Each document in this collection, denoted as the d-th document, is essentially a sequence of N_d words, which can be expressed as $(w_{d,1}, \ldots, w_{d,N_d})^\intercal$. Here, $w_{d,n}$ indicates the n-th word of the d-th document, with d ranging from 1 to D and n spanning from 1 to N_d. The distinct words present in the corpus constitute the "vocabulary", represented by the set $\mathcal{V}$ containing V unique words. Topic modeling methods work on the principle that any word within a document is generated from one of T possible topics. Consequently, the representation of the d-th document can be viewed as a vector $\boldsymbol{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,T})^\intercal$, where $\theta_{d,t}$ depicts the ratio of words in the d-th document that are derived from topic t, with $t = 1, \ldots, T$. It is evident that $\boldsymbol{\theta}_d$ falls within the T-dimensional simplex, $\mathcal{S}^T = \{\boldsymbol{\theta} : \theta_t > 0, \sum_{t=1}^{T} \theta_t = 1\}$. In a parallel fashion, each topic is denoted by a discrete probability distribution, $\boldsymbol{\phi}_t$, across the vocabulary $\mathcal{V}$. For every topic, from $t = 1$ to T, $\boldsymbol{\phi}_t$ resides in $\mathcal{S}^V$.

The Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is one of the most widely

used statistical tools for topic modeling. LDA uses word probabilities to represent topics, coherently with the framework sketched above. For each topic, the set of the most probable words with the highest probabilities typically gives a good representation of what the topic is about. The LDA improved the way of approaching topic models, when compared to the Non-Negative Matrix Factorization (NMF), proposed by Lee and Seung (1999). NMF factorizes the document-term matrix, i.e. a matrix that describes the frequency of terms that occur in a collection of documents, into two lower-dimensional matrices, one representing topics and the other representing the distribution of topics in documents. Another way to detect topics within a text using decomposition is the Latent Semantic Analysis (LSA) (Deerwester et al., 1990). LSA, also known as Latent Semantic Indexing (LSI), uses singular value decomposition to reduce the dimensionality of the term-document matrix and discover latent topics. A predecessor to LDA, based on LSA, was presented by Hofmann (1999) with the name of Probabilistic Latent Semantic Analysis (pLSA). It models the likelihood of words in documents given topics directly. However, it may suffer from overfitting when the number of topics is large relative to the size of the dataset. While pLSA was an important step in the development of probabilistic topic modeling, it has been largely superseded by Latent Dirichlet Allocation, which addresses its limitations, like overfitting, and provides a more principled probabilistic framework for topic modeling.

In the past two decades, there has been a proliferation of models built upon LDA. A comprehensive review of the principal methods for topic modeling derived from LDA from 2003 to 2016 can be found in Jelodar et al. (2019). More recent developments in Bayesian graphical models and related probabilistic topic modeling, as well as some noteworthy applications, are presented in Wood (2014). It is worth mentioning the Dynamic Topic Models (DTM), introduced by Blei and Lafferty (2006), which extends LDA to model how topics evolve over time in a corpus, making it suitable for analyzing temporal data. The online LDA (Hoffman et al., 2010) is an efficient and scalable variant of LDA, designed for processing large datasets in an online manner. Benton and Dredze (2018) introduced the deep Dirichlet Multinomial regression. It combines LDA with Dirichlet-Multinomial regression, allowing to capture topic associations with covariates or metadata, incorporating arbitrary document-level features to inform topic priors.

However, a typical problem that is not addressed by all these methodologies, based on LDA, is the incapacity to model topic correlation as well as to allow positive correlations between topics. These limitations are attributable to the stiffness of the Dirichlet distribution, which is the standard prior for the topic distributions involved in the LDA. Blei and Lafferty (2007) presented a solution using the logistic-Normal distribution. A critic to this

approach is that the interpretability of the topic is set aside in favor of easy correlation computation. Indeed, in this technique the logistic-Normal distribution is employed as the distribution for the categorical probability parameters after they have been translated from a Euclidean space to the simplex. It presents a notable challenge in interpreting parameters. This complexity stems from the log-ratio transformation that enables the model to be defined on the space of real numbers. Our proposed model has the goal of overcoming all these problems. It is fully interpretable since the distributions are defined on the proper bounded domain, i.e. the simplex. Furthermore, it still maintains many of the properties of the Dirichlet distribution.

A first extension of the Dirichlet distribution in this direction was presented by Ongaro and Migliorati (2013), where only negative correlations are allowed. Successively, Ongaro et al. (2020) presented an extension with more parameters, which helps in modeling compositional data in a more flexible way. This distribution is the basis of the extension of the LDA presented in Section 4.3. The advantage of employing the extended flexible Dirichlet (EFD) distribution lies in its enhanced parameterization, which ensures increased flexibility in the covariance matrix. As a specific instance of the EFD, the traditional Dirichlet distribution can be reclaimed through an appropriate choice of its parameters.

4.2 Corpus generating mechanisms

In this section, the two main techniques to generate documents for topic modelling analyses are introduced. In particular, we are going to explain in detail the LDA and Correlated Topic Model (CTM). They require to specify a distribution for the topics, which inherits the properties and the limits of the chosen distribution.

4.2.1 Latent Dirichlet Allocation

The LDA is a generative probabilistic model composed of a three-level hierarchical Bayesian structure introduced by Blei et al. (2003). In the LDA, documents are represented as random mixtures over latent topics, and each topic has its own distribution over a set of words $\mathcal{V}$.

The LDA model assumes that documents belonging to a corpus are composed by T common topics, each document being thought of as a probability distribution $\boldsymbol{\theta}_d$ over the topics, $d = 1, \dots, D$. Moreover, a topic is represented as a probability distribution $\boldsymbol{\phi}_t$ over the vocabulary $\mathcal{V}$, $t = 1, \dots, T$. Thus, a document and a topic may be depicted as a point

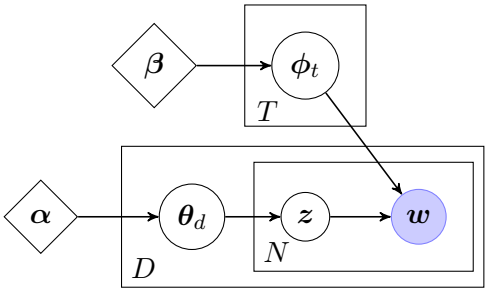


Figure 4.2: *DAG describing LDA model. The unobserved variables are drawn as circles whereas the observed ones are filled by blue color. Rhombuses represent hyperparameters. The outer rectangle represents documents, while the inner one represents the repeated topics and words within a document.*

in the T-part ($\mathcal{S}^T$) and V-part ($\mathcal{S}^V$) simplices, respectively. Under the LDA model, the generative process of the d-th document in a corpus is composed of three steps:

(i) generate a vector of topics distribution $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$;

(ii) for each word $w_{d,n}$, generate a topic z_n from a $Z_n \sim \text{Categorical}(1, \boldsymbol{\theta}_d)$;

(iii) generate a word $w_{d,n}$ from the specific distribution over $\mathcal{V}$ for the topic z_n, that is $W_{d,n}|z_n = t \sim \text{Categorical}(1, \boldsymbol{\phi}_t)$, $\boldsymbol{\phi}_t \sim Dir(\boldsymbol{\beta})$.

A graphical representation of the model can be seen in Figure 4.2. The Direct Acycled Graph (DAG) shows how to generate a corpus with LDA model implementing the aforementioned steps. The LDA further assumes that both topics and words are randomly drawn from Dirichlet distributions. LDA exploits all the properties of interpretability, identifiability and conjugacy of the specified priors. Moreover, it allows each document to be a mixture of topics with different proportions. Heterogeneity in grouped data displaying multiple underlying patterns is taken into account with this specification. However, the LDA may encounter difficulties with short or noisy documents and may not effectively capture nuanced topic shifts within documents. The constraint of allowing only negative correlations between topics might impact the outcomes of this model. In practice, the quality of the results can be sensitive to the choice of hyperparameters, and these may require careful tuning.

4.2.2 Correlated Topic Model

The CTM is a versatile probabilistic graphical model used for the analysis of document collections, particularly in the field of natural language processing. Introduced by Blei and Lafferty (2007), CTM extends the popular LDA model by incorporating possible positive correlations among topics. CTM has found applications in various domains, including text analysis, content recommendation systems, and information retrieval (e.g., Aznag et al., 2013). The relationship between the finite support (i.e., the simplex) and the natural parameterization is defined as follows:

$$\eta_t = \log\left(\frac{\theta_t}{\theta_T}\right), \quad t = 1, \dots, T.$$

However, various values of $\boldsymbol{\eta} = (\eta_1, \dots, \eta_T)$ can produce the same mean parameter and this representation does not provide the most concise exponential family for the categorical distribution. The generative process of the d-th document in a corpus is quite similar to the LDA's one, with the main difference of the distribution for $\boldsymbol{\eta}$. The logistic-Normal distribution makes the assumption that $\boldsymbol{\eta}$ follows a multivariate normal distribution and is subsequently transformed into the simplex through the inverse of this mapping, i.e. $f(\eta_t) = \exp(\eta_t)/\big(1 + \sum_j^{T-1} \exp(\eta_j)\big)$. The main steps of the generative process are

(i) generate a topics distribution $\boldsymbol{\eta}_d | \mu, \Sigma \sim N(\mu, \Sigma)$;

(ii) for each word $w_{d,n}$, generate a topic z_n from a $Z_n \sim \text{Categorical}(1, f(\boldsymbol{\eta}_d))$;

(iii) generate a word $w_{d,n}$ from the specific distribution over $\mathcal{V}$ for the topic z_n, that is $W_{d,n} | z_n = t \sim \text{Categorical}(1, \boldsymbol{\phi}_t)$, $\boldsymbol{\phi}_t \sim Dir(\boldsymbol{\beta})$.

Figure 4.3 shows the graphical representation of the CTM model. One of the key properties of CTM is its ability to capture topic correlations, which allows it to model complex dependencies among topics in documents. Unlike the LDA, which assumes that topics are negatively correlated, CTM allows the topics to be even positively correlated through a shared Gaussian distribution. This modeling choice makes the CTM particularly suitable for tasks where topics often co-occur or exhibit patterns of positive association, such as in analyzing large-scale text corpora. Nevertheless, CTM is not without drawbacks. One notable limitation is its increased computational complexity compared to the LDA, as it involves estimating additional parameters for modeling topic correlations. This complexity can make CTM less scalable for very large datasets. Additionally, interpreting the learned correlations among topics can be challenging, as the model's parameters may not have

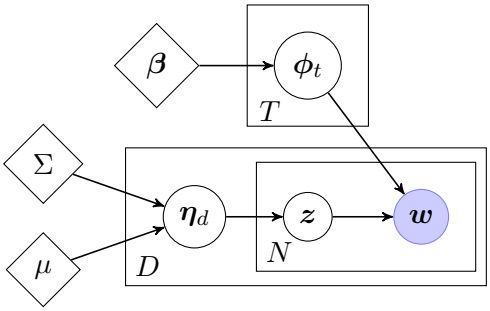


Figure 4.3: *DAG describing CTM method. The unobserved variables are drawn as circles whereas the observed ones are filled by blue color. Rhombuses represent hyperparameters. The outer rectangle represents documents, while the inner one represents the repeated topics and words within a document.*

straightforward semantic interpretations. This issue is due to the log-ratio transformation which allows the model to be specified on the real number space.

Notwithstanding the previously mentioned issues concerning both LDA and CTM, a solution is provided by a flexible and intrepretable distribution defined on the simplex. In the next section, we propose to assume that topics are distributed according to a sound distribution that maintains the good properties of the Dirichlet: the extended flexible Dirichlet distribution (Ongaro et al., 2020). The resulting generalization of the LDA is named extended flexible LDA (EFLDA).

4.3 Extended flexible latent Dirichlet allocation

Before introducing the EFLDA, it is important to see in detail the advantage of having a new distribution for the topic priors. This novelty improves the LDA model while still preserving its good properties.

4.3.1 Extended flexible Dirichlet

The EFD is an identifiable finite mixture with Dirichlet components, that is

$$\text{EFD}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{t=1}^{T} p_t \text{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha} + \tau_t \mathbf{e}_t), \tag{4.1}$$

where $\text{Dir}(\cdot;\boldsymbol{\alpha})$ denotes the probability density function of the Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$ with $\alpha_t > 0$, $\boldsymbol{\theta}$ and $\mathbf{p}$ lie in $\mathcal{S}^T$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)$, $\tau_t > 0$, $t = 1, \ldots, T$, and $\mathbf{e}_t$ is a vector of zeros except for the t-th element which is equal to one. Its probability density function (p.d.f.) can be written as:

$$f(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \Big(\prod_{t=1}^{T} \frac{\theta_t^{\alpha_t - 1}}{\Gamma(\alpha_t)} \Big) \sum_{i=1}^{T} p_i \frac{\Gamma(\alpha^+ + \tau_i)\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau_i)} \theta_i^{\tau_i}, \tag{4.2}$$

where $\alpha^+ = \sum_{t=1}^{T} \alpha_t$.

The Dirichlet is obtained as a special case of the EFD distribution by imposing the following constraints on the parameter space:

$$\begin{cases} \tau_t = 1 \\ p_t = \frac{\alpha_t}{\alpha^+} \end{cases} \qquad t = 1, \ldots, T. \tag{4.3}$$

Furthermore, by imposing $\alpha_t = 1$, $t = 1, \ldots, T$, to conditions in (4.3), it is possible to recover a uniform distribution on $\mathcal{S}^T$.

It is noteworthy that the EFD distribution retains many good properties of the Dirichlet one (e.g., identifiability, explicit expressions of joint moments and closure under many relevant operations on the simplex). Moreover, from Equation 4.3 it emerges that the EFD is characterized by a set of additional parameters, if compared with the Dirichlet. This richer parameterization guarantees flexibility in modeling dependences, though still preserving interpretability to a large extent (see Ongaro et al., 2020). In particular, Ongaro et al. (2020) showed that, despite the simplex space naturally induces negative associations, the EFD distribution may admit positive correlations among a subset of its elements. Positive correlations may be obtained by considering vectors $\boldsymbol{\tau}$ and $\mathbf{p}$ such that the discrete random variable Q, which takes values τ_t with probability p_t ($t = 1, \ldots, T$), is characterized by large variability.

As shown in Ongaro et al. (2020), it is straightforward to obtain the joint moments of $\boldsymbol{X} \sim EFD(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$ from the Dirichlet moments thanks to the mixture representation:

$$\mathbb{E}\Big[X_t\Big] = \alpha_t k_1 + \tau_t \frac{p_t}{\alpha^+ + \tau_t},$$

$$
\begin{aligned}
\text{Var}\Big(X_t\Big) &= \alpha_t^2\Big(k_2 - k_1^2\Big) + \frac{p_t\tau_i\Big(2\alpha_t + \tau_t + 1\Big)}{\Big(\alpha^+ + \tau_t\Big)\Big(\alpha^+ + \tau_t + 1\Big)} \\
&\quad + \alpha_t k_2 - \frac{p_t^2\tau_t^2}{\Big(\alpha^+ + \tau_t\Big)^2} - k_1\frac{2\alpha_t p_t\tau_t}{\alpha^+ + \tau_t}, \\
\text{Cov}\Big(X_t, X_l\Big) &= \alpha_t\alpha_l\Big(k_2 - k_1^2\Big) - \frac{p_t p_l\tau_t\tau_l}{\Big(\alpha^+ + \tau_t\Big)\Big(\alpha^+ + \tau_l\Big)} \\
&\quad + \frac{\alpha_t p_l\tau_l}{\alpha^+ + \tau_l}\Big(\frac{1}{\alpha^+ + \tau_l + 1} - k_1\Big) \\
&\quad + \frac{\alpha_l p_t\tau_i}{\alpha^+ + \tau_t}\Big(\frac{1}{\alpha^+ + \tau_t + 1} - k_1\Big),
\end{aligned}
$$

$(t, l = 1, \ldots, T, t \neq l)$ where:

$$
k_1 = \sum_{r=1}^{T}\frac{p_r}{\alpha^+ + \tau_r} \text{ and } k_2 = \sum_{r=1}^{T}\frac{p_r}{\Big(\alpha^+ + \tau_r\Big)\Big(\alpha^+ + \tau_r + 1\Big)}.
$$

The positive correlation implied by the EFD can be attributed to the arrangement of cluster means. In particular, it is allowed by the possibility that two clusters might align along a straight line with a positive slope. Focusing on the cluster structure under the EFD framework, the mean of the generic r-th Dirichlet mixture component (i.e., cluster) is a linear convex combination of a common barycenter $\bar{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}}{\alpha^+}$ and the r-th simplex vertex $\mathbf{e}_r$:

$$
\boldsymbol{\mu}_r^{EFD} = \frac{\boldsymbol{\alpha} + \tau_r\mathbf{e}_r}{\alpha^+ + \tau_r} = w_r\bar{\boldsymbol{\alpha}} + (1 - w_r)\,\mathbf{e}_r,
$$

where $w_r = \Big(\frac{\alpha^+}{\alpha^+ + \tau_r}\Big)$. This formulation results in the mean vector of the generic r-th component having its r-th element higher than the corresponding element of the other component means. This introduces a straightforward, yet fairly intricate, form of differentiation among components/clusters. Additionally, each mean $\boldsymbol{\mu}_r^{\text{EFD}}$ lies on the segment connecting the barycenter and the r-th simplex vertex. Each τ_r exclusively influences the corresponding r-th cluster mean. By increasing τ_r, this mean vector varies componentwise from $\bar{\boldsymbol{\alpha}}$ to the r-th vertex $\mathbf{e}_r$ in a continuous and monotonic manner. An example of the cluster structure with symmetric and asymmetric barycenter is shown in Figure 4.4.

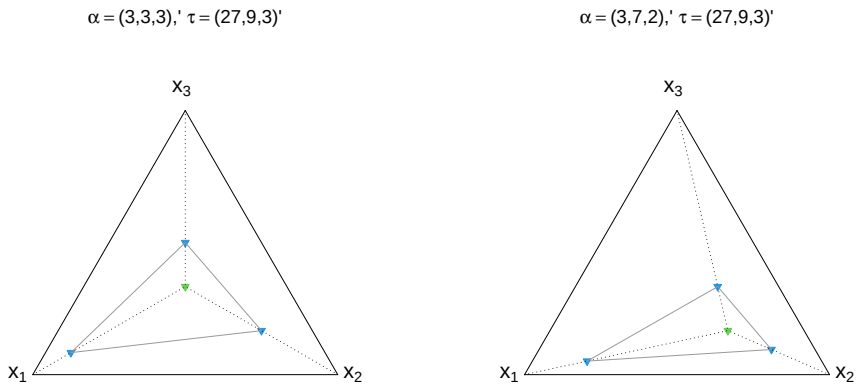


Figure 4.4: *Graphical representation of the EFD's mixture structure with ternary diagrams. Green triangles represent the common barycentre $\bar{\boldsymbol{\alpha}}$, while blue triangles represent component-specific mean vectors. Left panel: equal α_t values. Right panel: different α_t values.*

4.3.2 Extended flexible latent Dirichlet allocation

Let us consider a corpus composed of D documents. Under the EFLDA model, the generative process of the d-th document in a corpus $\mathcal{C}$, is composed of three steps:

(i) generate a topics distribution $\boldsymbol{\theta}_d \sim EFD(\boldsymbol{\alpha}, \mathbf{p}, \boldsymbol{\tau}) \in \mathcal{S}^T$;

(ii) for each word $w_{d,n}$, generate a topic z_n from a $Z_n \sim$ Categorical$(1, \boldsymbol{\theta}_d)$;

(iii) generate a word $w_{d,n}$ from the specific distribution over $\mathcal{V}$ for the topic z_n, that is $W_{d,n}|z_n = t \sim$ Categorical$(1, \boldsymbol{\phi}_t)$, $\boldsymbol{\phi}_t \sim Dir(\boldsymbol{\beta})$.

Figure 4.6 represents the graphical model to generate data from EFLDA. The hyperparameters $\boldsymbol{\alpha}.\boldsymbol{\tau}, \mathbf{p}$, and $\boldsymbol{\beta}$ have to be chosen. An example of data generated from this model, with $T = 3$ topics, can be seen in Figure 4.5. The two panels show the ternary plots of the distribution of the topics on the simplex, which are equilateral triangles allowing for a convenient and common representation of 3-part simplex elements (see Aitchison, 2003). Noteworthily, it is possible to see how the distribution is changing with different values of $\boldsymbol{\tau}$. In particular, different values of the barycenter, e.g. the magnitude of $\boldsymbol{\tau}$ with respect to $\boldsymbol{\alpha}$, lead to more separated cluster.

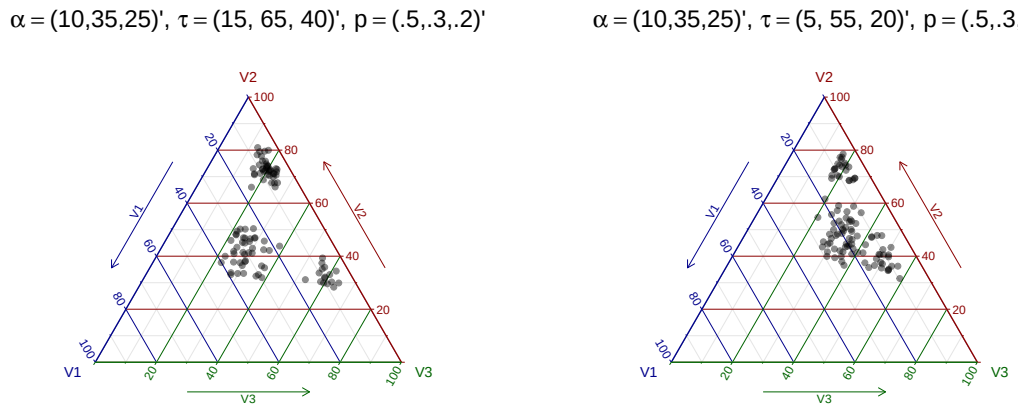


Figure 4.5: *Ternary diagrams showing data generated from EFLDA with different choice of the hyperparameters. V_t represents topic t for $t = 1, 2, 3$.*

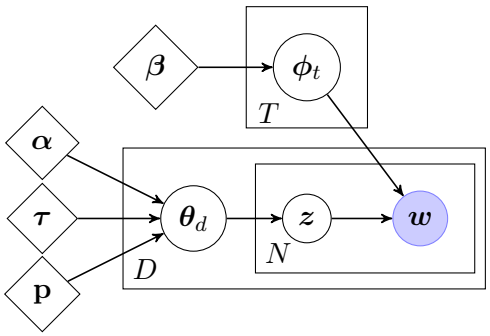


Figure 4.6: *Direct Acycled Graph (DAG) describing EFLDA method. The unobserved variables are drawn as circles, whereas the observed ones are filled by blue color. Rhombuses represent hyperparameters. The outer rectangle represents documents, while the inner one represents the repeated topics and words within a document.*

4.3.3 A special case: the flexible Dirichlet

As a special case of the EFD distribution, we can obtain the flexible Dirichlet (FD), introduced by Ongaro and Migliorati (2013). In the following, we will present the key formulas, as the FD serves as the topic distribution for all the results discussed in Section 4.5

The FD is a distribution defined on the T-part simplex and it is characterized by the following p.d.f.:

$$\mathrm{FD}(\mathbf{x}; \boldsymbol{\alpha}, \tau, \mathbf{p}) = \frac{\Gamma\Big(\alpha^+ + \tau\Big)}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \Big(\prod_{t=1}^{T} x_t^{\alpha_t - 1}\Big) \sum_{i=1}^{T} p_i \frac{\Gamma\Big(\alpha_i\Big)}{\Gamma\Big(\alpha_i + \tau\Big)} x_i^{\tau}, \tag{4.4}$$

where $\mathbf{p} \in \mathcal{S}^T$ is the vector of mixing weights, $\boldsymbol{\alpha} \in \mathbb{R}_+^T$, $\alpha^+ = \sum_{t=1}^T \alpha_t$, and $\tau > 0$. The FD allows also for a representation as a (structured) finite mixture of Dirichlet distributions, that is

$$\mathrm{FD}(\mathbf{x}; \boldsymbol{\alpha}, \tau, \mathbf{p}) = \sum_{t=1}^{T} p_t \mathrm{Dir}(\mathbf{x}; \boldsymbol{\alpha} + \tau \mathbf{e}_t). \tag{4.5}$$

It is also noteworthy to mention that the FD includes the Dirichlet distribution as a special case: Equation 4.4 coincides with the Dirichlet distribution if $\tau = 1$ and $p_t = \alpha_t/\alpha^+$ for $t = 1, \ldots, T$.

The FD distribution is characterized by D additional parameters if compared with the Dirichlet, while, having fewer parameters than the EFD, e.g., unique elements in $\mathbf{p}$ and τ. As a consequence, the FD still allows a more flexible modelization of the covariance matrix than the Dirichlet. Indeed, if $\mathbf{X} \sim \mathrm{FD}(\boldsymbol{\alpha}, \tau, \mathbf{p})$, then the following expressions hold for its first two-order moments:

$$\begin{aligned}
\mathbb{E}[X_t] &= \mu_t = \frac{\alpha_t + \tau p_t}{\alpha^+ + \tau}, \\
\mathrm{Var}(X_t) &= \frac{\mu_t(1-\mu_t)}{\alpha^+ + \tau + 1} + \frac{\tau^2 p_t(1-p_t)}{(\alpha^+ + \tau)(\alpha^+ + \tau + 1)}, \\
\mathrm{Cov}(X_t, X_l) &= -\frac{\mu_t \mu_l}{\alpha^+ + \tau + 1} - \frac{\tau^2 p_t p_l}{(\alpha^+ + \tau)(\alpha^+ + \tau + 1)}
\end{aligned}$$

$(t, l = 1, \ldots, T; t \neq l)$.

To better grasp the mixture structure of the FD, we can inspect the form of its component-specific barycentres $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T$:

$$\boldsymbol{\lambda}_t = (1-\delta)\boldsymbol{\alpha}/\alpha^+ + \delta \mathbf{e}_t, \quad \text{where } \delta = \frac{\tau}{\alpha^+ + \tau}. \tag{4.6}$$

Equation 4.6 shows that $\boldsymbol{\lambda}_t$ can be expressed as a weighted average of the vector $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+$ and the simplex's vertex $\mathbf{e}_t$ with a weight depending on τ. Thus, the vector $\bar{\boldsymbol{\alpha}}$ can be thought of as a "common" barycentre. Figure 4.7 illustrates the component-specific barycentres' behavior by means of ternary diagrams. Specifically, by joining the $\boldsymbol{\lambda}_t$'s (red triangles) with a segment, an equilateral triangle with edges parallel to the simplex's edges is obtained.

Looking at the formulas of the moments we can notice how the EFD allows for relevant generalizations of the FD dependence structure. The cluster structure under the FD framework has that only one parameter, τ, determines the distance of the D cluster means

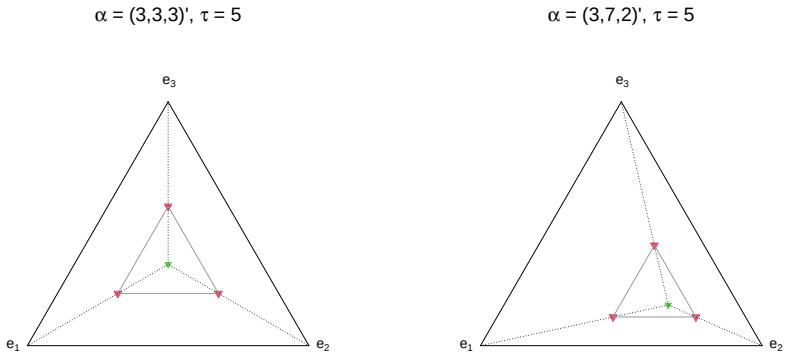


Figure 4.7: *Ternary diagrams showing the FD's mixture structure. Green triangles represent the common barycentre $\bar{\boldsymbol{\alpha}}$, while red triangles represent component-specific mean vectors $\boldsymbol{\lambda}_t$. Left panel: equal α_t values. Right panel: different α_t values.*

from each other and from the common barycenter $\bar{\alpha}$ in a symmetric manner. Specifically, as $\tau \to 0$, all clusters tend to have the same Dirichlet distribution, whereas when $\tau \to \infty$, they converge onto the corresponding vertices of the simplex. The D mean vectors of the FD form the vertices of a D-dimensional shifted and scaled simplex that is strictly contained within the original D-dimensional simplex, with edges parallel and proportional to those of the latter (Migliorati et al., 2017). In contrast, under the EFD there are no constraints on the edges. Specifically, EFD allows even strong positive correlations. This type of correlation can be found in real compositional data, even if the unit sum constraint naturally induces negative correlations as well.

4.3.4 Flexible latent Dirichlet allocation

For the sake of completeness, the flexible LDA (FLDA) model has the following generative process:

(i) generate a topics distribution $\boldsymbol{\theta}_d \sim FD(\boldsymbol{\alpha}, \mathbf{p}, \tau) \in \mathcal{S}^T$;

(ii) for each word $w_{d,n}$, generate a topic z_n from a $Z_n \sim \text{Categorical}(1, \boldsymbol{\theta}_d)$;

(iii) generate a word $w_{d,n}$ from the specific distribution over $\mathcal{V}$ for the topic z_n, that is $W_{d,n}|z_n = t \sim \text{Categorical}(1, \boldsymbol{\phi}_t)$, $\boldsymbol{\phi}_t \sim Dir(\boldsymbol{\beta})$.

As for the EFD, it is possible to prove that the FD is conjugate to the multino-

mial/categorical scheme, thus we can define the conditional distribution as

$$\boldsymbol{\theta}_d|\mathbf{z},\mathcal{C},\boldsymbol{\alpha},\tau,\mathbf{p} \;\sim\; FD\Big(\boldsymbol{\alpha}+\mathbf{c}_d,\tau,\mathbf{p}_d^*/p_d^{*+}\Big), \quad \mathbf{c}_d=(c_{1,d,\cdot},\ldots,c_{T,d,\cdot})^\intercal, \tag{4.7}$$

where $c_{t,d,\cdot}$ represents the number of words generated by topic t in document d, $p_d^{*+}=\sum_{t=1}^{T}p_{d,t}^*$, and the generic element of $\mathbf{p}_d^*=\left(p_{d,1}^*,\ldots,p_{d,T}^*\right)^\intercal$ is given by

$$p_{d,t}^*=p_t\frac{(\alpha_t+\tau)^{[c_{t,d,\cdot}]}}{(\alpha_t)^{[c_{t,d,\cdot}]}}, \quad t=1,\ldots,T, \tag{4.8}$$

with $x^{[n]}=x(x+1)\cdot\cdots\cdot(x+n-1)$ denoting the rising factorial function.

Figure 4.6 with equal τ also represents the DAG for the FLDA. As for EFLDA, the hyperparameters $\boldsymbol{\alpha},\tau,\mathbf{p}$ and $\boldsymbol{\beta}$ have to be chosen. Noticeable, in Figure 4.8, it is possible to see that increasing τ creates better separated clusters.

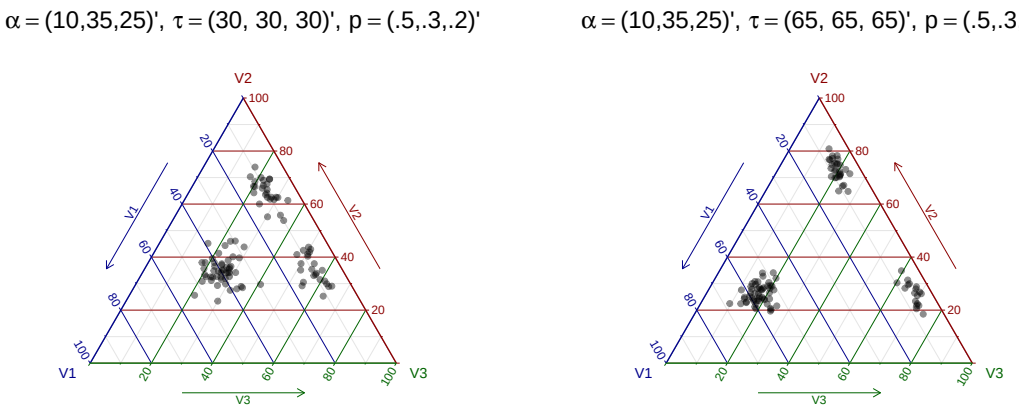


Figure 4.8: *Ternary diagrams showing data generated from FLDA with different choices of the hyperparameters. V_t represents topic t for $t=1,2,3$.*

4.4 Posterior Inference

The estimation procedure for the parameters involved in the LDA and EFLDA models, specifically the elements in $\Theta=\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_D\}$ and $\Phi=\{\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_T\}$, cannot be fulfilled by using standard techniques such as the maximum likelihood. This is because of the presence of the topic labels, which are latent variables complicating the complete-data likelihood function. Blei et al. (2003) showed that the posterior distribution of the hidden variables

given a document is intractable to compute. So, they provided an alternative variational-inference approach for estimating the quantities of interest. However, this approach requires the derivation of some challenging equations that may be difficult to compute, especially for the EFLDA model, and above all it gives only approximate solutions.

Thus, we prefer an approach based on MCMC techniques, such as the collapsed Gibbs sampling (CGS). Here, the basic idea is to generate one parameter at a time, by drawing it from its full conditional distribution. The main difference with respect to a standard Gibbs sampler is that full conditionals are computed by marginalizing some parameters out; the estimates of the dropped parameters are computed by means of closed-form expressions (e.g., by taking advantage of some conjugacy properties). Griffiths and Steyvers (2004) proposed a CGS approach to estimate the parameters of the LDA model by computing the full conditionals of the topic labels $Z_{d,n}$, and by deriving a closed-form expression for estimates of the elements in Θ and Φ. To sketch the CGS, it is useful to define some quantities:

- $c_{t,d,w} = \sum_{n=1}^{N_d} \mathbb{1}(z_{d,n} = t,\ w_{d,n} = w)$: number of times that word w is assigned to topic t in document d;
- $c_{t,d,\cdot} = \sum_{v=1}^{V} c_{t,d,v}$: number of words assigned to topic t in document d;
- $c_{t,\cdot,w} = \sum_{d=1}^{D} c_{t,d,w}$: number of times word w is assigned to topic t across documents;
- $c_{t,\cdot,\cdot} = \sum_{v=1}^{V} c_{t,\cdot,v}$ is the total number of words assigned to topic t.

In the LDA model, the full conditional distribution for $Z_{d,n}$, namely the probability that $w_{d,n}$ is assigned to topic t ($t = 1, \ldots, T$) given all the other topic assignments $\mathbf{z}_{-(d,n)}$ (where $\mathbf{z}_{-(d,n)}$ represents the vector of topic assignments $\mathbf{z}$ with the exclusion of the topic assignment $z_{d,n}$), takes the following form:

$$p(t) = p\Big(Z_{d,n} = t | \mathbf{z}_{-(d,n)}, \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\beta}\Big) \propto \frac{\Big(\alpha_t + c^{-}_{t,d,\cdot}\Big)\Big(\beta_{w_{d,n}} + c^{-}_{t,\cdot,w_{d,n}}\Big)}{\Big(\beta^{+} + c^{-}_{t,\cdot,\cdot}\Big)} \tag{4.9}$$

($t = 1, \ldots, T$), where the notation c^{-} refers to the previously defined counts excluding the n-th word of document d, and $w_{d,n}$ indicates which term of the vocabulary is associated

with the n-th word in document d (i.e., $w_{d,n} = v$ means that the n-th word of document d is the v-th element of the vocabulary). We can sample from these full conditionals by simply generating values from a discrete random variable with support $\{1, 2, \dots, T\}$ and probability mass function (p.m.f.) whose kernel is given by (4.9).

Thanks to the conjugacy property of the Dirichlet to the multinomial/categorical sampling, we can derive the following posterior expressions:

$$\boldsymbol{\theta}_d|\mathbf{z}, \mathcal{C}, \boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha} + \mathbf{c}_d), \quad \mathbf{c}_d = (c_{1,d,\cdot}, \dots, c_{T,d,\cdot})^\intercal; \tag{4.10}$$

$$\boldsymbol{\phi}_t|\mathbf{z}, \mathcal{C}, \boldsymbol{\beta} \sim Dir(\boldsymbol{\beta} + \mathbf{c}_t), \quad \mathbf{c}_t = (c_{t,\cdot,1}, \dots, c_{t,\cdot,V})^\intercal. \tag{4.11}$$

These posteriors allow estimating the elements in Θ and Φ. Indeed, by having a sample of size B for the topic labels, namely $\mathbf{z}^{(b)}$, $b = 1, \dots, B$, we can estimate $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$ as the posterior mean of (4.10) and (4.11), respectively:

$$\hat{\boldsymbol{\theta}}_d^{(b)} = \frac{\boldsymbol{\alpha} + \mathbf{c}_d^{(b)}}{\alpha^+ + N_d}, \qquad \hat{\boldsymbol{\phi}}_t^{(b)} = \frac{\boldsymbol{\beta} + \mathbf{c}_t^{(b)}}{\beta^+ + c_{t,\cdot,\cdot}^{(b)}}.$$

4.4.1 Full conditionals - FLDA

The CGS scheme for the flexible LDA (FLDA) is similar to the previously described LDA scheme. Indeed, the main difference is related to the expressions of the full conditionals, which vary given the FD distribution imposed on $\boldsymbol{\theta}_d$. It is possible to prove that

$$\begin{aligned} p(t) = p(Z_{d,n} = t|\mathbf{z}_{-(d,n)}, \mathcal{C}, \boldsymbol{\alpha}, \tau, \mathbf{p}, \boldsymbol{\beta}) &\propto \\ &\propto \frac{\left(\alpha_t + c_{t,d,\cdot}^-\right)\left(\beta_{v_{d,n}} + c_{t,\cdot,w_{d,n}}^-\right)}{\left(\beta^+ + c_{t,\cdot,\cdot}^-\right)} \Big\{ \sum_{h=1}^{T} p_{d,h}^* + p_{d,t}^* \Big(\frac{\tau}{\alpha_t + c_{t,d,\cdot}^-} \Big) \Big\}, \end{aligned} \tag{4.12}$$

$t = 1, \dots, T$, where $p_{d,t}^*$ is defined as in Equation 4.8. According to the posterior distribution of $\boldsymbol{\theta}_d$ reported in (4.7), an estimate is obtained by computing the posterior mean of the FD distribution, namely

$$\hat{\boldsymbol{\theta}}_d^{(b)} = \frac{\boldsymbol{\alpha} + \mathbf{c}_d^{(b)} + \tau \mathbf{p}_d^{*(b)}/p_+^{(b)}}{\alpha^+ + \tau + N_d}.$$

4.4.2 Full conditionals - EFLDA

For EFLDA, we have a similar CGS as for the FLDA one. The only difference relies on the specification of $p(t)$ that is as follows,

$$p(t) = p(Z_{d,n} = t|\mathbf{z}_{-(d,n)}, \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\beta}) \propto \frac{\left(\alpha_t + c^{-}_{t,d,\cdot}\right)\left(\beta_{w_{d,n}} + c^{-}_{t,\cdot,w_{d,n}}\right)}{\left(\beta^{+} + c^{-}_{t,\cdot,\cdot}\right)} \times \left\{ \sum_{h=1}^{T} \frac{p^{*}_{d,h}}{(\alpha^{+} + \tau_h + c^{-}_{\cdot,d,\cdot})(\alpha^{+} + \tau_h)^{[c^{-}_{\cdot,d,\cdot}]}} + \frac{p^{*}_{d,t}}{(\alpha^{+} + \tau_t + c^{-}_{\cdot,d,\cdot})(\alpha^{+} + \tau_t)^{[c^{-}_{\cdot,d,\cdot}]}} \left(\frac{\tau_t}{\alpha_t + c^{-}_{t,d,\cdot}} \right) \right\},$$

where $p^{*}_{d,h} = p_{d,h} \frac{(\alpha_h + \tau_h)^{[c_{h,d,\cdot}]}}{\alpha_h^{[c_{h,d,\cdot}]}}$. Noteworthily, unlike Equation 4.9, the second factor in Equation 4.13 depends on the cluster structure. Consequently, the posterior mean of $\boldsymbol{\theta}_d$ is

$$\hat{\theta}^{(b)}_{d,t} = \frac{\alpha_t + c^{(b)}_{d,t}/p^{(b)}_{+}}{\alpha^{+} + \boldsymbol{\tau} + N_d} + \frac{\tau_t p^{*}_{d,t}{}^{(b)}}{\alpha^{+} + \tau_t + N_d}.$$

Finally, the CGS for the LDA, the FLDA, and the EFLDA models can be summarized by Algorithm 2.

Algorithm 2 Collapsed gibbs sampling for topic modelling

```
Initialize the vector z (randomly) and compute the counts c_{t,d,v};
for b = 1, ..., B do:
    for each word in the corpus do:
        sample sample a new topic z^{(b)}_{d,n} for w_{d,n} from p(z);
        update the counts c_{t,d,v};
    end for
    Use z^{(b)} to compute the estimates θ̂^{(b)}_d and φ̂^{(b)}_t.
end for
end
```

The full conditionals $p(t)$ are given by (4.9) for the LDA, by (4.12) for the FLDA, and by (4.13) for the EFLDA respectively. More details on the full conditional distributions can be found in Appendix 4.A.

4.5 A focus on the FLDA

In this section, we present some simulation studies and a real data application concerning the FLDA model. The former is an extensive simulation study presented in Giampino et al. (2023) where LDA and FLDA are compared in a synthetic setting. The latter shows an application to real data on a corpus of books, where the true number of topic represents the number of books involved in the analysis (Ascari and Giampino, 2023).

4.5.1 Simulation study

In this section, we compare the LDA and the FLDA models through a simulation study involving three distinct corpus-generating schemes. All the considered scenarios share the number of topics and the dimension of the corpus. Indeed, we consider corpora containing $D = 250$ documents with an average length of $\varepsilon = 600$ words. Documents have been populated by considering an artificial vocabulary of $V = 300$ unique words which may be generated according to $T = 3$ latent topics.

The three simulating scenarios differ in their corpus-generating mechanism. In the first scenario, we generate each element $\boldsymbol{\theta}_d$ from a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (10, 25, 15)$. In the second scenario, each $\boldsymbol{\theta}_d$ is generated from an FD with parameters $\boldsymbol{\alpha} = (10, 25, 15), \mathbf{p} = (0.5, 0.3, 0.2)$, and $\tau = 30$. Lastly, we consider a logistic-Normal with parameters $\boldsymbol{\mu} = (-2, 3)$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$ as distribution for generating each $\boldsymbol{\theta}_d$ in the third scenario. Thus, the underlying true data-generating mechanisms are the LDA (first scenario), the FLDA (second scenario), and the CTM (third scenario) models.

Figure 4.9 shows the generated vectors $\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_D^*$ in the three scenarios by means of a ternary diagram. It is worth noting that three very different patterns are recovered, characterized by one main group of $\boldsymbol{\theta}_d^*$ values (first scenario), three distinct groups (second scenario), and most $\boldsymbol{\theta}_d^*$ values with at least one null component (third scenario). These $\boldsymbol{\theta}_d^*$ vectors are going to be considered as the *true* vectors of topic proportions generating the corpora.

Finally, to define the $T = 3$ topics, we generated vectors $\boldsymbol{\phi}_1^*, \boldsymbol{\phi}_2^*$, and $\boldsymbol{\phi}_3^*$ by sampling from a Dirichlet distribution with V-dimensional parameter $\boldsymbol{\beta} = (1, \dots, 1)$.

We replicate each scenario $R = 100$ times, by generating R different corpora and fitting both the LDA and the FLDA models. The aim of the simulation study is to compare the estimates of the two models with the real parameters, and establish the most reliable

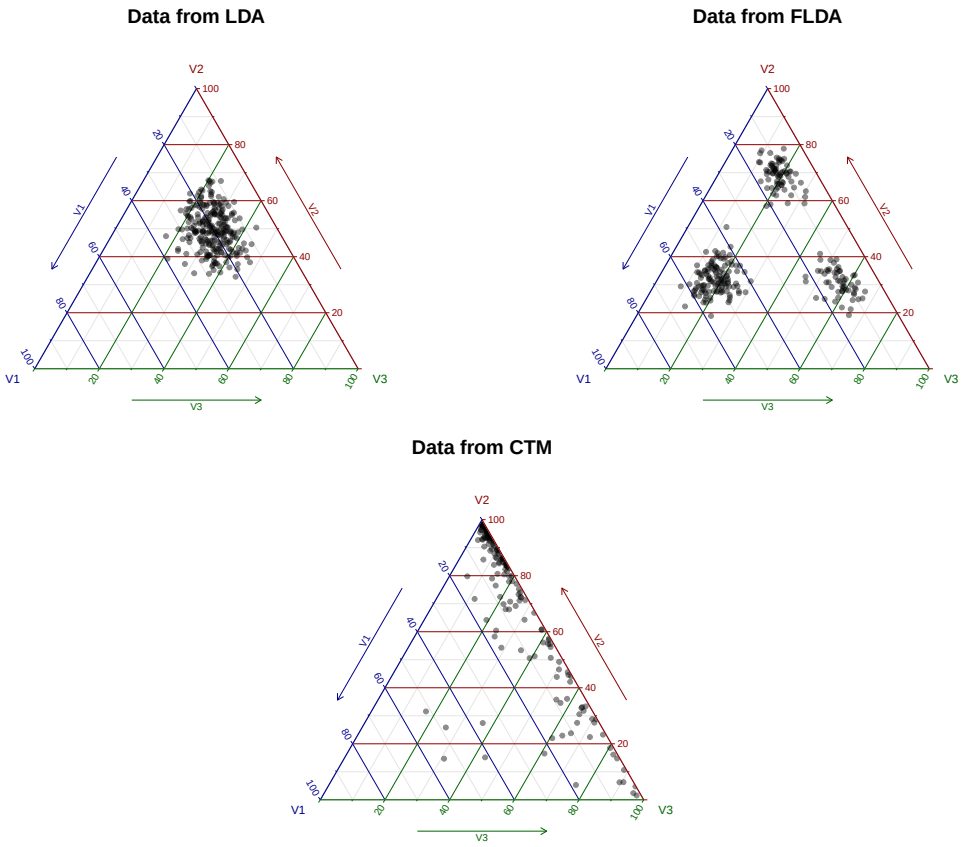


Figure 4.9: *Ternary diagrams representing the vectors of topic proportions generated from Dirichlet for LDA (top-left), from FD for FLDA (top-right), and from logistic-Normal for CTM (bottom) models.*

method.

We fit the models by considering the CGS illustrated in the previous section. More specifically, for each replica we run a single chain composed of $B = 2000$ iterations and discard 1000 iterations as a warm-up (see Figure 4.10 for traceplots).

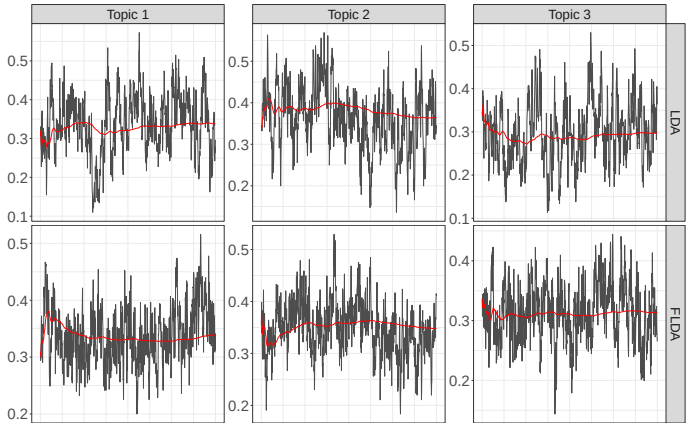
To estimate both the LDA and the FLDA parameters, one needs to specify the model hyper-parameters. In the LDA, the only quantity to specify is the T-dimensional vector $\boldsymbol{\alpha}$, which is typically chosen with equal elements $\alpha_1 = \cdots = \alpha_T = \alpha$ so as to adopt a symmetric Dirichlet distribution in a weakly-informative fashion. Many authors proposed some default values for the common α, such as $\alpha = 0.1$ (Teh et al., 2006) or $\alpha = 50/T$ (Qiu, 2014; Steyvers and Griffiths, 2007; Xia et al., 2013). In this simulation study, we prefer a value of $\alpha = 1$ so to select a uniform distribution on the simplex $\mathcal{S}^T$. For the FLDA, the implementation of some default choices is more tricky. Indeed, we have to specify the value of $\boldsymbol{\alpha}, \mathbf{p}$, and τ, and these parameters have a direct impact on the cluster structure of the FD distribution, which cannot (and should not) be treated by adopting a default choice. In a scenario with no prior information available, we decided to implement a preliminary step aimed at obtaining values for the $\boldsymbol{\alpha}, \mathbf{p}$, and τ. More precisely, by having the estimates from the chains $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_D$ from the LDA model, we use them as data points to estimate the parameters of a FD distribution. This inferential problem may be handled both via an EM-based procedure (Migliorati et al., 2017) or by a Bayesian approach implemented through the Stan language (Stan Development Team, 2017). Both these approaches provide initial estimates $\hat{\boldsymbol{\alpha}}, \hat{\mathbf{p}}$, and $\hat{\tau}$ that can be used to run the CGS described in Section 3.3. We implemented the second approach. Topic assignments have been initialized randomly.

To compare the estimates provided by the LDA and the FLDA models, we take advantage of three measures, namely the Kullback-Leibler (KL) divergence, and its symmetrized (SKL) version (Kullback and Leibler, 1951), and the Aitchison AIT distance (Aitchison, 1990). These metrics measure the discrepancy existing between two compositional vectors, thus larger values are associated with "farther" vectors. In particular, let $\mathbf{x}$ and $\mathbf{y}$ be two compositional vectors defined on the same simplex space $\mathcal{S}^T$. Then, the above metrics are defined as:

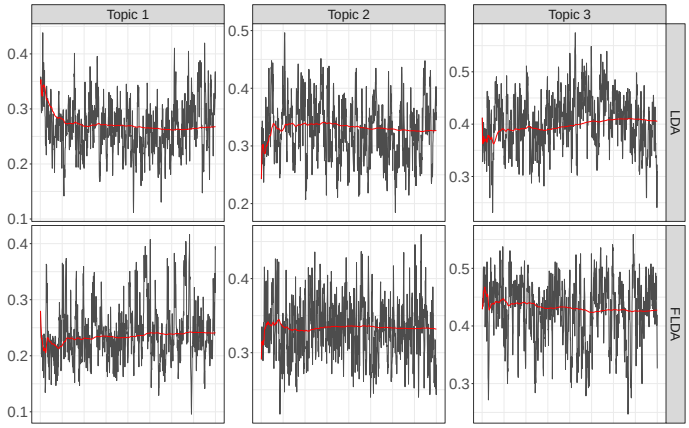
$$d_{KL}(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} x_t \log\left(\frac{x_t}{y_t}\right),$$

$$d_{SKL}(\mathbf{x}, \mathbf{y}) = d_{KL}(\mathbf{x}, \mathbf{y}) + d_{KL}(\mathbf{y}, \mathbf{x}),$$

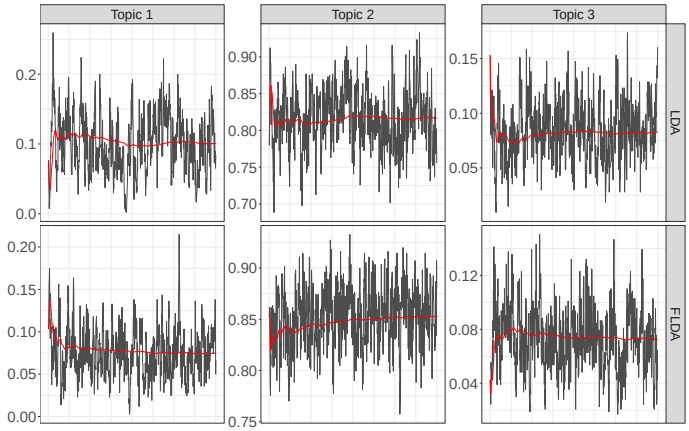
and



(a) *Corpus generating from the Dirichlet distribution.*



(b) *Corpus generating from the flexible Dirichlet distribution.*



(c) *Corpus generating from the logistic-Normal distribution.*

Figure 4.10: *Traceplots of the MCMC chains for each topic and for each model. In red the iterative means.*

$$d_{AIT}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{t=1}^{T} \left(\log\left(\frac{x_t}{\mu_0(\mathbf{x})}\right) - \log\left(\frac{y_t}{\mu_0(\mathbf{y})}\right)\right)^2},$$

where $\mu_0(\mathbf{x}) = \left(\prod_{t=1}^{T} x_t\right)^{1/T}$ denotes the geometric mean of the elements of $\mathbf{x}$.

In this framework, we can use these measures to compare an estimated vector of parameters (i.e., $\hat{\boldsymbol{\theta}}_d$ or $\hat{\boldsymbol{\phi}}_t$) with the true vector ($\boldsymbol{\theta}_d^*$ or $\boldsymbol{\phi}_t^*$).

In particular, Figure 4.11 shows the mean of the distances computed between $\boldsymbol{\theta}_d^*$ and its estimate $\hat{\boldsymbol{\theta}}_d^{(r)}$ in the r-th simulating replicate. For example, each point in the boxplots in the first column of Figure 4.11 represents the value

$$d_{AIT}^{(r)} = \frac{1}{D} \sum_{d=1}^{D} d_{AIT}\left(\hat{\boldsymbol{\theta}}_d^{(r)}, \boldsymbol{\theta}_d^*\right), \quad r = 1, \ldots, R.$$

It is easy to see that the FLDA model provides estimates for the topic proportions that are closer to the corresponding true values than the LDA's estimates in all the three considered scenarios and according to all the considered metrics. Therefore, the FLDA, contrarily to the LDA, shows robustness with respect to the data generating mechanism.

Interestingly, the estimates provided by the FLDA are better than the ones from the LDA model independently from the corpus-generating mechanism. This is probably due to the additional step of estimating the FD's parameters using the results provided by the LDA model and, consequently, improving it.

Similar considerations hold for the estimation of the topics, namely the parameters $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2$, and $\boldsymbol{\phi}_3$. Figure 4.12 shows the boxplots of the distances between the r-th estimate $\hat{\boldsymbol{\phi}}_t^{(r)}$ and the true $\boldsymbol{\phi}_t^*$, e.g. $d_{AIT,t}^{(r)} = d_{AIT}\left(\hat{\boldsymbol{\phi}}_t^{(r)}, \boldsymbol{\phi}_t^*\right)$. In all the plots we can note a general superiority of the FLDA model with respect to the classical LDA. The only partial exception is the estimation of the third topic (i.e., $\boldsymbol{\phi}_3$) when the corpus is generated according to a correlated topic model (CTM), that is a scenario where we expect both the LDA and the FLDA models to perform badly. In that case, the two approaches seem to provide similar results. Please note that in estimating the two remaining vectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ in the CTM scenario, the FLDA still provides better estimates.

Lastly, since the CGS estimation procedure for both the LDA and FLDA models provides Monte Carlo samples of $\mathbf{z}$, we can compare them with the set of true topics associated with each word in the corpus. Such a comparison is performed by considering the normalized Pearson's χ^2 statistic (Agresti, 2012). Figure 4.13 summarizes these normalized

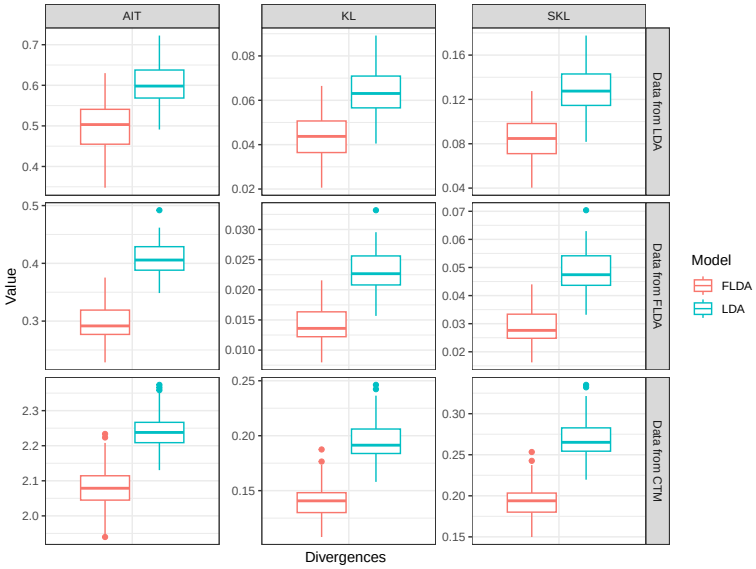


Figure 4.11: *Boxplot representing the mean distance between the estimated $\boldsymbol{\theta}_d$s and the true $\boldsymbol{\theta}_d^*$s. Each column refers to a metric, and each row corresponds to one of the simulated scenarios.*

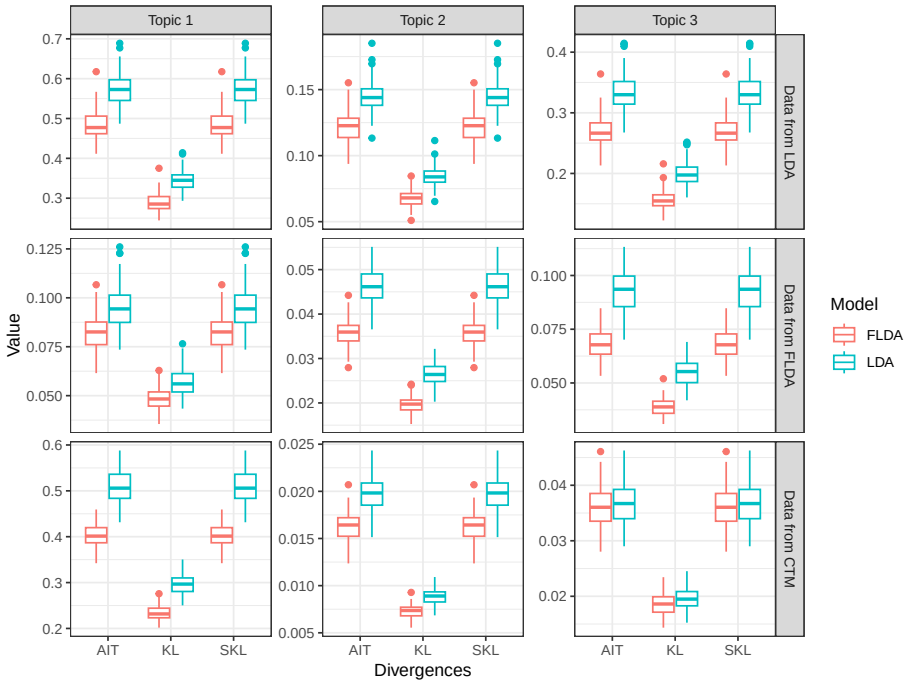


Figure 4.12: *Boxplot of distances between the estimated $\boldsymbol{\phi}_t$s and the true $\boldsymbol{\phi}_t^*$s.*

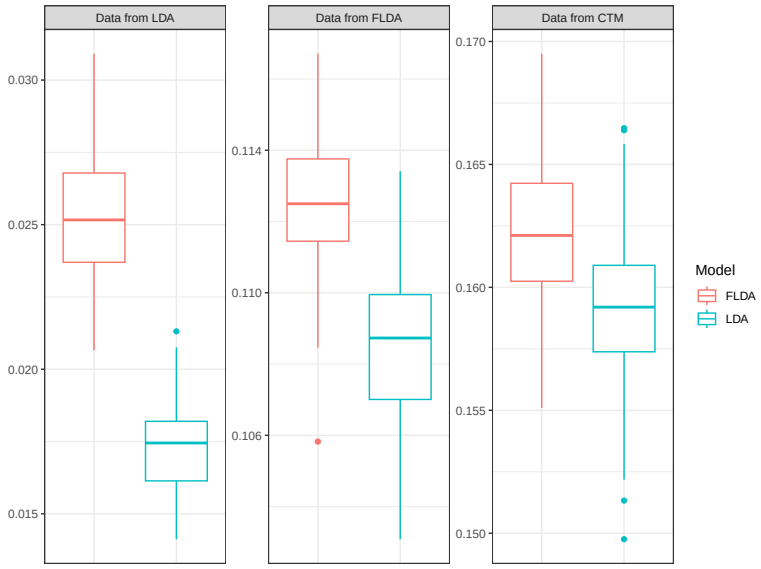


Figure 4.13: *Normalized Pearson's χ^2 statistic evaluating the association between real and predicted words' topic.*

statistics for the two models in the three simulative scenarios. The FLDA still shows better performance (i.e., predicted topics more associated with real topics) than the LDA.

4.5.2 Real data application: The Great Library Heist

This application is related to "The Great Library Heist", which is a popular topic model application. Supposed that during the night, a vandal broke into their professor's study and tore three books into single chapters. The single chapters are not labeled, so the professor is not able to cluster them so to restore the original books. In the following, we consider three books "Great Expectations", "20000 Leagues Under the Sea", and "Pride and Prejudice", and use their $D = 166$ chapters as documents forming the corpus. We will consider $T = 3$ latent topics, each of them hopefully representing one of the destroyed books. Words in the corpus $\mathcal{C}$ compose a vocabulary $\mathcal{V}$ of $V = 16531$ unique terms. We run both the LDA and the FLDA models for $B = 5000$ iterations. Figure 4.14 displays the estimated topic proportions $\boldsymbol{\theta}_d$ for all the documents, by conditioning on the true topic (i.e., the original book).

We can note from Table 4.1 that both the LDA and FLDA models represent chapters from "Great Expectations" as mainly composed of terms arising from topic 1. The FLDA, thanks to the flexible covariance matrix of the FD, improves the LDA performance by

providing $\boldsymbol{\theta}_d$'s more concentrated towards 0 or 1 than those provided by the LDA. Similar conclusions hold true for chapters from "20000 Leagues Under the Sea" and "Pride and Prejudice", which are characterized by high proportions of words from topics 2 and 3, respectively. Topics generated by the FLDA are represented by illustrating the 20 most probable words (Figure 4.15).

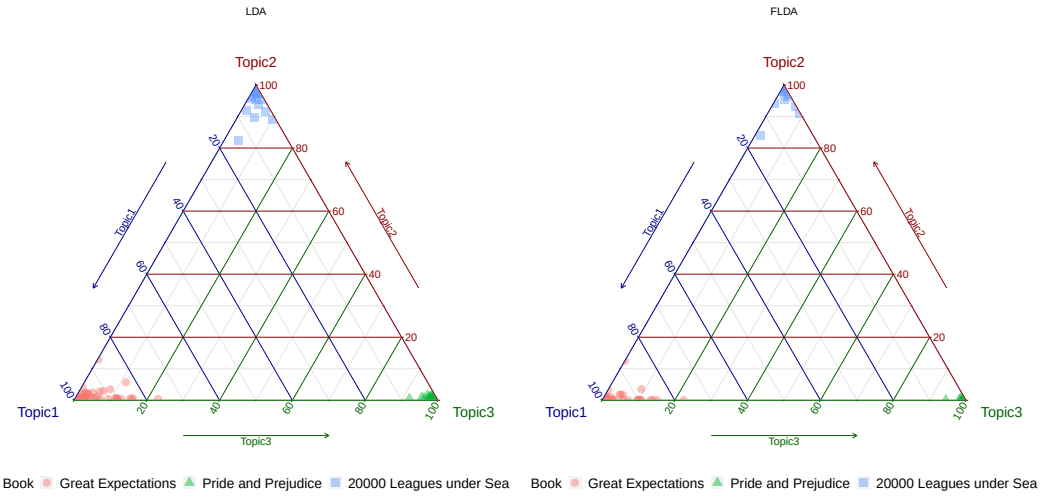


Figure 4.14: *Ternary plots of the elements of $\boldsymbol{\theta}_d$ estimates by the LDA (left panel) and the FLDA (right panel) conditioning on the true topic (i.e., the original book).*

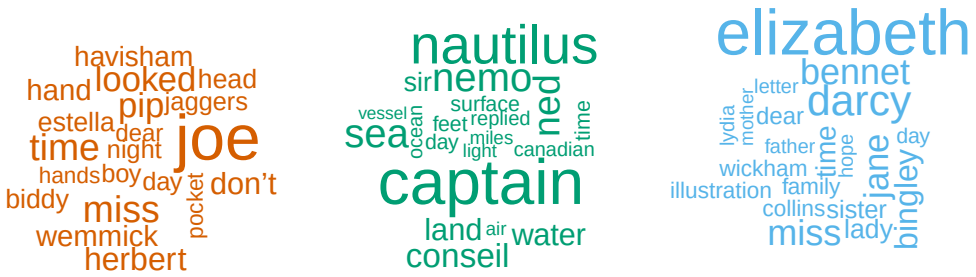


Figure 4.15: *Word clouds representing the 20 most probable words for each topic detected by the FLDA.*

	Great Expectation		Pride & Prejudice		20000 Leagues under the Sea	
	LDA	FLDA	LDA	FLDA	LDA	FLDA
θ_1	0.9663	0.9892	0.0076	0.0019	0.0669	0.002
θ_2	0.0053	0.0012	0.0045	0.001	0.9876	0.9967
θ_3	0.0203	0.0064	0.9878	0.9967	0.0054	0.0015

Table 4.1: *Element-wise median of the fitted $\hat{\boldsymbol{\theta}}_d$ stratified by original book and model.*

4.6 Future developments

Topic modeling techniques have gained widespread popularity due to their remarkable capacity to offer clear interpretations of latent topics within large textual datasets. These methods have enabled researchers and analysts to uncover meaningful patterns and insights from vast troves of unstructured textual data. In recent years, the field of topic modeling has seen significant advancements, particularly with the emergence of more flexible mixture structures like the extended flexible Dirichlet distribution and its subsequent evolution, the extended flexible latent Dirichlet allocation model. These innovations represent a leap forward in the realm of topic modeling, providing a more nuanced and powerful approach to uncovering hidden topics within documents, especially if characterized by strong positive correlation.

Moreover, the EFLDA model offers a richer parametrization, which plays a crucial role in the optimization and efficiency of its implementation. The MCMC runs in 0.18 seconds per iteration in a scenario with a vocabulary of 300 unique words, 200 documents, and over 120,000 words. This parametric richness empowers researchers to fine-tune the model's performance to better suit their specific objectives and datasets. It allows for greater control and adaptability in the topic modeling process, ultimately leading to more accurate and insightful results. This model exploits the EFD distribution that, as the Dirichlet distribution, is defined on the simplex. It also maintains the good properties of the Dirichlet distribution, e.g. conjugacy with the multinomial/categorical distribution, identifiability and recovering of moments and full conditionals in closed form.

Currently, our research efforts are focused on the exploration of various synthetic data scenarios, each generated using distinct schemes. Within this scope, we are meticulously examining the model's responsiveness to different numbers of topics, and concurrently, we are conducting a comprehensive performance analysis, juxtaposing the EFLDA against traditional LDA models.

In terms of goals, the overarching objective in the realm of topic modeling remains to develop models that can provide deeper and more nuanced insights into the underlying structure of text data. This includes not only improving the accuracy of topic identification but also enhancing the interpretability of these topics. In conclusion, topic modeling techniques like EFLDA have revolutionized our ability to extract meaningful information from unstructured text. Their flexibility and parametric richness make them powerful tools for uncovering latent topics within documents.

Bibliography

Agresti, A. (2012). *Categorical data analysis*, volume 792. John Wiley & Sons.

Aitchison, J. (1990). On coherence in parametric density estimation. *Biometrika*, 77(4):905–908.

Aitchison, J. (2003). *The Statistical Analysis of Compositional data*. The Blackburn Press, London.

Aldous, D. J., Ibragimov, I. A., Jacod, J., and Aldous, D. J. (1985). *Exchangeability and related topics*. Springer.

Ascari, R. and Giampino, A. (2023). A flexible topic model. In *CLADAG 2023 - Book of abstracts and short papers - 14th Scientific Meeting of the Classification and Data Analysis Group, Salerno, September 11-13 (pp. 334-337)*. Milano: Pearson.

Aznag, M., Quafafou, M., and Jarir, Z. (2013). Correlated topic model for web services ranking. *International Journal of Advanced Computer Science and Applications*, 4(6).

Benton, A. and Dredze, M. (2018). Deep Dirichlet multinomial regression. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 365–374.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, 1(1):17–35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Carpenter, B. (2010). Integrating out Multinomial parameters in Latent Dirichlet Allocation and Naive Bayes for collapsed Gibbs sampling. *Rapport Technique*, 4:464.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Giampino, A., Ascari, R., and Migliorati, S. (2023). A flexible generalization of the Latent Dirichlet Allocation. In *Data Analysis Book Series*. ISTE-Wiley. To be published.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Hoffman, M., Bach, F., and Blei, D. (2010). Online learning for latent Dirichlet allocation. *advances in neural information processing systems*, 23.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Migliorati, S., Ongaro, A., and Monti, G. S. (2017). A structured Dirichlet mixture model for compositional data inferential and applicative issues. *Statistics and Computing*, 27:963–983.

Ongaro, A. and Migliorati, S. (2013). A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis*, 114(1):412–426.

Ongaro, A., Migliorati, S., and Ascari, R. (2020). A new mixture model on the simplex. *Statistics and Computing*, 30:749–770.

Qiu, L. (2014). Gibbs collapsed sampling for latent Dirichlet allocation on spark. In *Proceedings of the 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, pages 17–28. PMLR.

Stan Development Team (2017). Stan Modeling Language Users Guide and Reference Manual.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Teh, Y., Newman, D., and Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Advances in neural information processing systems*, 19.

Wood, I. (2014). Recent advances and applications of probabilistic topic models. In *AIP Conference Proceedings*, volume 1636, pages 124–130. American Institute of Physics.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063.

Appendix

4.A Full conditionals

The derivation of the full conditional distributions is based on Carpenter (2010). We use the same notation provided along Chapter 4.

A.1 LDA

$$
\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \theta, \phi \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= p(\phi \mid \boldsymbol{\beta}) p(\theta \mid \boldsymbol{\alpha}) p(\mathbf{z} \mid \theta) p(\mathbf{w} \mid \phi, \mathbf{z}) \\
&= \prod_{t=1}^{T} p\Big(\boldsymbol{\phi}_t \mid \boldsymbol{\beta}\Big) \prod_{d=1}^{D} p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(z_{d,n} \mid \boldsymbol{\theta}_d\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(w_{d,n} \mid \boldsymbol{\phi}_{z_{d,n}}\Big)
\end{aligned}
$$

We want to compute the probability that topic $z_{a,b}$ is assigned to $w_{a,b}$ (i.e., the b-th word of document a) given $\mathbf{z}_{-(a,b)}$, all the other topic assignments to all the other words.

$$
\begin{aligned}
p\big(z_{a,b} \mid \mathbf{z}_{-(a,b)}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\big) &= \frac{p\big(z_{a,b}, \mathbf{z}_{-(a,b)}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\big)}{p\big(\mathbf{z}_{-(a,b)}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\big)} \\
&\propto p\big(z_{a,b}, \mathbf{z}_{-(a,b)}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\big) \\
&= p(\mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \iint p(\mathbf{w}, \mathbf{z}, \theta, \phi \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\theta d\phi \\
&= \iint p(\phi \mid \boldsymbol{\beta}) p(\theta \mid \boldsymbol{\alpha}) p(\mathbf{z} \mid \theta) p(\mathbf{w} \mid \phi, \mathbf{z}) d\theta d\phi \\
&= \int p(\mathbf{z} \mid \theta) p(\theta \mid \boldsymbol{\alpha}) d\theta \int p(\mathbf{w} \mid \phi, \mathbf{z}) p(\phi \mid \boldsymbol{\beta}) d\phi \\
&= \int \prod_{d=1}^{D} p\Big(\mathbf{z}_d \mid \boldsymbol{\theta}_d\Big) p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}\Big) d\theta \int \prod_{t=1}^{T} p\Big(\boldsymbol{\phi}_t \mid \boldsymbol{\beta}\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(w_{d,n} \mid \boldsymbol{\phi}_{z_{d,n}}\Big) d\phi \\
&= \underbrace{\prod_{d=1}^{D} \int p\Big(\mathbf{z}_d \mid \boldsymbol{\theta}_d\Big) p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}\Big) d\boldsymbol{\theta}_d}_{(I)} \times \underbrace{\prod_{t=1}^{T} \int p\Big(\boldsymbol{\phi}_t \mid \boldsymbol{\beta}\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(w_{d,n} \mid \boldsymbol{\phi}_{z_{d,n}}\Big) d\boldsymbol{\phi}_t}_{(II)}
\end{aligned}
$$

We can further define counts $c^{-(a,b)}$ that are defined in the same way as the other counts

described in Chapter 4, only without the counts for position (a, b). This means that:

- $c_{t,d,\cdot} = \begin{cases} c_{t,d,\cdot}^{-(a,b)} + 1 & \text{if } t = z_{a,b} \text{ and } d = a \\ c_{t,d,\cdot}^{-(a,b)} & \text{otherwise} \end{cases}$
- $c_{t,\cdot,j} = \begin{cases} c_{t,\cdot,j}^{-(a,b)} + 1 & \text{if } t = z_{a,b} \text{ and } j = w_{a,b} \\ c_{t,\cdot,j}^{-(a,b)} & \text{otherwise} \end{cases}$

Let's compute term (I) assuming that $\theta_d \sim \text{Dir}(\boldsymbol{\alpha})$:

$$
\begin{aligned}
(I) = \prod_{d=1}^{D} \int p\Big(\mathbf{z}_d \mid \boldsymbol{\theta}_d\Big) p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}\Big) d\boldsymbol{\theta}_d &= \prod_{d=1}^{D} \int \prod_{n=1}^{N_d} \theta_{d,z_{d,n}} \frac{\Gamma\Big(\alpha^+\Big)}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \prod_{t=1}^{T} \theta_{d,t}^{\alpha_t - 1} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \int \frac{\Gamma\Big(\alpha^+\Big)}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \prod_{t=1}^{T} \theta_{d,t}^{c_{t,d,\cdot}} \prod_{t=1}^{T} \theta_{d,t}^{\alpha_t - 1} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \int \frac{\Gamma\Big(\alpha^+\Big)}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \prod_{t=1}^{T} \theta_{d,t}^{c_{t,d,\cdot} + \alpha_t - 1} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \int \frac{\Gamma\Big(\alpha^+\Big)}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \frac{\prod_{t=1}^{T} \Gamma\Big(c_{t,d,\cdot} + \alpha_t\Big)}{\prod_{t=1}^{T} \Gamma\Big(c_{t,d,\cdot} + \alpha_t\Big)} \frac{\Gamma\Big(\sum_{t=1}^{T} c_{t,d,\cdot} + \alpha_t\Big)}{\Gamma\Big(\sum_{t=1}^{T} c_{t,d,\cdot} + \alpha_t\Big)} \\
&\qquad \times \prod_{t=1}^{T} \theta_{d,t}^{c_{t,d,\cdot} + \alpha_t - 1} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \frac{\Gamma\Big(\alpha^+\Big)}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \frac{\prod_{t=1}^{T} \Gamma\Big(c_{t,d,\cdot} + \alpha_t\Big)}{\Gamma\Big(\sum_{t=1}^{T} c_{t,d,\cdot} + \alpha_t\Big)} \\
&\qquad \times \underbrace{\int \frac{\Gamma\Big(\sum_{t=1}^{T} c_{t,d,\cdot} + \alpha_t\Big)}{\prod_{t=1}^{T} \Gamma\Big(c_{t,d,\cdot} + \alpha_t\Big)} \prod_{t=1}^{T} \theta_{d,t}^{c_{t,d,\cdot} + \alpha_t - 1} d\boldsymbol{\theta}_d}_{=1} \\
&= \prod_{d=1}^{D} \frac{\Gamma\Big(\alpha^+\Big)}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \frac{\prod_{t=1}^{T} \Gamma\Big(c_{t,d,\cdot} + \alpha_t\Big)}{\Gamma\Big(\sum_{t=1}^{T} c_{t,d,\cdot} + \alpha_t\Big)}
\end{aligned}
$$

Let's compute term (II) assuming that $\phi_t \sim \text{Dir}(\boldsymbol{\beta})$:

$$
\begin{aligned}
(II) &= \prod_{t=1}^{T} \int p\Big(\boldsymbol{\phi}_t \mid \boldsymbol{\beta}\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(w_{d,n} \mid \phi_{z_{d,n}}\Big) d\boldsymbol{\phi}_t \\
&= \prod_{t=1}^{T} \int \frac{\Gamma\Big(\beta^{+}\Big)}{\prod_{j=1}^{V} \Gamma\Big(\beta_j\Big)} \prod_{j=1}^{V} \phi_{t,j}^{\beta_t - 1} \prod_{d=1}^{D} \prod_{n=1}^{N_d} \phi_{z_{d,n}, w_{d,n}} d\boldsymbol{\phi}_t \\
&= \prod_{t=1}^{T} \int \frac{\Gamma\Big(\beta^{+}\Big)}{\prod_{j=1}^{V} \Gamma\Big(\beta_j\Big)} \prod_{j=1}^{V} \phi_{t,j}^{\beta_j - 1} \prod_{j=1}^{V} \phi_{t,j}^{c_{t,\cdot,j}} d\phi_t \\
&= \prod_{t=1}^{T} \int \frac{\Gamma\Big(\beta^{+}\Big)}{\prod_{j=1}^{V} \Gamma\Big(\beta_j\Big)} \prod_{j=1}^{V} \phi_{t,j}^{c_{t,j} + \beta_j - 1} d\phi_t \\
&= \prod_{t=1}^{T} \int \frac{\Gamma\Big(\beta^{+}\Big)}{\prod_{j=1}^{V} \Gamma\Big(\beta_j\Big)} \frac{\prod_{j=1}^{V} \Gamma\Big(c_{t,\cdot,j} + \beta_j\Big)}{\prod_{j=1}^{V} \Gamma\Big(c_{t,\cdot,j} + \beta_j\Big)} \frac{\Gamma\Big(\sum_{j=1}^{V} c_{t,\cdot,j} + \beta_j\Big)}{\Gamma\Big(\sum_{j=1}^{V} c_{t,\cdot,j} + \beta_j\Big)} \prod_{j=1}^{V} \phi_{t,j}^{c_{t,\cdot,j} + \beta_j - 1} d\phi_t \\
&= \prod_{t=1}^{T} \frac{\Gamma\Big(\beta^{+}\Big)}{\prod_{j=1}^{V} \Gamma\Big(\beta_j\Big)} \frac{\prod_{j=1}^{V} \Gamma\Big(c_{t,\cdot,j} + \beta_j\Big)}{\Gamma\Big(\sum_{j=1}^{V} c_{t,\cdot,j} + \beta_j\Big)} \underbrace{\int \frac{\Gamma\Big(\sum_{j=1}^{V} c_{t,\cdot,j} + \beta_j\Big)}{\prod_{j=1}^{V} \Gamma\Big(c_{t,\cdot,j} + \beta_j\Big)} \prod_{j=1}^{V} \phi_{t,j}^{c_{t,\cdot,j} + \beta_j - 1} d\phi_t}_{=1} \\
&= \prod_{t=1}^{T} \frac{\Gamma\Big(\beta^{+}\Big)}{\prod_{j=1}^{V} \Gamma\Big(\beta_j\Big)} \frac{\prod_{j=1}^{V} \Gamma\Big(c_{t,\cdot,j} + \beta_j\Big)}{\Gamma\Big(\sum_{j=1}^{V} c_{t,\cdot,j} + \beta_j\Big)}
\end{aligned}
$$

So, the full-conditionals have the form

$$
\begin{aligned}
&p\Big(z_{a,b} \,|\mathbf{z}_{-(a,b)}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\Big) \\
&\qquad = \prod_{d=1}^{D} \int p\Big(\mathbf{z}_d \mid \boldsymbol{\theta}_d\Big) p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}\Big) d\boldsymbol{\theta}_d \times \prod_{t=1}^{T} \int p\Big(\boldsymbol{\phi}_t \mid \boldsymbol{\beta}\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(w_{d,n} \mid \boldsymbol{\phi}_{z_{d,n}}\Big) d\boldsymbol{\phi}_t \\
&\qquad = \prod_{d=1}^{D} \frac{\Gamma\Big(\alpha^+\Big)}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \frac{\prod_{t=1}^{T} \Gamma\Big(c_{t,d,\cdot} + \alpha_t\Big)}{\Gamma\Big(\sum_{t=1}^{T} c_{t,d,\cdot} + \alpha_t\Big)} \times \prod_{t=1}^{T} \frac{\Gamma\Big(\beta^+\Big)}{\prod_{j=1}^{V} \Gamma\Big(\beta_j\Big)} \frac{\prod_{j=1}^{V} \Gamma\Big(c_{t,,j} + \beta_j\Big)}{\Gamma\Big(\sum_{j=1}^{V} c_{t,,j} + \beta_j\Big)} \\
&\qquad \propto \prod_{d=1}^{D} \frac{\prod_{t=1}^{T} \Gamma\Big(c_{t,d,\cdot} + \alpha_t\Big)}{\Gamma\Big(\sum_{t=1}^{T} c_{t,d,\cdot} + \alpha_t\Big)} \times \prod_{t=1}^{T} \frac{\prod_{j=1}^{V} \Gamma\Big(c_{t,,j} + \beta_j\Big)}{\Gamma\Big(\sum_{j=1}^{V} c_{t,,j} + \beta_j\Big)} \\
&\qquad = \Big[\prod_{d \neq a} \frac{\prod_{t=1}^{T} \Gamma\Big(c_{t,d,\cdot} + \alpha_t\Big)}{\Gamma\Big(\sum_{t=1}^{T} c_{t,d,\cdot} + \alpha_t\Big)}\Big] \cdot \frac{\prod_{t=1}^{T} \Gamma\Big(c_{t,a,\cdot} + \alpha_t\Big)}{\Gamma\Big(\sum_{t=1}^{T} c_{t,a,\cdot} + \alpha_t\Big)} \\
&\qquad\qquad\qquad \times \prod_{t=1}^{T} \frac{\Big[\prod_{j \neq w_{a,b}} \Gamma\Big(c_{t,\cdot,j} + \beta_j\Big)\Big] \Gamma\Big(c_{t,\cdot,w_{a,b}} + \beta_{w_{a,b}}\Big)}{\Gamma\Big(\sum_{j=1}^{V} c_{t,\cdot,j} + \beta_j\Big)} \\
&\qquad \propto \frac{\prod_{t=1}^{T} \Gamma\Big(c_{t,a,\cdot} + \alpha_t\Big)}{\Gamma\Big(\sum_{t=1}^{T} c_{t,a,\cdot} + \alpha_t\Big)} \times \prod_{t=1}^{T} \frac{\Gamma\Big(c_{t,\cdot,w_{a,b}} + \beta_{w_{a,b}}\Big)}{\Gamma\Big(\sum_{j=1}^{V} c_{t,\cdot,j} + \beta_j\Big)} \\
&\qquad \propto \frac{\Big(\prod_{k \neq z_{a,b}} \Gamma\Big(c_{t,a,\cdot}^{-(a,b)} + \alpha_t\Big)\Big) \Gamma\Big(c_{z_{a,b},a,\cdot}^{-(a,b)} + \alpha_{z_{a,b}} + 1\Big)}{\Gamma\Big(1 + \alpha^+ + \sum_{t=1}^{T} c_{t,a,\cdot}^{-(a,b)}\Big)} \cdot \\
&\qquad\qquad\qquad \cdot \left(\prod_{t \neq z_{a,b}} \frac{\Gamma\Big(c_{t,\cdot,w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^+ + \sum_{j=1}^{V} c_{t,\cdot,j}^{-(a,b)}\Big)}\right) \frac{\Gamma\Big(c_{z_{a,b},\cdot,w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}} + 1\Big)}{\Gamma\Big(1 + \beta^+ + \sum_{j=1}^{V} c_{z_{a,b},',j}^{-(a,b)}\Big)}
\end{aligned}
$$

Recalling that $\Gamma(x+1) = x\Gamma(x)$, we can rewrite the previous formula as

$$
\begin{aligned}
&\propto \frac{\Big(\prod_{t\neq z_{a,b}} \Gamma\Big(c_{t,a,\cdot}^{-(a,b)} + \alpha_t\Big)\Big)\Gamma\Big(c_{z_{a,b},a,\cdot}^{-(a,b)} + \alpha_{z_{a,b}}\Big)\Big(c_{z_{a,b},a,\cdot}^{-(a,b)} + \alpha_{z_{a,b}}\Big)}{\Gamma\Big(1 + \alpha^{+} + \sum_{t=1}^{T} c_{t,a,\cdot}^{-(a,b)}\Big)} \\
&\qquad \times \left(\prod_{t\neq z_{a,b}} \frac{\Gamma\Big(c_{t,\cdot,w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^{+} + \sum_{j=1}^{V} c_{t,\cdot,j}^{-(a,b)}\Big)}\right) \frac{\Gamma\Big(c_{z_{a,b},w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\Big)\Big(c_{z_{a,b},\cdot w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^{+} + \sum_{j=1}^{V} c_{z_{a,b},\cdot,j}^{-(a,b)}\Big)\Big(\beta^{+} + \sum_{j=1}^{V} c_{z_{a,b},\cdot,j}^{-(a,b)}\Big)} \\
&= \frac{\Big(\prod_{t=1}^{T} \Gamma\Big(c_{t,a,\cdot}^{-(a,b)} + \alpha_t\Big)\Big)\Big(c_{z_{a,b},a,\cdot}^{-(a,b)} + \alpha_{z_{a,b}}\Big)}{\Gamma\Big(1 + \alpha^{+} + \sum_{t=1}^{T} c_{t,a,\cdot}^{-(a,b)}\Big)} \left(\prod_{t=1}^{T} \frac{\Gamma\Big(c_{t,\cdot,w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^{+} + \sum_{j=1}^{V} c_{t,\cdot,j}^{-(a,b)}\Big)}\right) \frac{\Big(c_{z_{a,b},\cdot,w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\Big)}{\Big(\beta^{+} + \sum_{j=1}^{V} c_{z_{a,b},\cdot,j}^{-(a,b)}\Big)} \\
&\propto \frac{\Big(c_{z_{a,b},a,\cdot}^{-(a,b)} + \alpha_{z_{a,b}}\Big)\Big(c_{z_{a,b},\cdot,w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\Big)}{\Big(\beta^{+} + \sum_{j=1}^{V} c_{z_{a,b},\cdot j}^{-(a,b}\Big)}
\end{aligned}
$$

- $c_{z_{a,b},a,}^{-(a,b)}$ is the number of other words in document a that have been assigned to topic $z_{a,b}$
- $c_{z_{a,b},\cdot,w_{a,b}}^{-(a,b)}$ is the number of times the current word $w_{a,b}$ has been assigned to topic $z_{a,b}$

A.2 EFLDA

$$
\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \theta, \phi \mid \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\beta}) &= p(\phi \mid \boldsymbol{\beta}) p(\theta \mid \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) p(\mathbf{z} \mid \theta) p(\mathbf{w} \mid \phi, \mathbf{z}) \\
&= \prod_{t=1}^{T} p\Big(\boldsymbol{\phi}_t \mid \boldsymbol{\beta}\Big) \prod_{d=1}^{D} p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}\Big) \\
&\quad \times \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(z_{d,n} \mid \boldsymbol{\theta}_d\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(w_{d,n} \mid \boldsymbol{\phi}_{z_{d,n}}\Big)
\end{aligned}
$$

We want to compute the probability that topic $z_{a,b}$ is assigned to $w_{a,b}$ (i.e., the b-th word of document a) given $\mathbf{z}_{-(a,b)}$, all the other topic assignments to all the other words.

$$
\begin{aligned}
p\Big(z_{a,b} \mid \mathbf{z}_{-(a,b)}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\Big) &= \frac{p\Big(z_{a,b}, \mathbf{z}_{-(a,b)}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\Big)}{p\Big(\mathbf{z}_{-(a,b)}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\Big)} \\
&\propto p\Big(z_{a,b}, \mathbf{z}_{-(a,b)}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\Big) \\
&= p(\mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \iint p(\mathbf{w}, \mathbf{z}, \theta, \phi \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\theta d\phi \\
&= \iint p(\phi \mid \boldsymbol{\beta}) p(\theta \mid \boldsymbol{\alpha}) p(\mathbf{z} \mid \theta) p(\mathbf{w} \mid \phi, \mathbf{z}) d\theta d\phi \\
&= \int p(\mathbf{z} \mid \theta) p(\theta \mid \boldsymbol{\alpha}) d\theta \int p(\mathbf{w} \mid \phi, \mathbf{z}) p(\phi \mid \boldsymbol{\beta}) d\phi \\
&= \int \prod_{d=1}^{D} p\Big(\mathbf{z}_d \mid \boldsymbol{\theta}_d\Big) p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}\Big) d\theta \int \prod_{t=1}^{T} p\Big(\boldsymbol{\phi}_t \mid \boldsymbol{\beta}\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(w_{d,n} \mid \boldsymbol{\phi}_{z_{d,n}}\Big) d\phi \\
&= \underbrace{\prod_{d=1}^{D} \int p\Big(\mathbf{z}_d \mid \boldsymbol{\theta}_d\Big) p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}\Big) d\boldsymbol{\theta}_d}_{(I)} \times \underbrace{\prod_{t=1}^{T} \int p\Big(\boldsymbol{\phi}_t \mid \boldsymbol{\beta}\Big) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\Big(w_{d,n} \mid \boldsymbol{\phi}_{z_{d,n}}\Big) d\boldsymbol{\phi}_t}_{(II)}
\end{aligned}
$$

We can further define counts $c^{-(a,b)}$ that are defined in the same way as the above counts, only without the counts for position (a, b). This means that:

- $c_{t,d,\cdot} = \begin{cases} c_{t,d,\cdot}^{-(a,b)} + 1 & \text{if } t = z_{a,b} \text{ and } d = a \\ c_{t,d,\cdot}^{-(a,b)} & \text{otherwise} \end{cases}$
- $c_{t,\cdot,j} = \begin{cases} c_{t,\cdot,j}^{-(a,b)} + 1 & \text{if } t = z_{a,b} \text{ and } j = w_{a,b} \\ c_{t,\cdot,j}^{-(a,b)} & \text{otherwise} \end{cases}$

Let's compute term (I) assuming that $\theta_d \sim \text{EFD}(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$:

$$
\begin{aligned}
(I) &= \prod_{d=1}^{D} \int p\Big(\mathbf{z}_d \mid \boldsymbol{\theta}_d\Big) p\Big(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}\Big) d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \int \prod_{n=1}^{N_d} \theta_{d,z_{d,n}} \frac{1}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \Big(\prod_{t=1}^{T} \theta_{d,t}^{\alpha_t - 1} \Big) \cdot \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + \tau_h\Big)} \theta_{d,h}^{\tau_h} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \int \Big(\prod_{t=1}^{T} \theta_{d,t}^{c_{t,d,}} \Big) \frac{1}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \Big(\prod_{t=1}^{T} \theta_{d,t}^{\alpha_t - 1} \Big) \cdot \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + \tau_h\Big)} \theta_{d,h}^{\tau_h} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \int \Big(\prod_{t=1}^{T} \theta_{d,t}^{c_{t,d,\cdot} + \alpha_t - 1} \Big) \frac{1}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \cdot \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + \tau_h\Big)} \theta_{d,h}^{\tau_h} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \int \frac{1}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \cdot \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + \tau_h\Big)} \Big(\prod_{t=1}^{T} \theta_{d,t}^{c_{t,d,} + \alpha_t - 1} \Big) \theta_{d,h}^{\tau_h} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \frac{1}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \cdot \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + \tau_h\Big)} \int \Big(\prod_{t=1}^{T} \theta_{d,t}^{c_{t,d,} + \alpha_t - 1} \Big) \theta_{d,h}^{\tau_h} d\boldsymbol{\theta}_d \\
&= \prod_{d=1}^{D} \frac{1}{\prod_{t=1}^{T} \Gamma\Big(\alpha_t\Big)} \cdot \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + \tau_h\Big)} \frac{\prod_{t=1}^{T} \Gamma\Big(\alpha_t + c_{t,d,\cdot}\Big)\Gamma\Big(\alpha_h + \tau_h + c_{h,d,\cdot}\Big)}{\Gamma\Big(\alpha^+ + \tau_h + \sum_{t=1}^{T} c_{t,d,\cdot}\Big)\Gamma\Big(\alpha_h + c_{h,d,\cdot}\Big)} \\
&= \prod_{d=1}^{D} \Big(\prod_{t=1}^{T} \frac{\Gamma\Big(\alpha_t + c_{t,d,\cdot}\Big)}{\Gamma\Big(\alpha_t\Big)} \Big) \cdot \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)}{\Gamma\Big(\alpha_h + c_{h,d,\cdot}\Big)} \frac{\Gamma\Big(\alpha_h + \tau_h + c_{h,d,\cdot}\Big)}{\Gamma\Big(\alpha_h + \tau_h\Big)} \frac{\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha^+ + \tau_h + \sum_{t=1}^{T} c_{t,d,\cdot}\Big)}
\end{aligned}
$$

(II) coincides with (II) of the LDA:

$$
(II) = \prod_{t=1}^{T} \frac{\Gamma\Big(\beta^+\Big)}{\prod_{j=1}^{V} \Gamma\Big(\beta_j\Big)} \frac{\prod_{j=1}^{V} \Gamma\Big(c_{t,,j} + \beta_j\Big)}{\Gamma\Big(\beta^+ + \sum_{j=1}^{V} c_{t,,j}\Big)}.
$$

So, the full-conditionals have the form

$$
\begin{aligned}
&p\Big(z_{a,b} \mid \mathbf{z}_{-(a,b)}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\beta}\Big) \\
&\quad \propto \Big\{ \prod_{d=1}^{D} \Big(\prod_{t=1}^{T} \frac{\Gamma\Big(\alpha_t + c_{t,d,\cdot}\Big)}{\Gamma\Big(\alpha_t\Big)} \Big) \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h + \tau_h + c_{h,d,\cdot}\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + c_{h,d,\cdot}\Big)\Gamma\Big(\alpha_h + \tau_h\Big)\Gamma\Big(\alpha^+ + \tau_h + c_{\cdot,d,\cdot}\Big)} \Big\}
\end{aligned}
$$

$$
\begin{aligned}
&\times \prod_{t=1}^{T} \frac{\Gamma\Big(\beta^{+}\Big)}{\prod_{j=1}^{V}\Gamma\Big(\beta_j\Big)} \frac{\prod_{j=1}^{V}\Gamma\Big(c_{t,\cdot,j}+\beta_j\Big)}{\Gamma\Big(\beta^{+}+c_{t,\cdot,\cdot}\Big)} \\
&\propto \Big\{ \prod_{d=1}^{D} \Big[\Big(\prod_{t=1}^{T} \Gamma\Big(\alpha_t + c_{t,d,\cdot}\Big)\Big) \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h+\tau_h+c_{h,d,\cdot}\Big)\Gamma\Big(\alpha^{+}+\tau_h\Big)}{\Gamma\Big(\alpha_h+c_{h,d,}\Big)\Gamma\Big(\alpha_h+\tau_h\Big)\Gamma\Big(\alpha^{+}+\tau_h+c_{\cdot,d,\cdot}\Big)} \Big] \Big\} \\
&\quad \times \prod_{t=1}^{T} \frac{\prod_{j=1}^{V}\Gamma\Big(c_{t,\cdot,j}+\beta_j\Big)}{\Gamma\Big(\beta^{+}+c_{t,\cdot,\cdot}\Big)} \\
&\propto \Big\{ \prod_{d\neq a} \Big[\Big(\prod_{t=1}^{T} \Gamma\Big(\alpha_t + c_{t,d,\cdot}\Big)\Big) \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h+\tau_h+c_{h,d,\cdot}\Big)\Gamma\Big(\alpha^{+}+\tau_h\Big)}{\Gamma\Big(\alpha_h+c_{h,d,\cdot}\Big)\Gamma\Big(\alpha_h+\tau_h\Big)\Gamma\Big(\alpha^{+}+\tau_h+c_{,d,\cdot}\Big)} \Big] \Big\} \\
&\quad \times \Big[\Big(\prod_{t=1}^{T} \Gamma\Big(\alpha_t + c_{t,a,\cdot}\Big)\Big) \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h+\tau_h+c_{h,a,\cdot}\Big)\Gamma\Big(\alpha^{+}+\tau_h\Big)}{\Gamma\Big(\alpha_h+c_{h,a,\cdot}\Big)\Gamma\Big(\alpha_h+\tau_h\Big)\Gamma\Big(\alpha^{+}+\tau_h+c_{,a,\cdot}\Big)} \Big] \\
&\quad \times \prod_{t=1}^{T} \frac{\Big[\prod_{j\neq w_{a,b}}\Gamma\Big(c_{t,\cdot,j}+\beta_j\Big)\Big]\Gamma\Big(c_{t,\cdot,w_{a,b}}+\beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^{+}+c_{t,\cdot,\cdot}\Big)} \\
&\propto \Big[\Big(\prod_{t=1}^{T} \Gamma\Big(\alpha_t + c_{t,a,\cdot}\Big)\Big) \sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h+\tau_h+c_{h,a,\cdot}\Big)\Gamma\Big(\alpha^{+}+\tau_h\Big)}{\Gamma\Big(\alpha_h+c_{h,a,\cdot}\Big)\Gamma\Big(\alpha_h+\tau_h\Big)\Gamma\Big(\alpha^{+}+\tau_h+\sum_{t=1}^{T}c_{t,a,\cdot}\Big)} \Big] \\
&\quad \times \prod_{t=1}^{T} \frac{\Gamma\Big(c_{t,\cdot,w_{a,b}}+\beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^{+}+c_{t,\cdot,\cdot}\Big)} \\
&\propto \Big(\prod_{t\neq z_{a,b}} \Gamma\Big(\alpha_t + c_{t,a,\cdot}^{-(a,b)}\Big)\Big)\Gamma\Big(\alpha_{z_{a,b}} + c_{z_{a,b},a,\cdot}^{-(a,b)} + 1\Big) \\
&\quad \times \Big[\prod_{k\neq z_{a,b}} \frac{\Gamma\Big(c_{t,\cdot,w_{a,b}}^{-(a,b)}+\beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^{+}+c_{t,\cdot,\cdot}^{-(a,b)}\Big)} \Big] \frac{\Gamma\Big(c_{z_{a,b},\cdot w_{a,b}}^{-(a,b)}+\beta_{w_{a,b}}+1\Big)}{\Gamma\Big(1+\beta^{+}+c_{z_{a,b},\cdot,\cdot}^{-(a,b)}\Big)} \\
&\quad \times \Big[\sum_{h\neq z_{a,b}} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h+\tau_h+c_{h,a,\cdot}^{-(a,b)}\Big)\Gamma\Big(\alpha^{+}+\tau_h\Big)}{\Gamma\Big(\alpha_h+c_{h,a,\cdot}^{-(a,b)}\Big)\Gamma\Big(\alpha_h+\tau_h\Big)\Gamma\Big(\alpha^{+}+\tau_h+1+c_{\cdot,a,\cdot}^{-(a,b)}\Big)} \\
&\quad + p_{z_{a,b}} \frac{\Gamma\Big(\alpha_{z_{a,b}}\Big)\Gamma\Big(\alpha_{z_{a,b}}+\tau_{z_{a,b}}+c_{z_{a,b},a,\cdot}^{-(a,b)}+1\Big)\Gamma\Big(\alpha^{+}+\tau_{z_{a,b}}\Big)}{\Gamma\Big(\alpha_{z_{a,b}}+c_{z_{a,b},a,\cdot}^{-(a,b)}+1\Big)\Gamma\Big(\alpha_{z_{a,b}}+\tau_{z_{a,b}}\Big)\Gamma\Big(\alpha^{+}+\tau_{z_{a,b}}+1+c_{\cdot,a,\cdot}^{-(a,b)}\Big)} \Big]
\end{aligned}
$$

Recalling that $\Gamma(x+1) = x\Gamma(x)$ and let $N_a^{-} = c_{,a,\cdot}^{-(a,b)} = \sum_{t=1}^{T} c_{t,a,\cdot}^{-(a,b)}$ is the number of words contained in document a without considering word b and $x^{[n]} = x(x+1)\dots(x+n-1)$ is

the rising factorial function. We can rewrite the previous formula as

$$
\begin{aligned}
&\propto \Big\{\Big(\prod_{t\neq z_{a,b}} \Gamma\Big(\alpha_t + c^{-(a,b)}_{t,a,\cdot}\Big)\Big)\Gamma\Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)\Big\}\Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big) \\
&\quad\times \Big[\prod_{t\neq z_{a,b}} \frac{\Gamma\Big(c^{-(a,b)}_{t,\cdot,w_{a,b}} + \beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^+ + c^{-(a,b)}_{t,\cdot,\cdot}\Big)}\Big]\frac{\Gamma\Big(c^{-(a,b)}_{z_{a,b},\cdot w_{a,b}} + \beta_{w_{a,b}}\Big)}{\Gamma\Big(\beta^+ + c^{-(a,b)}_{z_{a,b},\cdot,\cdot}\Big)}\frac{\Big(c^{-(a,b)}_{z_{a,b},\cdot w_{a,b}} + \beta_{w_{a,b}}\Big)}{\Big(\beta^+ + c^{-(a,b)}_{z_{a,b},\cdot,\cdot}\Big)} \\
&\quad\times \Big[\sum_{h\neq z_{a,b}} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h + \tau_h + c^{-(a,b)}_{h,a,\cdot}\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + c^{-(a,b)}_{h,a,\cdot}\Big)\Gamma\Big(\alpha_h + \tau_h\Big)\Gamma\Big(\alpha^+ + \tau_h + 1 + c^{-(a,b)}_{\cdot,a,\cdot}\Big)} \\
&\quad\quad + p_{z_{a,b}} \frac{\Gamma\Big(\alpha_{z_{a,b}}\Big)\Gamma\Big(\alpha_{z_{a,b}} + \tau_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)\Gamma\Big(\alpha^+ + \tau_{z_{a,b}}\Big)}{\Gamma\Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)\Gamma\Big(\alpha_{z_{a,b}} + \tau_{z_{a,b}}\Big)\Gamma\Big(\alpha^+ + \tau_{z_{a,b}} + 1 + c^{-(a,b)}_{\cdot,a,\cdot}\Big)} \\
&\quad\quad\times \frac{\Big(\alpha_{z_{a,b}} + \tau_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)}{\Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)}\Big] \\
&\propto \Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)\cdot\frac{\Big(c^{-(a,b)}_{z_{a,b},\cdot,w_{a,b}} + \beta_{w_{a,b}}\Big)}{\Big(\beta^+ + c^{-(a,b)}_{z_{a,b},\cdot,\cdot}\Big)} \\
&\quad\times \Big[\sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h + \tau_h + c^{-(a,b)}_{h,a,\cdot}\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + c^{-(a,b)}_{h,a,\cdot}\Big)\Gamma\Big(\alpha_h + \tau_h\Big)\Gamma\Big(\alpha^+ + \tau_h + 1 + c^{-(a,b)}_{\cdot,a,\cdot,}\Big)} \\
&\quad\quad + p_{z_{a,b}} \frac{\Gamma\Big(\alpha_{z_{a,b}}\Big)\Gamma\Big(\alpha_{z_{a,b}} + \tau_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)\Gamma\Big(\alpha^+ + \tau_{z_{a,b}}\Big)}{\Gamma\Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)\Gamma\Big(\alpha_{z_{a,b}} + \tau_{z_{a,b}}\Big)\Gamma\Big(\alpha^+ + \tau_{z_{a,b}} + 1 + c^{-(a,b)}_{\cdot,a,\cdot}\Big)} \\
&\quad\quad\times \left(\frac{\Big(\alpha_{z_{a,b}} + \tau_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)}{\Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)} - 1\right)\Bigg] \\
&\propto \Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)\frac{\Big(c^{-(a,b)}_{z_{a,b},w_{a,b}} + \beta_{w_{a,b}}\Big)}{\Big(\beta^+ + c^{-(a,b)}_{z_{a,b},\cdot,\cdot}\Big)} \\
&\quad\times \Big[\sum_{h=1}^{T} p_h \frac{\Gamma\Big(\alpha_h\Big)\Gamma\Big(\alpha_h + \tau_h + c^{-(a,b)}_{h,a,\cdot}\Big)\Gamma\Big(\alpha^+ + \tau_h\Big)}{\Gamma\Big(\alpha_h + c^{-(a,b)}_{h,a,\cdot}\Big)\Gamma\Big(\alpha_h + \tau_h\Big)\Gamma\Big(\alpha^+ + \tau_h + 1 + c^{-(a,b)}_{\cdot,a,\cdot}\Big)} \\
&\quad\quad + p_{z_{a,b}} \frac{\Gamma\Big(\alpha_{z_{a,b}}\Big)\Gamma\Big(\alpha_{z_{a,b}} + \tau_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Big)\Gamma\Big(\alpha^+ + \tau_{z_{a,b}}\Big)}{\Gamma\Big(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\Gamma\Big(\alpha_{z_{a,b}} + \tau_{z_{a,b}}\Big)\Gamma\Big(\alpha^+ + \tau_{z_{a,b}} + 1 + c^{-(a,b)}_{\cdot,a,\cdot,}\Big)}\Big(\frac{\tau_{z_{a,b}}}{\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}}\Big)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\propto \left(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}\right) \frac{\left(c^{-(a,b)}_{z_{a,b},\cdot,w_{a,b}} + \beta_{w_{a,b}}\right)}{\left(\beta^{+} + c^{-(a,b)}_{z_{a,b},\cdot,\cdot}\right)} \\
&\quad \times \left[\sum_{h=1}^{T} p_h \frac{\left(\alpha_h + \tau_h\right)^{\left[c^{-(a,b)}_{h,a,\cdot}\right]}}{\left(\alpha_h\right)^{\left[c^{-(a,b)}_{h,a,\cdot}\right]}} \frac{1}{\left(\alpha^{+} + \tau_h + N_a^{-}\right)\left(\alpha^{+} + \tau_h\right)^{\left[N_a^{-}\right]}} \right. \\
&\quad \left. + p_{z_{a,b}} \frac{\left(\alpha_{z_{a,b}} + \tau_{z_{a,b}}\right)^{\left[c^{-(a,b)}_{z_{a,b},a,\cdot}\right]}}{\left(\alpha_{z_{a,b}}\right)^{\left[c^{-(a,b}_{z_{a,b},a,\cdot}\right]}} \frac{1}{\left(\alpha^{+} + \tau_{z_{a,b}} + N_a^{-}\right)\left(\alpha^{+} + \tau_{z_{a,b}}\right)^{\left[N_a^{-}\right]}} \left(\frac{\tau_{z_{a,b}}}{\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,\cdot}}\right) \right]
\end{aligned}
$$

Aknowledgement

With immense pleasure and deep sense of gratitude, I wish to express my sincere thanks to my supervisor **Bernardo Nipoti**, without his continuous encouragement, this research would not have been successfully completed.

I express my genuine thanks to **Michele Guindani**, for his kind words of support and encouragement. I also want to thank him for motivating me to carry out my research and for hosting me in the U.S. for an entire year.

A special thanks also to **Sonia Migliorati** for a long time my guide and my tutor, always ready to spend some words to help me going on with this path.

I would like to express my profound gratitude for the unwavering support provided by **my friends** throughout my Ph.D. journey. Specifically, I owe a debt of thanks to **Lidia**, who has consistently been my source of meme-humor both within and beyond the academic sphere; **Roberto**, my steadfast friend from within Bicocca; and **Alessandro & Federico**, my cherished "cotutelle" colleagues. Additionally, I am deeply thankful to have crossed paths with **Federica, Ziyi**, and all **my Californian friends** on the other side of the globe.

Last but by no means least, I want to convey my profound appreciation to my boyfriend, **Giorgio**, for his consistent encouragement, unwavering moral support, exceptional patience, infinite love and profound understanding.

It was a roller-coaster, it was painful, it was tough but it also was a great adventure.

Alice