SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of
Economics, Management and Statistics

PhD program: **Economics, Statistics and Data Science** Cycle: **XXXVII**
Curriculum: **Statistics**

Bayesian learning of latent discrete structures: graphs, clusters and features allocation

Surname: **COLOMBI**
Name: **ALESSANDRO**
Registration number: **883875**

Supervisor: Prof. **RAFFAELE ARGIENTO**
Co-Supervisor: Prof. **LUCIA PACI**

Coordinator: Prof. **MATTEO MANERA**

Academic Year: 2024/2025

Abstract

Discrete random structures play a central role in modern statistics, as they provide powerful tools to represent hidden aspects of complex phenomena. This thesis develops Bayesian methodologies for the analysis of three fundamental classes of latent discrete structures: graphs, which describe the network of conditional dependencies among random variables; clusters, which represent groups of observations sharing common characteristics; and dynamic features, which correspond to collections of latent attributes that may evolve over time. Across these domains, a unifying challenge lies in the combinatorial explosion of possible configurations, which makes exhaustive enumeration and direct evaluation infeasible. Within the Bayesian framework, this difficulty is addressed through the careful design of prior distributions and through efficient inferential strategies that allow learning in high-dimensional discrete spaces. The thesis is organized into two main parts. The first focuses on graphical models, introducing a block-structured approach with an application to spectrometric data. The second part, on Bayesian nonparametrics, extends the analysis to infinite-dimensional settings, moving beyond the classical exchangeable framework to address more complex scenarios such as clustering and species sampling across multiple populations, as well as dynamic feature allocation, where latent characteristics can appear, disappear, and reappear over time.

Acknowledgements

First of all, I would like to thank my supervisors, Raffaele Argiento and Lucia Paci, for their invaluable and constant support throughout these years.
I also want to express my gratitude to Professors Garritt Page and Jeffrey Miller, whose insightful comments and suggestions during the review process have greatly enhanced both the quality and readability of this thesis.
I would also like to express my sincere gratitude to all those with whom I had the pleasure of collaborating directly in the preparation of the works included in this thesis. Thanks to Alessia Pini, Federico Camerlenghi, and Jim Griffin, whom I also wish to thank for welcoming me to University College London for three months.
Although not formally involved in the work I carried out during these years, I would also like to thank Mario Beraha, Lorenzo Ghilotti, Matteo Gianella and Alessandra Guglielmi for their valuable advice, suggestions, and exchange of ideas.

Contents

List of Figures

List of Tables

Introduction

Discrete random variables arise naturally in modern statistics, as they are often employed to describe hidden properties of complex phenomena. In this thesis, we focus on a Bayesian analysis of three fundamental classes of latent discrete structures: (i) graphs, which capture the unobservable network of conditional dependencies among the components of a multivariate random variable; (ii) clusters, which represent latent groupings of parameters and enable the identification of population subgroups that are internally homogeneous yet distinct from individuals in other groups; and (iii) features, that is, collections of specific characteristics for which we study their evolution over time.

A common challenge across these three domains lies in the size of the discrete configuration spaces: the number of possible graphs, partitions, or feature allocations grows so quickly that exhaustive enumeration and direct evaluation are computationally infeasible. Within the Bayesian framework, this difficulty can be addressed in two complementary ways. First, by carefully specifying prior distributions that encode our intuition about which structures are more plausible, we can guide the exploration of the discrete space toward its most promising regions. Second, by designing efficient computational methods, we can perform inference and learning even in such high-dimensional, combinatorial settings.

This thesis is organized into two main parts, entitled *Graphical Models* and *Bayesian Nonparametrics*. The first part is devoted to the study of graphical structures, with a particular focus on an application involving spectrometric data which represent the spectrum of absorbance of strawberry purees. Through a suitable representation of the spectrum using B-spline basis, the problem of studying the structure of dependencies among the chemical-groups that form the purees is translated into the identification of dependencies among the spline coefficients. The task is further complicated by the fact that our scientific interest lies not only in connections among individual coefficients but also in connections among subsets of spectral bands, and consequently among groups of nodes in the graph. This naturally leads to the study of block-structured graphs, where the adjacency matrix is endowed with a block structure reflecting such grouped dependencies.

In Chapter 2, we work conditionally on a known and fixed partition of the nodes, provided by domain experts, and propose a novel class of prior distributions on the graph space that explicitly incorporates this grouping information. The resulting priors enforce a block structure on the adjacency matrix of the graph. Our construction proceeds by mapping the original graph into a block-structured multigraph representation, where each block of potential edges between two groups of nodes is collapsed into a single edge in the multigraph representation. Independent Bernoulli priors are then assigned to the edges of this multigraph, so that nodes belonging to different groups are either fully connected or completely disconnected. We refer to this new class of priors as block graph priors. By construction, they impose structural constraints that drastically reduce the size of the discrete space in which the latent graph lives, thereby making inference more tractable. To explore this space efficiently, we develop a tailored

Reversible Jump Markov Chain Monte Carlo sampler (Green, 1995), defined jointly over the graph and its associated precision matrix, and designed to exploit the block structure encoded by the prior. In particular, we extend the procedure of Lenkoski (2013) so that each move of the chain modifies an entire block of edges at once, ensuring that the block structure of the adjacency matrix is preserved throughout the sampling process. Furthermore, we combine this scheme with the Exchange algorithm (Murray et al., 2006), which introduces a secondary reversible move that circumvents the intractable normalizing constant of the G-Wishart distribution. We refer to the resulting method as the Block Double Reversible Jump algorithm.

Chapter 3 addresses the main limitation of the previous chapter, namely the assumption of a fixed partition of the nodes based on expert knowledge. While such expert-driven partitions are valuable, they are not directly supported by the data and do not allow us to quantify the uncertainty in key structural features such as the number of groups. This motivates the need for a data-driven partitioning approach, grounded in a probabilistic model and capable of learning both the grouping of nodes and its associated uncertainty. Building on the ideas of van den Boom et al. (2022b), we propose a graphical model that not only allows the estimation of individual edges but also provides a more expressive description of the graph by employing a Stochastic Block Model (Holland et al., 1983) as a prior on the graph structure. However, our setting introduces a crucial additional constraint: the spectral bands are inherently ordered, since each corresponds to a specific wavelength interval. This ordering constraint prevents the usage of standard strategies as in Legramanti et al. (2022) in the Stochastic Block Models literature for learning the random partition of the nodes. To address this challenge, we draw inspiration from the rich literature on changepoint models (Smith, 1975; Green, 1995; Barry and Hartigan, 1993). Although infrared spectra are not temporal processes, wavelengths are naturally ordered in the same way as time, making changepoint models a natural methodological analogue. Among the many approaches available for multiple changepoint detection, we adopt the model of Martínez and Mena (2014), which offers an elegant extension of Bayesian nonparametric tools to the case of ordered partitions. From a computational perspective, this model requires simultaneous learning of two latent discrete structures: the graph and the partition of its nodes. To this end, we employ suitable adaptations of the adaptive split-merge sampler for changepoint models proposed by Benson and Friel (2018) to learn the ordered partition and the Birth-Death chain by Mohammadi and Wit (2015) to sample over the space of graphs.

In Chapter 4, we provide a concise review of the fundamental concepts of Bayesian nonparametrics in the exchangeable setting, with a particular emphasis on their connection with point processes. In addition, we outline the basic ideas for moving beyond exchangeability, which constitutes the main goal of the three chapters that follow. Specifically, in Chapter 5, we present a mixture model based on a novel family of Bayesian priors designed for multilevel data and obtained by normalizing a finite point process. In this case, the latent discrete structure we are interested in is the partition of individuals both within and between the groups. To achieve this, we study a Bayesian nonparametric prior that allows for a random number of atoms to be shared among groups, while assigning group-specific mixture weights to accommodate heterogeneity between them. This construction combines flexibility at the individual level with the ability to capture structural differences across groups. Methodologically, our approach follows the line of Argiento and De Iorio (2022) and extends the popular Mixture of Finite Mixtures model (Miller and Harrison, 2018) to a hierarchical framework. A full distribution theory for this new family and the induced clustering is developed, including the marginal, posterior, and

predictive distributions. From a computational perspective, we propose both marginal and conditional Gibbs samplers, carefully designed to provide efficient posterior inference. Our experiments show that these algorithms significantly outperform existing implementations of the Hierarchical Dirichlet Process, which currently represents the state-of-the-art approach for modeling multilevel data.

In Chapter 6, we apply the Bayesian nonparametric prior introduced in the previous chapter to an ecological setting, studying the problem of unseen distinct and shared species when samples are collected in two different areas. Here, the discrete structure of interest is the classification of observations into species identified during sampling. Unlike in previous chapters, this structure is observable, and the goal is not to infer a latent partition but to predict the behavior of future observations, namely whether they will belong to already detected species or reveal new, previously unseen ones. More specifically, our approach answers the question: "How many species not yet observed in the two areas will be discovered in a future sample?" To this end, we derive closed-form distributional results for both in-sample characterizations and out-of-sample predictions, including the number of unseen species in future samples of arbitrary finite size. These results provide a principled basis for sample size determination in biodiversity studies, as well as for evaluating sample coverage (Good, 1953), both within the observed data and in future collections. A notable contribution is that our theory extends beyond the classical frequentist framework, which typically only handles one-step-ahead predictions, by offering results valid for any finite sample sizes in the two areas.

Finally, in Chapter 7, we address the challenge of modeling time-varying data within a Bayesian nonparametric framework. Indeed, while most Bayesian nonparametric models have been developed for static settings, many modern applications involve temporal dynamics that require priors capable of evolving over time. To this end, we introduce a novel and unified framework for time-dependent Bayesian nonparametric modeling, applicable to a broad class of statistical problems whose likelihood belongs to the exponential family. Specifically, we construct a sequence of time-dependent random measures with arbitrary marginal distributions. To achieve this, we build upon the work of Pitt et al. (2002), extending their parametric construction to the nonparametric setting. This extension is made possible by leveraging the conjugacy property ensured by the broad class of exponential completely random measures introduced by Broderick et al. (2018). Such a property guarantees mathematical tractability and enables the definition of flexible and analytically convenient transition kernels that can be efficiently evaluated. The resulting process admits a simple yet powerful representation as a quasi-autoregressive process of order one, offering both interpretability and a foundation for generalizations to more intricate forms of temporal dependence. These processes are particularly suitable for real-world applications, among which we focus on dynamic feature allocation models, where time-evolving features are captured by our construction, which naturally exhibits well-understood properties such as the appearance, disappearance, and possible reappearance of latent features over time.

The thesis concludes with a brief discussion and an outline of possible future research directions for each chapter, which we chose to present collectively at the end rather than at the conclusion of each individual chapter.

Part I

Graphical models

Chapter 1

Introduction to Gaussian Graphical Models for Spectrometric Data Analysis

This chapter, based on Codazzi et al. (2022), is a joint work with Laura Codazzi, Matteo Gianella, Raffaele Argiento, Lucia Paci, and Alessia Pini. Most of the work was carried out before the start of my Ph.D. program. The method and data, however, are instrumental for the development of Chapters 2 and 3, which were instead conducted during the Ph.D.

1.1 Motivating problem

The analysis of the interaction of infrared radiation with matter, through absorption, emission, or reflection, is accomplished by *infrared spectroscopy*. This technique is used to study and identify chemical substances in solid, liquid, or gaseous forms. For instance, mid-infrared spectroscopy coupled with chemometrics or statistical techniques have been used in the literature to study the substance composition and detect the presence of adulterants in food (Kemsley et al., 1996; Downey et al., 1997; Holland et al., 1998; Meza-Márquez et al., 2010).

From a mathematical point of view, a spectrum is a continuous function of the wavelength. The dependence structure between the signal at different wavelength bands is important to understand which bands are related to the different components: if two different bands of the spectrum are associated, we can conclude that they refer to the same components, or to closely related components. Our goal in this paper is then to investigate the structure of conditional dependence among different portions of an absorbance spectrum. The importance of assessing the dependence structure between infrared emission bands is a rather new topic in mid-infrared spectroscopy of chemical substances, even though a few recent works focus on this topic. We mention Casa et al. (2022), who focus on assessing the correlation structure via latent variable models. The problem is however not new in other fields like astronomy, where correlation between emission bands is often studied (e.g., Cohen et al., 1986).

Since infrared spectra are continuous functions of the wavelength, we embed this problem in the framework of functional data analysis (see e.g., Lee et al., 2002; Ramsay and Silverman, 2005; Xiao et al., 2016). In the Bayesian setting, Yang et al. (2016) proposed a hierarchical model with Gaussian-Wishart processes for simultaneously smoothing multiple functional observations and estimating mean-covariance functions. However, like any Gaussian process based approach, their model suffers serious

computational burden when functional data are observed on high-dimensional grids. To address this computational issue, Yang et al. (2017) proposed to approximate the underlying true functional data with basis functions, and derive the induced Bayesian hierarchical model for the associated smoothing coefficients based on a Gaussian-Wishart prior.

Following the latter approach, we interpret the analysis of spectrometric data as a smoothing problem of functional data, followed by inference on the smoothing coefficients. Specifically, we use a B-spline basis expansion to represent the functional data: indeed, a generic B-spline basis function has a compact support, meaning that it is different from zero only in a portion of the domain, i.e., on a defined portion of the spectrum. If the number of basis functions is large, this portion is very small compared to the whole domain, and identifies a very precise band of the spectrum. Through the basis expansion, a coefficient is associated with every basis function, and the whole function can be represented through the vector of associated coefficients $\boldsymbol{\beta}$. As a consequence, we can detect a hidden association between different bands of the spectrum through an association between the corresponding smoothing coefficients.

Inference on B-spline basis coefficients has been used in the functional data analysis literature for assessing the mean structure of functional data along the domain (see e.g., Pini and Vantini, 2016). Here, instead, we focus on the dependence structure of the smoothing coefficients. With a Bayesian perspective, this concerns the specification of a prior distribution for the coefficients that encodes a dependence structure. Usually, Bayesian penalized splines or random walk priors (Lang and Brezger, 2004; Telesca and Inoue, 2008; Crainiceanu and Goldsmith, 2010) are employed to induce smoothness on the $\boldsymbol{\beta}$'s, translating into a banded precision matrix that is fixed and depends only on a smoothing parameter. However, in our application, assuming that the dependence structure is known and restricted to the structure induced by a random walk prior appears inadequate. Indeed, coefficients associated with closer bands of the spectrum are not necessarily expected to be similar since they may relate to different components. Rather, long range interactions may also exist between far apart portions of the spectrum. Hence, a more flexible modeling framework, encompassing the random walk case, is advocated.

A widely used approach for modeling dependence in samples of functional data is functional principal component analysis (FPCA). We mention here Crainiceanu and Goldsmith (2010), studying FPCA in the Bayesian framework. However, information about the dependency structure obtained by this method is sometimes based on qualitative considerations on the shape of the principal components.

In this Chapter, we move a step forward and develop a method to learn the dependence structure among portions of the absorbance spectrum based on graphical modeling. Graphical models describe a mapping between a graph and a family of multivariate probability models (Lauritzen, 1996). Namely, they embody conditional independence relationships among a set of random variables which can be read off from the graph. Usually, the structure of the underlying graph is unknown and needs to be estimated on the basis of the available data: this is referred to as graph structural learning.

Graphical models have been widely applied to infer several types of networks, particularly under the Bayesian framework, see among others Dobra et al. (2004); Peterson et al. (2015); Tan et al. (2017); Ni et al. (2017); Cremaschi et al. (2019); Paci and Consonni (2020).

All of the cited works assume a graphical model as a model for the observed data, meaning that the graph represents the conditional independence relationships among the measured variables. In contrast, here we introduce a graphical model as a prior specification for the smoothing coefficients, so that the graph encodes the dependence structure among portions of the absorbance spectrum. Beyond the

advantage in terms of interpretation, which remains central, another benefit of using the graphical model as a prior for the spline coefficients is that this approach exploits the functional representation of each observation. This enables dimensionality reduction and more effective filtering of observational noise, thereby improving both inference and interpretability.

Graph-based priors for regression coefficients have been studied in a variable selection setting by, among others, Liu et al. (2014), Chakraborty and Lozano (2019), and Cai et al. (2020), who proposed Bayesian regularization priors to favor sparsity and clustering based on the graph Laplacian. However, these approaches are not suitable to deal with functional data where curve-specific coefficients are needed to simultaneously smooth the curves.

In addition, recent works focused on graphical modeling for multivariate functional data analysis, see among others Zhu et al. (2016), Li and Solea (2018), Qiao et al. (2019),Qiao et al. (2020). These approaches, known as functional graphical models, aim at depicting the conditional dependence structure among multiple random functions observed over a set of individuals, i.e., they assume a graphical model where the vertices of the underlying graph are the multiple functions and the edges are estimated from the conditional covariance function extended to the functional domain. Rather, in our setting, we work with univariate functional data modeled through a B-spline basis expansion, and we assume a graphical model where the nodes of the graph are the smoothing coefficients and the edges reflect the conditional dependence structure among the portions of the functional domain associated with such coefficients. Graphical models and splines have been jointly studied in the literature, although in a different context, see e.g., Morrissey et al. (2011) and Ni et al. (2015).

Summarizing, the contribution of Codazzi et al. (2022) reviewed in this chapter is to introduce a Bayesian hierarchical model for simultaneously smoothing functional data and learning the independence structure along the domain of the curves. Absorbance spectra are modeled as continuous functional data through a cubic B-spline basis expansion such that the dependence between two bands of the spectrum is reflected in the relationship among the corresponding smoothing coefficients. A Gaussian graphical model is then assumed as a prior for basis expansion coefficients to enable structural learning of frequency bands of the spectrum. The Bayesian p-spline model is also encompassed as a special case by our specification when the precision matrix is assumed to be fixed. On the computational side, we design an efficient sampling strategy to approximate the joint posterior distribution of the graph and model parameters. We illustrate our method on simulated datasets as well as on a real data set, studying the infrared absorbance spectra of strawberry purees.

1.2 Bayesian graphical model for smoothing functional data

1.2.1 Smoothing procedure

We reinterpret our task as a traditional smoothing procedure of functional data in a Bayesian framework. Let n be the number of curves, r the number of grid points, i.e., the wavelengths at which absorbance is measured, and p be the number of basis functions chosen for the smoothing. Let $y_i(s)$ be the absorbance at wavelength $s \in [d, u]$ for spectrum i, $i = 1, \ldots, n$, where $[d, u] \subset \mathbb{R}$ is the common domain of all curves. The whole curve $y_i(s)$ is then the absorbance spectrum of unit i. According to the usual

smoothing technique, the model we assume for the i-th spectrum is:

$$y_i(s) = \sum_{j=1}^{p} \beta_{ij}\varphi_j(s) + \varepsilon_i(s),$$

where $\varphi_1, \ldots, \varphi_p$ are suitable cubic B-spline basis functions, $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^\top$ is the spline coefficients' vector specific to the i-th curve and $\varepsilon_i(s) \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2_\varepsilon)$. We introduce the cubic B-spline design matrix $\boldsymbol{\Phi} \in \mathbb{R}^{r\times p}$ as:

$$\boldsymbol{\Phi} = \begin{bmatrix} \varphi_1(s_1) & \varphi_2(s_1) & \ldots & \varphi_p(s_1) \\ \varphi_1(s_2) & \varphi_2(s_2) & \ldots & \varphi_p(s_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(s_r) & \varphi_2(s_r) & \ldots & \varphi_p(s_r) \end{bmatrix}. \tag{1.1}$$

So, the element lj-th of $\boldsymbol{\Phi}$ is j-th basis function $\varphi_j(\cdot)$ evaluated at the l-th grid point s_l, $l = 1, \ldots, r$. Note that for ease of notation the smoothing model is described here for data evaluated at the same grid of r points. However, the smoothing model can be also used in the case where curves are measured at different wavelengths, by modifying $\boldsymbol{\Phi}$ accordingly, which would be particularly useful for sparse functional data.

The right panel of Figure 1.1 shows an example of a 10-dimensional cubic B-spline basis for the curve in the left panel. Note that the support of each basis function is a small portion of the domain, which decreases its size when the number of basis functions increases. For instance, in Figure 1.1, the support of each basis function overlaps with only four other functions' supports. To facilitate the interpretation of the results of our model, we will consider each B-spline basis coefficient as representative of the central part of each B-spline basis function only.

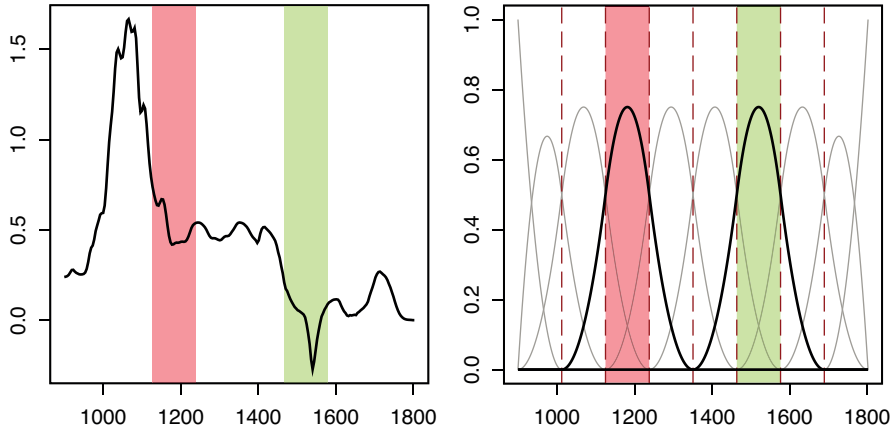

Figure 1.1: Example of a curve (left panel) and the corresponding 10-dimensional cubic B-spline basis (right panel). The colored regions in the two panels highlight the corresponding portions of the functional domain associated with the B-spline basis coefficients.

Let $\boldsymbol{Y}_i = (y_i(s_1), \ldots, y_i(s_r))^\top$ be the absorbance spectrum at all observed wavelengths of curve i. The smoothing model assumes the observation to follow a conditional normal sampling distribution, that is

$$\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i,\ \tau^2_\varepsilon \overset{\text{ind}}{\sim} \mathrm{N}_r\left(\boldsymbol{\Phi}\boldsymbol{\beta}_i,\ \tau^2_\varepsilon \mathbf{I}_r\right), \tag{1.2}$$

where τ^2_ε is the error variance and $\boldsymbol{I}_r$ is the $r \times r$ identity matrix.

In this work, we rely on a hierarchical Bayesian approach to smooth all functional observations simultaneously, enabling the borrowing of strength across all curves. To accomplish that, we place a prior distribution for the $\boldsymbol{\beta}_i$ coefficients given by

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n \mid \boldsymbol{\mu}, \boldsymbol{\Omega} \overset{\text{iid}}{\sim} \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}), \tag{1.3}$$

where $\boldsymbol{\mu}$ is the prior mean vector and $\boldsymbol{\Omega}$ is the precision matrix. Yang et al. (2017) derived the conjugate prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}^{-1}$ from the assumption that the underlying true functional data come from a Gaussian process, and thus resulting in a Normal-Wishart distribution that depends on the basis functions. In this work, instead, we frame the prior model for the $\boldsymbol{\beta}$'s into the Gaussian graphical modeling setting. Indeed, given the peculiar form of each B-spline function, the dependence between two bands of the spectrum is reflected in the relationship among the corresponding smoothing regression parameters. Hence, our goal boils down to the study of the conditional dependence structure of the $\boldsymbol{\beta}$'s encoded in the precision matrix $\boldsymbol{\Omega}$.

1.2.2 Graph structural learning

Let $G = (V, E)$ be an undirected graph defined by the node set $V = \{1, \ldots, p\}$ representing the coefficients and by the undirected edge set $E \subset V \times V$. We assume the precision matrix $\boldsymbol{\Omega}$ to be Markov with respect to G. Namely, dropping the subscript i for simplicity in the exposition, coefficients β_j and β_k are conditionally independent given all the remaining variables, $\boldsymbol{\beta}_{\setminus\{j,k\}}$, whenever $\{j, k\} \notin E$, if and only if the corresponding entry in matrix $\boldsymbol{\Omega}$ is zero, i.e., $\beta_j \perp \beta_k \mid \boldsymbol{\beta}_{\setminus\{j,k\}} \Leftrightarrow \boldsymbol{\Omega}_{jk} = 0$; see Figure 1.2 for an example. The graph G summarizes the conditional independence structure of the coefficients and

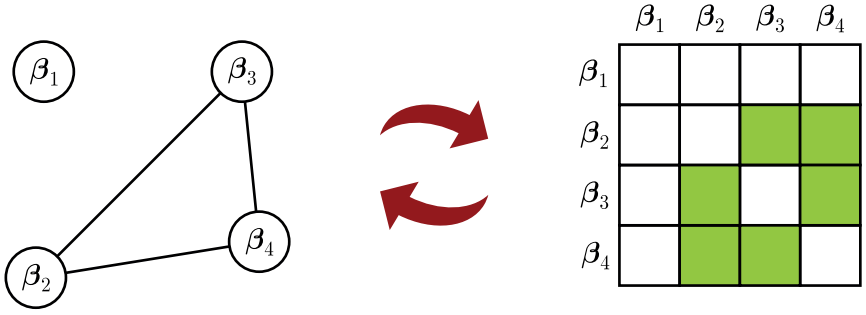

Figure 1.2: Example of a Gaussian graphical model with four nodes. On the left, the graph with four nodes and, on the right, the corresponding precision matrix where green cells represent nonzero entries.

represents the main goal of our inference. Indeed, the graph is unknown and must be estimated. Note that this is different from estimating the conditional network among the predictors of a regression model as in Peterson et al. (2016). Unlike measurement-error models in which predictors and response are both measured with error, spline basis expansion is fixed and known. Hence, we infer a sparse network on the associated smoothing coefficients, which provides information on which bases have similar slopes.

Given $\mathbb{P}_G$ the space of $p \times p$ positive definite matrices that are Markov with respect to graph G, we assume the precision matrix $\boldsymbol{\Omega} \in \mathbb{P}_G$ to be distributed, a priori, as a G-Wishart$(d, \boldsymbol{D})$ (Roverato, 2002; Letac and Massam, 2007) with density

$$P(\boldsymbol{\Omega} \mid G, d, \boldsymbol{D}) = I_G(d, \boldsymbol{D})^{-1} |\boldsymbol{\Omega}|^{(d-2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{D})\right\}, \tag{1.4}$$

where $d > 2$ is the shape parameter, $\boldsymbol{D}$ is the $p \times p$ positive definite inverse scale matrix, $\text{tr}(\cdot)$ denotes the trace operator and $I_G(d, \boldsymbol{D})$ is the normalizing constant,

$$I_G(d, \boldsymbol{D}) = \int_{\mathbb{P}_G} |\boldsymbol{\Omega}|^{(d-2)/2} \exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}\boldsymbol{D}\right)\right\} \text{d}\boldsymbol{\Omega}. \tag{1.5}$$

The main challenge associated with density in Equation (1.4) arises from the normalizing constant in Equation (1.5); although an analytic form does exist (Uhler et al., 2018), the expression is mathematically complex and very difficult to compute in practice. The sampling strategy described in Section 1.3 is then designed to avoid the calculation of such intractable normalizing constant.

The last ingredient to fully specify the model is the prior distribution on the graph G. Several alternatives have been proposed in the literature, see among others Dobra et al. (2004); Jones et al. (2005); Scott and Carvalho (2008); Paci and Consonni (2020). A common choice is to consider a discrete uniform distribution over the space of all possible graphs $\mathcal{G}$, that is

$$\pi(G) = \frac{1}{|\mathcal{G}|}, \text{ for each } G \in \mathcal{G}. \tag{1.6}$$

Alternatively, it is possible to induce a prior on the graph by assigning to each possible edge $f \in E$ an independent Bernoulli prior with parameter $\theta_f \in (0, 1)$, that yields

$$\pi(G) \propto \prod_{f \in E} \theta_f \prod_{f \in \bar{E}} (1 - \theta_f), \tag{1.7}$$

where $\bar{E}$ is the complement of E. In a general setting, the Bernoulli parameters θ_f can be different from edge to edge. Conversely, when $\theta_f = \theta$, $\forall f \in E$, we get $\pi(G) \propto \theta^{|E|}(1-\theta)^{\binom{p}{2}-|E|}$, where $|E|$ is the cardinality of E. Notice that, when $\theta_f = 0.5$, $\forall f \in E$, the prior Equation (1.7) collapses to the uniform prior in Equation (1.6). In this work, we also consider θ to be random with a conjugate prior, i.e., $\theta \sim \text{Beta}(\alpha_1, \alpha_2)$. Hence, integrating out θ, we obtain a *multiplicity-correction prior* for G (Scott and Carvalho, 2008), given by

$$\pi(G) \propto \binom{p(p-1)/2}{2} \Gamma\left(\alpha_1 + |E|\right) \Gamma\left(\alpha_2 + p(p-1)/2 - |E|\right), \tag{1.8}$$

where $p(p-1)/2$ is the maximum number of links in an undirected graph having p nodes.

For the remaining parameters, we choose independent semi-conjugate vague priors. Summing up, our Bayesian hierarchical model is defined as follows:

$$\begin{aligned}
\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i,\ \tau_\varepsilon^2 &\overset{\text{ind}}{\sim} \text{N}_r\left(\boldsymbol{\Phi}\boldsymbol{\beta}_i, \tau_\varepsilon^2 \mathbf{I}_r\right), \\
\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n \mid \boldsymbol{\mu},\ \boldsymbol{\Omega} &\overset{\text{iid}}{\sim} \text{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}\right), \\
\boldsymbol{\Omega} \mid G &\sim \text{G-Wishart}\left(d, \boldsymbol{D}\right), \\
G &\sim \pi\left(G\right), \\
\boldsymbol{\mu} &\sim \text{N}_p\left(\mathbf{0}, \sigma_\mu^2 \mathbf{I}_p\right) \text{ and} \\
\tau_\varepsilon^2 &\sim \text{Inv-gamma}\left(a, b\right),
\end{aligned} \tag{1.9}$$

where Inv-gamma(a, b) denotes the Inverse Gamma distribution with shape parameter a and rate parameter b. The Bayesian model in Equation (1.9) enables joint modeling of all the curves with borrowing strength of information across functional data as well as to fully account for uncertainty over the graph.

1.3 Sampling strategy

To deliver Bayesian posterior inference we need to approximate the joint posterior distribution of the $\boldsymbol{\beta}$'s, τ^2_ε, $\boldsymbol{\mu}$, $\boldsymbol{\Omega}$ and G. Here, one of the main challenges arises from the graph G that belongs to a huge discrete space endowed with a complex topology. Indeed, there are $2^{p(p-1)/2}$ possible graphs representing the conditional independence structure of a p-dimensional vector of coefficients. For instance, with $p = 7$ B-spline basis functions only, we end up with more than 2 million graphs. As a consequence, any simulation algorithm struggles to explore the space of all possible graphs.

A customary approach to explore such space is by means of a Markov Chain Monte Carlo (MCMC) algorithm. In this framework, a reversible Markov chain is built such that its limiting distribution is the joint posterior of the graph and the precision matrix. Usually, the algorithm proceeds by proposing at each step a new graph which differs from the current one for only one edge, i.e., local moves. This procedure is called add-delete Metropolis Hastings, see among others Giudici and Castelo (2003); Bhadra and Mallick (2013). Then, the Markov chain moves over the graph space by comparing different graphs, in order to find the one having the highest marginal posterior. However, this procedure has some drawbacks. For instance, Jones et al. (2005) showed that the add-delete Metropolis Hastings method suffers from lack of convergence in high-dimensional settings.

As an alternative, Mohammadi and Wit (2015) proposed a birth and death MCMC algorithm (BDMCMC) based on a continuous time Markov process where jumps between birth and death events are taken to be random variables with specific rates. The authors showed that the BDMCMC outperforms alternative Bayesian approaches in terms of convergence and mixing in the graph space as well as computing time.

We build a sampling strategy that blends Gibbs-sampler steps for updating the smoothing components of the Bayesian model with a BDMCMC step for updating the precision matrix $\boldsymbol{\Omega}$ and the graph G, as described in Algorithm 1. As concerns the Gibbs-sampler steps, we sample the $\boldsymbol{\beta}$ coefficients from their full conditional distribution given by

$$\boldsymbol{\beta}_i \mid \text{rest} \sim \mathrm{N}_p\left(\boldsymbol{b}_{n_i}, \boldsymbol{B}_n\right), \tag{1.10}$$

where $\boldsymbol{B}_n = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}/\tau^2_\varepsilon + \boldsymbol{\Omega}\right)^{-1}$ and $\boldsymbol{b}_{n_i} = \boldsymbol{B}_n\left(\boldsymbol{\Phi}^T\boldsymbol{Y}_i/\tau^2_\varepsilon + \boldsymbol{\Omega}\boldsymbol{\mu}\right)$ for all $i = 1 \dots n$. The full conditional distribution of $\boldsymbol{\mu}$ is given by

$$\boldsymbol{\mu} \mid \text{rest} \sim \mathrm{N}_p\left(\boldsymbol{m}, \boldsymbol{M}\right), \tag{1.11}$$

where $\boldsymbol{M} = \left(\boldsymbol{I}_p/\sigma^2_\mu + n\boldsymbol{\Omega}\right)^{-1}$ and $\boldsymbol{m} = \boldsymbol{M}\boldsymbol{\Omega}\sum_{i=1}^n \boldsymbol{\beta}_i$. For the τ^2_ε, we sample from

$$\tau^2_\varepsilon \mid \text{rest} \sim \mathrm{IG}\left(a + \frac{nr}{2}, \frac{1}{2}\left(2b + \sum_{i=1}^{n}\left(\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{\beta}_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{\beta}_i\right)\right)\right). \tag{1.12}$$

Finally, since only $\boldsymbol{\beta}$'s are directly influenced by the graphical component, we can rewrite the joint full

conditional of $(G, \mathbf{\Omega})$ as

$$\mathbb{P}(G, \mathbf{\Omega} | \boldsymbol{Y}_1, ..., \boldsymbol{Y}_n, \tau^2_\varepsilon, \boldsymbol{\mu}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_n) = \mathbb{P}\left(G, \mathbf{\Omega} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_n, \boldsymbol{\mu}\right). \tag{1.13}$$

As introduced in Section 1.2, even when the graph is known, sampling from a G-Wishart distribution poses computational issues. As a consequence, any method based on the computation of the marginal full conditional of the graph,

$$\mathbb{P}(G \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_n) \propto \pi(G) \frac{I_G\left(d + n, D + U\right)}{I_G\left(d, D\right)}, \tag{1.14}$$

where $U = \sum_{i=1}^{n} (\boldsymbol{\beta}_i - \boldsymbol{\mu})(\boldsymbol{\beta}_i - \boldsymbol{\mu})^T$, is not efficient because it requires the computation of both the prior and posterior normalizing constant of the density in Equation (1.4).

To sample from Equation (1.13) while avoiding computation of Equation (1.14), we employ the BDMCMC approach developed by Mohammadi and Wit (2015) for Gaussian graphical model selection. The BDMCMC algorithm is based on a continuous time Markov process, where edges are added or removed via birth or death events and the time between jumps is taken to be an exponential random variable with specific rates. Birth and death events are characterized as follows:

- Death: Each edge $e = (j, k)$ between nodes j and k, where $j < k$, dies independently of the others as a Poisson process with rate $\delta_e(\mathbf{\Omega})$. The overall death rate is then $\delta(\mathbf{\Omega}) = \sum_{e \in E} \delta_e(\mathbf{\Omega})$. When the death of an edge e occurs, then the process jumps to a new state $(G^{-e}, \mathbf{\Omega}^{-e})$, where $G^{-e} = (V, E \backslash \{e\})$ and $\mathbf{\Omega}^{-e} \in P_{G^{-e}}$.

- Birth: A new edge $e = (j, k)$ between nodes j and k, where $j < k$, is born independently of the others as a Poisson process with birth rate $\lambda_e(\mathbf{\Omega})$. The overall birth rate is then $\lambda(\mathbf{\Omega}) = \sum_{e \in \bar{E}} \lambda_e(\mathbf{\Omega})$. If the birth of an edge $e \in \bar{E}$ occurs, then the process jumps to a new state $(G^{+e}, \mathbf{\Omega}^{+e})$, where $G^{+e} = (V, E \cup \{e\})$ and $\mathbf{\Omega}^{+e} \in P_{G^{+e}}$.

Birth and death processes are assumed to be independent. Hence, the intensity of the jump process is given by $w(\mathbf{\Omega}) = 1/(\lambda(\mathbf{\Omega}) + \delta(\mathbf{\Omega}))$. We recall that this implies that the time between two consecutive jumps is exponentially distributed with mean $w(\mathbf{\Omega})$ that we refer to as the weight of the state. The choice of birth and death rates is made such that the balance conditions given by Preston (1977) hold, so that the stationary distribution of the birth-death process is the full conditional distribution in Equation (1.13). The last step is then updating the precision matrix $\mathbf{\Omega}$ using the exact sampler described in Mohammadi and Wit (2015) and originally developed by Lenkoski (2013). The BDMCMC for customary Gaussian graphical model selection is implemented by the R package `BDgraph` (Mohammadi and Wit, 2019).

To sum up, our sampling strategy is given in Algorithm 1. As far as Step 1 is concerned, it consists in sampling from conjugate full conditionals described above. Then we exploit the `BDgraph` function to complete Step 2. We call the function for only one iteration using as argument the parameters updated at Step 1. In other words, we run the `BDgraph` function as if data were $\boldsymbol{\beta_1} - \boldsymbol{\mu}, ..., \boldsymbol{\beta_n} - \boldsymbol{\mu}$. We also create an auxiliary structure to store all the explored graphs and its associated expected holding time $w(\mathbf{\Omega})$. The R code implementing Algorithm 1 is available at the link `https://github.com/TeoGiane/FGM`.

Algorithm 1: MCMC sampler.

For each iteration:

Step 1. Sample the regression parameters

1.1 Sample the parameters β's from their full conditional in Equation (1.10).

1.2 Sample the parameter μ from its full conditional in Equation (1.11).

1.3 Sample the parameter τ_ε^2 from its full conditional in Equation (1.12).

Step 2. Sample $(G, \mathbf{\Omega})$ from their full conditional using the BDMCMC algorithm and save the expected holding time $w(\mathbf{\Omega})$.

1.3.1 Posterior graph

Given a MCMC output of size T, posterior summaries of the precision matrix and the graph need to take into account the fact that Step 2 of Algorithm 1 involves sampling from a continuous time Markov process. In particular, the posterior estimate of $\mathbf{\Omega}$ is based on the Rao-Blackwell theorem (Cappé et al., 2003), yielding a posterior estimate that is

$$\hat{\mathbf{\Omega}} = \frac{\sum_{t=1}^{T} w_t(\mathbf{\Omega})\mathbf{\Omega}_t}{\sum_{t=1}^{T} w_t(\mathbf{\Omega})}, \tag{1.15}$$

where $\mathbf{\Omega}_t$ is the t iteration of the MCMC sample of the precision matrix.

As far as the posterior graph is concerned, one approach for selecting the graph is to use the maximum a posteriori estimate, i.e., the highest posterior probability graph. However, this approach is not generally reliable since the space of possible graphs is quite large and any particular graph may be encountered only a few times in the course of the MCMC sampling. A more practical solution is instead to estimate the marginal posterior edge inclusion probabilities. Namely, given the MCMC output, the posterior inclusion probabilities are estimated as

$$\hat{p}_{jk} = \frac{\sum_{t=1}^{T} \mathbf{1}\left((j,k) \in E_t\right) w(\mathbf{\Omega}_t)}{\sum_{t=1}^{T} w(\mathbf{\Omega}_t)}, \tag{1.16}$$

where $\mathbf{1}\left((j,k) \in E_t\right)$ is the indicator function representing the inclusion of the edge linking nodes j and k and drawn at iteration t. Here, we select the graph containing all edges whose posterior inclusion probability in Equation (1.16) exceeds a given threshold s. In particular, we compare two different thresholds. The first one is $s = 0.5$, in analogy with the median probability model of Barbieri and Berger (2004), originally proposed in the linear regression setting. The second is based on the Bayesian False Discovery rate (BFDR; Müller et al. 2007; Peterson et al. 2015)

$$\text{BFDR} = \frac{\sum_{j<k}(1-\hat{p}_{jk})\mathbf{1}\left(\hat{p}_{jk} \geq s\right)}{\sum_{j<k}\mathbf{1}\left(\hat{p}_{jk} \geq s\right)}, \tag{1.17}$$

where s is selected so that BFDR in Equation (1.17) is below 0.05.

1.4 Analysis of fruit purees data

Food safety and authenticity are extremely important to guarantee access to good-quality and healthy products. According to the Food Authenticity Assurance Organisation, "food authenticity is the process of irrefutably proving that a food or food ingredient is in its original, genuine, verifiable and intended form as declared and represented". Although authenticity is not a novel concept, it has recently received particular attention due to the increasing number of food frauds and adulterations. As an example, the UK National Food Crime Unit has increased from 796 cases in 2015 to 1, 193 in 2019, with 364 notices in the first three months of 2019[1].

Traditionally, all strategies and techniques developed to detect food adulterants and, more in general, to study the composition of substances, were based on chemistry. The amount of marker compounds in a test material was determined and then compared with the values obtained for previously documented authentic material. The resulting test is often time consuming and expensive. Moreover, this approach is prone to errors, since food adulterers are becoming increasingly complex to detect. As a consequence, the pressing demand for rapid and inexpensive tests leads to the adoption of novel techniques. Nowadays, strategies based on mid-infrared spectroscopy coupled with chemometrics or statistical techniques have been used to investigate the substance composition and detect, for instance, the presence of adulterants in coffee (Downey et al., 1997), fruit purees (Kemsley et al., 1996; Holland et al., 1998) or minced beef (Meza-Márquez et al., 2010).

In this Chapter, we analyze the spectrum of absorbance of 351 fruit purees prepared using exclusively fresh whole strawberries (without the addition of adulterants) measured on an equally-spaced grid of 235 wavelengths and then normalized with respect to the area under the curve; see Figure 1.3. Data are publicly available at `https://data.mendeley.com/datasets/frrv2yd9rg`. The shape of the

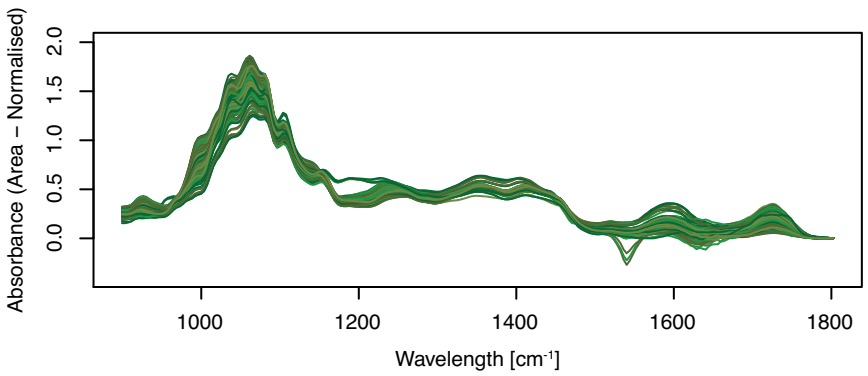

Figure 1.3: Plot of 351 spectra of absorbance of pure fruit purees, measured at 235 different wavelengths of the middle-infrared spectra.

spectra is very similar for all the curves: they are all defined by a recognizable peak at wavelengths 1000-1200 nm^{-1} and a few secondary others, like the ones around 1600 nm^{-1} and 1700 nm^{-1}. A further description of the data can be found in Holland et al. (1998). The dataset has recently been analyzed by Waghmare and Panaretos (2024)

The chemical analysis of such a complex spectrum is not trivial. Indeed, we are not dealing with a pure substance that would absorb at a specific wavelength, but those purees have a more heterogeneous

[1] National Food Crime Unit of the Food Standards Agency.

composition, which leads to overlapping effects. The goal of the analysis is to provide useful insights about which wavelength bands are related to the different components of the purees, and more specifically, about the conditional dependence structure between wavelength bands.

We fit the model in Equation (1.9) by assuming $p = 40$ cubic B-spline basis functions. Different values of p were explored, obtaining similar results with respect to the ones described in this section, so they are not reported here for brevity. The final choice of $p = 40$ faced the trade off between good fitting of the smoothed curves and limited computational burden. As a prior for the graph, we consider the same multiplicity correction prior. We mention also that, in a different experiment (not discussed here), we have adopted a uniform prior for the graph. We obtained results completely in line with the ones we are going to present.

Figure 1.4 presents an example of two smoothed curves (dotted lines) compared to the original ones (solid lines). Notice that the smoothed curves follow the shape of the original ones, smoothing away some pointwise variability. The left panel of Figure 1.5 shows the posterior mean of the coefficients

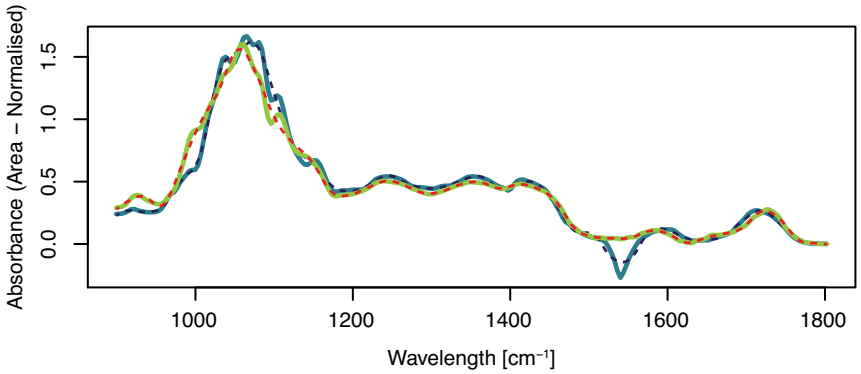

Figure 1.4: An example of two smoothed curves through the model in Equation (1.9); solid lines represent original curves while dotted lines are the smoothed curves.

while the right panel displays the estimated precision matrix $\mathbf{\Omega}$.

The top panel of Figure 1.6 displays the posterior estimate of the graph obtained using the BFDR criterion. Specifically, the top-right panel shows the estimated graph with nodes numbered and colored according to different portions of the spectra as highlighted in the bottom panel of the figure. The different colors are associated with different portions of the domain that correspond to the peaks of the signal. The graph offers an interpretation of the dependence structure among the wavelength bands and so the components of the purees. The estimated graph is quite sparse and characterized by a block structure. The blocks are located mainly close to the diagonal, with some extra-diagonal exceptions. The prevalence of blocks close to the diagonal is expected, since close wavelength bands are likely to be associated with similar components. The blocks far away from the diagonal are of interest, since they suggest that some components associated with very different wavelengths might be related.

The large block in the range 899.3 - 1493.6 nm^{-1} corresponds to the main peak of the curves. The big size of this block suggests that the interactions between the corresponding bands are not only due to the overlap of spline supports, rather they may also depend on the absorbance effects of several species overlay. We also notice two small off-diagonal blocks that correspond to the interactions between the bands at the highest peak and the others corresponding to lower peaks at around 1564.9 nm^{-1} and 1700 nm^{-1}, respectively. Indeed, with the infrared spectrometer, the same species can exhibit peaks at different

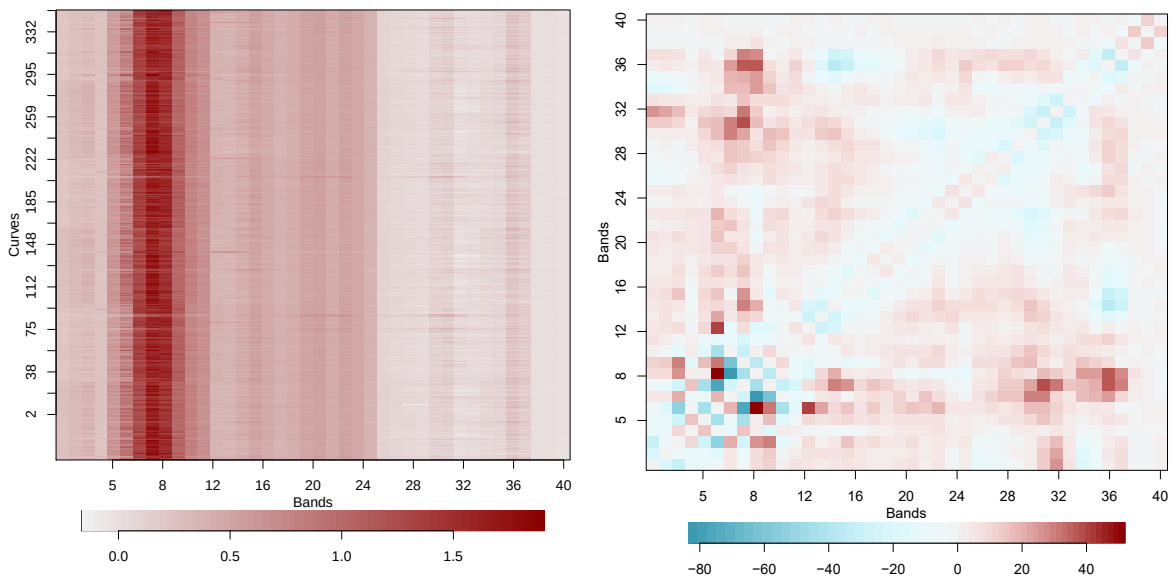

Figure 1.5: Posterior mean of all $\boldsymbol{\beta}$ coefficients (left panel) and the estimated precision matrix (right panel); in the latter plot, the main diagonal has been omitted.

wavelengths, and our analysis identifies which peaks should be analyzed together. Finally, looking at the graph representation, we note that the isolated nodes in the graph are associated with intervals where the curves are rather flat (e.g., nodes 17-18-19), which means that the concentration of the corresponding substances is negligible. On the other hand, most of the identified links connect different peaks. In particular, the main peak, which represents the most present substance, is the center of the network; all other peaks are connected to this one.

Interestingly, the block structure of the graph in Figure 1.6 was not explicitly built into the model. Nonetheless, it naturally emerges, as the observed connections tend to occur between groups of nodes, reflecting the fact that each peak is composed of more than a single band. In the following chapters, we develop two new methodologies aimed at explicitly incorporating the learning of block structures in the latent, unobserved graph. In Chapter 2, we condition on a given partition of the nodes and focus on learning connections between entire blocks of edges rather than individual edges. In Chapter 3, we go a step further and infer the latent partition of the nodes, rather than assuming it to be known from empirical evidence.

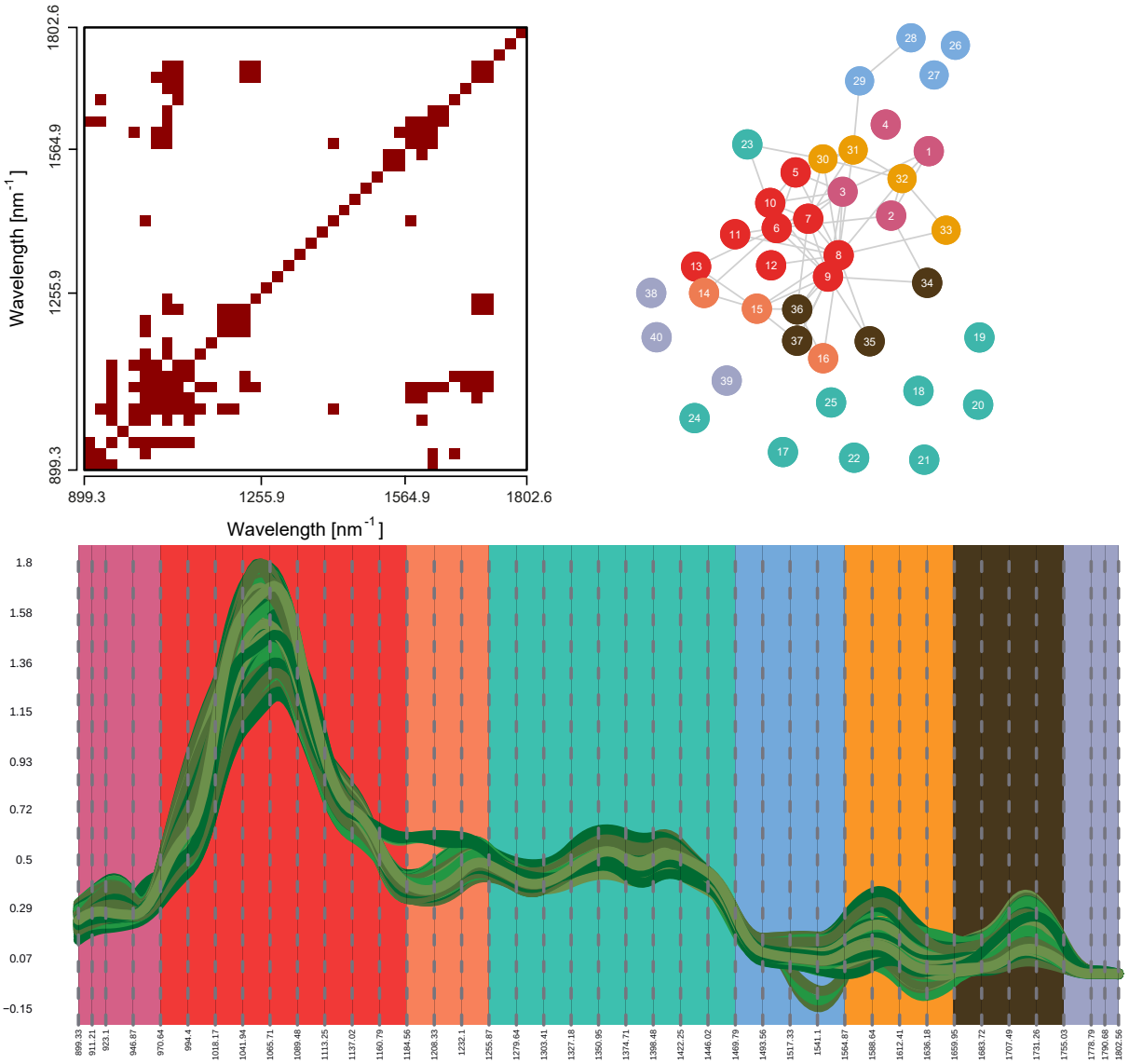

Figure 1.6: Top panels: posterior estimate of the graph obtained using the BFDR criterion. Nodes are colored according to different portions of the spectra highlighted in the bottom panel.

Chapter 2

Learning Block Structured Graphs in Gaussian Graphical Models

This chapter is based on Colombi et al. (2024b) and it is a joint work with Raffaele Argiento, Lucia Paci and Alessia Pini.

2.1 Introduction

The graphical modeling framework introduced in Chapter 1 provides a powerful tool for studying the dependence structure among variables. This approach relies on the concept of conditional independence, expressed through a correspondence between a graph and a family of multivariate probability models. As in the previous chapter, we work under the assumption of normality, leading to Gaussian graphical models (Lauritzen, 1996). Specifically, let $G = (V, E)$ be an undirected graph, where $V = \{1, \dots, p\}$ denotes the set of p nodes and E the set of undirected edges. Each node represents one of the variables of interest, while the edges encode the pattern of nonzero entries in the precision matrix. The absence of an edge between two vertices corresponds to conditional independence of the associated variables given all the others, which in turn implies that the relevant entries of the precision matrix are zero.

Usually, the structure of the underlying graph is unknown and needs to be estimated based on the available data: this is referred to as (graph) structural learning. In a Bayesian framework, the specification of a prior on the graph space and, conditionally on the graph, a prior on the precision matrix is required, see Section 1.2.2. A common practice is to choose a discrete uniform distribution over the graph space $\mathcal{G}$, i.e., the space of all possible undirected graphs with p nodes. This is appealing for its simplicity, but it is not a convenient choice to encourage sparsity, as it assigns most of the probability mass to graphs with a "medium" number of edges (Jones et al., 2005). One way to mitigate this drawback is to first assign a prior distribution to the number of edges $|E|$, for example a truncated Poisson distribution, and then condition on $|E|$ by placing a uniform prior over all graphs with that fixed number of edges. A second alternative, which is the one we pursue in this thesis, consists in inducing the prior over $\mathcal{G}$ by assuming independent Bernoulli(θ_e) priors for each edge. The Bernoulli parameters θ_e may differ from edge to edge, but usually, a common value θ is assigned. For example, Jones et al. (2005) suggested setting $\theta = 2/(p-1)$ to encourage sparsity in the graph; Scott and Carvalho (2008) placed instead a Beta hyperprior on θ, a solution known as the multiplicity correction prior. Similarly, Scutari (2013)

described a multivariate Bernoulli distribution where edges are not necessarily independent.

A common feature of the graph priors mentioned above is that they are non-informative since the only type of prior belief they elicit in the model is the expected sparsity rate of the graph. In this Chapter, instead, we develop a prior on the graph space that aims to be informative, according to prior information available for the application at hand. Since the graph describes the conditional dependence structure of variables involved in complex and, usually, high-dimensional phenomena, it is unrealistic to assume that prior knowledge is available for any one-to-one relationships between the observed quantities. Rather, it is more reasonable to envision that variables are grouped into smaller subsets. This is common in biological applications where the groups may be families of bacteria (Osborne et al., 2021), or genomics, where groups of genes are known to be part of a common process (Yook et al., 2004). Also in market basket analysis, products and customers can be easily grouped (Giudici and Castelo, 2003).

Additionally, in some applications, the variables of interest may have a natural ordering, leading to groups of nodes that are contiguous with respect to such ordering. In this case, if prior information is available about a possible partition of the nodes, the modeling assumptions must reflect that the nodes can not be relabeled. In the spectrometric data analysis setting we introduced in Chapter 1, the goal is usually to investigate relationships among the substances within a compound by observing its spectrum, which can be represented as a continuous function of the wavelength. To this end, a nonparametric regression coupled with a Gaussian graphical model on basis expansion coefficients can be employed for smoothing the data, providing an estimate of their conditional independence structure. Since the elements of a B-spline basis have compact support, the conditional independence structure of the smoothing coefficients is reflected in portions of the spectrum that are known a priori to be grouped in intervals of chemical interest. Here, the nodes representing the spline coefficients are naturally ordered, and the fixed groups of nodes turn out to be contiguous. Thus, a prior on the graph should elicit such relevant features.

In this Chapter, we propose a class of prior distributions on the graph space that leverages information on the groups of nodes and encodes a block structure in the adjacency matrix associated with the graph. Our approach consists of mapping the graph G into a block structured multigraph representation G_B, where blocks of edges are represented by a single edge between two groups of nodes. Then, independent Bernoulli priors are assumed for the edges of the multigraph G_B. In other words, we allow nodes in different groups to be only fully connected or not connected at all. As a result, posterior learning aims at discovering the underlying pattern between groups of nodes, based on the available data. We call this novel class of priors *block graph priors*.

Bayesian posterior inference on the graph is usually performed through MCMC under the conjugate G-Wishart prior distribution on the precision matrix (Roverato, 2002; Atay-Kayis and Massam, 2005). However, posterior computation is expensive for general non-decomposable graphs for two main reasons. Firstly, note that the cardinality of the graph space is $|\mathcal{G}| = 2^{\binom{p}{2}}$, i.e., it is large even if a moderate number of variables p is included. In practice, it can not be explicitly enumerated, but one needs to rely on search algorithms to explore it and learn which edges should be included or not. However, even when the number of nodes is limited, it may be difficult to identify high posterior probability regions of the graph space.

A second challenge for developing efficient methods for structural learning is due to the presence of the G-Wishart prior distribution, which is defined conditionally on a graph G, and it is known only up to an intractable normalizing constant. Explicit formulas do exist for special cases such as complete or

decomposable graphs (Dawid and Lauritzen, 1993) which, however, are hard to justify from an applied side and increasingly restrictive as the number of nodes increases. Uhler et al. (2018) provided a recursive expression for the normalizing constant, but the procedure is computationally efficient only for some specific types of graphs. In practice, the normalizing constant is usually evaluated through numerical approximations such as the importance sampler (Roverato, 2002; Dellaportas et al., 2003), the Monte Carlo approximation (Atay-Kayis and Massam, 2005), and the Laplace approximation (Moghaddam et al., 2009; Lenkoski and Dobra, 2011). Unfortunately, these methods become unstable with an increasing number of nodes (see Jones et al. 2005 and Wang and Li 2012 for further details). Recent solutions have been proposed in the literature; for example, Wang and Li (2012) leverages on the partial analytical structure of the G-Wishart distribution while Mohammadi and Wit (2019) rely on an approximation of the ratio of two normalizing constants arising when two models are compared. Also, van den Boom et al. (2022a) used a delayed acceptance MCMC (Christen and Fox, 2005) coupled with an informed proposal distribution (Zanella, 2020) on the graph space to enable embarrassingly parallel computation. However, all these approaches are suited for comparing models whose graphs differ by a single edge, and so they are inappropriate to address block structural learning. Rather, an MCMC method that modifies more than one link at a time is needed in our setting.

In this Chapter, we introduce a Reversible Jump MCMC sampler (Giudici and Green, 1999; Dobra et al., 2011), defined over the joint space of the graph and the precision matrix, that leverages the structure induced by the block graph prior. In particular, we generalize the procedure of Lenkoski (2013) so that the algorithm modifies an entire block of edges at each step of the chain to guarantee a block structure of the adjacency matrix associated with the graph. Moreover, the Reversible Jump algorithm is coupled with the Exchange algorithm (Murray et al., 2006), consisting of a second reversible move, to avoid the calculation of the G-Wishart normalizing constant. We refer to the novel sampling method as the *Block Double Reversible Jump* (BDRJ) algorithm. As a result, the algorithm builds a Markov chain that visits only the subspace of block structured graphs, which is, in general, much smaller relative to the original graph space and allows us to infer the relationships among the fixed groups of nodes.

2.2 Block Structured Graph prior

2.2.1 From a graph to a block multigraph

Let $\boldsymbol{\beta}$ be a p-random vector distributed as a multivariate normal distribution with zero mean and precision matrix $\boldsymbol{\Omega}$, i.e., $\mathrm{N}_p\left(\mathbf{0}, \boldsymbol{\Omega}^{-1}\right)$; without loss of generality, we assume here $\boldsymbol{\beta}$ to be centered around zero. Let $G = (V, E)$ be an undirected graph, where $V = \{1, \ldots, p\}$ is the set of p nodes and $E \subset \{\, (i, j) \mid i < j,\, i, j \in V \,\}$ is the set of undirected edges between the nodes. $\boldsymbol{\beta}$ is said to be Markov with respect to G if, for any edge (i, j) that does not belong to E, the i-th and j-th variables of $\boldsymbol{\beta}$ are conditionally independent given all the others, i.e., $\beta_i \perp\!\!\!\perp \beta_j \mid \boldsymbol{\beta}_{-(ij)}$, where $\boldsymbol{\beta}_{-(ij)}$ is the random vector containing all elements in $\boldsymbol{\beta}$ except the i-th and the j-th. Under the normality assumption, the conditional independence relationship between variables has an equivalent representation in terms of the null elements of the precision matrix $\boldsymbol{\Omega}$. Specifically, each node is associated with one of the variables of interest, and edges describe the structure of the non-zero elements of the precision matrix. The absence of a link between two nodes means that the two corresponding variables are conditionally independent, given all the others, and the corresponding entry of the precision matrix is zero. Hence, the following

equivalence provides an interpretation of the graph

$$\beta_i \perp\!\!\!\perp \beta_j \,|\, \boldsymbol{\beta}_{-(ij)} \iff (i,j) \notin E \iff \boldsymbol{\Omega}_{ij} = 0,$$

where $\boldsymbol{\Omega}_{ij}$ is the entry (i,j) of the precision matrix $\boldsymbol{\Omega}$. The graph G is usually unknown and must be learned from the data. In a Bayesian framework, it is considered as a random variable having values in $\mathcal{G}$, i.e., the space of all possible undirected graphs with p nodes. The starting point for our proposed model assumes that the p observed variables are grouped in K mutually exclusive groups that are known a priori. Each group has cardinality n_i and $\sum_{i=1}^{K} n_i = p$. We admit the possibility of having some $n_i = 1$, as long as $K < p$.

In our setting, the groups are given, and the adjacency matrix of the underlying graph has to satisfy a precise block structure. To accomplish that, we define a new space of undirected graphs whose nodes represent the groups of variables and edges represent the relationships between them. Namely, let $V_B = \{B_1, \ldots, B_K\}$ be a partition of V in K groups that is available a priori. Then, we define $G_B = (V_B, E_B)$ to be an undirected graph whose nodes are the sets $B_k, k = 1, \ldots, K$ and that allows for a self-loop for node k if $n_k > 1$. Namely, the set of edges E_B is given by

$$E_B \subset \Big\{(l,m) \,|\, l,m \in V_B, \wedge\, l < m, \quad (l,l) \,|\, l \in V_B, \wedge\, n_l > 1\Big\}.$$

Graphs that have self-loops are called *multigraphs*. We denote by $\mathcal{G}_B$ the set of all possible multigraphs G_B having V_B as a set of nodes.

To clarify the relationship between $\mathcal{G}_B$ and $\mathcal{G}$, consider $G_B \in \mathcal{G}_B$ and $G \in \mathcal{G}$. By definition, the set of nodes of the multigraph G_B is obtained by grouping the nodes of the graph G. The following map defines a relationship between the two sets of edges. Let $\rho : \mathcal{G}_B \to \mathcal{G}$, such that $G_B = (V_B, E_B) \mapsto G = (V, E)$ by the following transformations:

$$\begin{aligned} &V = \{1, \ldots, p\} = \bigcup_{k=1}^{K} B_k \\ &\text{if } (h,k) \in E_B \Rightarrow\ (i,j) \in E\ \forall i \in B_h,\ \forall j \in B_k \\ &\text{if } (h,k) \notin E_B \Rightarrow\ (i,j) \notin E\ \forall i \in B_h,\ \forall j \in B_k \end{aligned} \tag{2.1}$$

A visual representation of this mapping is given in Figure 2.1. Once ρ is set we can associate each G_B in $\mathcal{G}_B$ to one and only one G in $\mathcal{G}$ since ρ is injective. We refer to G_B as the multigraph form of G.

Figure 2.1: The map from multigraph $G_B \in \mathcal{G}_B$ (left panel) to its block structured form $G \in \mathcal{B}$ (right panel).

Nevertheless, ρ is not surjective, which implies that there are graphs G that do not have a representative

in $\mathcal{G}_B$. Indeed, only those graphs whose adjacency matrix has a particular block structure can be represented in a multigraph form. The non-surjective map is the key ingredient to define a subset of $\mathcal{G}$ of block structured graphs that satisfy the modeling assumptions. Let $\mathcal{B}$ be the image of ρ, i.e., the subset of $\mathcal{G}$ containing all the graphs having p nodes and a block structure consistent with V_B. Moreover, $\rho : \mathcal{G}_B \to \mathcal{B}$ is a bijection, which means that every graph $G \in \mathcal{B}$ is associated to its representative $G_B \in \mathcal{G}_B$ via ρ^{-1}. We say that $G \in \mathcal{B}$ is the block graph representation of the multigraph $G_B \in \mathcal{G}_B$. This representation of block graphs allows us to work in a space where we can use standard tools of graphical analysis.

In a different setting, Cremaschi et al. (2022) employs a block structure of the graph that is similar to ours. In their work, the multigraph is used to describe the conditional dependence structure across Markov processes, while the larger graph is used to capture the dependence at the observation level. However, unlike our approach, the self-loops in the multigraph are assumed to be known and not learned from the data, due to the nature of their application. Hence, they end up with a larger graph which always assumes the presence of K cliques.

2.2.2 Prior on the graph space

The map described in Section 2.2.1 allows us to introduce a class of priors on $\mathcal{G}$ that encodes the knowledge about the partition of the nodes. A customary choice in the literature is to assume independent Bernoulli priors for edges in $\mathcal{E}$, the set of all possible edges. Such an assumption seems reasonable only if one may assume a priori independence between the edges. Nevertheless, to induce a block structure in the adjacency matrix of the graph, independent Bernoulli priors for the edges would not be appropriate.

The class of priors we propose is built on two assumptions: (i) the nodes have a natural ordering and their groups are known a priori; (ii) a zero mass probability must be placed on those graphs in $\mathcal{G}$ where nodes in the same groups are not fully connected or not connected at all, i.e., to all graphs that belong to $\mathcal{G}\backslash\mathcal{B}$. Based on these assumptions, we need to specify the probability of graphs that are in $\mathcal{B}$, which can be represented through a multigraph $G_B \in \mathcal{G}_B$ using Equation (2.1). Each graph in $\mathcal{G}_B$ can be thought of as an undirected graph having K nodes and possible self-loops. Therefore, it is reasonable to assume that a prior on $\mathcal{G}_B$ can be defined by assigning independent Bernoulli priors to its edges. Finally, the prior probability of each graph in $\mathcal{B}$ is set to be equal to the probability of its representative in $\mathcal{G}_B$, which can be obtained by applying the ρ^{-1} map. Namely,

$$\pi(G) \propto \begin{cases} \pi_B\left(\rho^{-1}(G)\right), & \text{if } G \in \mathcal{B}, \\ 0, & \text{if } G \in \mathcal{G}\backslash\mathcal{B}, \end{cases} \tag{2.2}$$

where $\pi_B\left(\rho^{-1}(G)\right) = \pi_B\left(G_B\right) = \theta^{|E_B|}(1-\theta)^{\binom{K}{2}-|E_B|}$ is the Bernoulli prior over the set $\mathcal{G}_B$, where each link has prior probability of inclusion θ, which is fixed a priori. We refer to the prior distribution in Equation (2.2) as *block-Bernoulli prior*. In particular, the prior is composed of two ingredients: $\pi_B(\cdot)$, that is, a prior on the space $\mathcal{G}_B$, and ρ^{-1}, which is deterministic. This implies that the same construction is valid even if $\pi_B(\cdot)$ is replaced by any other prior distribution for graphical models, such as the combination of a truncated Poisson on the number of edges and a uniform distribution on the graphs having the prescribed size. The only constraint is that it must be a probability distribution over $\mathcal{G}_B$, not over $\mathcal{G}$. We define the resulting class of priors as the *block graph priors*.

2.3 Block Double Reversible Jump algorithm

A popular choice as a prior for the precision matrix $\mathbf{\Omega}$, conditional on the graph G, is the G-Wishart distribution, introduced by Roverato (2002) to deal with non-decomposable graphs. Following Mohammadi and Wit (2015), we work with a Shape-Inverse Scale parametrization of the G-Wishart distribution, that is, we say that $\mathbf{\Omega} \mid G \sim \text{G-Wishart}(b, D)$ if its density is given by

$$\mathbb{P}\left(\mathbf{\Omega} \mid G\right) = I_G\left(b, D\right)^{-1} |\mathbf{\Omega}|^{\frac{b-2}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Omega} D\right)\right\} \mathbf{1}_{\mathbb{P}_G},$$

where tr is the trace operator, $b > 2$ is the shape parameter, D is the inverse scale matrix, which is symmetric and positive definite, and $\mathbb{P}_G$ is the space of all $p \times p$ symmetric and positive definite matrices that are Markov with respect to G. Finally,

$$I_G(b, D) = \int_{\mathbb{P}_G} |\mathbf{\Omega}|^{\frac{b-2}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Omega} D\right)\right\} \, \mathrm{d}\mathbf{\Omega},$$

is the normalizing constant. In this work, b and D are fixed hyperparameters. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$ be an i.i.d. sample of size n from a $\text{N}_p\left(\mathbf{0}, \mathbf{\Omega}^{-1}\right)$, where the precision matrix $\mathbf{\Omega}$ is Markov with respect to the graph G. Thanks to conjugacy, the full conditional distribution of $\mathbf{\Omega}$ is $\mathbf{\Omega} \mid G, \boldsymbol{\beta} \sim$ G-Wishart $(b + n, D + U)$, where $U = \boldsymbol{\beta}^T \boldsymbol{\beta}$.

The goal of Bayesian structural learning is to compute the posterior distribution

$$\mathbb{P}(G \mid \mathbf{y}) = \int_{\mathbb{P}_G} p(\mathbf{y} \mid \mathbf{\Omega}) p(\mathbf{\Omega} \mid G) \pi(G) \mathrm{d}\mathbf{\Omega} \propto \pi(G) \frac{I_G(b + n, D + U)}{I_G(b, D)},$$

which depends on the ratio of the posterior and the prior G-Wishart normalizing constants.

As anticipated in the Introduction, posterior inference with the G-Wishart distribution is challenging since the joint posterior distribution of the graph and the precision matrix is doubly intractable (Murray et al., 2006). Indeed, the normalizing constant $I_G(b, D)$ does not have a simple analytical form for general non-decomposable graphs, making the computation of a Metropolis-Hastings acceptance probability not feasible. To address this issue, Monte Carlo (Atay-Kayis and Massam, 2005) and Laplace approximations (Moghaddam et al., 2009; Lenkoski and Dobra, 2011) have been introduced. Alternatively, Mohammadi et al. (2021) proposed an approximation of the ratio $I_G(b, D)/I_{G'}(b, D)$, which is, in practice, the quantity required in the computation of the acceptance-rejection probability of a proposed graph G'. However, all the aforementioned methods are not able to exploit prior information about the block structure of the graph since they are suited to modify only one edge at each step of the MCMC algorithm. Rather, to ensure a block structure of the graph compatible with our prior beliefs, edges can not be modified at will, at least not in the space $\mathcal{B}$. To our knowledge, there are no theoretically grounded methods available in the literature to compute the G-Wishart normalizing constant ratio in our setting.

For this reason, we move a step forward and develop a Block Reversible Jump Markov Chain Monte Carlo sampler (Giudici and Green, 1999; Dobra et al., 2011) defined over the joint space of graph and precision matrix. It generalizes the procedure by Lenkoski (2013) in such a way that it modifies an entire block of edges at each step of the chain to guarantee that the visited graphs always belong to the space of block structured graphs $\mathcal{B}$. By doing so, the search is limited to the subset of graphs whose structure is consistent with V_B. Moreover, exploiting a trans-dimensional version of the Exchange algorithm (Murray

et al., 2006), our algorithm avoids any calculation of the G-Wishart normalizing constant.

We denote the current state of the chain by $\left(\mathbf{\Omega}^{[s]}, G^{[s]}\right)$, where $\mathbf{\Omega}^{[s]} \in \mathbb{P}_{G^{[s]}}$ and $G^{[s]} \in \mathcal{B}$. Since the graph is constraining the support of the precision matrix, the Reversible Jump technique is needed to handle trans-dimensional moves due to a different number of unknown entries of the precision matrix in subsequent iterations. In the first step of the algorithm, the state $(\mathbf{\Omega}', G')$ is proposed, and the graph G' is accepted or rejected. It consists of three parts: (i) a new graph G' is proposed in the neighborhood of the multigraph representative $G_B^{[s]}$ of the current graph $G^{[s]}$; (ii) a matrix $\mathbf{\Omega}'$ compatible with the constraints imposed by G' is constructed; (iii) the acceptance-rejection probability is computed by exploiting a modified Exchange algorithm (Murray et al., 2006). Note that, in part (ii), the proposed matrix $\mathbf{\Omega}'$ must be guaranteed to be a precision matrix. Differently to Giudici and Green (1999), who proposed a Reversible Jump sampler that limits itself to visit decomposable graphs and requires checking $\mathbf{\Omega}'$ positive definiteness, we follow Dobra et al. (2011) and Lenkoski (2013) and adopt a reparametrization based on the Cholesky decomposition, so that $\mathbf{\Omega}'$ is positive definite by construction. The algorithm requires a double reversible move, leading to a Double Reversible Jump sampling strategy. In the following, each of these three parts is described.

(i) Proposing a new graph G'

A common feature of existing MCMC methods for graphical models is to build Markov chains such that the proposed graph $G' = (V, E')$ belongs to the one-edge-away neighborhood of G, which is defined as

$$nbd_p(G) := nbd_p^+(G) \cup nbd_p^-(G), \tag{2.3}$$

where $nbd_p^+(G)$ and $nbd_p^-(G)$ are the sets of undirected graphs having p nodes that can be obtained by adding or removing an edge to $G \in \mathcal{G}$, respectively. A step in an MCMC algorithm that selects $G' \in nbd_p(G^{[s]})$ is said to be a local move. The proposed BDRJ approach is innovative because it proposes moves that modify an entire block of edges instead of just a single one. In other words, our moves are local in the space $\mathcal{G}_B$ but not in the space $\mathcal{G}$.

Our procedure leverages the definition of the map ρ and generalizes standard graphical modeling tools to the space $\mathcal{B}$. Hence, suppose $G^{[s]} \in \mathcal{B}$ and we aim to construct a new graph $G' \in \mathcal{B}$. Firstly, we map the current graph into its multigraph representative $G_B^{[s]} \in \mathcal{G}_B$, where $G_B^{[s]} = \rho^{-1}\left(G^{[s]}\right)$. Then, with probability α_G, we add a new edge or, with probability $(1 - \alpha_G)$, we remove one of its existing edges. Namely, the new multigraph representation $G'_B \in \mathcal{G}_B$ is drawn from

$$q\left(G'_B | G^{[s]}\right) = \alpha_G \,\text{Unif}\left(nbd_K^{\mathcal{B},+}\left(\rho^{-1}\left(G^{[s]}\right)\right)\right) + (1 - \alpha_G)\,\text{Unif}\left(nbd_K^{\mathcal{B},-}\left(\rho^{-1}\left(G^{[s]}\right)\right)\right), \tag{2.4}$$

where, similarly to Equation (2.3), $nbd_K^{\mathcal{B}}\left(G_B^{[s]}\right)$ is the one-edge-away neighborhood of $G_B^{[s]}$ with respect to the space of multigraphs $\mathcal{G}_B$. Finally, ρ is applied again to map the resulting multigraph back in $\mathcal{B}$ to obtain G', i.e., we set $G' = \rho\left(G'_B\right)$.

If $\alpha_G = 0.5$, Equation (2.4) gives equal probability to addition and deletion moves. To lighten the notation, we always refer to this case. The proposal distribution in Equation (2.4) is preferred over choosing uniformly in the whole neighborhood as in Madigan and York (1995). Indeed, Dobra et al. (2011) noticed that in a simple uniform edge sampling, the probability of proposing a move that adds (or

deletes) an edge is too small if the current graph has a very large (or small) number of edges. Therefore, Equation (2.4) guarantees a better mixing in the resulting Markov chain. Furthermore, Equation (2.4) reveals how the multigraph representation enables us to use standard tools of structural learning in the space $\mathcal{G}_B$ to get a non-standard proposal in the original space $\mathcal{G}$.

(ii) Constructing the precision matrix $\mathbf{\Omega}' \mid G'$

Once the graph is selected, we need to specify a method to construct a proposed precision matrix $\mathbf{\Omega}'$ that satisfies the constraints imposed by the new graph G'. In principle, the method of Wang and Li (2012), based on the partial analytical structure of the G-Wishart, appears to be an efficient choice. However, this solution strongly relies on the possibility of writing down an explicit formula of the full conditional distribution of the elements of $\mathbf{\Omega}$. Such a result, presented in Roverato (2002), can be handled in practice only if one edge of the graph is modified at each step of the algorithm. Instead, the proposal distribution presented in Section 2.3 modifies an arbitrary number of edges. Differently, we build on a generalization of the Reversible Jump mechanism of Lenkoski (2013) and exploit the Cholesky decomposition of $\mathbf{\Omega}^{[s]}$ to guarantee the positive definiteness of $\mathbf{\Omega}'$ and the zero constraints imposed by G'.

Indeed, $\mathbf{\Omega}^{[s]} \in \mathbb{P}_{G^{[s]}}$ implies that it is possible to compute its Cholesky decomposition, $\mathbf{\Omega}^{[s]} = \left(\mathbf{\Phi}^{[s]}\right)^T \mathbf{\Phi}^{[s]}$, where $\mathbf{\Phi}^{[s]}$ is an upper triangular matrix. This is appealing because the zero constraints imposed by $G^{[s]}$ on the off-diagonal elements of $\mathbf{\Omega}^{[s]}$ induce a precise structure and properties on $\mathbf{\Phi}^{[s]}$. In particular, let $\nu\left(G^{[s]}\right) = \left\{(i,j) \mid i,j \in V, i = j \vee (i,j) \in E^{[s]}\right\}$ be the set of edges belonging to $G^{[s]}$ plus the diagonal entries of its adjacency matrix. Hence, $\mathbf{\Phi}^{\nu\left(G^{[s]}\right)} = \left\{\mathbf{\Phi}_{ij} \mid i,j \in \nu\left(G^{[s]}\right)\right\}$ is said to be the set of *free elements* of $\mathbf{\Phi}^{[s]}$. The remaining entries are uniquely determined through the completion operation (Atay-Kayis and Massam, 2005, Proposition 2) as a function of the free elements. We refer to these elements as the *non-free elements*. See Roverato (2002) and Atay-Kayis and Massam (2005) for an exhaustive overview.

Suppose that the proposed graph G' is obtained from Equation (2.4) by adding the edge (l,m) to the multigraph representation of $G^{[s]}$. The set of edges that are changing in $\mathcal{G}$ is then $L = \left\{(i,j) \mid i,j \in V, i < j, (i,j) \in E', (i,j) \notin E^{[s]}\right\}$. The cardinality $|L|$ is arbitrary and, in general, greater than one. We call $V(L) = B_l \cup B_m$ the set of the vertices involved in the change. Note that $\nu(G') = \nu\left(G^{[s]}\right) \cup L$. Our solution to define the new free elements is to maintain the same value for all the ones that are not involved in the change and to set the new ones by perturbing the current, non-free elements, independently and with constant variance σ_g^2. Namely, draw $\eta_h \overset{\text{ind}}{\sim} \mathrm{N}\left(\mathbf{\Phi}_h^{[s]}, \sigma_g^2\right)$ and set $\mathbf{\Phi}'_h = \eta_h$ for each $h \in L$. Then, all non-free elements of $\mathbf{\Phi}'$ are derived through the completion operation (Atay-Kayis and Massam, 2005) and the proposed precision matrix $\mathbf{\Omega}' = (\mathbf{\Phi}')^T \mathbf{\Phi}'$ is then obtained. Note that, we are generating a random variable $\boldsymbol{\eta}$ of length $|L|$ that matches the dimension gap between $\mathbf{\Omega}^{[s]}$ and $\mathbf{\Omega}'$. In the case of dimension reduction, say $\mathbf{\Omega}^{[s]} \to \mathbf{\Omega}'$, the move is deterministic since it is defined in terms of the opposite move $\mathbf{\Omega}' \to \mathbf{\Omega}^{[s]}$, where the extra elements do not need to be sampled. Then, the acceptance probability is the reciprocal acceptance probability of the corresponding increasing move.

(iii) Computing the acceptance-rejection probability

Finally, we frame the previous mechanism in the Exchange algorithm paradigm (Murray et al., 2006) to eliminate the presence of the G-Wishart normalizing constants. Specifically, we employ the Double Reversible Jump procedure (Lenkoski, 2013), which is the trans-dimensional equivalent of the Exchange algorithm.

Let $\widetilde{\boldsymbol{W}}$ be a latent $p \times p$ symmetric and positive definite matrix that is Markov with respect to G', i.e., $\widetilde{\boldsymbol{W}} \in \mathbb{P}_{G'}$. The matrix $\widetilde{\boldsymbol{W}}$ is sampled from a G-Wishart(b, D) distribution using the exact sampler of Lenkoski (2013). The BDRJ considers switching between $\left(\boldsymbol{\Omega}^{[s]}, G^{[s]}, \widetilde{\boldsymbol{W}}, G'\right)$ to the alternative $\left(\boldsymbol{\Omega}', G', \boldsymbol{W}^0, G^{[s]}\right)$, with $\boldsymbol{W}^0 \in \mathbb{P}_{G^{[s]}}$, by performing two reversible jump moves: (i) a dimension increasing jump from $\left(\boldsymbol{\Omega}^{[s]}, G^{[s]}\right)$ to $(\boldsymbol{\Omega}', G')$ according to the posterior parameters $b + n$ and $D + U$ of the G-Wishart distribution; (ii) a dimension decreasing jump from $\left(\widetilde{\boldsymbol{W}}, G'\right)$ to $\left(\boldsymbol{W}^0, G^{[s]}\right)$ according to the prior parameters b and D of the G-Wishart distribution. Thus, the augmented target is the joint distribution $p\left(\boldsymbol{\Omega}^{[s]}, G^{[s]}, \widetilde{\boldsymbol{W}}, G' \mid \mathbf{y}\right)$ and the proposed graph G' is accepted with probability $\min(1, R^+)$, with

$$
\begin{aligned}
R^+ &= \frac{p\left(\boldsymbol{\Omega}', G', \boldsymbol{W}^0, G^{[s]} \mid \mathbf{y}\right)}{p\left(\boldsymbol{\Omega}^{[s]}, G^{[s]}, \widetilde{\boldsymbol{W}}, G' \mid \mathbf{y}\right)} \frac{q\left(\boldsymbol{\Omega}' \mid \boldsymbol{\Omega}^{[s]}\right)}{q\left(\boldsymbol{W}^0 \mid \widetilde{\boldsymbol{W}}\right)} \frac{\mathcal{J}\left(\boldsymbol{\Omega}', \boldsymbol{W}^0\right)}{\mathcal{J}\left(\boldsymbol{\Omega}^{[s]}, \widetilde{\boldsymbol{W}}\right)} \\
&= \frac{p\left(\mathbf{y} \mid \boldsymbol{\Omega}', G'\right)}{p\left(\mathbf{y} \mid \boldsymbol{\Omega}^{[s]}, G^{[s]}\right)} \frac{p\left(\boldsymbol{\Omega}' \mid G'\right)}{p\left(\boldsymbol{\Omega} \mid G^{[s]}\right)} \frac{p\left(\boldsymbol{W}^0 \mid G^{[s]}\right)}{p\left(\widetilde{\boldsymbol{W}} \mid G'\right)} \frac{q\left(G^{[s]} \mid G'\right)}{q\left(G' \mid G^{[s]}\right)} \frac{\pi(G')}{\pi\left(G^{[s]}\right)} \\
&\quad \times \frac{q\left(\boldsymbol{\Omega}' \mid \boldsymbol{\Omega}^{[s]}\right)}{q\left(\boldsymbol{W}^0 \mid \widetilde{\boldsymbol{W}}\right)} \frac{\mathcal{J}\left(\boldsymbol{\Omega}', \boldsymbol{W}^0\right)}{\mathcal{J}\left(\boldsymbol{\Omega}^{[s]}, \widetilde{\boldsymbol{W}}\right)},
\end{aligned}
\tag{2.5}
$$

where $q(\cdot, \cdot)$ denotes the density of the proposal distribution, and $\mathcal{J}(\cdot, \cdot)$ denotes the Jacobian of the transformations involved in the reversible moves, as detailed in Appendix 2.5. Note that the normalizing constant ratio in the acceptance-rejection probability in Equation (2.5) cancels out. The MCMC algorithm is then completed with a second step consisting in sampling the precision matrix $\boldsymbol{\Omega}^{[s+1]}$ from its full conditional distribution, i.e., $\boldsymbol{\Omega}^{[s+1]} \mid G^{[s]}, \mathbf{y} \sim$ G-Wishart$(b + n, D + U)$. The resulting algorithm is summarized in Algorithm 2. Appendix 2.5 reports the details on the derivation of the acceptance-rejection probability in Equation (2.5). The `R` package `BGSL`, implementing the BDRJ algorithm, is available at `github.com/alessandrocolombi/BGSL`.

Our proposed method BDRJ borrows the skeleton of the Double Reversible Jump but modifies the proposal distribution to guarantee that $G' \in \mathcal{B}$ and, as a consequence, that multiple elements of the precision matrix are updated accordingly. Since only proposal distributions and priors have been changed, we still have a valid MCMC scheme that can now infer relationships in block structured graphs.

2.3.1 Posterior inference

Ideally, we would like to approximate its posterior distribution with the relative frequency of each sampled graph. Then, one way of providing a pointwise estimate of the graph structure is to use the maximum a posteriori strategy, which represents the mode of the posterior distribution. As noticed by Jones et al. (2005), for problems with even a moderate number of nodes p, the space to be explored is so large

Algorithm 2: Block Double Reversible Jump

Suppose the chain to be in state $\left(\mathbf{\Omega}^{[s]}, G^{[s]}\right)$, with $\mathbf{\Omega}^{[s]} = \left(\mathbf{\Phi}^{[s]}\right)^T \left(\mathbf{\Phi}^{[s]}\right) \in \mathbb{P}_{G^{[s]}}$ and $G^{[s]} \in \mathcal{B}$.

For each iteration:

Step 1. Updating the graph G

1.1. Sample G'_B from $q\left(G'_B \mid G^{[s]}\right)$ given by (2.4). Set $G' = \rho^{-1}\left(G^{[s]}\right)$. Suppose an additional move is selected. Call L the set of new edges.

1.2. Draw $\widetilde{\boldsymbol{W}} \mid G' \sim$ G-Wishart(b, D) from an exact sampler (Lenkoski, 2013).

1.3. For each $h \in L$, draw $\eta_h \sim N\left(\mathbf{\Phi}_h^{[s]}, \sigma_g^2\right)$

1.4. Set $(\mathbf{\Phi}')^{\nu(G^{[s]})} = \left(\mathbf{\Phi}^{[s]}\right)^{\nu(G^{[s]})}$ and $\mathbf{\Phi}'_h = \eta_h \ \ \forall h \in L$.
Derive the remaining elements by the completion operation and define $\mathbf{\Omega}' = (\mathbf{\Phi}')^T \mathbf{\Phi}'$.

1.5. Set $(\mathbf{\Phi}^0)^{\nu(G^{[s]})} = (\widetilde{\mathbf{\Phi}})^{\nu(G^{[s]})}$. Derive the remaining elements by completion operation and define $\boldsymbol{W}^0 = \left(\mathbf{\Phi}^0\right)^T \mathbf{\Phi}^0$.

1.6. Compute $\gamma\left(\left(\mathbf{\Omega}^{[s]}, G^{[s]}\right) \to (\mathbf{\Omega}', G')\right) = \min\{1, R^+\}$ where

$$R^+ = \frac{\exp\left\{-\frac{1}{2}\left\langle \mathbf{\Omega}' - \mathbf{\Omega}^{[s]},\ D + U\right\rangle\right\}}{\exp\left\{-\frac{1}{2}\left\langle \widetilde{\boldsymbol{W}} - \boldsymbol{W}^0,\ D\right\rangle\right\}} \prod_{i \in V(L)} \left(\frac{\mathbf{\Phi}_{ii}^{[s]}}{\mathbf{\Phi}_{ii}^0}\right)^{\nu_i^{G'} - \nu_i^{G^{[s]}}} \times \exp\left\{\frac{1}{2\sigma_g^2} \sum_{h \in L}\left[\left(\mathbf{\Phi}'_h - \mathbf{\Phi}_h^{[s]}\right)^2 - \left(\mathbf{\Phi}_h^0 - \widetilde{\mathbf{\Phi}}_h\right)^2\right]\right\} \frac{\pi(G')}{\pi\left(G^{[s]}\right)}.$$

1.7. Draw $c \sim$ Unif$[0, 1]$. if $c < \gamma$ then set $G^{[s+1]} = G'$.

Step 2. Updating the precision matrix $\mathbf{\Omega}$
Draw $\mathbf{\Omega}^{[s+1]} \mid G^{[s+1]}, \mathbf{y} \sim$ G-Wishart$(b + n, D + U)$.

that the graph frequency can not be viewed as a good estimate of its posterior probability because each particular graph may be encountered only a few times in the MCMC sampling (Peterson et al., 2015). A more practical and stable solution is instead to estimate the posterior edge inclusion marginally. Let S be the size of the MCMC output, then the posterior inclusion probabilities are estimated as

$$\hat{p}_{ij} = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\left((i,j) \in E^{[s]}\right), \tag{2.6}$$

where $\mathbf{1}\left((i,j) \in E^{[s]}\right)$ is the indicator function representing the inclusion of the edge between nodes i and j in the graph $G^{[s]} = \left(V, E^{[s]}\right)$ visited during the s-th iteration. We call $\widehat{\mathbf{P}}$ the upper triangular matrix having elements $\hat{p}_{ij}$, for $i = 1, \ldots, p$ and $j = i, \ldots, p$, that are the proportion of MCMC iterations, after the burn-in, in which the edge (i, j) has been selected to be part of the graph. Since $\widehat{\mathbf{P}}$ contains the posterior probabilities of edge inclusion, the matrix represents the uncertainty of including or not including an edge in the graph.

A pointwise graphical estimate $\widehat{G}$ is carried out by selecting all edges whose posterior inclusion probability in Equation (2.6) exceeds a given threshold τ. Following the same approach introduced in Chapter 1, we exploit the Bayesian False Discovery Rate (BFDR; Müller et al. 2007; Peterson et al. 2015) criterion instead of relying on the median graph, which arbitrarily sets $\tau = 0.5$. See Section 1.3.1 for further details.

For what concerns the precision matrix, we simply average over the MCMC samples (Cremaschi et al., 2019; Wang and Li, 2012) to obtain the posterior mean $\widehat{\mathbf{\Omega}}$

$$\widehat{\mathbf{\Omega}} = \frac{1}{S} \sum_{s=1}^{S} \mathbf{\Omega}^{[s]}.$$

Note that, even if each $\mathbf{\Omega}^{[s]}$ has a block structure induced by $G^{[s]}$, $\widehat{\mathbf{\Omega}}$ does not share the same structure of the selected graph $\widehat{G}$.

2.4 Simulation study

We carry out a simulation study to evaluate the ability of our methodology to recover the structure of the generating graph. We compare our performance to the Birth and Death approach (BDgraph for short) proposed by Mohammadi and Wit (2015) for Gaussian graphical models with a standard non-informative prior on the graph space. We rely on the implementation provided by the corresponding `R` package `BDgraph` (Mohammadi and Wit, 2019). In the latter, authors derived an efficient MCMC method where moves are decided according to specific birth and death transition kernels for the edges. Every proposed move is accepted, so the chain converges very quickly. Moreover, using an approximation of the G-Wishart normalizing constants ratio approximation (Mohammadi et al., 2021) allows for speeding up calculations. In addition, we employ the proposed BDRJ algorithm assuming a partition of nodes where each node forms a single block; this boils down to the exact sampler called Double Reversible Jump (DRJ; Lenkoski 2013). Graph posterior estimates from both BDRJ, DRJ, and BDgraph approaches have been obtained by cutting the posterior probability of inclusion of each edge with the threshold chosen via the BFDR in Equation (1.17).

We consider two different simulation scenarios to compare the ability of the aforementioned methods to learn the structure of conditional dependencies. In the first experiment, we generate data using an underlying graph whose adjacency matrix has a block structure, i.e., with fixed groups of nodes. In the second experiment, we investigate the performance of the block graph prior model when the true graph has incomplete blocks and some isolated edges.

2.4.1 Performance evaluation

To assess the performance of recovering the graph structure, we compute the standardized Structural Hamming Distance (Std-SHD, Tsamardinos et al. 2006) from the underlying graph, which is, in the case of undirected graphs, equal to the number of wrongly estimated edges, standardized with respect to the number of all possible ones, i.e.,

$$\text{Std-SHD} \;=\; \frac{\text{FP} + \text{FN}}{\binom{p}{2}}, \tag{2.7}$$

where FP and FN are the number of false positives and false negatives, respectively. Following Osborne et al. (2021), we also take into consideration the F_1-score, defined as

$$\text{F}_1\text{-score} \;=\; \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{2.8}$$

where TP is the number of true positives. Both indices lie between 0 and 1: for the Std-SHD lower values are preferred (0 value stands for perfect match), while for F_1-score higher values correspond to better performance (1 value stands for perfect match). The main difference between the two indices is that Std-SHD equally weights errors due to false positives or negatives, while F_1-score places higher importance on the number of correct discoveries that are the true positives. To visualize their difference, consider the following simple example: set the true graph to have a sparsity index equal to 0.1 and consider a trivial estimator, i.e., the empty graph. The resulting Std-SHD score would equal 0.1, which seems reasonably good even if the estimated graph is not capturing any significant information. On the other hand, the F_1-score score is not deceived as it would be equal to 0. In addition to the previous two indices, we compute the sensitivity index, i.e., TP/(TP+FN) and the specificity index, i.e., TN/(TN+FP), where TN is the number of true negatives.

2.4.2 Results

Experiment 1 **- Complete Blocks**

We set $n = 500$, $p = 40$, and $K = p/2$ groups of equal size, which leads to off-diagonal blocks of size 2×2. The underlying blocked structure graphs have been randomly generated by sampling from $\pi(G \mid \theta)$ with different sparsity indices θ, uniformly distributed in the interval $[0.2, 0.6]$. Given the graph, the true precision matrix has been sampled from a G-Wishart $\left(3, \boldsymbol{I}_p\right)$. For this study, σ_g^2 was set equal to 0.5, after a little tuning phase. The MCMC sample comprises $400,000$ iterations plus $100,000$ extra iterations that were discarded as a burn-in period. Figure 2.2 shows an example of the true graph and compares the final estimates of the forecited methods; green squares mean that there is an edge between the corresponding nodes. The second, third, and fourth panels from the left display the graph estimated by BDRJ, DRJ, and by BDgraph, respectively. Clearly, the visual inspection suggests that BDRJ provides a better estimate of the underlying graph than the competitors.

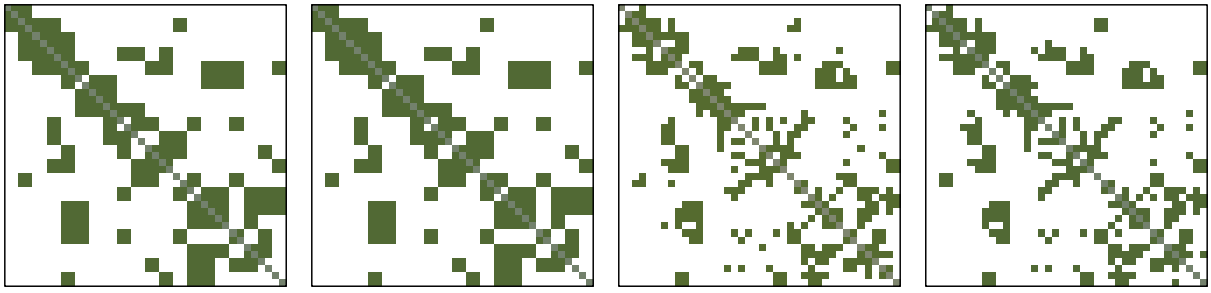

Figure 2.2: From the left: the true graph (first panel) and the estimated graph obtained using BDRJ (second panel), DRJ (third panel), and BDgraph (fourth panel) for Experiment 1. Dark squares represent the included edges.

Figure 2.3 shows the boxplot of the sensitivity index, the specificity index, the Std-SHD and the F_1-scores over 50 simulated datasets. The sensitivity and specificity indices of BDRJ are almost similar, being both centered around 0.9. This denotes a good balance between including or not including blocks of edges, without any preference between being conservative or not. Rather, the specificity of DRJ and BDgraph is close to one, i.e., a much higher value than the sensitivity index that is around 0.7. Therefore, we conclude that DRJ and BDRJ tend to be more reliable in terms of discovered conditional independence relationships.

We note that, overall, BDRJ outperforms the competitors in terms of both the Std-SHD and the F1-score. The number of misclassified edges is rather low for BDRJ, with a median value of Std-SHD equal to 0.0455. Many true discoveries are achieved, and indeed the median F_1-score under BDRJ is 0.885. BDgraph and DRJ perform worse with respect to both indices; the medians Std-SHD are equal to 0.0622 and 0.0647 while the F_1-scores are both equal to 0.807. The reason for these differences is that our approach takes advantage of the block structure of the true graph, which is not incorporated into the models used by the other methods. Instead, these methods attempt to estimate every possible link independently, leading to more errors in the final estimate and a less interpretable graph structure. It is difficult to explain why certain edges are missing within grouped structures.

Experiment 2 - Incomplete blocks

In this experiment, we analyze the performance of our model when the underlying graph has incomplete blocks. To simulate the data under this scenario, we first draw a block structured graph from $\pi(G \mid \theta)$, where $\theta = 0.2$. Then, edges are removed within each block with a probability equal to 0.25. By doing so, the block structure is incomplete but still recognizable. We set $n = 500$, $p = 30$, and $K = p/2$ groups of equal size; given the graph, the true precision matrix has been sampled from G-Wishart$(3, \boldsymbol{I}_p)$.

Figure 2.4 shows an example of the true graph (first panel from the left) and the estimated one using BDRJ (second panel), DRJ (third panel), and BDgraph (right panel). A simple visual inspection of the figure suggests that our approach tends to include incomplete blocks rather than discard them, leading to some false discoveries, as expected. On the other hand, both BDgraph and DRJ do not make any assumptions about the graph structure, so in principle, they should be able to recover the graph correctly. In practice, as soon as the dimension of the graph increases, this is unlikely to happen. Instead, the tendency is to be more conservative, which leads to fewer false discoveries.

To clarify, we report in Figure 2.5 the boxplots of the sensitivity and specificity indexes computed

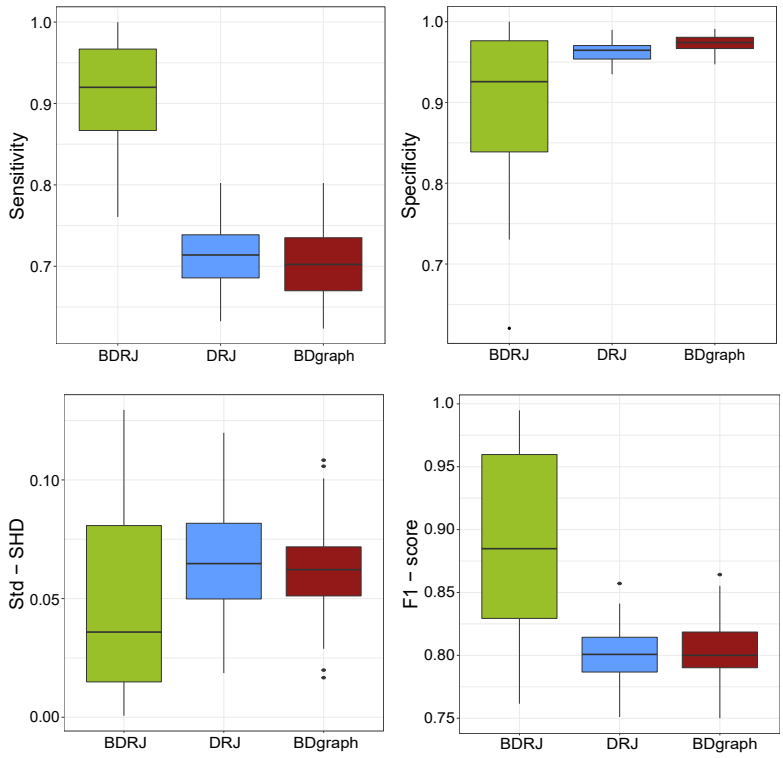

Figure 2.3: Boxplot of the sensitivity index (top-left panel), specificity index (top-right panel), Std-SHD (bottom-left panel), and F_1-score (bottom-right panel) over the simulated datasets of Experiment 1.

over 50 simulated datasets. Our approach outperforms the competitors in terms of sensitivity since it provides more true discoveries and fewer false negatives. This means that it is unlikely that a missing edge is instead present in the underlying graph. On the other hand, the BDgraph solution is preferable to BDRJ and DRJ in terms of specificity, i.e., an included edge likely represents an actual connection in the underlying graph. As expected, DRJ and BDgraph show similar performances, still with slightly different behavior. As observed also in Experiment 1, the effect of the exact sampler used in DRJ seems to be an increase of sensitivity at the price of reduced specificity. The specificity and sensitivity indices provide a clear picture of the differences between BDRJ and the two competitor approaches, which is no longer true looking at the Std-SHD and the F_1-score values in Figure 2.5. The Std-SHD is slightly higher for BDRJ, coherently with the fact that the prior information assumed in this case is misspecified, but overall the difference between the three methods is limited. The F_1-score, in particular, is very similar for the three methods, meaning that the three approaches are almost equivalent in terms of misclassified edges. In other words, when dealing with a structured graph with incomplete blocks, the expected findings from a stochastic search in the complete or block graph space are comparable, but with the latter depicting more interpretable results.

2.5 Analysis of fruit purees

We illustrate with a real-world application how our model is able to exploit prior information to enrich the data analysis and provide more interpretable results. The motivating problem is the analysis of spectrometric data of fruit purees introduced in Chapter 1.

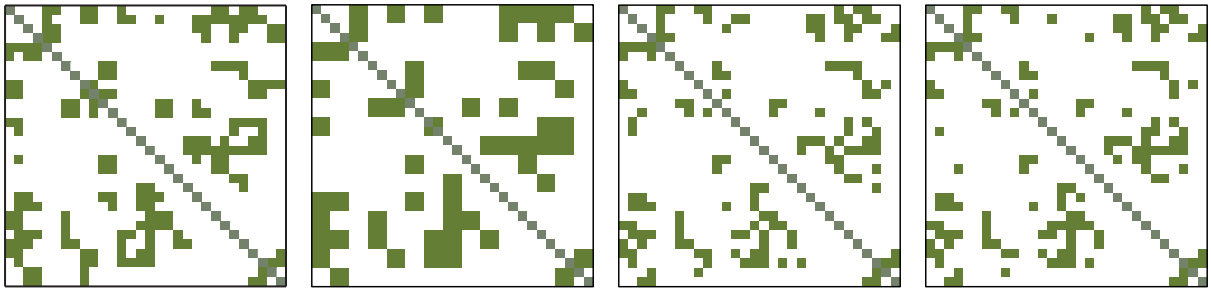

Figure 2.4: From the left: the true graph (first panel) and the estimated graph obtained using BDRJ (second panel), DRJ (third panel), and BDgraph (fourth panel) for Experiment 2. Dark squares represent the included edges.

From a mathematical point of view, a spectrum is a continuous function of the wavelength, so we framed it within the functional data analysis setting. The classical smoothing strategy was enriched by placing a Gaussian graphical model on the basis expansion coefficients, providing an estimate of their conditional independence structure. Since the elements of a B-spline basis have compact support, the conditional independence structure is reflected in well-defined portions of the domain. The Bayesian hierarchical formulation enables the borrowing of strength among different curves, and the graphical model allows the sharing of information along different subintervals of the functional datum. Finally, note that, in this application, the support of the spline basis functions coincides with the spectrum bands. Therefore, the problem of studying interactions between different substances simply translates into studying the dependencies between the basis expansion coefficients, which can be read from the graph.

In Chapter 1, we used independent Bernoulli prior distributions on the graph edges and the BDgraph method to provide posterior inference on the graph and the precision matrix. As already discussed, this is a general non-informative setting that does not allow for including further prior information even when they are available. This is one of those situations: for infrared spectrometric data, peaks of the signals are associated with the vibrational modes of the different molecules present in the substance (Atkins and De Paula, 2013). This is why the signals can be decomposed into different parts, corresponding to the peaks observed along the domain. As a matter of fact, domain experts identified nine intervals of the spectrum of chemical interest associated with the most significant peaks of the signal (Defernez et al., 1995).

The partition can be visualized in Figure 2.7, where different colors highlight the nine groups. Clearly, the nodes representing the basis expansion coefficients are ordered, so the groups are contiguous in the functional domain. The figure also shows the support of $p = 40$ spline functions.

Let $\boldsymbol{y}_t = (y_t(s_1), \ldots, y_t(s_r))^T$ be the absorbance spectrum at all observed wavelengths of curve t, with $t = 1, \ldots, T = 351$. We employ the block structured Gaussian graphical model described in Section 2.2 to smooth the functional data and accommodate prior knowledge on the subintervals of the spectrum. In particular, we assume $\pi_B(G_B \mid \theta) = \theta^{|E_B|}(1-\theta)^{\binom{K}{2}-|E_B|}$, where $\theta = 2/(K-1) = 0.25$, that is the choice suggested by Jones et al. (2005) but taking into account that the number of nodes in the multigraph space is $K = 9$, not $p = 40$. All the other prior distributions and hyperparameters are set as

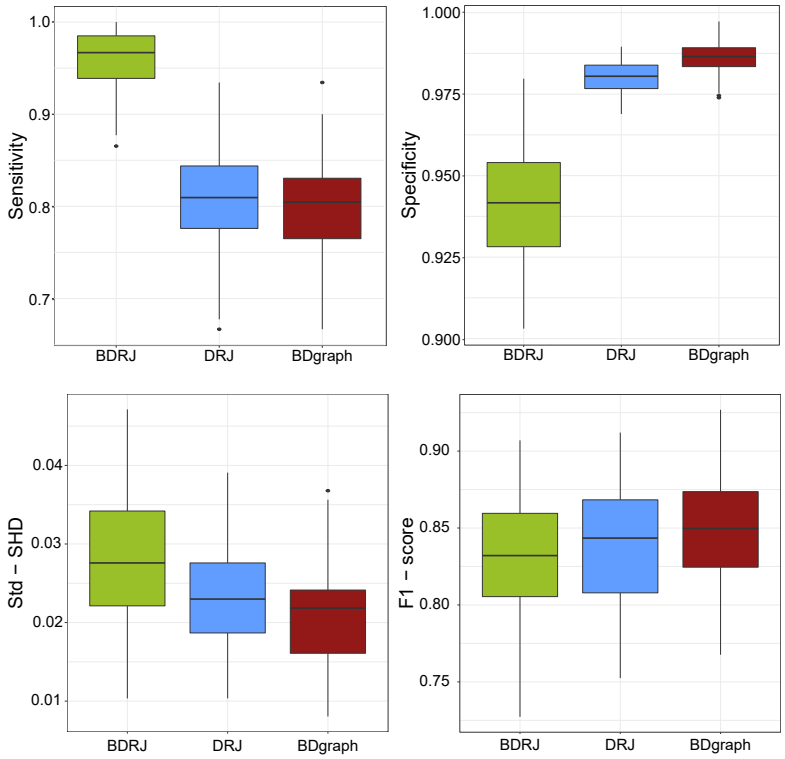

Figure 2.5: Boxplot of the sensitivity index (top-left panel), specificity index (top-right panel), Std-SHD (bottom-left panel), and F_1-score (bottom-right panel) over the simulated datasets of Experiment 2.

in Chapter 1. Summing up, the Bayesian hierarchical model is defined as follows:

$$
\begin{aligned}
\boldsymbol{y}_t \mid \boldsymbol{\beta}_t,\ \tau_\varepsilon^2 &\overset{\text{ind}}{\sim} \mathrm{N}_r\left(\boldsymbol{\Omega}\boldsymbol{\beta}_t, \tau_\varepsilon^2 \mathbf{I}_r\right), \\
\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T \mid \boldsymbol{\mu},\ \boldsymbol{\Omega} &\overset{\text{iid}}{\sim} \mathrm{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}\right), \\
\boldsymbol{\Omega} \mid G &\sim \text{G-Wishart}\left(d, \boldsymbol{D}\right), \\
G &\sim \pi\left(G\right), \\
\boldsymbol{\mu} &\sim \mathrm{N}_p\left(\mathbf{0}, \sigma^2 \mathbf{I}_p\right), \\
\tau_\varepsilon^2 &\sim \text{Inv-gamma}\left(a, b\right),
\end{aligned}
\tag{2.9}
$$

where the lj-th element of the matrix $\boldsymbol{\Omega}$ is the j-th basis function evaluated at the l-th grid point s_l, $l = 1, \ldots, r = 235$, $\mathbf{I}_r$ denotes the identity matrix of size r, $\pi(G)$ is defined in Equation (2.2) and Inv-gamma(a, b) denotes the Inverse-Gamma distribution with shape parameter a and rate parameter b. Posterior inference of the model in Equation (2.9) is obtained via the BDRJ algorithm run for $450,000$ iterations after a burn-in of $50,000$ and a thinning value of 25. After some tuning, we set $\sigma_g^2 = 1$. The algorithm runs on a laptop having an Intel(R) Core(TM) i7-1065G7 CPU 1.30GHz processor with 16GB RAM. The running time per iteration is around 0.02 seconds.

The left panel of Figure 2.7 shows the adjacency matrix estimated using the BFDR in Equation (1.17), where filled boxes represent the selected edges. The right panel displays the corresponding network; nodes are colored according to their group membership. The posterior adjacency matrix is characterized by two large diagonal blocks. The first one represents interactions within the main peak (the three red

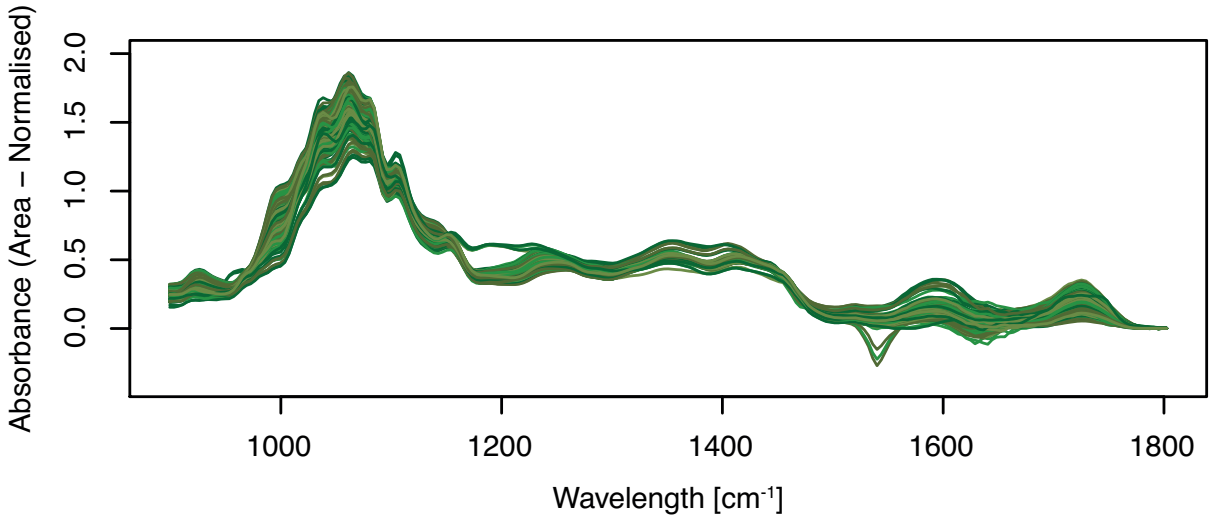

Figure 2.6: Plot of 351 spectra of absorbance of pure fruit purees, measured at 235 different wavelengths of the middle-infrared spectra.

groups), also revealing short-term interactions between the first group and the one immediately after. Similarly, the second diagonal block represents short-term interactions between the two peaks in the spectrum's tail and their connections.

However, a distinctive feature of graphical models is the possibility of investigating long-term interactions, and we indeed found some off-diagonal blocks. The most connected group is the red one, $[1018.17 - 1089.48]nm^{-1}$, which is the central part of the main peak. It has four long-term interactions, and, in particular, it is connected to both the final peaks. Another connection is the one between the first peak and the second-to-last peak (in yellow). It is the only long-term connection that does not involve the main peak or the last one. Finally, note that four nodes in the interval $[1303.41 - 1398.48]nm^{-1}$ are completely disconnected, which is probably due to the fact that curves are flat in that interval, meaning that no particular substance is absorbing in that region. Therefore, it makes sense that it is not correlated with the others.

Our estimated structure of dependencies and the one reported in Chapter 1 are similar, even though there are some differences due to the different modeling assumptions. In the previous chapter, we relied on the BDgraph method, which does not look for block structured graphs. Therefore the estimated graph is more fragmented, i.e., there are incomplete extra diagonal blocks and some isolated edges. Moreover, the graph is sparser than the one shown in Figure 2.7. This is coherent with the simulation experiments performed in Section 2.4, where we empirically showed that BDgraph is more parsimonious in including edges, which leads to a lower number of false positives but also to a lower number of discoveries with respect to BDRJ. Moreover, the approximation of the G-Wishart normalizing constants used in BDgraph can also be a cause of increased sparsity, as pointed out by the authors (Mohammadi et al., 2021). On the other hand, our method is, by construction, prone to false positives. This difference is clearly visible when analyzing the dependencies in the tail of the spectrum, $[1541.1 - 1802.56]nm^{-1}$. The groups are large, with five and seven nodes, respectively. According to BDgraph, some connections among the nodes in such groups are present, but only a few are selected. BDRJ is also able to recognize such interactions. However, it includes a large diagonal block of nodes due to the size of the involved groups. We also notice that the contrary is possible: BDgraph found some isolated edges that, according to

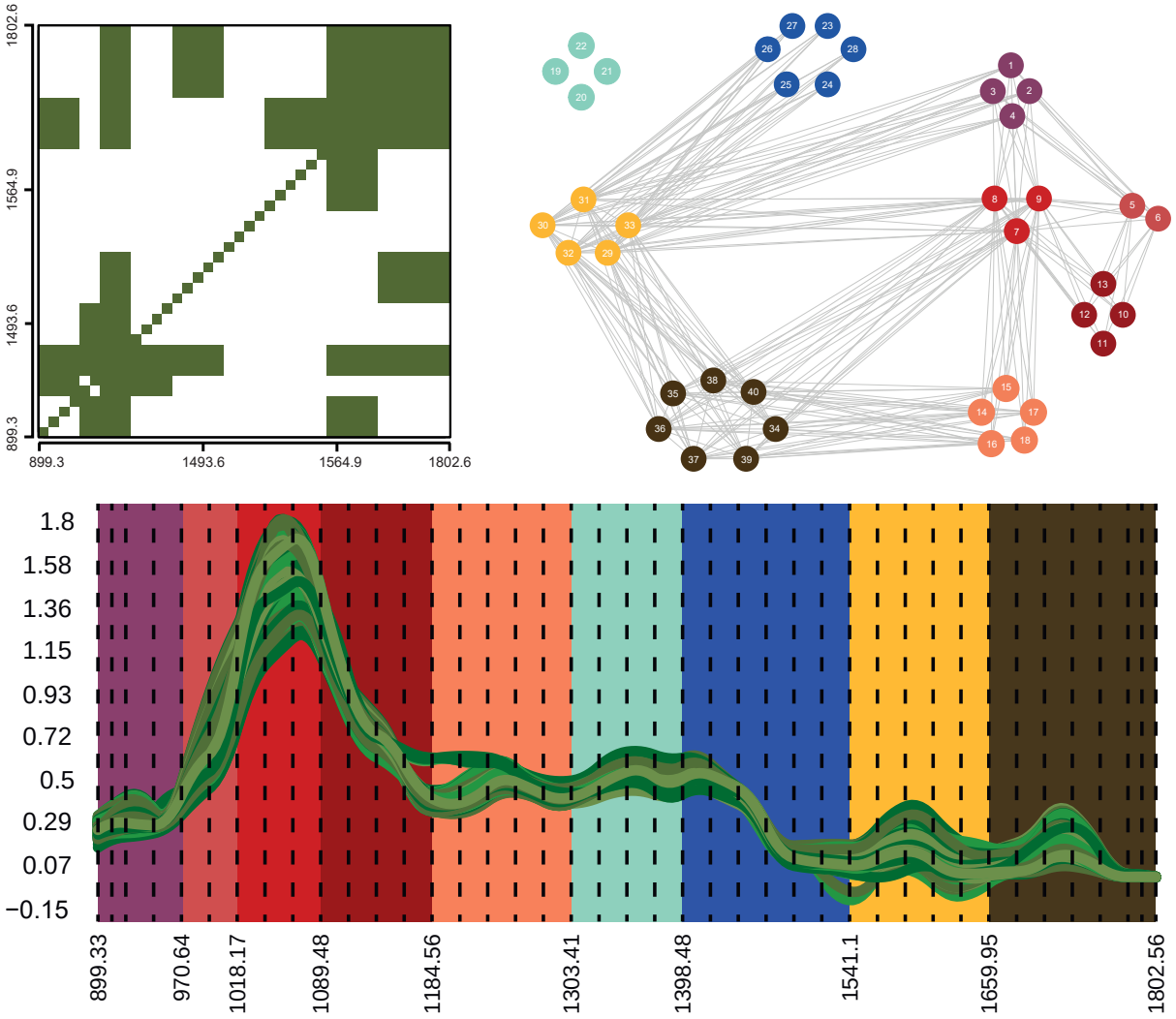

Figure 2.7: Top panels: a posterior graph estimate obtained using the BFDR criterion. Nodes are colored according to different portions of the spectra highlighted in the bottom panel, where different colors represent the nine groups.

BDRJ, are not strong enough to justify the inclusion of a whole block and therefore are filtered away.

Overall, our approach improves the interpretability of the research findings. We recall that in an absorbance spectrum, each chemical group absorbs light at a specific wavelength; therefore, the main goal is to study the interactions between the peaks of the spectrum, as suggested by the experts. As a consequence, the BDgraph solution is overly detailed since incomplete blocks do not provide information about the relationships between the peaks but only about each portion of the peaks; isolated edges are even more difficult to explain. On the other hand, our model takes into account the prior knowledge and balances the mathematical exactness and the interpretability of the results.

Appendix of Chapter 2

A.1 Deriving the acceptance-rejection probability

The BDRJ algorithm considers switching between $(\mathbf{\Omega}^{[s]}, G^{[s]}, \widetilde{W}, G')$ to the alternative $(\mathbf{\Omega}', G', \boldsymbol{W}^0, G^{[s]})$, where $\boldsymbol{W}^0 \in \mathbb{P}_{G^{[s]}}$, by performing two Reversible Jump moves: (i) a dimension increasing step from $(\mathbf{\Omega}^{[s]}, G^{[s]})$ to $(\mathbf{\Omega}', G')$ according to posterior parameters $b+n$ and $D+U$ and (ii) a dimension decreasing step from $(\widetilde{W}, G')$ to $(\boldsymbol{W}^0, G^{[s]})$ according to prior parameters b and D.

As mentioned in Section 2.3, the proposed graph G' is obtained from Equation (2.4) by adding the edge (l, m) to the multigraph representation of $G^{[s]}$. Regarding the precision matrix, the double reversible jump is performed by leveraging on the change of variables $\mathbf{\Omega} \mapsto \mathbf{\Phi}$, where $\mathbf{\Phi}$ is an upper triangular matrix such that $\mathbf{\Omega} = \mathbf{\Phi}^T\mathbf{\Phi}$, see Section 2.3. We set $\mathbf{\Phi}'_{ij} = \mathbf{\Phi}_{ij}$ for all $(i, j) \in \nu\left(G^{[s]}\right)$. The free elements are the ones in the set L that are proposed by perturbing the old values independently and with the same variance σ_g^2, which is a tuning parameter. Namely, we draw $\eta_h \overset{\text{ind}}{\sim} \mathrm{N}\left(\mathbf{\Phi}_h^{[s]}, \sigma_g^2\right)$ and set $\mathbf{\Phi}'_h = \eta_h$ for each $h \in L$. This defines all the free elements of $\mathbf{\Phi}'$, while the non-free elements are determined through the completion operation (Atay-Kayis and Massam, 2005). Hence, $\mathbf{\Phi}'$ is well defined as well as $\mathbf{\Omega}' = (\mathbf{\Phi}')^T\mathbf{\Phi}'$.

The probability to accept the proposed values of $(\mathbf{\Omega}', G')$ is equal to $\min(1, R^+)$, where

$$R^+ = \frac{p\left(\mathbf{\Omega}', G', \boldsymbol{W}^0, G^{[s]} \mid \mathbf{y}\right)}{p\left(\mathbf{\Omega}^{[s]}, G^{[s]}, \widetilde{W}, G' \mid \mathbf{y}\right)} \frac{J\left(\mathbf{\Omega}' \to \mathbf{\Phi}'\right) J\left(\boldsymbol{W}^0 \to \mathbf{\Phi}^0\right)}{J\left(\mathbf{\Omega}^{[s]} \to \mathbf{\Phi}^{[s]}\right) J\left(\widetilde{W} \to \widetilde{\mathbf{\Phi}}\right)} \frac{q\left(\mathbf{\Omega}' \mid \mathbf{\Omega}^{[s]}\right)}{q\left(\boldsymbol{W}^0 \mid \widetilde{W}\right)}, \tag{2.10}$$

where $J(A \to B)$ denotes the Jacobian of the transformation from A to B. As usual with discrete spaces, in Equation (2.10), the Jacobian needed for matching the dimensions of the compared states has been omitted since it reduces to the determinant of the identity matrix.

First, we recall that the Cholesky decomposition of the precision matrix discussed in Section 2.3 allows us to easily compute the determinant of $\mathbf{\Omega}$, see Roverato (2002), that is

$$\det(\mathbf{\Omega}) = \prod_{i=1}^{p} \mathbf{\Phi}_{ii}^2. \tag{2.11}$$

Note that this formulation involves only diagonal values of $\mathbf{\Phi}$, which are free elements by definition. Hence, Equation (2.11) implies that $\det(\mathbf{\Omega}') = \det\left(\mathbf{\Omega}^{[s]}\right)$.

The first ratio in Equation (2.10) can be factorized as follows:

$$\begin{aligned}
\frac{p\left(\boldsymbol{\Omega}', G', \boldsymbol{W}^0, G^{[s]} \mid \mathbf{y}\right)}{p\left(\boldsymbol{\Omega}^{[s]}, G^{[s]}, \widetilde{\boldsymbol{W}}, G' \mid \mathbf{y}\right)} &= \frac{p\left(\mathbf{y} \mid \boldsymbol{\Omega}', G'\right)}{p\left(\mathbf{y} \mid \boldsymbol{\Omega}^{[s]}, G^{[s]}\right)} \frac{p(\boldsymbol{\Omega}' \mid G')}{p\left(\boldsymbol{\Omega} \mid G^{[s]}\right)} \frac{q\left(G^{[s]} \mid G'\right)}{q\left(G' \mid G^{[s]}\right)} \\
&\times \frac{p\left(\boldsymbol{W}^0 \mid G^{[s]}\right)}{p\left(\widetilde{\boldsymbol{W}} \mid G'\right)} \frac{\pi(G')}{\pi\left(G^{[s]}\right)} \\
&= \frac{\sqrt{|\boldsymbol{\Omega}'|}}{\sqrt{|\boldsymbol{\Omega}|}} \exp\left\{-\frac{1}{2}\left\langle \boldsymbol{\Omega}' - \boldsymbol{\Omega}^{[s]},\, U\right\rangle\right\} \\
&\times \frac{I_{G^{[s]}}(b, D)}{I_{G'}(b, D)} \exp\left\{-\frac{1}{2}\left\langle \boldsymbol{\Omega}' - \boldsymbol{\Omega}^{[s]},\, D\right\rangle\right\} \frac{\left|nbd_K^{\mathcal{B},+}\left(G_B^{[s]}\right)\right|}{\left|nbd_K^{\mathcal{B},-}\left(G'_B\right)\right|} \\
&\times \frac{I_{G'}(b, D)}{I_{G^{[s]}}(b, D)} \frac{1}{\exp\left\{-\frac{1}{2}\left\langle \widetilde{\boldsymbol{W}} - \boldsymbol{W}^0,\, D\right\rangle\right\}} \frac{\pi(G')}{\pi(G^{[s]})} \\
&= \frac{\exp\left\{-\frac{1}{2}\left\langle \boldsymbol{\Omega}' - \boldsymbol{\Omega}^{[s]},\, D + U\right\rangle\right\}}{\exp\left\{-\frac{1}{2}\left\langle \widetilde{\boldsymbol{W}} - \boldsymbol{W}^0,\, D\right\rangle\right\}} \frac{\left|nbd_K^{\mathcal{B},+}\left(G_B^{[s]}\right)\right|}{\left|nbd_K^{\mathcal{B},-}\left(G'_B\right)\right|} \frac{\pi(G')}{\pi\left(G^{[s]}\right)},
\end{aligned} \tag{2.12}$$

where $\langle A, B\rangle$ denotes the trace of the product between A and B. Note that the two ratios of G-Wishart densities allow us to eliminate the presence of their normalizing constants. Also, note that, thanks to Equation (2.11), all the determinants of the matrices in Equation (2.12) canceled out.

For what concerns the change of variable from a precision matrix to its Cholesky decomposition, the Jacobian of such a transformation is

$$J(\boldsymbol{\Omega} \mapsto \boldsymbol{\Phi}) = 2^p \prod_{i=1}^{p} \boldsymbol{\Phi}_{ii}^{\nu_i^G}, \tag{2.13}$$

where $\nu_i^G = |\{j : j > i \text{ and } (i, j) \in E\}|$ is the sum of elements in i-th row of the adjacency matrix, from position $i + 1$ up to the end. Then, the ratio of the Jacobians appearing in Equation (2.10) is readily computed using Equation (2.13). That is,

$$\frac{J(\boldsymbol{\Omega}' \to \boldsymbol{\Phi}')}{J(\boldsymbol{\Omega}^{[s]} \to \boldsymbol{\Phi}^{[s]})} = \frac{2^p}{2^p} \frac{\prod_{i=1}^{p} \left(\boldsymbol{\Phi}'_{ii}\right)^{\nu_i^{G'}+1}}{\prod_{i=1}^{p} \left(\boldsymbol{\Phi}_{ii}^{[s]}\right)^{\nu_i^{G}+1}} = \prod_{i \in V(L)} \left(\boldsymbol{\Phi}_{ii}^{[s]}\right)^{\nu_i^{G'} - \nu_i^{G^{[s]}}}. \tag{2.14}$$

The last equality follows by noticing that $\nu_i(G) = \nu_i(G')$ for all $i \neq V(L)$ and those diagonal elements are not modified by construction. Analogously, one can show that

$$\frac{J(\boldsymbol{W}^0 \to \boldsymbol{\Phi}^0)}{J(\widetilde{\boldsymbol{W}} \to \widetilde{\boldsymbol{\Phi}})} = \prod_{i \in V(L)} \left(\boldsymbol{\Phi}_{ii}^0\right)^{\nu_i^{G'} - \nu_i^{G^{[s]}}}.$$

Under the assumption that G' is obtained by adding edge (l, m) to the multigraph representation of $G^{[s]}$, the exponent in (2.14) reduces to

$$\nu_i^{G'} - \nu_i^{G^{[s]}} = |\{j \in B_k : j > i\}|,$$

which is equal to the number of nodes in group k whose index is greater than i.

Finally, the last ratio in Equation (2.10) is due to the randomness in the construction of the proposed and the auxiliary matrices. By definition, each term is just the ratio of independent multivariate Gaussian densities, i.e.,

$$q\left(\boldsymbol{\Omega}' \mid \boldsymbol{\Omega}^{[s]}\right) = \left(\frac{1}{\sqrt{2\pi\sigma_g^2}}\right)^{|L|} \exp\left\{-\frac{1}{2\sigma_g^2}\sum_{h\in L}\left(\boldsymbol{\Phi}'_h - \boldsymbol{\Phi}_h\right)^2\right\},$$

where, for sake of clarity, we explicitly wrote the ratio in terms of $\boldsymbol{\Phi}'$. Similarly, we obtain the quantity $q\left(\boldsymbol{W}^0 \mid \widetilde{\boldsymbol{W}}\right) \propto \exp\left\{-\frac{1}{2\sigma_g^2}\sum_{h\in L}\left(\boldsymbol{\Phi}^0_h - \widetilde{\boldsymbol{\Phi}}_h\right)^2\right\}$.

Wrapping everything together, we end up with

$$R^+ = \frac{\exp\left\{-\frac{1}{2}\left\langle \boldsymbol{\Omega}' - \boldsymbol{\Omega}^{[s]},\ D + U\right\rangle\right\}}{\exp\left\{-\frac{1}{2}\langle \widetilde{\boldsymbol{W}} - \boldsymbol{W}^0,\ D\rangle\right\}} \prod_{i\in V(L)}\left(\frac{\boldsymbol{\Phi}^{[s]}_{ii}}{\boldsymbol{\Phi}^0_{ii}}\right)^{\nu_i^{G'} - \nu_i^{G^{[s]}}}$$
$$\times \exp\left\{\frac{1}{2\sigma_g^2}\sum_{h\in L}\left[\left(\boldsymbol{\Phi}'_h - \boldsymbol{\Phi}^{[s]}_h\right)^2 - \left(\boldsymbol{\Phi}^0_h - \widetilde{\boldsymbol{\Phi}}_h\right)^2\right]\right\}\frac{\pi(G')}{\pi\left(G^{[s]}\right)}.$$

Chapter 3

Product Partition Models to Learn Random Blocks in Gaussian Graphical Models

This chapter extends the paper of Colombi et al. (2025b) and is a joint work with Raffaele Argiento, Lucia Paci, and Alessia Pini.

3.1 Introduction

In this Chapter, we return to the fruit purees dataset introduced in Chapter 1 and further analyzed in Chapter 2 as well as in Waghmare and Panaretos (2024). Despite the differences in methodology, these works consistently point to comparable graph structures, characterized by strong associations between spectral bands that are not necessarily adjacent. A key feature emerging from these results is that the adjacency matrix of the graph exhibits a block structure, meaning that connections arise primarily between groups of nodes rather than within them. This behavior is illustrated in Figure 3.1, where we display the estimated graph obtained in Section 1.4 (left panel) together with its corresponding network representation (right panel). In the network plot, nodes are grouped according to the same expert-defined partition used in Chapter 2. It is worth emphasizing that this partition is not derived from the data but instead reflects expert knowledge, corresponding to the most relevant peaks in the signal. For this reason, we refer to it as the *expert-defined partition* of the nodes. Although the estimated graph is overall sparse, the network is far from being homogeneous. Indeed, connections are primarily grouped in blocks: certain clusters of nodes are closely interconnected, while others exhibit several edges across different groups. Looking at the adjacency matrix, we observe that these blocks are not simply concentrated along the diagonal with only a few sparse links elsewhere; i.e., the network does not exhibit only short-term interactions between nearby wavelets. Rather, there is a main block on the diagonal, and highly structured off-diagonal connections with dense links among some groups of nodes.

Despite the clear block structure, previous studies, including our model in Chapter 1 and Waghmare and Panaretos (2024), ignore the presence of a latent structure among the nodes in the statistical model. In Chapter 2, such a structure was considered, but it was assumed to be known and fixed according to the partition provided by experts. However, this partition is not data-driven: it reflects domain

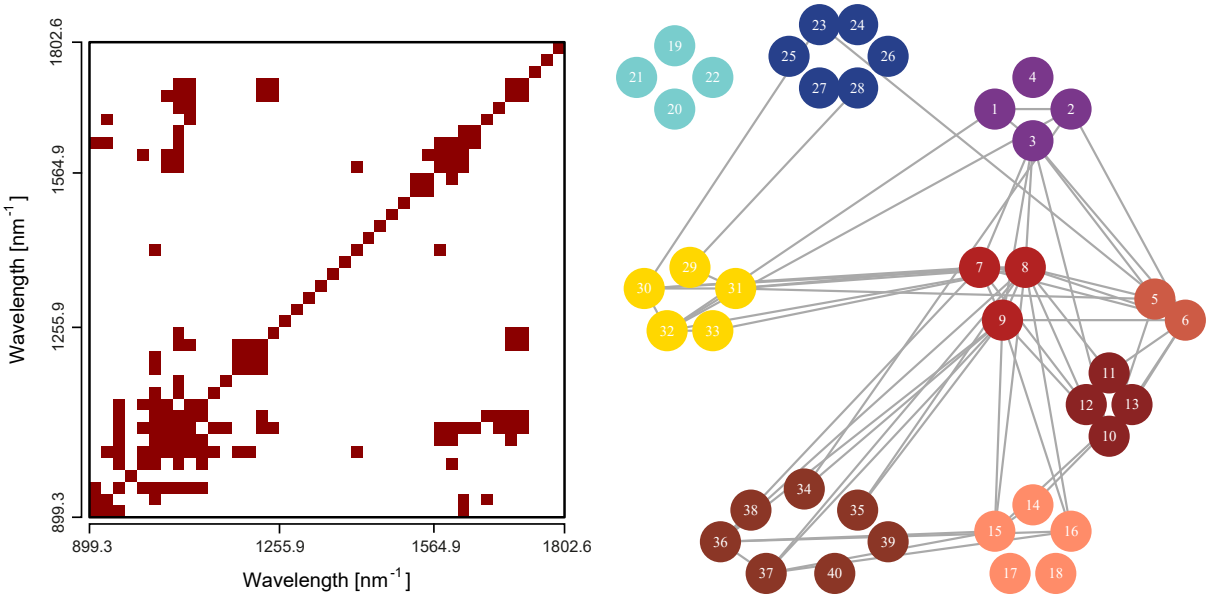

Figure 3.1: Left panel: posterior estimate of the graph obtained using the method of Codazzi et al. (2022), see Chapter 1. Right panel: network representation of the estimated graph, with nodes grouped according to the expert-defined partition described in Chapter 2.

knowledge rather than being supported by statistical evidence. As a consequence, both the number and the composition of the groups of nodes remain uncertain. For example, should nodes 5 through 13 be split into multiple groups or treated as a single one? From a chemical perspective, they all correspond to substances associated with the main peak, see Figure 2.7. However, examining the inferred connections in the left panel of Figure 3.1 reveals only weak community structure: group boundaries are fuzzy, many edges cross between groups, and the contrast between within-group and between-group connections is weak. Similarly, in the central portion of the spectrum, connections are so sparse that it is unclear whether nodes 19 through 28 form one group or two. Note that such observations on the potential structures in Figure 3.1 refer only to a point estimate of the underlying graph. A two-step approach, in which the graph is first estimated (as in Chapter 1) and then the partition is analyzed conditionally on that estimate, fails to account for the uncertainty inherent in the graph itself. In contrast, van den Boom et al. (2022b) argues that a joint modeling framework where both the random graph and its latent structure are estimated simultaneously is more suitable to better propagate the uncertainty from single-edge estimation to large-scale structural learning.

This chapter is therefore motivated by the need for a data-driven partition of nodes, grounded in a statistical model, which enables us to quantify uncertainty in the partition. Following van den Boom et al. (2022b), we propose a graphical model that not only enables the estimation of individual edges but also provides a richer description of the graph, including the identification of latent node structures and block patterns in the corresponding edges. Some works have been proposed in this direction, both in the frequentist (Ambroise et al., 2009; Tan et al., 2015; Kumar et al., 2020) and Bayesian literature (Palla et al., 2012; Sun et al., 2015). The common idea in these works is to induce a clustering of the nodes such that variables are more likely to be connected within each group than to variables in different groups. Conversely, the main focus of the method proposed in this chapter is to investigate the dependencies among different groups. In other words, discovering edges among different groups is our primary interest, while these edges are discouraged in the existing literature.

Within this line of research, van den Boom et al. (2022b) propose borrowing ideas and models from network analysis and using them to define priors for random graphs. Mathematically, networks and graphs are the same object: a set of nodes connected by edges, which may have different features such as being directed, undirected, or weighted. However, the terminology reflects two distinct research traditions that have developed largely in parallel, with only limited overlap. We briefly pause here to clarify the differences.

Random graphs are the mathematical foundation of graphical models, which we have already introduced in Chapters 1 and 2. In this setting, the set of edges is not observed but instead represents the collection of conditional dependencies among multivariate data, whether continuous, discrete, or mixed. The primary inferential goal is to determine whether a given edge is present, a task commonly referred to as *individual edge estimation*. Within this framework, typical priors on the graph include the uniform prior and the independent Bernoulli prior, introduced in Section 1.2.2. Both of these choices, however, completely ignore large-scale or latent structures in the graph.

By contrast, network analysis is a research area in which the set of edges is typically observed. Examples abound: nodes may represent users in a social network with edges denoting friendships, or they may correspond to web pages linked through hyperlinks. The range of applications is so wide that it is difficult to summarize the goals of network analysis in a single sentence. Restricting ourselves to the statistical description of networks, the most common objective is to develop generative models that, among other tasks, describe the evolution of the network as new nodes arrive, identify the most influential nodes, or recover latent structures (Newman, 2018; Kolaczyk, 2009).

The Erdős–Rényi models (Erdös and Rényi, 1959) are the cornerstone of statistical generative models for networks. There are two closely related formulations of such models. In the first, a network with p nodes is chosen uniformly at random from the collection of all possible networks with a fixed number of edges. In the second, the network is generated by connecting nodes independently, with each edge included with probability θ. It is straightforward to notice the close analogy with the uniform and independent Bernoulli priors described in Chapter 1. This highlights the strong connection between generative statistical models for networks and prior distributions for random graphs. In particular, the analogy with the most widely studied network model in the literature also makes clear the limitations of the prior choice used in Chapter 1. It is well known that Erdős–Rényi models generate homogeneous networks with no specific structure, as every node is statistically equivalent to all others. As a result, these models are not suitable for uncovering and describing the latent structures that characterize many real-world networks. More importantly, they fail to capture the type of block-like behavior observed in Figure 3.1. It follows that, if our goal is to model latent structures in the graph, we must move beyond the independent Bernoulli prior distribution. Among the generative models for learning groups of nodes in network data, the Stochastic Block Model (SBM) (Holland et al., 1983) is the most common and well-established formulation.

In a SBM, the nodes are partitioned into mutually exclusive and exhaustive groups. Conditionally on this partition, the set of edges is generated according to the *stochastic equivalence* assumption (Nowicki and Snijders, 2001), meaning that the probability of an edge depends only on the group memberships of the nodes involved. Both the clustering of the nodes and the number of groups are typically unknown and must be inferred from the data. In the classical SBM formulation (Nowicki and Snijders, 2001), however, the number of groups is fixed in advance, thereby ruling out any assessment of uncertainty

on this quantity. The same limitation applies to approaches that rely on state-of-the-art model selection procedures to determine the number of groups. See Lee and Wilkinson (2019) for a review. More recently, Bayesian nonparametric methods have become the standard tool for addressing this issue, as they provide data-driven strategies to determine the number of clusters while simultaneously quantifying the associated uncertainty, see the second part of this thesis. At first sight, practitioners may feel overwhelmed by the wide variety of Bayesian nonparametric priors available for this purpose. Yet, the Extended Stochastic Block Model (ESBM) framework introduced in Legramanti et al. (2022) offers a unifying perspective, as it allows one to employ a wide range of Bayesian nonparametric priors within the class of Gibbs-type priors De Blasi et al. (2015). This framework encompasses several popular models, such as the Infinite Relational Model (Kemp et al., 2006), and the Mixture of Finite Mixtures SBM (Geng et al., 2019). Importantly, by framing the problem within a well-studied area of Bayesian nonparametric theory, the ESBM framework also opens the door to clustering mechanisms that had not previously been explored in network analysis. These include the Gnedin model (Gnedin, 2010) or the use of Product Partition Models (Hartigan, 1990), which allows for the integration of external information, enabling groups to form according to specific external attributes. In this latter case, the stochastic equivalence assumption is relaxed. A related approach, although not part of the ESBM framework, is the Degree-Corrected Stochastic Block Model (Tan and Iorio, 2019), where the probability of an edge is further modulated by the popularity of the involved nodes—a node-specific rather than group-specific attribute.

The ESBM framework is a natural choice for specifying a prior on the random, unobserved graph G, as it moves beyond the independent Bernoulli setting and allows for learning a random number of latent block structures. However, the nature of our problem introduces an additional constraint: the spectral bands are inherently ordered since they correspond to precise wavelength intervals. This ordering constraint is not accommodated by Gibbs-type priors, which are built on the assumption of exchangeability. Namely, the model is invariant under arbitrary relabeling of the nodes. In our context, this assumption is inconsistent with the data: it is physically wrong to think that we can permute portions of the spectrum and end up with an equivalent object. To address this limitation, we draw inspiration from the literature on changepoint models (Smith, 1975; Green, 1995; Barry and Hartigan, 1993), which has a long and well-established tradition in detecting the most likely clustering of time points while preserving their natural ordering. Although infrared spectra are not functions of time, the wavelengths are ordered in exactly the same way, which makes the analogy with changepoint problems both intuitive and compelling.

Within the extensive literature on multiple changepoint detection, we adopt the approach of Martínez and Mena (2014) because it allows us to remain within the ESBM framework while generalizing it to ordered partitions. The idea is straightforward: the Gibbs-type prior is modified to assign zero probability to partitions that violate the ordering constraint. The mass removed in this way is then equally redistributed among all partitions that share the same number of groups and group sizes. Thanks to a combinatorial result by Pitman (2006), the resulting prior on ordered partitions is simply a rescaled version of the original Gibbs-type prior, thereby inheriting its key properties, including analytical tractability. This ordered extension of the ESBM framework forms the foundation of the model we introduce in the remainder of the Chapter.

3.2 Model developments

For the smoothing of functional data, we adopt the same modeling framework introduced in Chapter 1, which we briefly recall here for completeness. We consider an independent sample of n spectra, each evaluated on a common grid of r points corresponding to the wavelengths at which absorbance is measured. Every spectrum is expressed through a representation based on p basis functions. Let $y_i(s)$ denote the absorbance spectrum of unit i at wavelength s. The model assumed for the i-th spectrum is

$$y_i(s) = \sum_{j=1}^{p} \beta_{ij}\varphi_j(s) + \varepsilon_i(s), \quad i = 1, \dots, n,$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})^T$ is the regression coefficient vector associated with the i-th curve, and $\varepsilon_i(s) \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_\varepsilon^2)$. Moreover, we denote by $\Phi \in \mathbb{R}^{r\times p}$ the cubic B-spline design matrix, whose entries are given by $\varphi_j(s_k)$, i.e., the evaluation of the j-th basis function at the k-th grid point s_k, for $j = 1, \dots, p$ and $k = 1, \dots, r$.

Let $\boldsymbol{Y}_i = (y_i(s_1), \dots, y_i(s_r))^T$ denote the absorbance spectrum of the i-th curve observed at all wavelengths. The statistical model for $\boldsymbol{Y}_i$ is given by

$$\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i, \tau_\varepsilon^2 \overset{\text{ind}}{\sim} \mathrm{N}_r\left(\boldsymbol{\Phi}\boldsymbol{\beta}_i, \tau_\varepsilon^2 \mathbf{I}_r\right).$$

Independence is assumed across different curves. As for the measurement variance τ_ε^2, we place a Inv-gamma prior, i.e., $\tau_\varepsilon^2 \sim \text{Inv-gamma}(a_\varepsilon, b_\varepsilon)$. The basis expansion coefficients, denoted as $\boldsymbol{\beta}$'s, are interpreted as in Chapter 1. Since each spline has compact support and dominates over the other basis functions only on a specific portion of the domain, the associated spline coefficient can be interpreted as capturing the contribution of that local region. In the application of interest, each portion of the domain corresponds to a specific band of the spectrum and, consequently, to a chemical group in the strawberry purees. Therefore, learning the dependency structure among the chemical substances that generate the absorbance spectrum is equivalent to inferring the dependency structure among the spline coefficients $\boldsymbol{\beta}$'s. To this end, we adopt the same strategy introduced in Chapter 1, and place a Gaussian graphical model prior on the basis coefficients. Specifically,

$$\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n \mid \boldsymbol{\mu}, \boldsymbol{\Omega} \overset{\text{iid}}{\sim} \mathrm{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}\right),$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$ is the prior mean vector on which we place a Normal hyperprior, $\boldsymbol{\mu} \sim \mathrm{N}_p\left(0, \sigma_\mu^2 \mathbf{I}_p\right)$. Finally, we recall that $\boldsymbol{\Omega}$ is the precision matrix.

The graphical modeling component arises in the specification of the prior distribution for the precision matrix $\boldsymbol{\Omega}$. Rather than adopting the conjugate Wishart prior, we employ a G-Wishart prior, which enforces sparsity by constraining to zero those entries of $\boldsymbol{\Omega}$ that correspond to conditionally independent spline coefficients. Specifically, we set

$$\boldsymbol{\Omega} \mid G \sim \text{G-Wishart}(b, D), \tag{3.1}$$

where $b > 2$ is the shape parameter and D is a $p \times p$ positive definite matrix, which we refer to as inverse-scale matrix. The analytical form of the G-Wishart density has been recalled in Equation (1.4).

3.2.1 Graphical model prior

The prior distribution in Equation (3.1) is specified conditionally on the random graph $G = (V, E)$, an undirected graph with node set $V = \{1, \dots, p\}$, where each node corresponds to a spectral band, and edge set $E \subset V \times V$. Thus, the unobserved graph G encodes the conditional independence structure among the spline coefficients $\boldsymbol{\beta}$ and constitutes the main object of inference.

As outlined in Section 3.1, our goal is to formulate a graphical model that jointly captures both individual edge relationships and large-scale block structures. To achieve this, we introduce a generic partition $\rho_p = \{C_1, \dots, C_K\}$ of the p nodes into $K \leq p$ disjoint groups, with group sizes $n_k = |C_k|$, for $k = 1, \dots, K$. For convenience, each partition ρ_p can be equivalently represented by a node membership vector $\boldsymbol{z} = (z_1, \dots, z_p)$, where $z_j \in \{1, \dots, K\}$, according to the equivalence relationship

$$j \in C_k \iff z_j = k$$

for all nodes $j = 1, \dots, p$ and all possible groups $k = 1, \dots, K$.

The proposed graphical model is defined by specifying a joint prior distribution on the precision matrix $\boldsymbol{\Omega}$, the random graph G, and the random partition $\boldsymbol{z}$. Following van den Boom et al. (2022b), we adopt the factorization

$$\mathbb{P}(\boldsymbol{\Omega}, G, \boldsymbol{z}) \propto \mathbb{P}(\boldsymbol{\Omega} \mid G)\, \mathbb{P}(G \mid \boldsymbol{z})\, \mathbb{P}(\boldsymbol{z}) ,$$

where $\mathbb{P}(\boldsymbol{\Omega} \mid G)$ is taken to be the G-Wishart prior distribution introduced in Equation (3.1). Finally, we note that prior uncertainty on the number of groups is naturally induced by $\mathbb{P}(\boldsymbol{z})$, since K is simply the number of distinct labels appearing in the membership vector $\boldsymbol{z}$.

Regarding the choice of $\mathbb{P}(G \mid \boldsymbol{z})$, we employ a SBM prior for G in order to accommodate the discovery of potential latent block structures. The stochastic equivalence assumption is incorporated through a latent symmetric matrix $\boldsymbol{Q}$ of size $K \times K$, whose entries $Q_{r,s} \in [0, 1]$ represent the prior probability of an edge between any node in group r and any node in group s. Formally,

$$\mathbb{P}((i, j) \in E \mid \boldsymbol{z}, \boldsymbol{Q}) \overset{\text{ind.}}{=} Q_{z_i, z_j} , \tag{3.2}$$

for all $i, j \in 1, \dots, p$ with $i < j$. This definition is reminiscent of the independent Bernoulli prior but introduces two crucial distinctions. First, the Bernoulli probabilities are constant within blocks, thereby enforcing homogeneity among nodes that belong to the same group and facilitating the formation of edge blocks. Second, the model is conditional on the partition $\boldsymbol{z}$, which is itself a random quantity, and thus naturally allows the number and composition of groups to be inferred from the data.

The stochastic equivalence assumption is a highly flexible concept that can characterize a wide variety of network structures. For instance, when $K = 2$ and the probabilities satisfy $Q_{11} < Q_{12} \ll Q_{22}$, the resulting graph exhibits a core–periphery structure (Borgatti and Everett, 2000): a subset of nodes (the core) is densely connected, while the remaining nodes (the periphery) are only connected to the core but not among themselves. See Amongero and Blasi (2025) for a novel Bayesian nonparametric approach for assortative SBM. On the other hand, in our ongoing analysis of the spectra of strawberry purees, the hypothesis of stochastic equivalence plays a crucial role. In the estimated graph shown in Figure 3.1, we are interested in identifying blocks of edges that may occur far from the main diagonal of the adjacency matrix, rather than focusing solely on within-group connections.

Therefore, we complete the prior specification by assuming i.i.d. Beta priors on the elements of $\boldsymbol{Q}$, that is

$$\boldsymbol{Q}_{r,s} \overset{\text{iid}}{\sim} \text{Beta}(\alpha_1, \alpha_2)\,,$$

for any $1 \leq r \leq s \leq K$. This choice is consistent with the notion of stochastic equivalence, as it does not favor any particular block of edges a priori. As a result, the prior expected graph G is homogeneous and unstructured, with density $\alpha_1/(\alpha_1 + \alpha_2)$. Nonetheless, the data subsequently drive the posterior learning of block structures, which in turn induces homogeneity within blocks of edges and heterogeneity across groups. We refer to Gopalan et al. (2012) and Lu and Szymanski (2019) for community detection approaches that impose stronger assumptions, such as different priors on-diagonal and off-diagonal entries of $\boldsymbol{Q}$.

Although quantifying uncertainty in the block probabilities is important, the primary inferential goal is the recovery of the node partition. For this reason, $\boldsymbol{Q}$ is commonly treated as a nuisance parameter and integrated out of the model. Exploiting beta–binomial conjugacy, we obtain

$$\mathbb{P}\left(G \mid \boldsymbol{z}\right) = \prod_{r=1}^{K} \prod_{s=r}^{K} \frac{B\left(\alpha_1 + S_{r,s}, \alpha_2 + S^{\star}_{r,s}\right)}{B(\alpha_1, \alpha_2)}\,, \tag{3.3}$$

where $S_{r,s}$ and $S^{\star}_{r,s}$ denote the number of edges and non-edges between the nodes in groups r and s, respectively. For brevity, we will write $G \mid \boldsymbol{z} \sim \text{SBM}(\alpha_1, \alpha_2)$ to refer to the distribution in Equation (3.3). Marginalizing out $\boldsymbol{Q}$ substantially reduces the computational burden. Indeed, the dimension of $\boldsymbol{Q}$ depends on the number of groups K, which in our case is unknown and inferred from the data. If $\boldsymbol{Q}$ were explicitly included in the sampling strategy, then the number of unknowns would change across MCMC iterations, forcing us to resort to reversible-jump moves, as we did in Chapter 2. By integrating $\boldsymbol{Q}$ out, we avoid dealing with this problem.

3.2.2 Ordered random partition prior

Prior distributions for a random partition ρ_p have been extensively studied in the model-based clustering literature, where the distributions for random partitions are commonly employed as priors for the underlying clustering. In the Bayesian nonparametric literature, these priors are implicitly induced by the model through the marginalization of a random probability measure. A detailed construction is deferred to Section 4. Here, we limit ourselves to say that, under the exchangeability assumption, Gnedin and Pitman (2006) proposed the following parametric form for $\mathbb{P}(\boldsymbol{z})$, which is known as the class of *Gibbs-type* partitions:

$$\mathbb{P}\left(\boldsymbol{z}\right) = V_n^K \prod_{k=1}^{K} (1-\sigma)_{n_k - 1}\,, \tag{3.4}$$

where $(x)_n$ denotes the rising factorial coefficient, $n_k = \sum_{j=1}^{p} \mathbf{1}\left(z_j = k\right)$ is the number of nodes in the k-th group and the prior parameters are the *discount parameter* $\sigma < 1$ and the set of weights $\{V_j^k : 1 \leq k \leq j \leq p\}$ which must be nonnegative and satisfy the following recurrence relationship:

$$V_1^1 = 1\,, \quad V_{j+1}^{k+1} = V_j^k - (j - k\sigma)V_{j+1}^k\,.$$

In the following, we only consider values of σ in the range $[0, 1)$, which is equivalent to assume that if we let the number of B-spline basis grows indefinitely, then the number of groups grows accordingly. We refer to De Blasi et al. (2015) for a more detailed interpretation of how σ regulates the asymptotic behavior of the clustering.

Although the prior distribution in Equation (3.4) has support over all possible partitions of p objects, it depends only on the group cardinalities n_k. In other words, all partitions with identical cluster sizes, regardless of which specific nodes are assigned to each group, have the same probability. This property directly arises from the underlying assumption of *node exchangeability*, i.e., the invariance of the model under arbitrary relabeling of the nodes. In such cases, the prior can be written as $\mathbb{P}(\boldsymbol{z}) = p(n_1, \dots, n_K)$, where $p(n_1, \dots, n_K)$ is symmetric in its arguments and is referred to as the *Exchangeable Partition Probability Function* (EPPF). Every Gibbs-type prior induces an EPPF, although the converse is not true. The ESBM framework proposed in Legramanti et al. (2022) was the first attempt to apply Equation (3.4) within the context of the SBM, but, in principle, it is possible to extend the ESBM framework to the broader class of all possible EPPFs.

In Section 3.1, we anticipated that assuming node exchangeability and, consequently, exchangeability of the spectral bands, is not appropriate for our application. Instead, we require the node partition ρ_p to respect the natural ordering of the wavelengths, thereby restricting attention to a smaller subset of all possible partitions. To formalize this idea, we rely on the notion of *admissible partitions* introduced in Martínez and Mena (2014). Specifically, a partition $\rho_p = (C_1, \dots, C_K)$ is said to be admissible if and only if for every node $i \in C_k$ and $j \in C_h$, the condition $k < h$ implies $i < j$. Note that in this definition, both the ordering of the nodes and the ordering of the groups matter. For this reason, we now use parentheses to denote ρ_p. Following Martínez and Mena (2014), the collection of all admissible partitions is referred to as the class of set compositions of V, denoted by $\mathcal{C}_V$. Moreover, the authors showed that $|\mathcal{C}_V| = 2^{p-1}$.

Since $\mathcal{C}_V$ is a strict subset of the space of all partitions of p objects, a natural way to define a prior distribution on admissible partitions ρ_p is to assign zero prior probability to all non-admissible partitions. Rather, for admissible partitions ρ_p, one can set their probabilities to be proportional to those induced by any Gibbs-type prior with the same cluster sizes. As shown in Pitman (2006), the normalizing constant of this operation is straightforward to compute, leading to the following result:

$$\mathbb{P}(\rho_p) = \begin{cases} \binom{n}{n_1, \dots, n_K} \frac{1}{K!} p(n_1, \dots, n_K), & \rho_p \text{ admissible} \\ 0, & \rho_p \text{ not admissible}, \end{cases} \tag{3.5}$$

where $p(n_1, \dots, n_K)$ can be any EPPF, for instance a Gibbs-type prior as in Equation (3.4). The prior distribution $\mathbb{P}(\rho_p)$ in Equation (3.5) depends only on the group size n_k's, under the implicit assumption that the ordering of the nodes is preserved. Hence, following the same notation as in Martínez and Mena (2014), we write $\mathbb{P}(\rho_p) = p'(n_1, \dots, n_K)$. Furthermore, $p'(n_1, \dots, n_K)$ is symmetric in its arguments, just like any EPPF. In practice, this means that $p'(\cdot)$ is balanced in the sense that it does not favor, a priori, either larger or smaller groups at the beginning or in the tail of the spectrum. On the other hand, this is also a slight abuse of notation, since by writing $p'(n_1, \dots, n_K)$ we omit the fact that the corresponding partition must be admissible and therefore the nodes are not exchangeable. Equivalently, the prior in Equation (3.5) can be interpreted as a Product Partition Model (Hartigan, 1990; Barry and Hartigan, 1993), where the cohesion function is defined as a transformation of the EPPF. In this construction, the

probability mass of all partitions with cluster sizes $(n_1, \ldots, n_K)$ is concentrated exclusively on those partitions with the same $(n_1, \ldots, n_K)$ but preserving the order of the data.

As a specific choice for the EPPF, we employ the two parameters generalization of the celebrated Ewens' sampling formula (Ewens, 1972), known as the Poisson-Dirichlet distribution (Perman et al., 1992). Under this choice, Martínez and Mena (2014) show that Equation (3.5) simplifies to

$$p'(n_1, \ldots, n_K) = \frac{p!}{K!} \frac{\prod_{k=1}^{K-1} (\theta + k\sigma)}{(\theta+1)_{(p-1)}} \prod_{k=1}^{K} \frac{(1-\sigma)_{(n_k-1)}}{n_k!}, \tag{3.6}$$

where $\sigma \in [0, 1)$ and $\theta > -\sigma$. We refer to Equation (3.6) as *ordered Poisson-Dirichlet* prior and we write $\boldsymbol{z} \sim \text{OPD}(\sigma, \theta)$. Finally, if we let $\sigma \to 0$, then we derive the ordered version of the Ewens' sampling formula.

Martínez and Mena (2014) proved that an appealing feature of this prior construction is that the induced prior distribution on the number of groups K is preserved with respect to the same quantity computed in the exchangeable case, a result that has been thoroughly studied in the literature. For the general case of Gibbs-type partitions, we refer to De Blasi et al. (2015). Instead, in the specific case of Equation (3.6), the prior distribution for the number of groups is given by

$$\mathbb{P}(K = k) = \frac{1}{\sigma^k (\theta+1)_{p-1}} \prod_{i=1}^{k-1} (\theta + i\sigma) (-1)^k C(p, k; \sigma),$$

where $C(p, k; \sigma)$ is the central generalized factorial coefficient Charalambides (2002), whose detailed description is deferred to Section E.1. The prior expected value of the number of groups is given by

$$E(K) = \frac{(\theta+\sigma)_p}{\sigma(\theta+1)_{p-1}} - \frac{\theta}{\sigma}. \tag{3.7}$$

Equation (3.7) can be used to guide the selection of prior hyperparameters, enabling us to align the expected number of groups with prior beliefs.

Wrapping up, the Bayesian hierarchical formulation of the proposed graphical model for a nonparametric regression on spectrometric data, which allows the learning of random blocks, is expressed as follows:

$$\begin{aligned}
\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i, \tau_\varepsilon^2 &\overset{\text{ind}}{\sim} \text{N}_r\left(\boldsymbol{\Phi}\boldsymbol{\beta}_i, \tau_\varepsilon^2 \mathbf{I}_r\right), \\
\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n \mid \boldsymbol{\mu}, \boldsymbol{\Omega} &\overset{\text{iid}}{\sim} \text{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}\right), \\
\boldsymbol{\Omega} \mid G &\sim \text{G-Wishart}(d, \boldsymbol{D}), \\
G \mid \boldsymbol{z} &\sim \text{SBM}(\alpha_1, \alpha_2), \\
\boldsymbol{z} &\sim \text{OPD}(\sigma, \theta).
\end{aligned}$$

We complete the model placing the following hyperpriors for the model hyperparameters:

$$\begin{aligned} \boldsymbol{\mu} &\sim \mathrm{N}_p\left(\mathbf{0}, \sigma_\mu^2 \mathbf{I}_p\right), \\ \tau_\varepsilon^2 &\sim \text{Inv-Gamma}\,(a, b), \\ \theta \mid \sigma &\sim \text{Gamma}_{(-\sigma)}(c, d), \\ \sigma &\sim \text{Beta}\,(a, b)\ . \end{aligned} \tag{3.8}$$

where $\text{Gamma}_{(-\sigma)}$ is $(-\sigma)$-shifted Gamma distributed with parameters (c, d), i.e., $\theta + \sigma$ is gamma distributed with parameters (c, d). The hyperprior for (σ, θ) was first proposed by Martínez and Mena (2014).

3.3 Sampling strategy

The sampling strategy follows the approach presented in Chapter 1, with the only differences being an adjusted step for updating the graph and an additional step for updating the partition. Here, we detail only this additional step and refer to Algorithm 3 for the pseudo-code of the complete sampling strategy. In Section B.4, we provide the updates for the hyperparameters σ and θ, as described in Martínez and Mena (2014).

The target conditional distribution is given by

$$\mathbb{P}\left(\boldsymbol{\Omega}, G, \boldsymbol{z} \mid \text{rest}\right) \propto \mathbb{P}\left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p \mid \boldsymbol{\Omega}\right) \mathbb{P}\left(\boldsymbol{\Omega} \mid G\right) \mathbb{P}\left(G \mid \boldsymbol{z}\right) \mathbb{P}\left(\boldsymbol{z}\right).$$

To sample from this distribution, we employ a blocked Gibbs sampling strategy: in the first step, we sample the graph G and the precision matrix $\boldsymbol{\Omega}$, conditional on the random partition and the $\boldsymbol{\beta}$ coefficients; in the second step, we update the allocation variables $\boldsymbol{z}$, conditional on the graph.

Concerning the first update, we adapt the Birth-and-Death approach of Mohammadi and Wit (2015), already employed in Chapter 1. At this stage of the blocked Gibbs sampler, the partition induced by the vector $\boldsymbol{z}$ is treated as fixed. The only adjustment required with respect to the algorithm described in Section 1.3 is to incorporate the new prior distribution on the graph defined in Equation (3.3). In this setting, the birth rate is proportional to

$$\frac{\mathbb{P}(G^{+e} \mid \boldsymbol{z})}{\mathbb{P}(G \mid \boldsymbol{z})} = \frac{S_{rs} + \alpha_1}{S^\star_{rs} - 1 + \alpha_2},$$

while the death rate is proportional to

$$\frac{\mathbb{P}(G^{-e} \mid \boldsymbol{z})}{\mathbb{P}(G \mid \boldsymbol{z})} = \frac{S^\star_{rs} + \alpha_2}{S_{rs} - 1 + \alpha_1},$$

where $G^{\pm e}$ denotes graph G with the added/removed edge $e = (i, j)$, such that $z_i = r$ and $z_j = s$.

After the graph update, the random partition is sampled conditionally on the graph G from the distribution $\mathbb{P}(\boldsymbol{z} \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \boldsymbol{\Omega}, G) \propto \mathbb{P}(G \mid \boldsymbol{z})\mathbb{P}(\boldsymbol{z})$. This full conditional distribution, as well as the prior $\mathbb{P}(\boldsymbol{z})$, only takes values on the space of admissible partitions. Such partitions can be interpreted as a static analogue of a changepoint model, where the ordering of the data is preserved. For this reason, we rely on a modified version of the adaptive split–merge algorithm for changepoint analysis proposed

in Benson and Friel (2018). A non-adaptive version of this algorithm was used in Corradin et al. (2022). See also Matechou and Argiento (2023) for a more recent application in the Bayesian nonparametric setting.

As already highlighted in the previous section, integrating out $\boldsymbol{Q}$ allows us to avoid transdimensional moves and focus solely on exploring the space of all possible ordered partitions. Although this assumption significantly reduces the dimensionality compared to the case of unrestricted partitions, the resulting space is still potentially very large and grows exponentially with the number of nodes in the graph. Hence, even though it is discrete and finite, brute-force enumeration of all $\boldsymbol{z}$ is intractable. Consequently, we explore this space by relying on the adaptive split-merge algorithm for changepoint models proposed by Benson and Friel (2018). The main idea behind adaptive MCMC is to refine the proposal distribution online, using information from the past history of the chain to improve mixing and achieve convergence more rapidly. Early attempts in this direction include Haario et al. (2001), who in a continuous setting proposed to adapt the covariance matrix in a multivariate normal proposal distribution. More recently, adaptive methods for discrete state spaces have emerged, see Liang et al. (2022) and Liang et al. (2023b) for variable selection problems, and Liang et al. (2023a) for exploring the space of DAGs.

Before describing the details of the algorithm, we introduce an equivalent parametrization of the group membership vector $\boldsymbol{z}$ that is more convenient under the assumption of ordered nodes. Specifically, let $\boldsymbol{r} = (r_1, \ldots, r_p) \in \{0,1\}^p$, where $r_j = 1$ if and only if the j-th node is the right endpoint of a group of nodes, and $r_j = 0$ otherwise. By definition, the final node is always the end of a group, hence $r_p = 1$. With this representation, the number of groups is easily recovered from $\boldsymbol{r}$ as $K = \sum_{j=1}^{p} r_j$. By analogy to temporal models, where this representation was first introduced, we refer to nodes with $r_j = 1$ as *changepoint nodes*. For instance, consider $p = 8$ and $K = 4$, with the ordered partition $\rho_p = (\{1\}, \{2,3,4\}, \{5,6\}, \{7,8\})$. In this case, $\boldsymbol{z} = (1,2,2,2,3,3,4,4)$ and $\boldsymbol{r} = (1,0,0,1,0,1,0,1)$ are equivalent representations of ρ_p. The corresponding changepoint nodes are $\{1,4,6,8\}$.

Let $\boldsymbol{z}'$ denote the proposed state and $\boldsymbol{r}'$ its equivalent representation in terms of changepoint nodes. Let $q(\boldsymbol{z}' \mid \boldsymbol{z})$ be the proposal distribution from $\boldsymbol{z}$ to $\boldsymbol{z}'$. The proposal consists of split–merge moves, where we first decide whether to attempt a split or a merge, with probabilities p_{split} and $1 - p_{\text{split}}$, respectively. In practice, we always use the symmetric choice $p_{\text{split}} = 0.5$.

A split move proposes to turn a non-changepoint node into a changepoint, thereby increasing the number of groups. Specifically, suppose node j is selected: then $r_j = 0$ and $r'_j = 1$, while all other nodes remain unchanged, i.e., $r_i = r'_i$ for all $i \neq j$. The new partition $\boldsymbol{z}'$ is then deterministically determined from $\boldsymbol{r}'$. Conversely, a merge move deletes an existing changepoint, replacing $r_j = 1$ with $r'_j = 0$, again leaving all other entries of $\boldsymbol{r}$ unchanged. Because the space of all possible vectors $\boldsymbol{z}$ is large, it is important to preferentially add changepoints at locations with high posterior changepoint probability and remove them in regions with low posterior probability. The adaptive nature of the MCMC arises from controlling the probabilities with which candidate nodes are selected for a split or a merge. To this end, we introduce two p-dimensional, iteration-specific parameter vectors, which represent unnormalized probabilities used to select the node at which a split or merge move is proposed. Specifically, we let

$$\boldsymbol{a}^{(t)} = \left(a_1^{(t)}, \ldots, a_p^{(t)}\right), \quad \boldsymbol{d}^{(t)} = \left(d_1^{(t)}, \ldots, d_p^{(t)}\right),$$

where $a_j^{(t)}$ and $d_j^{(t)}$ are the unnormalized probabilities that node j is chosen as a candidate for splitting

or merging a group at iteration t, respectively. Moreover, we define

$$a^\star = \sum_{j:r_j=0} a_j^{(t)}, \quad d^\star = \sum_{j:r_j=1} d_j^{(t)},$$

so that if node j is not a changepoint, an addition move proposes it with probability $a_j^{(t)}/a^\star$, while if it is a changepoint, a deletion move proposes to remove it with probability $d_j^{(t)}/d^\star$. Note that both $a^\star$ and $d^\star$ depend on the current iteration, although we omit this dependence in the notation for simplicity. Since, by definition, $r_p = 1$, we set $a_p^{(t)} = d_p^{(t)} = 0$ for all iterations t.

We now present the acceptance probability of a split move. Suppose that the j-th node has been selected to split the current cluster into two groups, and further assume that $z_j = s$. Recall that the number of groups increases by one after a split. Due to the admissibility assumption, there is only one valid way to split the corresponding cluster: nodes with indices less than or equal to j are assigned to the first group, while the remaining nodes form the second group. We denote the cardinalities of the two new groups by ℓ and $n_s - \ell$, respectively. The proposed move is accepted with probability $\alpha_{\text{split}} = \min(1, R^+)$, where

$$R^+ = \frac{\mathbb{P}(G \mid z')}{\mathbb{P}(G \mid z)} \frac{\mathbb{P}(z')}{\mathbb{P}(z)} \frac{q(z \mid z')}{q(z' \mid z)}. \tag{3.9}$$

In what follows, we refer to the first ratio as the *target ratio*, the second as the *prior ratio*, and the third as the *proposal ratio*. The target ratio can be computed using Equation (3.3). Computation can be accelerated by analytically simplifying common factors; see Equation (3.17) for the resulting expression. After canceling common factors, the ratio of admissible prior distributions in Equation (3.9) reduces to

$$\frac{\mathbb{P}(z')}{\mathbb{P}(z)} = \frac{(\theta + K\sigma)}{K} \binom{n_s}{\ell} \frac{\Gamma(\ell - \sigma)\Gamma(n_s - \ell - \sigma)}{\Gamma(1 - \sigma)\Gamma(n_s - \sigma)}. \tag{3.10}$$

See Section B.2.1 for the detailed derivation of Equation (3.10). Finally, the proposal ratio distributions appearing in Equation (3.9) equals

$$\frac{q(z \mid z')}{q(z' \mid z)} = \frac{1 - p_{\text{split}}}{p_{\text{split}}} \frac{d_j^{(t)}/(d^\star + d_j^{(t)})}{a_j^{(t)}/a^\star}. \tag{3.11}$$

The first factor in (3.11) accounts for the relative probabilities of selecting a split versus a merge move. The second factor compares the probability of choosing node j for the split (denominator) with the probability of selecting the same node for the reverse merge (numerator), conditional on the split having been performed. Symmetrically, a merge move at node j can be viewed as the reverse of a split move occurring at node j, transitioning from the proposed partition z' back to the current partition z. We therefore define the acceptance probability of a merge move as $\alpha_{\text{merge}} = \min(1, R^-)$.

The adaptive step consists of updating the two weights vectors $\boldsymbol{a}^{(t)}$ (in case of a split move) and $\boldsymbol{d}^{(t)}$ (in case of a merge move) at each iteration t using the following scheme:

- If a split move at node j has been accepted, then update:

$$\log\left(a_j^{(t+1)}\right) = \log\left(a_j^{(t)}\right) + \frac{h}{t/p}(\alpha_{\text{split}} - \alpha_{\text{target}}).$$

- If a merge move at node j has been accepted, then update:

$$\log\left(d_j^{(t+1)}\right) = \log\left(d_j^{(t)}\right) + \frac{h}{t/p}(\alpha_{\text{merge}} - \alpha_{\text{target}}),$$

where $h > 0$ is the initial adaptation, which is then mitigated as the number of iterations increases through the ratio t/p. See Benson and Friel (2018) for further details. Finally, α_{target} is the target Metropolis-Hastings acceptance rate.

To conclude the update of the latent membership variables $\boldsymbol{z}$, we follow the approach of Corradin et al. (2022) and introduce a third acceleration step, namely a shuffle move. This move preserves the number of changepoints but modifies the allocation of nodes across adjacent clusters by shifting the changepoint forward or backward. In this way, only the cluster cardinalities are altered, while their total number remains unchanged. Specifically, assume $K > 1$, since the shuffle move is not feasible otherwise. We first select uniformly at random a group $s \in \{1, \dots, M-1\}$ to be shuffled with its neighbor, the $(s+1)$-th group. Then, we draw uniformly at random the number of nodes to remain in group s, say $\ell \in \{1, \dots, n_s + n_{s+1} - 1\}$. The cluster sizes in the proposed partition are $(n_1, \dots, n_{s-1}, \ell, n_s + n_{s+1} - \ell, n_{s+2}, \dots, n_K)$. The move is accepted with probability $\alpha_{\text{shuffle}} = \min(1, R)$, with

$$R = \frac{\mathbb{P}(\boldsymbol{G} \mid \boldsymbol{z}')}{\mathbb{P}(\boldsymbol{G} \mid \boldsymbol{z})} \frac{\mathbb{P}(\boldsymbol{z}')}{\mathbb{P}(\boldsymbol{z})}. \tag{3.12}$$

Similarly to the split and merge step, the first ratio can be computed using Equation (3.3) or it is possible to analytically eliminate some terms, see Equation (B.3). Moreover, the ratio of the partition priors further simplify with respect to Equation (3.10) since the number of groups K does not change in a shuffle move. Specifically, it equals

$$\frac{\mathbb{P}(\boldsymbol{z}')}{\mathbb{P}(\boldsymbol{z})} = \frac{n_s! n_{s+1}!}{\ell!(n_s + n_{s+1} - \ell)!} \frac{\Gamma(l - \sigma)\Gamma(n_s + n_{s+1} - \ell - \sigma)}{\Gamma(n_s - \sigma)\Gamma(n_{s+1} - \sigma)}.$$

Algorithm 3: MCMC sampler with random blocks.

For each iteration:

Step 1. Sample the regression parameters

- **1.1** Sample the parameters $\boldsymbol{\beta}$'s from their full conditional in Equation (1.10).
- **1.2** Sample the parameter μ from its full conditional in Equation (1.11).
- **1.3** Sample the parameter τ_ε^2 from its full conditional in Equation (1.12).

Step 2. Sample $(G, \boldsymbol{\Omega})$ conditionally on the current random partition $\boldsymbol{z}$ and the $\boldsymbol{\beta}$'s coefficients using a modified version of the Birth-and-Death algorithm where the birth and death rates are given in Equations (3.3) and (3.3), respectively.

Step 3. Update $\boldsymbol{z}$ conditionally on $(G, \boldsymbol{\Omega})$ using the adaptive split-merge algorithm.

Step 4. Sample (σ, θ) from their full conditional distributions in Equation (3.18) and (B.4).

3.4 Analysis of fruit purees dataset

We apply the proposed methodology to the fruit purees dataset introduced in Chapter 1. Recall that the dataset consists of a sample of $n = 351$ strawberry purees measured on an equally spaced grid of 235 different wavelengths over the interval $I = [899.327, 1802.564]\,\text{nm}^{-1}$. Following the analyses in Chapters 1 and 2, we fit the model using $p = 40$ basis functions, which also determines the size of the graph. The hyperparameters are set equal to the previous chapters: $b = 3$, $D = \mathbf{I}r$, $a_\varepsilon = 10$, $b_\varepsilon = 0.001$, and $\sigma^2_\mu = 100$. The SBM prior on the graph is specified with $\alpha_1 = 12$ and $\alpha_2 = 3$, yielding an expected block density of 0.8 and a corresponding variance of 0.01. Finally, we set $\sigma = 0.0001$ and $\theta = 2$ without employing hyperpriors. From Equation (3.7), this choice corresponds to a prior expected number of groups almost equal to 7, which is slightly lower than the number of clusters identified in the expert-defined partition used in the previous chapter. The Gibbs sampler was run for a total of 50,000 iterations, with the first 10,000 iterations discarded as burn-in. The algorithm has been implemented in Rcpp (Eddelbuettel and François, 2011), which is a language extension of `R` that allows us to combine `C++` and `R`. The total running time is about 19 minutes, which corresponds to 45 iterations per second. The machine specifications are reported in Section 2.5.

The left panel of Figure 3.2 displays the posterior distribution of the number of groups, while Figure 3.4 shows the corresponding traceplot. The effective sample size is 1195.6. Interestingly, none of the posterior mass is assigned to the case of a single group. This corresponds to the situation in which the model presented in this chapter reduces to that of Chapter 1, which is not supported by the data. Consequently, this result provides evidence in favor of our hypothesis that a latent structure exists in the unobserved network of dependence. The posterior distribution is approximately symmetric around its mode, which is nine. The cases of eight and ten groups are also well supported by the data. Nonetheless, the final estimated partition consists of ten groups. Overall, these findings are in close agreement with the expert-defined partition, which contained nine groups in total. The final partition, shown in Figure 3.3 together with the smoothed curves, was obtained using Binder's loss function, adapted to the case of ordered partitions as described in Corradin et al. (2022). The difference in the number of groups between the data-driven and expert-defined partitions lies in the fact that the data favor a solution in which the two final peaks of the spectrum are assigned to two distinct groups, whereas in the expert-defined partition they were aggregated into a single macro-group. In the central portion of the spectrum, the two partitions exhibit only minor discrepancies, both being characterized by five groups with comparable sizes. More notable differences arise around the main peak: in Figure 3.3, we observe a large group of seven nodes, while in the expert-defined partition this same region was split into multiple smaller groups.

The estimated graph, shown in the right panel of Figure 3.2, exhibits a structure comparable to both the findings of Chapter 1 and those in Waghmare and Panaretos (2024), thereby aligning closely with the existing literature. The traceplot of the graph size is shown in Figure 3.4. Remarkably, the SBM prior enables us to simultaneously provide single-edge estimation as well as to uncover the large-scale structure of the spectrum partitioned into multiple groups.

We conclude by acknowledging that this study remains a preliminary analysis, and further investigation is warranted. For instance, in this study, we set σ and θ to fixed values rather than treating them as random with hyperpriors. Indeed, understanding the role of these parameters when transitioning from the two-parameter Poisson–Dirichlet distribution to its ordered version is needed. In particular, σ and θ are typically interpreted through the (generalized) Chinese Restaurant Process metaphor (see Chapter 4).

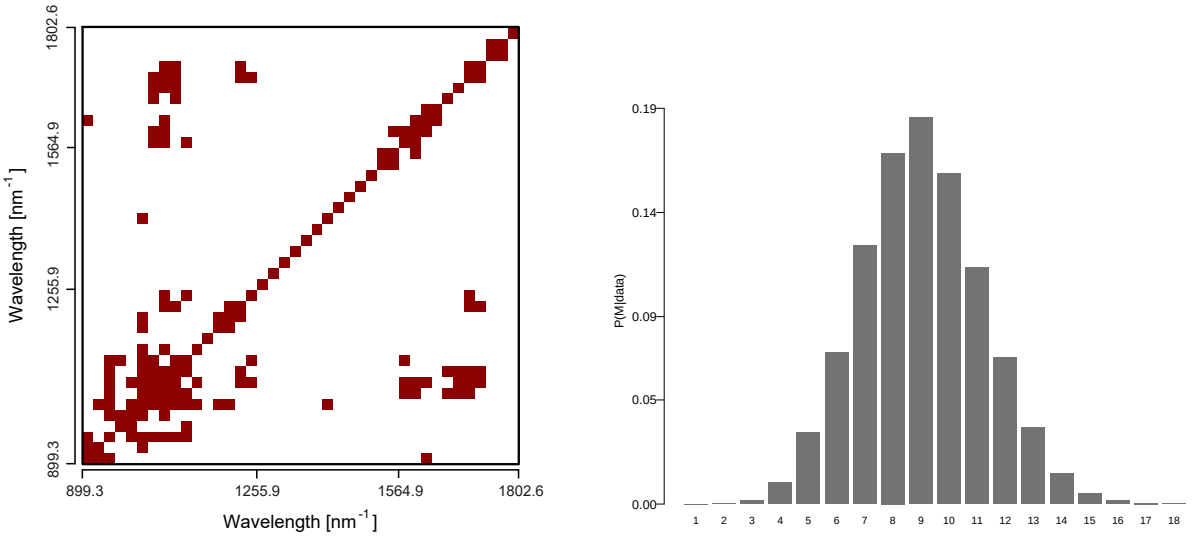

Figure 3.2: Left: posterior graph estimate; Right: posterior distribution of the number of groups.

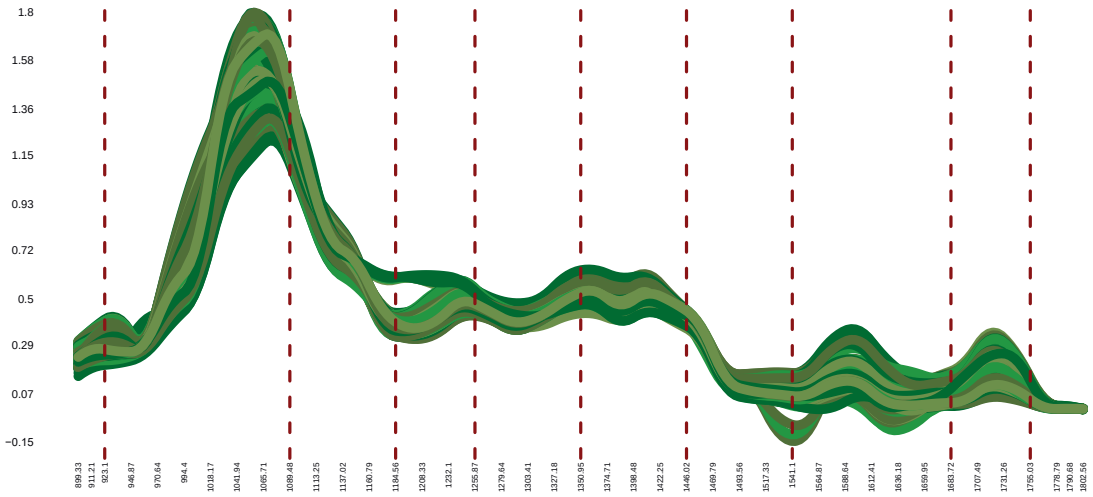

Figure 3.3: Plot of the 351 smoothed curves along with the estimated partition of the spectra. Groups are delimited using vertical red dotted lines.

However, the product partition model under consideration breaks exchangeability, loses the projectivity property, and therefore no longer admits the interpretation in terms of predictive distributions.

3.5 Extension to informed random partition

In Chapter 2, we relied on the expert-defined partition to introduce an informative prior that allowed us to study the connections between groups of nodes. However, in that setting, the partition was assumed to be known and fixed in advance. Motivated by the interest in learning the partition of nodes directly from the data, thereby accounting for uncertainty in the latent graph structure, this chapter develops a graphical model that jointly estimates the graph while searching for possible large-scale structures in the network of connections. To achieve this, however, we completely discard the expert-defined partition that had been externally provided.

An ideal approach would instead be to center the prior on all possible ordered partitions of the nodes around the expert-defined partition, thereby informing the model with knowledge-domain information while still allowing the final partition to be derived directly from the data. Recently, Smith and Allenby (2020) and Paganin et al. (2021) introduced, for the first time, methods that allow prior information on partitions to be incorporated by including an initial guess on the clustering. Their approach is based on adjusting partition probabilities according to the distance from the initial partition, as defined by a suitable loss function. This framework is available for any random partition distribution, but it requires the evaluation of a normalizing constant that becomes intractable even for relatively small sample sizes. While appealing, this distance-based approach was not suitable for our setting, since in our motivating problem the learning of the partition is only one component of a larger model that is entirely grounded in a precise probabilistic framework.

By contrast, the more recent contributions of Page et al. (2022) and Paganin et al. (2025) move in the desired direction, as they introduce a fully probabilistic framework that can accommodate available prior information on a given partition. Similarly to what we presented in Section 3.2.2, we need to extend the setting proposed in Paganin et al. (2025) to adapt to the case of ordered partitions. To this end, let ρ_0 denote the ordered expert-defined partition and let $\boldsymbol{z}_0$ and $\boldsymbol{r}_0$ be the corresponding vector with cluster memberships and changepoint representation. We let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ be a latent indicator vector such that $\gamma_j = 1$ if the j-th node is forced to be, or not to be, a changepoint node according to its status in the expert-defined partition. Namely, $\gamma_j = 1$ guarantees that $r_j = r_{0,j}$ while $\gamma_j = 0$ allows the possibility of r_j to be different from the corresponding quantity in the expert-defined partition. Prior knowledge about the nature of the j-th node can be included in the model by assuming $\boldsymbol{\gamma}$ to be random with independent components. Namely, $\gamma_j \overset{\text{ind}}{\sim} \text{Be}(\eta_j)$ so that values of η_j close to one force the posterior of r_j to be equal to $r_{0,j}$ while small values of η_j indicate that we are unsure if the j-th node is changepoint or not.

The prior distribution for the random partition proposed in Paganin et al. (2025) is

$$p(\rho_p \mid \rho_0) = \sum_{\boldsymbol{\gamma} \in \Gamma} p(\rho_p \mid \rho_0, \boldsymbol{\gamma})\, p(\boldsymbol{\gamma}), \tag{3.13}$$

where $\Gamma = \{0,1\}^{p-1}$ is the space of all possible configurations for $\boldsymbol{\gamma}$. Note that $|\Gamma| = 2^{p-1}$ since, by construction, the final node is such that $r_p = 1$ and therefore we force $\gamma_p = 1$. The partition probability $p(\rho_p \mid \rho_0, \boldsymbol{\gamma})$ in Equation (3.13) relies on the definition of *compatibility* of ρ_p with respect to ρ_0

according to γ introduced in Page et al. (2022). The key idea is that only those partitions ρ_p that can be reached from ρ_0 by suitably reallocating items as indicated by γ receive positive probability mass. Specifically, Page et al. (2022) proposes

$$p(\rho_p \mid \rho_0, \gamma) \propto p(n_1, \ldots, n_K)\mathbf{1}\left(\rho_p \text{ is compatible}\right) , \tag{3.14}$$

where $p(n_1, \ldots, n_K)$ can be any EPPF evaluated in the same cluster cardinalities n_k defined according to ρ_p. Equation (3.14) is similar to Equation (3.5), where the notion of admissibility is replaced by the one of compatibility, and the way in which the remaining mass is distributed is different.

It would be of interest to study a generalization of Equation (3.14) that accounts simultaneously for both admissibility and compatibility. A first step in this direction is to inform the partition only about the presence of changepoint nodes at specific positions. That is, consider a γ vector such that γ_j can equal one only if $r_{0,j} = 1$. We still allow the possibility of setting $\gamma_j = 0$ even when $r_{0,j} = 1$, but enforce $\gamma_j = 0$ whenever $r_{0,j} = 0$. As usual, we set $\gamma_p = 1$. This information implies the presence of $H = \sum_{j=1}^{p} \gamma_j$ subgroups of nodes, i.e., each subgroup is made of those nodes that fall among two entries equal to one in the γ vector. For example, consider $p = 8$ and let $\boldsymbol{z}_0 = (1, 1, 2, 2, 3, 4, 4, 4)$ be the initial guess, or, equivalently, let $\boldsymbol{r}_0 = (0, 1, 0, 1, 1, 0, 0, 1)$ be its changepoint representation. If we consider $\gamma = (0, 1, 0, 0, 1, 0, 0, 1)$, then, the $H = 3$ subgroups of nodes are $\{(1, 2), (3, 4, 5), (6, 7, 8)\}$. The subgroup representation allows us to easily set up a prior for any admissible and compatible partition ρ_p as the product of smaller, independent partitions. To this end, let $\tilde{n}_h$ denote the number of nodes in the h-th subgroup, with $\sum_{h=1}^{H} \tilde{n}_h = p$. Moreover, let K_h, with $h = 1, \ldots, H$, denote the number of clusters in the h-th subgroup of nodes, and let $n_{h,m}$ represent the corresponding cluster cardinalities. Then, the following expression defines a proper distribution over the partitions ρ_p that are both admissible and compatible with ρ_0 according to γ:

$$p(\rho_p \mid \rho_0, \gamma) = \prod_{h=1}^{H} \frac{1}{K_h!} \binom{\tilde{n}_h}{n_{h,1}, \ldots, n_{h,K_h}} p\left(n_{h,1}, \ldots, n_{h,K_h}\right) . \tag{3.15}$$

Equation (3.15) naturally generalizes Equation (3.5) and, indeed, the latter is recovered in the case when all γ_j's are equal to zero, except the final entry. On the other hand, informing the partition about the absence of a changepoint at a specific node, i.e., setting $\gamma_j = 1$ when $r_{0,j} = 0$ is more challenging and requires further investigation. Intuitively, the difficulty lies in the fact that the prior can no longer be decomposed into smaller, independent subgroups.

Appendix of Chapter 3

B.1 Deriving the birth rate

Assume the edge to be included in the graph is $e = (i, j)$ such that $z_i = r$ and $z_j = s$. Since we are considering a birth move, $e \notin E$, hence G^{+e} is such that the number of connections from groups r and s is increased by one unit and, consequently, the number of non-edges between the same groups decreases by one. Then,

$$\begin{aligned}\frac{\mathbb{P}(G^{+e} \mid \boldsymbol{z})}{\mathbb{P}(G \mid \boldsymbol{z})} &= \frac{B(S_{rs} + 1 + \alpha_1, S^{\star}_{rs} - 1 + \alpha_2)}{B(\alpha_1, \alpha_2)} \frac{B(\alpha_1, \alpha_2)}{B(S_{rs} + \alpha_1, S^{\star}_{rs} + \alpha_2)} \\ &= \frac{\Gamma(S_{rs} + 1 + \alpha_1)\Gamma(S^{\star}_{rs} - 1 + \alpha_2)}{\Gamma(S_{rs} + \alpha_1 + S^{\star}_{rs} + \alpha_2)} \frac{\Gamma(S_{rs} + \alpha_1 + S^{\star}_{rs} + \alpha_2)}{\Gamma(S_{rs} + \alpha_1)\Gamma(S^{\star}_{rs} + \alpha_2)} \\ &= \frac{S_{rs} + \alpha_1}{S^{\star}_{rs} - 1 + \alpha_2}\,.\end{aligned}$$

The death rate can be derived analogously.

B.2 Deriving the acceptance probability in a split move

Consider the split-merge algorithm presented in Section 3.3 and focus on split moves only, as the probability of merge moves can be derived accordingly. We now present how to derive the target and prior ratio needed to evaluate the acceptance probability R^{+}, as defined in Equation (3.9).

Firstly, we set the notation and main relationships among the quantities involved in the split move. Suppose that the j-th node has been selected as a position for splitting the current cluster into two groups and let $z_j = s$. We recall that the number of groups after the split is incremented by one unit. Because of the admissibility assumption, there is only one possible way of splitting the corresponding cluster, that is setting the nodes smaller or equal than j in the first group and the remaining ones in the other group. We let ℓ to be the number of nodes in the left-most group. Namely, the cluster cardinalities after the split move are $\left(n'_1, \ldots, n'_s, n'_{s+1}, \ldots, n'_{K+1}\right)$, where

$$\begin{aligned} n'_k &= n_k\,, \quad 1 \leq k < s\,, \\ n'_s &= \ell\,, \\ n'_{s+1} &= n_s - \ell\,, \\ n'_k &= n_{k-1}\,, \quad s+1 \leq k \leq K+1\,. \end{aligned} \tag{3.16}$$

Moreover, let $\boldsymbol{S}$ and $\boldsymbol{S}^{\star}$ be the $K \times K$ symmetric matrices such that their elements are the number of

edges S_{hv} and non-edges $S^{\star}_{hv}$ among the group h and v, respectively. Similarly, we let $\boldsymbol{S}'$ and $\boldsymbol{S}'^{,\star}$ be the corresponding quantities after the split. The size of such matrices is $K+1$. The following blocks of elements are unchanged during the split:

$$\begin{aligned} S'_{hv} &= S_{hv}\,, \quad 1 \le h \le v < s\,, \\ S'_{hv} &= S_{h,v-1}\,, \quad 1 \le h < s\,, s+1 < v \le K+1 \\ S'_{hv} &= S_{h-1,v-1}\,, \quad s+1 < h \le v \le K+1\,. \end{aligned}$$

The same relationships holds for $S'^{,\star}_{hv}$.

B.2.1 Target ratio

We first consider the target ratio. To this hand, we rewrite $\mathbb{P}(G \mid \boldsymbol{z})$, first defined in Equation (3.3), to highlight the product terms where the number of connections among the two groups changes after the split. We have that

$$\begin{aligned} \mathbb{P}(G \mid \boldsymbol{z}) &= \left(\frac{1}{B(\alpha_1,\alpha_2)}\right)^{K(K+1)/2} \\ &\times B(\alpha_1 + S_{ss}, \alpha_2 + S^{\star}_{ss}) \prod_{h=1}^{s-1} B(\alpha_1 + S_{hs}, \alpha_2 + S^{\star}_{hs}) \prod_{v=s+1}^{K} B(\alpha_1 + S_{sv}, \alpha_2 + S^{\star}_{sv}) \\ &\times \prod_{h=1}^{s-1}\prod_{u=h}^{s-1} B(\alpha_1 + S_{hu}, \alpha_2 + S^{\star}_{hu}) \prod_{v=s+1}^{K}\prod_{u=v}^{K} B(\alpha_1 + S_{vu}, \alpha_2 + S^{\star}_{vu}) \prod_{h=1}^{s-1}\prod_{v=s+1}^{K} B(\alpha_1 + S_{hv}, \alpha_2 + S^{\star}_{hv}) \end{aligned}$$

while $\mathbb{P}(G \mid \boldsymbol{z}')$ can be written as

$$\begin{aligned} \mathbb{P}(G \mid \boldsymbol{z}') &= \left(\frac{1}{B(\alpha_1,\alpha_2)}\right)^{(K+1)(K+2)/2} \\ &\times B(\alpha_1 + S'_{ss}, \alpha_2 + S'^{,\star}_{ss})\, B(\alpha_1 + S'_{s,s+1}, \alpha_2 + S'^{,\star}_{s,s+1})\, B(\alpha_1 + S'_{s+1,s+1}, \alpha_2 + S'^{,\star}_{s+1,s+1}) \\ &\times \prod_{h=1}^{s-1} B(\alpha_1 + S'_{hs}, \alpha_2 + S'^{,\star}_{hs}) B(\alpha_1 + S'_{h,s+1}, \alpha_2 + S'^{,\star}_{h,s+1}) \\ &\times \prod_{v=s+2}^{K+1} B(\alpha_1 + S'_{sv}, \alpha_2 + S'^{,\star}_{sv}) B(\alpha_1 + S'_{s+1,v}, \alpha_2 + S'^{,\star}_{s+1,v}) \\ &\times \prod_{h=1}^{s-1}\prod_{u=h}^{s-1} B(\alpha_1 + S'_{hu}, \alpha_2 + S'^{,\star}_{hu}) \prod_{v=s+1}^{K+1}\prod_{u=v}^{K+1} B(\alpha_1 + S'_{vu}, \alpha_2 + S'^{,\star}_{vu}) \prod_{h=1}^{s-1}\prod_{v=s+2}^{K+1} B(\alpha_1 + S'_{hv}, \alpha_2 + S'^{,\star}_{hv})\,. \end{aligned}$$

The double products in the final lines of Equations (B.2.1) and (B.2.1) simplify because of Equation (B.2). Consequently, the target ratio reduces to

$$\begin{aligned}\frac{\mathbb{P}(G \mid z')}{\mathbb{P}(G \mid z)} &= \left(\frac{1}{B(\alpha_1,\alpha_2)}\right)^{K+1} \\ &\times \frac{B(\alpha_1+S'_{ss},\alpha_2+S'^{,\star}_{ss})\, B(\alpha_1+S'_{s,s+1},\alpha_2+S'^{,\star}_{s,s+1})\, B(\alpha_1+S'_{s+1,s+1},\alpha_2+S'^{,\star}_{s+1,s+1})}{B(\alpha_1+S_{ss},\alpha_2+S^{\star}_{ss})} \\ &\times \prod_{h=1}^{s-1} \frac{B(\alpha_1+S'_{hs},\alpha_2+S'^{,\star}_{hs})B(\alpha_1+S'_{h,s+1},\alpha_2+S'^{,\star}_{h,s+1})}{B(\alpha_1+S_{hs},\alpha_2+S^{\star}_{hs})} \\ &\times \prod_{v=s+2}^{K+1} \frac{B(\alpha_1+S'_{sv},\alpha_2+S'^{,\star}_{sv})B(\alpha_1+S'_{s+1,v},\alpha_2+S'^{,\star}_{s+1,v})}{B(\alpha_1+S_{s,v-1},\alpha_2+S^{\star}_{s,v-1})}\,.\end{aligned} \tag{3.17}$$

B.2.2 Prior ratio

Moving to the prior ratio, we have that the proposed partition z' is, by construction, admissible. Hence, we are left with computing

$$\frac{\mathbb{P}(z')}{\mathbb{P}(z)} = \frac{p'\left(n'_1,\ldots,n'_s,n'_{s+1},\ldots,n'_{K+1}\right)}{p'\left(n_1,\ldots,n_s,\ldots,n_K\right)}\,.$$

The numerator is

$$\begin{aligned}&p'\left(n'_1,\ldots,n'_s,n'_{s+1},\ldots,n'_{K+1}\right) \\ &\quad= \frac{1}{K+1}(\theta+K\sigma)\frac{(1-\sigma)_{(n'_s-1)}}{n'_s!}\frac{(1-\sigma)_{(n'_{s+1}-1)}}{n'_{s+1}!} \\ &\qquad\times \frac{p!}{K!}\frac{\prod_{m=1}^{K-1}(\theta+m\sigma)}{(\theta+1)_{(p-1)}}\prod_{m=1}^{s-1}\frac{(1-\sigma)_{(n'_m-1)}}{n'_m!}\prod_{m=s+2}^{K+1}\frac{(1-\sigma)_{(n'_m-1)}}{n'_m!}\,,\end{aligned}$$

while the denominator is given in Equation 3.6. Then, the ration in Equation (B.2.2) simplifies to the expression reported in Equation (3.10) because of the relationships given in Equation (3.16).

B.3 Deriving the acceptance probability in a shuffle move

Consider a shuffle move as presented in Section 3.3. Let s be the label of the selected group and let ℓ be the number of nodes to keep in the s-th group. The cluster cardinalities prior and after the shuffle move are $(n_1,\ldots,n_{s-1},n_s,n_{s+1},\ldots,n_K)$ and $(n_1,\ldots,n_{s-1},\ell,n_s+n_{s+1}-\ell,\ldots,n_K)$. We recall that the number of groups K does not change in this case. Following the same approach and the same notation

as in Section B.2.1, the target ratio is given by

$$\begin{aligned}
&\frac{\mathbb{P}(\boldsymbol{G} \mid \boldsymbol{z}')}{\mathbb{P}(\boldsymbol{G} \mid \boldsymbol{z})} \\
&\quad = \frac{B(\alpha_1 + S'_{ss}, \alpha_2 + S'^{,\star}_{ss})\, B(\alpha_1 + S'_{s,s+1}, \alpha_2 + S'^{,\star}_{s,s+1})\, B(\alpha_1 + S'_{s+1,s+1}, \alpha_2 + S'^{,\star}_{s+1,s+1})}{B(\alpha_1 + S_{ss}, \alpha_2 + S^{\star}_{ss})\, B(\alpha_1 + S_{s,s+1}, \alpha_2 + S^{\star}_{s,s+1})\, B(\alpha_1 + S_{s+1,s+1}, \alpha_2 + S^{\star}_{s+1,s+1})} \\
&\quad \times \prod_{h=1}^{s-1} \frac{B(\alpha_1 + S'_{hs}, \alpha_2 + S'^{,\star}_{hs}) B(\alpha_1 + S'_{h,s+1}, \alpha_2 + S'^{,\star}_{h,s+1})}{B(\alpha_1 + S_{hs}, \alpha_2 + S^{\star}_{hs}) B(\alpha_1 + S_{h,s+1}, \alpha_2 + S^{\star}_{h,s+1})} \\
&\quad \times \prod_{v=s+2}^{K} \frac{B(\alpha_1 + S'_{sv}, \alpha_2 + S'^{,\star}_{sv}) B(\alpha_1 + S'_{s+1,v}, \alpha_2 + S'^{,\star}_{s+1,v})}{B(\alpha_1 + S_{sv}, \alpha_2 + S^{\star}_{sv}) B(\alpha_1 + S_{s+1,v}, \alpha_2 + S^{\star}_{s+1,v})} .
\end{aligned}$$

The prior ratio is simply computed as the ratio between p' evaluated at the corresponding cluster cardinalities. Namely,

$$\frac{\mathbb{P}(\boldsymbol{z}')}{\mathbb{P}(\boldsymbol{z})} = \frac{p'(n_1, \ldots, n_{s-1}, \ell, n_s + n_{s+1} - \ell, \ldots, n_K)}{p'(n_1, \ldots, n_{s-1}, n_s, n_{s+1}, \ldots, n_K)} .$$

The ratio in Equation (3.3) after noticing that the number of groups K does not change.

B.4 Update of σ and θ

In this section, we revise the updated of the hyperparameter σ and θ proposed in Martínez and Mena (2014). The corresponding prior distributions have been defined in Equation (3.8).

The full conditional distribution of σ is given by

$$p(\sigma \mid \theta, n_1, \ldots, n_K) \propto \sigma^{a-1}(1-\sigma)^{b-1}(\theta+\sigma)^{c-1} e^{-d\sigma} \prod_{m=1}^{K-1}(\theta + m\sigma) \prod_{m=1}^{K}(1-\sigma)_{(n_m-1)} . \quad (3.18)$$

with $\sigma \in (\max\{-\theta, 0\}, 1)$. We employ an adaptive Metropolis-Hastings step to sample from it.

For the posterior distribution of θ, we rely on the augmentation technique defined in (Martínez and Mena, 2014, Proposition 1). The latter, consists in introducing two latent variables $\xi \sim \text{gamma}(1, \theta+1)$, $\zeta \sim \text{Beta}(\theta+2, p)$ so that the full conditional of θ can be expressed as a mixture of $(-\sigma)$-shifted gamma distributions. Namely,

$$p(\theta \mid \sigma, \xi, \zeta, K) = \sum_{j=0}^{K+1} w_j \, \text{gamma}_{(-\sigma)}(c+j, d+\xi-\log(\zeta)) ,$$

where the mixing weights w_j, for $j = 0, \ldots, K+1$, are proportional to

$$w_j \propto \frac{\Gamma(c+j)}{(\sigma(d+\xi-\log(\zeta)))^j} \left[(p-\sigma)(p-\sigma+1)|s^j_{K-1}| + (2p-2\sigma+1)|s^{j-1}_{K-1}| + \sigma^2|s^{j-2}_{K-1}| \right] ,$$

where $|s^j_n|$ is the signless Stirling number of the first king, see Charalambides (2002). In particular, we recall that $|s^0_0| = 1$ while $|s^j_n| = 0$ for any $j \leq 0$ and $j > n$.

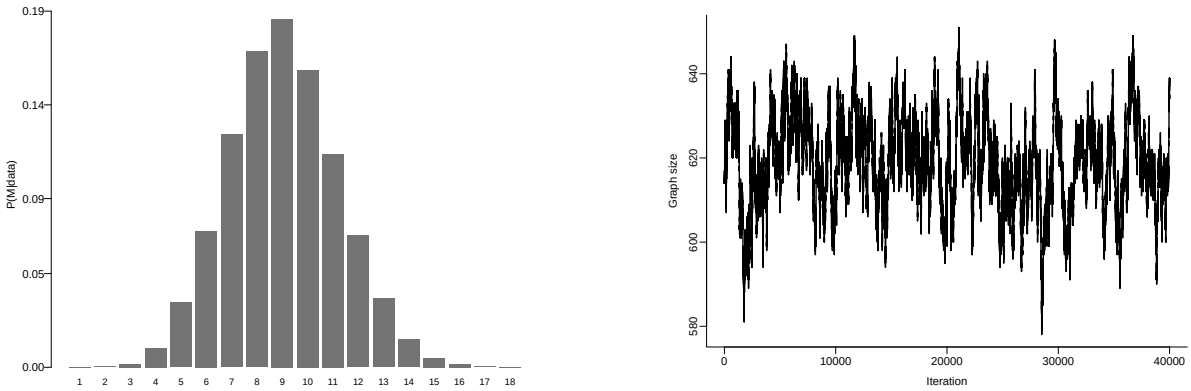

Figure 3.4: Traceplot of the number of clusters (left panel) and the graph size (right panel) after the burn-in phase.

B.5 Additional details on fruit purees dataset

Part II

Bayesian nonparametrics

Chapter 4

Introduction to Bayesian Nonparametrics

The inferential procedure requires assuming a form of homogeneity in the data-generating process. The Bayesian approach relies on the minimal assumption that the order in which the data are observed should not affect the inference. This naturally leads to the framework of exchangeable learning models (de Finetti, 1937; Hewitt and Savage, 1955), which embody the principle that the order of data collection is irrelevant to the underlying probabilistic structure. Formally, consider an infinite sequence of observations $(X_i)_{i\geq 1}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where each X_i takes values in a Polish space $\mathcal{X}$ endowed with its Borel σ-algebra $\mathcal{X}$. The sequence $(X_i)_{i\geq 1}$ is said to be exchangeable if its joint distribution remains invariant under any finite permutation of the indices. Specifically,

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

for any permutation π of the set $\{1, \ldots, n\}$ and for any $n \geq 1$. The de Finetti's representation theorem (de Finetti, 1937) establishes that the assumption of exchangeability is equivalent to the existence of a probability distribution $\mathscr{P}$ on the space $\mathcal{P}_{\mathbb{X}}$ of all probability measures on $\mathbb{X}$, such that the following Bayesian hierarchical model holds:

$$\begin{aligned} X_i \mid P &\overset{\text{iid}}{\sim} P, \quad i \geq 1 \\ P &\sim \mathscr{P}. \end{aligned} \tag{4.1}$$

where $\mathscr{P}$ is referred to as the prior distribution. The object P is a random probability measure, that is, a measurable map from $(\Omega, \mathcal{A})$ to $(\mathcal{P}_{\mathbb{X}}, \sigma(\mathcal{P}_{\mathbb{X}}))$, where $\sigma(\mathcal{P}_{\mathbb{X}})$ denotes a suitably defined σ-algebra on the space $\mathcal{P}_{\mathbb{X}}$. More explicitly, P can be viewed as a function of two variables, $\omega \in \Omega$ and $A \in \mathcal{X}$: for each fixed ω, $P(\omega, \cdot)$ is a probability measure on $(\mathbb{X}, \mathcal{X})$, while for every measurable set A, $P(\cdot, A)$ is a random variable.

De Finetti's theorem represents a fundamental cornerstone of the Bayesian framework, as it provides the theoretical foundation for the use of prior distributions. Moreover, it serves as the basis for distinguishing between what we refer to as Bayesian parametric and nonparametric inference. In the parametric setting, a specific functional form for the data-generating distribution is assumed, that is, $P = P_\theta$, so that $\mathscr{P}$ is supported on a finite-dimensional subspace of $\mathcal{P}_{\mathbb{X}}$. In this case, unlike in the frequentist paradigm, the parameter θ is treated as a random variable endowed with a prior distribution that encodes prior beliefs about its plausible values before the experiment is conducted. Conversely, Bayesian

nonparametrics (BNP) refers to the setting in which the distribution of the observations is left completely unspecified, with no parametric assumptions. Here, the parameter space is infinite-dimensional and is typically defined as the set of all possible solutions to a given learning problem. For instance, in a regression problem the parameter space may consist of all continuous functions, while in a density estimation problem it may include all probability densities. A BNP model relies on only a finite subset of the available parameter dimensions to explain a finite sample of data, allowing its effective complexity to grow with the sample size: as more observations are collected, more "dimensions" of the prior are activated, enabling the model to adapt automatically to the underlying structure (Orbanz and Teh, 2011).

In this context, following Walker (2010), the problem formulated in Equation (4.1) involves the construction of the prior distribution $\mathscr{P}$ on the space $(\mathcal{P}_{\mathbb{X}}, \sigma(\mathcal{P}_{\mathbb{X}}))$ of probability measure functions. The learning mechanism involves updating $\mathscr{P}$ as data arrive, so that after n observations beliefs about P are now encapsulated in the posterior distribution, given by

$$\mathscr{P}(\mathrm{d}P \mid X_1, \ldots, X_n) \;=\; \frac{\prod_{i=1}^{n} P(X_i)\, \mathscr{P}(\mathrm{d}P)}{\int \prod_{i=1}^{n} P(X_i)\, \mathscr{P}(\mathrm{d}P)}\,.$$

This in turn provides information about the future observation X_{n+1} via the predictive density

$$\mathbb{P}(X_{n+1} \mid X_1, \ldots, X_n) \;=\; \int P(X_{n+1})\, \mathscr{P}(\mathrm{d}P \mid X_1, \ldots, X_n)\,.$$

It is worth noting that some authors use the term nonparametric to indicate that no assumptions are made about parameters, but only about the observable variables. In essence, following de Finetti (de Finetti, 1937), rather than directly constructing the distribution $\mathscr{P}$, one assumes the exchangeability of the data and specifies the predictive laws $\mathbb{P}\left(X_{n+1} \mid X_1, \ldots, X_n\right)$, which are interpreted as learning rules, that is, as prescriptions for how to learn about future events given the available information. Nowadays, this perspective is referred to as the predictive approach or predictive modeling, and it is deeply rooted in Bayesian foundations, since the existence of $\mathscr{P}$ is still guaranteed by de Finetti's theorem. We refer to (Fortini and Petrone, 2025) for further reference on this emerging line of research.

In this thesis, however, we follow the former approach, that is, we aim to explicitly construct the probability measures $\mathscr{P}$. To this end, the cornerstone in the construction of BNP priors is the Dirichlet process, due to Ferguson (1973). Specifically, consider a concentration parameter $\alpha > 0$ and let P_0 be a diffuse probability distribution on $\mathbb{X}$, which we refer to as base measure. If $P \sim \mathscr{P}$ and $\mathscr{P}$ is the Dirichlet process, then P is an almost surely discrete probability measure Ferguson (1973); Blackwell (1973) which we represent as

$$P(\cdot) \overset{\mathrm{d}}{=} \sum_{j \geq 1} w_j\, \delta_{\tau_j}(\cdot)\,, \tag{4.2}$$

where the w_j's take values in $(0, 1)$ and sum up to one almost surely, $\sum_{j\geq 1} w_j = 1$, while the τ_j's are random variables taking values on $\mathbb{X}$. We write $P \sim \mathrm{DP}\left(\alpha, P_0\right)$. Ferguson presented two equivalent constructions of the Dirichlet process. The first defines it as a stochastic process indexed by the sets of the σ-algebra $\mathcal{X}$, rather than by time. In this framework, one first specifies the collection of finite-dimensional distributions, given by Dirichlet distributions, and then verifies Kolmogorov's consistency conditions to ensure the existence of the infinite-dimensional distribution.

The second construction, instead, arises from the intuition that just as the Dirichlet distribution can

be defined as the joint distribution of independent Gamma random variables normalized by their sum, then the Dirichlet process can be obtained as the normalization of a Gamma process with independent increments. In the following section, we develop this second construction in a more general framework and recover the Dirichlet process as a special case.

4.1 Bayesian nonparametric priors via Point processes

4.1.1 Completely Random Measures and their normalization

Over the past 25 years, an extensive body of literature has focused on constructing random probability measures that are almost surely discrete through the normalization of random measures. This seminal idea was first introduced in Regazzini et al. (2003) for random probabilities on the real line, under the name of Normalized Random Measures with independent Increments (NRMIs). We refer to Lijoi and Prünster (2010) for a comprehensive overview. This approach is deeply rooted in the theory of point processes, which, together with Palm calculus, constitute the main technical tools employed in Chapter 5 to define a novel random probability measure and to establish the associated theoretical results. Accordingly, in Section C.1, we review the key concepts underlying this theory, the most fundamental of which is the Poisson process. See Kingman (1993); Daley and Vere-Jones (2008); Baccelli et al. (2020) for detailed accounts.

A random measure G on $\mathbb{X}$ is said to be a Completely Random Measure (CRM) if, for any collection of disjoint sets $A_1, \dots, A_n$ in $\mathbb{X}$, the corresponding random variables $G(A_1), \dots, G(A_n)$ are mutually independent. An intuitive way to represent CRMs is due to Kingman (1993) and it decomposes them into the sum of three independent components: a non-random measure (any deterministic measure), a countable collection of non-negative random masses at non-random locations (constructed by a set of random weights at fixed locations, where the number of fixed locations can be either finite or infinite), and a countable collection of non-negative random masses at random locations. As in most of the literature, we discard the first deterministic component and deal with CRMs consisting solely of the second and/or third part. We refer to Kingman (1993) for the general definition of a CRM. Here, we consider a CRM G as an almost surely discrete random measure that can be represented as follows

$$G(\cdot) \overset{\mathrm{d}}{=} \sum_{l=1}^{K} s_l^* \, \delta_{\lambda_l^*}(\cdot). + \sum_{l \geq 1} s_l \, \delta_{\lambda_l}(\cdot) \tag{4.3}$$

where $\{s_1^*, \dots, s_K^*\}$ are a collection of random masses at K (which we assume to be finite) fixed points $\lambda_1^*, \dots, \lambda_K^*$ in $\mathbb{X}$ and $\{s_l, l \geq 1\}$ and $\{\lambda_l, l \geq 1\}$ are a collection of random masses and random locations respectively. Hence, the two sums differ in the nature of the locations (or atoms), which can be either fixed or random. We can simplify this by writing a random measure G as the sum of two independent random measures, one representing the fixed component and one representing the ordinary (random) component. In this case, we write $G \overset{\mathrm{d}}{=} G_{\text{fix}} + G_{\text{ord}}$. Finally, we distinguish between the random masses (or jumps) associated with the ordinary part, the s_l's, and those associated with the fixed part, the s_l^*'s.

The law of the random measure G is uniquely identified by the law of the ordinary and the fixed part. The latter is rather straightforward to specify since there can only be a finite number of jumps associated to the fixed atoms. In particular, these are taken to be independent random variables whose density with

respect to the Lebesgue measure is $\rho_{\text{fix},l}(s)\mathrm{d}s$ and supported on $\mathbb{R}^+$ or some subset of it. On the other hand, the law of the ordinary component is more involved as we are dealing with infinitely many pairs of jumps and locations. Here, we follow the classical approach of Kallenberg (2017) and define G_{ord} as a functional of a point process Ψ.

In particular, we let $\Psi = \sum_{l\geq 1} \delta_{\{s_l,\lambda_l\}}$ be a marked point process on the space $\mathbb{R}^+ \times \mathbb{X}$ with ground process Φ, see the definition in Section C.1. Specifically, Φ is a Poisson process on $\mathbb{R}^+$ with intensity $\rho_{\text{ord}}(s)\mathrm{d}s$ is considered. If intensity is such that

$$\int_{\mathbb{R}^+} \rho_{\text{ord}}(s)\mathrm{d}s = +\infty\,, \tag{4.4}$$

$$\int_{\mathbb{R}^+} \min\{1, s\}\rho_{\text{ord}}(s)\mathrm{d}s < +\infty\,, \tag{4.5}$$

then Ψ will have infinitely many points and we say that it is infinitely active. See Regazzini et al. (2003) for further details. Following the literature, we refer to ρ_{ord} as Lévy intensity of Φ. To obtain Ψ, the points of Φ are marked with i.i.d. marks $\lambda_l \overset{\text{iid}}{\sim} P_0$, where P_0 is some diffuse probability measure taking values in $\mathbb{X}$. It follows that Ψ is a Poisson Process on $\mathbb{R}^+ \times \mathbb{X}$ with intensity

$$\mathrm{M}_\Psi(\mathrm{d}s, \mathrm{d}\lambda) \;=\; \rho_{\text{ord}}(s)\mathrm{d}s\, P_0(\mathrm{d}\lambda)\,.$$

Finally, the CRM G_{ord} is obtained as the following functional of Ψ

$$G_{\text{ord}}(\mathrm{d}\lambda) \;=\; \int_{\mathbb{R}^+\times\mathbb{X}} s\,\mathbf{1}_{\{\mathrm{d}\lambda\}}(x)\,\Psi(\mathrm{d}s, \mathrm{d}x)\,.$$

In particular, the point process Ψ and the jumps associated with the fixed atoms are assumed to be independent. By construction, the locations in both the fixed and ordinary components are almost surely distinct, which makes CRMs particularly well suited for constructing almost surely discrete random measures. A priori, we assume that G does not have fixed atoms, i.e., $G \overset{d}{=} G_{\text{ord}}$. With a slight abuse of notation, in the following, we will refer to ρ_{ord} as the Lévy intensity of G.

The random probability measure P, as defined in Equation (4.2), is defined as

$$P(\cdot) \;=\; \frac{G(\cdot)}{G(\mathbb{X})} \;=\; \sum_{l\geq 1} \frac{s_l}{T}\,\delta_{\lambda_l}(\cdot)\,, \tag{4.6}$$

where $T = \sum_{l\geq 1} s_l$.

Notable examples of CRMs include the generalized gamma process Hougaard (1986); Brix (1999); Lijoi et al. (2007b) which corresponds to setting

$$\rho_{\text{ord}}(s) \;=\; \frac{\gamma}{\Gamma\,(1-\sigma)}\, s^{-(1+\sigma)}\, e^{-\theta s}\mathbf{1}\,(s \in (0,\infty))\;,$$

for $\gamma > 0$, $\sigma \in [0, 1)$ and $\theta > -\sigma$. This family includes the Gamma process as a special case, corresponding to $\theta = 1$ and $\sigma = 0$. When normalized as in Equation (4.6), the Gamma process yields the Dirichlet process with concentration parameter γ. In the following, we adopt the notation $G \sim \mathrm{GGP}(\theta, \sigma, \gamma, P_0)$.

Another example of a CRM we employ in Chapter 7 is the three-parameters Beta process Teh and

Gorur (2009) which corresponds to setting

$$\rho_{\text{ord}}(s) = \gamma \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\,\Gamma(c+\sigma)} s^{-(1+\sigma)} (1-\sigma)^{-(c+\sigma+1)} \mathbf{1}(s \in (0,1)) , \tag{4.7}$$

for $\gamma > 0, \sigma \in [0,1)$ and $c > -\sigma$. The special case of $\sigma = 0$ is known as the two-parameters Beta process and it has first been introduced in Hjort (1990). In the following, we write $G \sim \text{BetaP}(c, \sigma, \gamma, P_0)$.

4.1.2 Independent Finite Point Processes and their normalization

In Section 4.1.1, we noted that the non-integrability of the Lévy intensity, see Equation (4.4), ensures that the number of atoms in P is almost surely unbounded. However, the case in which P has a finite number of atoms almost surely is also of practical interest, as demonstrated in Chapters 5 and 6. In this setting, we say that P is finitely active.

From a probabilistic perspective, this case was first studied by Gnedin and Pitman (Gnedin and Pitman, 2006; Pitman, 2006) and later applied in a Bayesian nonparametric framework by De Blasi et al. (2015), where it appears as a special case of a Gibbs-type prior with a negative parameter. Its use in applications was popularized by Miller and Harrison (2018). Notably, Argiento and De Iorio (2022) introduced, for the first time, a construction of a random measure P with a finite, yet random, number of atoms via the normalization of a suitable point process. Their construction closely follows the approach presented in Section 4.1.1, thereby bridging the gap between finitely and infinitely active random measures. In particular, the construction in Argiento and De Iorio (2022) mirrors the steps outlined in Section 4.1.1, but replaces the infinitely active Poisson process with Lévy intensity ρ_{ord} with an Independent Finite Point Process (IFPP). Specifically,

$$\Phi \overset{\text{d}}{=} \sum_{l=1}^{M} \delta_{s_l} ,$$

where M is an integer valued random variable with probability mass function q_M, and the atoms, conditional on M, are i.i.d. draws from a probability distribution on the positive real numbers with density $h(\cdot)$; that is, $s_l \overset{\text{iid}}{\sim} h$. In particular, we assume $q_M(0) = 0$ to ensure that at least one atom exists almost surely. We refer to Section C.1 for further details.

Following the construction in Section 4.1.1, we define Ψ as a marked point process on $\mathbb{R}^+ \times \mathbb{X}$, with ground process Φ and i.i.d. marks λ_l drawn from P_0. The unnormalized random measure μ is then given as the following functional of Ψ:

$$\mu(\text{d}\lambda) = \int_{\mathbb{R}^+ \times \mathbb{X}} s\, \mathbf{1}_{\{\text{d}\lambda\}}(x)\, \Psi(\text{d}s, \text{d}x) .$$

In the following, we use μ to denote a finitely active random measure constructed via IFPPs, to distinguish it from G, which represents an infinitely active random measure constructed via CRMs. Finally, the random probability measure P is defined as

$$P(\cdot) = \frac{\mu(\cdot)}{\mu(\mathbb{X})} = \sum_{l=1}^{M} \frac{s_l}{T} \delta_{\lambda_l}(\cdot) , \tag{4.8}$$

where $T = \sum_{l=1}^{M} s_l$.

The most notable example in this class, introduced by Argiento and De Iorio (2022), is the Finite Dirichlet Process (FDP). This corresponds to the special choice of h as the gamma density, i.e., $s_l \overset{\text{iid}}{\sim} \text{Gamma}(\gamma, 1)$. If so, the joint probability of the normalized weights $w_l = s_l/T$, conditionally to M, is a symmetric Dirichlet distribution. This is,

$$(w_1, \dots, w_M) \mid M \sim \text{Dir}_M(\gamma, \dots, \gamma) .$$

In a slightly different approach, some authors do not consider a random number of atoms $M \sim q_M$ but they fix it to a large, conservative upper bound, i.e., $q_M = \delta_{\tilde{M}}$ for some large $\tilde{M}$. In this setting, one obtains the Dirichlet-Multinomial process (Muliere and Secchi, 1995) and its generalization, the Pitman-Yor multinomial process (Lijoi et al., 2020).

4.2 Applications of Bayesian nonparametrics

Let P be an almost surely discrete probability measure, which can either be finitely or infinitely active, as constructed in Section 4.1. These admit the representation $P = \sum_{l=1}^{M} w_j \delta_{\tau_l}$, with $M \leq \infty$ and such that the weights w_j's are independent on the location τ_l's. If so, we say that P is a proper species sampling model, a general class of random measures introduced and studied in Pitman (1996).

The use of the terminology species sampling is not arbitrary. Indeed, because of the almost surely discreteness of P, it is clear that any i.i.d. sample $\boldsymbol{X}_n = (X_1, \dots, X_n)$ from P contains ties with positive probability and can equivalently be summarized using the unique values $\{X_1^*, \dots, X_K^*\}$ and the corresponding frequencies $n_j = \#\left\{i : X_i = X_j^*\right\}$, for $j = 1, \dots, K$. Consequently, discrete nonparametric priors are not well suited for modeling directly data generated by a continuous distribution. However, as already noted in Pitman (1996), when the data come from a discrete distribution, as it happens for species sampling problems in ecology, biology and population genetics, it is natural to assign a discrete nonparametric prior to the unknown proportions. More precisely, suppose that a population consists of $M \leq \infty$ of species: one can think of w_l as the proportion of the l-th species in the population and τ_l is the label assigned to species l. Since the labels τ_l are generated by a non-atomic distribution they are almost surely distinct: hence, distinct species will have distinct labels attached. This approach extends beyond ecology and is relevant to many other disciplines. Thus, the term "species" can be interpreted in a broader sense. For instance, it can refer to fields such as topic modeling (Efron and Thisted, 1976), typos in texts (Nayak, 1988), bugs in computer code (Chao and Yang, 1993), or genomics (Mao, 2004). However, in the interest of Chapter 6, we mainly refer to ecology terminologies. An overview of foundational details and applications are discussed in Section 4.2.1, while in Chapter 6 we present a novel extension to a problem regarding species sampling in multiple areas.

For problems in which the observations are believed to come from an entirely continuous distribution, we have already noted that the almost surely discrete nature of species sampling models makes them unsuitable for directly modeling such data. Nonetheless, a very successful approach in Bayesian nonparametrics is to convolve species sampling models with a suitable kernel function, giving rise to the so-called Bayesian nonparametric mixture models. An overview of the foundational concepts and applications of these models is provided in Section 4.2.2, while in Chapter 5 we present a novel extension addressing the problem of clustering in multilevel data.

4.2.1 Species sampling problems

In this section, we focus on scenarios in which observed data are discrete, and species sampling models are used directly to represent the underlying generative process. In this setting, Pitman (1996) provides a formal description of the family of predictive distributions induced by a species sampling model that describe the evolution of the sampling process of the different species. Firstly, we draw $X_1^* \sim P_0$ and then, for any $n \geq 1$,

$$\begin{aligned} &\mathbb{P}\left(X_{n+1} \in \cdot \mid X_1, \ldots, X_n\right) \\ &\quad = \sum_{j=1}^{K} p_{K,n+1}(n_1, \ldots, n_j + 1, \ldots, n_K)\, \delta_{X_j^*}(\cdot) \;+\; p_{K+1,n+1}(n_1, \ldots, n_K, 1)\, P_0(\cdot)\,, \end{aligned} \tag{4.9}$$

where $\{p_{K,n}(n_1, \ldots, n_K) \;:\; 1 \leq K \leq n\,, n \geq 1\}$ is a collection of weights, which only depends on the observed species frequencies n_j's, providing the probability of observing either the already observed j-th species $p_{K,n}(n_1, \ldots, n_j + 1, \ldots, n_K)$ or the probability of observing a new, never seen before, species with probability $p_{K+1,n}(n_1, \ldots, n_K, 1)$. The main issue with Pitman's statement in Equation (4.9) lies in the fact that it guarantees the existence of the predictive weights $p_{K,n}(n_1, \ldots, n_K)$, but it does not provide any hint on their form. The Bayesian nonparametrics framework, however, provides a systematic way to compute them. To this end, we first introduce the so-called Exchangeable Partition Probability Function, which coincides with the probability of observing a specific sample of size n having $K \leq n$ distinct elements with frequencies $n_1, \ldots, n_k$:

$$\Pi_K^{(n)}\left(n_1, \ldots, n_K\right) \;=\; \int_{\mathbb{X}^k} E\left[\prod_{j=1}^{K} P^{n_j}(\mathrm{d}x_j)\right]. \tag{4.10}$$

The EPPF is a fundamental and well-studied tool for performing posterior inference in both species sampling and mixture models. In fact, it coincides with the marginal likelihood of the model in Equation (4.1). Because the data X_i are exchangeable, the function $\Pi_K^{(n)}\left(n_1, \ldots, n_K\right)$ must be symmetric in its arguments. Moreover, the collection $\left\{\Pi_K^{(n)}\left(n_1, \ldots, n_K\right) \;:\; 1 \leq K \leq n\,, n \geq 1\right\}$ must satisfy the following projectivity property:

$$\begin{aligned} &\Pi_1^{(1)}(1) \;=\; 1\,, \\ &\Pi_K^{(n)}\left(n_1, \ldots, n_K\right) \;=\; \sum_{j=1}^{K} \Pi_K^{(n+1)}\left(n_1, \ldots, n_j + 1, \ldots, n_K\right) \;+\; \Pi_{K+1}^{(n+1)}\left(n_1, \ldots, n_K, 1\right)\,, \quad n \geq 2\,. \end{aligned} \tag{4.11}$$

In particular, from Equation (4.10) we have a way to compute the EPPF from any species sampling model P. Interestingly, the viceversa also holds: for any function $\Pi_K^{(n)}\left(n_1, \ldots, n_K\right)$ that is symmetric and satisfies the projectivity property in Equation (4.11) there exists a species sampling model P such that Equation (4.10) holds. See Kingman (1975); Pitman (1995, 2006) for further details. Furthermore, from Equation (4.11), it follows that the predictive weights $p_{K,n}(n_1, \ldots, n_K)$ in Equation (4.9) can be

evaluated as

$$\begin{aligned} p_{K,n+1}(n_1,\ldots,n_j+1,\ldots,n_K) &= \frac{\Pi_K^{(n+1)}\left(n_1,\ldots,n_j+1,\ldots,n_K\right)}{\Pi_K^{(n)}\left(n_1,\ldots,n_K\right)}, \quad j=1,\ldots,K, \\ p_{K+1,n+1}(n_1,\ldots,n_K,1) &= \frac{\Pi_{K+1}^{(n+1)}\left(n_1,\ldots,n_K,1\right)}{\Pi_K^{(n)}\left(n_1,\ldots,n_K\right)}. \end{aligned} \tag{4.12}$$

In the case of a Dirichlet process prior, $P \sim \mathrm{DP}(\alpha, P_0)$, then Equation (4.9) reduces to

$$\mathbb{P}\left(X_{n+1} \in \cdot \mid X_1,\ldots,X_n\right) = \sum_{j=1}^{K} \frac{n_j}{\alpha+n}\,\delta_{X_j^*}(\cdot) + \frac{\alpha}{\alpha+n}P_0(\cdot). \tag{4.13}$$

The above expression shows that the analytical simplicity of the Dirichlet process comes at the cost of reduced modeling flexibility, since the predictive distribution fails to capture certain important features of the data. Indeed, the probability of discovering a new species is simply proportional to the concentration parameter α and the sample size n but it depends neither on the number of distinct elements in the sample, K, nor on their frequencies, the n_j's. This is rather unrealistic in applications and it has motivated the Bayesian nonparametrics literature to develop models beyond the Dirichlet process, aiming for a better trade-off between analytical tractability and richer data-driven learning. A prominent example is the two-parameters Poisson-Dirichlet (Pitman–Yor) process introduced in Pitman and Yor (1997).

Bayesian nonparametric applications to species sampling problems have become popular after the seminal work by Lijoi et al. (2007a). In this setting, the authors showed that, given a sample of n exchangeable observations $\boldsymbol{X}_n = (X_1,\ldots,X_n)$, the Bayesian nonparametrics approach is well suited to solve prediction problems regarding a future, unobservable, sample of m additional observations $(X_{m+1},\ldots,X_{n+m})$, with $m \geq 1$. In particular, Lijoi et al. (2007a) are both interested in the out-of-sample number of new distinct species in the additional sample, and the reverse problem, that is, how large should be the additional sampling effort m so that the probability of discovering at least a specified number of new distinct species exceeds a certain threshold. To answer such questions, they build on the results by Pitman, e.g., Pitman (2006), to derive analytical expressions for both the distribution of the in-sample and out-of-sample number of distinct species. The closed-form expressions of such quantities allows to perform Bayesian uncertainty quantification. If the latter is not of interest, one can resort to the pointwise Bayesian estimator by computing the corresponding expected value. Specifically, Lijoi et al. (2007a) studied priors $\mathscr{P}$ belonging to the broad class of Gibbs-type priors, with a specific focus on the case of the two-parameters Poisson-Dirichlet process. The derived expressions heavily relies on the computation of the generalized factorial coefficients, see Section E.1, which are computationally intense to evaluate for very large values of m. To circumvent this problem, in the case of a two parameter Poisson-Dirichlet model, Favaro et al. (2009) presented an asymptotic approximation, which, however, still requires simulation to be evaluated. Recently, the problem has also been addressed in Contardi et al. (2025) who instead presented an asymptotic normal approximation which is straightforward to evaluate. Regarding the estimation of species richness, Colwell et al. (2009), Lijoi et al. (2007a), and Favaro et al. (2009) approached the problem as an extrapolation of the unseen species problem. Indeed, as the sample size m varies, the out-of-sample number of distinct species forms the so-called extrapolation curve (Gotelli and Colwell, 2001), and species richness can be considered as the limit as $m \to \infty$.

However, this limit may be either finite or infinite, depending on whether $\mathscr{P}$ defines an infinitely or finitely active random measure. The choice of which family is more realistic depends on the specific application. Although the assumption of infinitely many species is purely abstract, it is often useful in practice to describe experiments in which the number of discovered species is very large, or when the rate of discovering new species is so high that only infinitely active CRMs can provide satisfactory predictions. The two-parameter Poisson–Dirichlet process belongs to this class of models. In contrast, finitely active models are characterized by an extrapolation curve that reaches an asymptote: no matter how much the sampling effort increases, the model does not predict the discovery of new species. Notable examples in this class include Gnedin's model (Gnedin, 2010) and the Finite Dirichlet Process (Argiento and De Iorio, 2022). It is worth mentioning that Zito et al. (2023) proposed a tunable Bayesian nonparametric approach that interpolates between these two regimes. For a comprehensive review of additional results and applications of Bayesian nonparametrics in species sampling problems, we refer to Balocchi et al. (2024b).

4.2.2 Bayesian nonparametric mixture models

We now turn our attention to problems in which the observations are believed to come from an entirely continuous distribution. Nonparametric mixture models were first introduced by Ferguson (1983) and Lo (1984). The approach gained widespread popularity after Escobar and West (1995) developed a simulation algorithm to approximate the posterior distribution, demonstrating the practical utility of using the Dirichlet process as the prior for the mixing measure in a mixture model. Since then, the so-called Dirichlet process mixture model has arguably become the most important and widely used model in Bayesian nonparametrics. The model for observations $Y_1, \dots, Y_n$ is

$$\begin{aligned} Y_1, \dots, Y_n \mid P &\overset{\text{iid}}{\sim} \int_{\mathbb{X}} f(\cdot \mid \theta) P(\mathrm{d}\theta)\,, \\ P &\sim \mathscr{P}\,, \end{aligned} \tag{4.14}$$

where $\{f(\cdot \mid \theta)\}_{\theta \in \mathbb{X}}$ is a parametric family of densities with respect to Lebesgue measure on $\mathbb{R}^q$ or counting measure on a countable subset of $\mathbb{R}^q$, and the specification of the parameter space Θ depends on the application. We refer to this parametric family as the kernel of the mixture model. The most popular example consists in letting $\theta = (\mu, \sigma^2)$ and $f(\cdot \mid \theta)$ to represent the probability density function of a Gaussian random variable with mean μ and variance σ^2. Because of the almost surely discreteness of P, the integral in Equation (4.14) can equivalently be expressed as

$$Y_1, \dots, Y_n \mid P \overset{\text{iid}}{\sim} \sum_{l=1}^{M} w_l\, f(\cdot \mid \tau_l)\,, \tag{4.15}$$

where $M \leq \infty$ is the number of mixture components, w_l's are here addressed as the mixing weights and the τ_l's are the component specific parameters. Mixture models as in Equation (4.15) are the building blocks for model-based clustering. Indeed, we can reformulate the model in Equation (4.14) as the

following hierarchical model

$$\begin{aligned} Y_i \mid \theta_i &\overset{\text{ind}}{\sim} f(\cdot \mid \theta_i), \quad i = 1, \dots, n, \\ \theta_i \mid P &\overset{\text{iid}}{\sim} P, \quad i = 1, \dots, n, \\ P &\sim \mathscr{P}. \end{aligned} \tag{4.16}$$

We note that the θ_i's in Equation (4.16) are i.i.d. realizations from P, just like the X_i's in Section 4.2.1. The difference in notation, which is maintained consistently in Chapters 5 and 6, is meant to emphasize that in Section 4.2.1 and Chapter 6, the species sampling model P is used directly to model the observed data, whereas in this section and in Chapter 5, it is used to generate latent parameters.

The almost-sure discreteness of P entails that there will be ties among the θ_i's and we let $\{\theta_1^*, \dots, \theta_K^*\}$ be the corresponding unique values. The clustering emerges since we can partition observations into clusters based on the latent parameters θ_i. In particular, we say that observation i and j belong to the same cluster if and only if $\theta_i = \theta_j$, or, equivalently, if there exists some $l \in \{1, \dots, K\}$ such that θ_i and θ_j are both equal to θ_l^*. We let $\rho = \{C_1, \dots, C_K\}$ be the clustering defined through the previous relationship and we let n_j be the corresponding cluster sizes, for $j = 1, \dots, K$. In particular, ρ is a random partition and its prior distribution, induced from model (4.14), is the EPPF introduced in Equation (4.10). A straightforward implication is that the law of ρ only depends on the cluster sizes n_j's.

MCMC schemes are necessary to perform inference for the model in Equation (4.14). As mentioned, the first successful attempt under a Dirichlet process prior was due to Escobar and West (1995), but it was Neal (2000) who proposed several efficient algorithms that became the state-of-the-art from that point onward. In general, to perform inference, we would like to draw samples from the posterior distribution $\mathscr{P}(\mathrm{d}P \mid Y_1, \dots, Y_n)$, which requires first computing this posterior. In the case of the Dirichlet process, Ferguson (1973) showed that the model is conjugate, meaning the posterior is still a Dirichlet process with updated parameters. However, this property does not hold for a general species sampling model. For more general priors, the posterior distribution has been derived for the class of NRMIs by James et al. (2009) and for the class of NIFPPs by Argiento and De Iorio (2022).

Even once the posterior is known, sampling from it may remain challenging when the underlying prior is infinitely active, since in principle this requires generating an infinite-dimensional object. This problem is typically addressed using truncation methods or slice samplers (Ishwaran and James, 2001; Walker, 2007; Papaspiliopoulos and Roberts, 2008; Griffin and Walker, 2011; Favaro and Teh, 2013). All algorithms that aim to sample directly from the posterior distribution $\mathscr{P}(\mathrm{d}P \mid Y_1, \dots, Y_n)$ are referred to as conditional samplers.

Alternatively, another class of algorithms leverages the tractable marginalization of P with respect to the mixing distribution, effectively removing the infinite-dimensional aspect of the problem. These are known as marginal samplers. For a detailed overview in the case of Dirichlet process mixture models, see Neal (2000), and for subsequent developments in more general settings, see Favaro and Teh (2013); Miller and Harrison (2018); Argiento and De Iorio (2022). For this second class of algorithms, the EPPF plays a crucial role. As noted by Ishwaran and James (2001), this approach can be applied to any mixture model for which the system of predictive distributions induced by P is known explicitly. Indeed, in Section 4.2.1, we showed that the derivation of the predictive distribution follows directly from the EPPF, see Equation (4.12).

The literature on this topic is vast, and a comprehensive review is beyond the scope of this thesis. We therefore limit ourselves to highlighting two recent review papers, Wade (2023); Grazian (2025).

4.2.3 Features allocation problems

More recent developments in Bayesian nonparametrics have extended this paradigm to classes of priors specifically designed for integer-valued random matrices with an infinite number of columns. Such random matrices can be used either as direct representations of data. For instance, Masoero et al. (2021) model genetic variation using binary matrices, where each row corresponds to a patient and each column encodes the presence or absence of a specific mutation relative to a reference genome. Their objective was to predict the emergence of new variants in a follow-up study. Building on the same binary-matrix framework, Ghilotti et al. (2025) applied these models to ecological settings, where each row represents a sampling site and each column indicates the presence or absence of a particular species. These models enable predictions about unseen species in new environments, providing valuable insights into biodiversity and species distribution. Similarly, Zhou et al. (2016) developed models for count matrices, applying them to text analysis, where each row corresponds to a document and each column records the frequency of a given word. Their focus was on predicting the occurrence of new words as additional documents are collected, an essential task in natural language processing (Manning and Schutze, 1999).

In this thesis, we do not focus on modeling data directly using random matrices; rather, these matrices are employed to represent latent structures that characterize the probabilistic law of complex data. Mathematically, each row of the latent matrix can be represented by the measure

$$\Theta_i \overset{\mathrm{d}}{=} \sum_{l \geq 1} \xi_{i,l}\, \delta_{\lambda_l} .$$

Hence, if $\xi_{i,j} \in 0, 1$, Θ_i indicates which features, here denoted as features, are present in the i-th individual among the infinitely many possible ones, represented by the atoms λ_j. Alternatively, if $\xi_{i,j}$ is a non-negative integer, Θ_i not only identifies which features the i-th individual possesses, but also encodes the degree or intensity with which each feature λ_j is present. Examples include: factor models (Thurstone, 1947; Anderson, 2003), tensor models (Chu and Ghahramani, 2009; Xiong et al., 2010), hidden Markov chains (Rabiner and Juang, 1986; Beal et al., 2001), network models (e.g., stochastic block models Holland et al. 1983; Legramanti et al. 2022), topic models (e.g., latent Dirichlet allocation model Blei et al. 2003 and Poisson factor analysis Dunson and Herring 2005; Zhou et al. 2012).

A distinctive feature of all these models is that they address situations in which the dimension of the latent structure must be chosen. For example, this can involve selecting the number of factors in a factor model or the number of topics in document analysis. Choosing this dimension is generally a challenging problem, which can be addressed using information criteria (Bai and Ng, 2002) or marginal likelihood approaches (Lee and Song, 2002). We refer to Owen and Wang (2016) for a comprehensive review. In contrast, Bayesian nonparametrics provide an elegant solution by allowing a finite number of classes (e.g., features or topics) to be selected from an infinite collection. Similar to Sections 4.2.1 and 4.2.2, the Bayesian nonparametric priors for this type of data are almost surely discrete measures. In the previous sections, we considered a prior P that was a probability measure, i.e., $P(\mathbb{X}) = 1$ almost surely. In this section, however, we focus on a class of priors G that are only constrained to be finite almost surely, i.e., $\mathbb{P}(G(\mathbb{X}) < \infty) = 1$. Specifically, G is defined as in Section 4.1.1, but unlike P, it is not normalized with

respect to its total mass.

The most notable example in this class of models is the infinite latent features model by Ghahramani and Griffiths (2005), which can be considered analogous to the Dirichlet process mixture model as a fundamental building block in this literature. We present it briefly here following Thibaux and Jordan (2007), who first highlighted the connection between the original model by Ghahramani and Griffiths (2005) and the CRMs framework introduced in Section 4.1.1.

Specifically, we let G be a Beta process as defined in Equation (4.7). Conditionally on G, the rows of the random matrix Θ_i are defined by

$$\xi_{i,l} \mid s_l \sim \text{Be}(s_l)\,, \quad l \geq 1\,. \tag{4.17}$$

with independence assumed across different rows, i.e., $\Theta_i \perp \Theta_j \mid G$ for any $i \neq j$. A crucial assumption for well-defined inference is that each Θ_i carries finite information. Otherwise, by the Borel–Cantelli lemma, all features would appear infinitely often even in a finite number of observations. Mathematically, the condition in Equation (4.5) ensures that this does not occur, guaranteeing that inference is well posed.

The model in Equation (4.17) is commonly referred to as a Bernoulli process, and we write $\Theta_i \mid G \sim \text{BeP}(G)$. In the infinite latent features model by Ghahramani and Griffiths (2005), the observable data Y_i, for $i = 1, \ldots, n$, are real valued vectors of length D, e.g., vectorized images, whose mean is specified as a functional of Θ_i:

$$\int_{\mathbb{X}} \lambda \Theta_i\,(\mathrm{d}\lambda) \;=\; \sum_{l \geq 1} \xi_{i,l}\, \lambda_l\,,$$

where $\mathbb{X} = \mathbb{R}^D$, and each feature λ_l represents a specific characteristic of the images that may appear multiple times across the sample. The corresponding base probability measure is taken as a multivariate normal distribution of dimension D with zero mean and covariance matrix $\sigma_A^2 \mathbf{I}_D$, for some $\sigma_A^2 > 0$.

Assuming independence across observations, the model can be expressed in the following hierarchical form:

$$\begin{aligned} Y_i \mid \Theta_i &\overset{\text{ind}}{\sim} \text{N}_D\left(\sum_{l \geq 1} \xi_{i,l} \lambda_l\,, \sigma_X^2\right)\,, \quad i = 1, \ldots, n\,, \\ \Theta_i \mid G &\overset{\text{iid}}{\sim} \text{BeP}\,(G)\,, \\ G &\sim \text{BetaP}\,(c, \sigma, \gamma, P_0)\,. \end{aligned} \tag{4.18}$$

In Chapter 7, we present a novel extension of the model in Equation (4.18) where the normal likelihood is replaced by a more general generalized linear model where we consider a time series of observations rather than n static and independent data.

4.3 Beyond exchangeability

In many applied problems, data are heterogeneous and exhibit dependence structures that go beyond simple exchangeability. Starting from the seminal contributions of MacEachern (1999) and MacEachern (2000), a large body of literature has developed to address inferential challenges arising from nonexchangeable observations within a Bayesian nonparametric framework. Three major lines of generalization break the standard assumption of exchangeability: the introduction of covariates, time-dependent

extensions, and models for multiple subpopulations.

Regarding the use of covariates in a Bayesian nonparametric setting, we only briefly mention the topic, as no chapter in this thesis is specifically dedicated to it. The seminal papers by MacEachern (1999, 2000) introduced the Dependent Dirichlet Process (DDP), in which covariates can be incorporated either in the atoms or the weights of the stick-breaking representation of the Dirichlet process, such that marginally, i.e., for any fixed value of the predictors, the random measure remains Dirichlet process distributed. This elegant idea has inspired many generalizations over the years, including the ANOVA-DDP (Iorio et al., 2004), the Spatial-DDP (Gelfand et al., 2005), and the Probit Stick-Breaking process (Chung and Dunson, 2009). This list is by no means exhaustive; for a more comprehensive and up-to-date review, we refer to Quintana et al. (2022).

Another strategy to incorporate covariates in a Bayesian nonparametric model is to build on the class of Product Partition Models (PPMs) (Hartigan, 1990), which we already encountered in Chapter 3. In this framework, the prior on an exchangeable partition is defined as proportional to the product of suitable functions of its clusters. Specifically, let $\rho_n = C_1, \ldots, C_K$ denote a partition of n elements into K clusters. Then,

$$p(\rho_n) \propto \prod_{k=1}^{K} c(C_k) , \tag{4.19}$$

where the cohesion function $c(\cdot)$ quantifies the degree of "closeness" among the elements within each cluster. The choice of $c(\cdot)$ can vary greatly depending on the application. For example, to ensure exchangeability, Lijoi et al. (2007b) showed that each Gibbs-type prior corresponds to a specific choice of the cohesion function. The most notable example is the Dirichlet process, whose EPPF is recovered by setting $c(C_k) = \alpha \times \Gamma(n_k)$, where α is the concentration parameter and n_k is the cluster size (Quintana and Iglesias, 2003). However, Gibbs-type priors represent only a subclass of PPMs, commonly referred to as exchangeable PPMs. In Chapter 3, we employed the non-exchangeable cohesion function proposed by Martínez and Mena (2014) to define a prior with positive mass only on admissible partitions. Müller and Quintana (2010) and Müller et al. (2011) extended the PPM framework to directly incorporate covariate information into the clustering process by introducing an additional factor, $g(\cdot)$, in the product of Equation (4.19). This similarity function is non-negative and formalizes the similarity among covariates within each cluster. Similar approaches have been applied to variable selection and spatial modeling; see, for example, Park and Dunson (2010), Quintana et al. (2015), Page and Quintana (2016), and Aiello et al. (2025).

4.3.1 Time dependent extensions

Although Bayesian nonparametric models are often applied to static problems, time-varying data frequently arise, calling for latent variable models that evolve dynamically. For instance, one may be interested in dynamic clustering problems, where cluster memberships can change over time and clusters may split or merge as new data arrive. Other examples include extracting features from video streams rather than from collections of still images, studying the temporal evolution of latent structures in networks, or tracking changes in the most relevant topics across time. In such cases, the static model introduced in Equation (4.1) is no longer suitable to capture the underlying temporal evolution. This calls for the definition of prior distributions over sequences of discrete random (probability) measures that can evolve dynamically over time.

These challenges have been explored across several research areas, often within distinct modeling frameworks. Within the extensive literature on dynamic random partitions, recent contributions include Page et al. (2022), De Iorio et al. (2023), and Franzolini et al. (2025a), among many others. However, this thesis does not focus on dynamic clustering. Instead, in Chapter 7, we turn our attention to dynamic factor models, where temporal evolution is driven by time-dependent random measures. Several works have pursued this direction, such as Williamson et al. (2010) and Perrone et al. (2017), who proposed time-dependent extensions of the Indian Buffet Process; Zuur et al. (2003) for dynamic factor analysis; Srebro and Roweis (2005) and Zhang et al. (2016) for dynamic topic modeling; and Caron and Teh (2012) for dynamic network modeling.

However, unlike the static setting presented in Section 4.2.3, the time-varying case lacks a simple, general, and unified framework that can be directly applied to evolving data while still capturing their essential structural properties. This gap motivates the contribution of Chapter 7, where we introduce a new Bayesian nonparametric approach capable of handling a wide range of time-varying statistical problems whose only requirement is that the likelihood belongs to the exponential family. As a latent process, we employ a highly flexible class of time-dependent integer valued matrices that naturally capture the appearance, disappearance, and potential reappearance of latent features over time.

The works by Pitt et al. (2002) and Pitt and Walker (2005) constitute the pivotal foundations upon which we build the temporal evolution of the random measures. In the parametric setting, Pitt et al. (2002) illustrate how to construct stationary time series with arbitrary parametric marginal distributions by introducing a suitable latent variable and defining a Markov chain that alternates between the emission and latent processes. A distinctive feature of their formulation lies in its use of the celebrated conjugate prior construction of Diaconis and Ylvisaker (1979), which ensures conjugacy within the exponential family. By exploiting this property, Pitt et al. (2002) achieve analytical tractability of the transition kernel, a key advantage that enhances both the theoretical understanding of temporal dynamics and the computational efficiency of inference.

Recently, Naik et al. (2022) employed the construction of Pitt et al. (2002) to define a sequence of time-dependent random measures $(\Theta_t)_{t\geq 1}$ following a Markov model, where each Θ_t is marginally distributed as a generalized gamma process (see Equation (4.1.1)). Their framework serves as a Bayesian nonparametric prior for a dynamic network model. In Chapter 7, we build on this idea and extend it to a more general framework in which the marginal distribution can be (almost) any desired completely random measure. Since conjugacy is the key property exploited by Pitt et al. (2002) to define the temporal evolution, we leverage the class of exponential completely random measures introduced by Broderick et al. (2018), thereby enabling a general construction for time-dependent sequences of random measures. The main differences from Naik et al. (2022) are twofold. First, while they consider a diffuse emission process and a discrete latent one, we focus on the opposite construction, where the emission process is discrete. Second, our model is more general not only because it allows for a wide range of time-dependent priors, but also because the corresponding likelihood is not restricted to dynamic networks and can accommodate the full class of generalized linear models.

Finally, it is worth mentioning that Mena and Walker (2005) were the first to combine Bayesian nonparametrics with the idea of Pitt et al. (2002). However, their notion of "nonparametric" differs from that of Naik et al. (2022) and from the approach developed in Chapter 7. Specifically, while our aim is to employ the framework of Pitt et al. (2002) to define a sequence of time-dependent random

measures, Mena and Walker (2005) focus instead on constructing sequences of time-dependent random variables whose transition kernel has a nonparametric form induced by a Dirichlet process mixture. Antoniano-Villalobos and Walker (2016) also operates in this setting.

4.3.2 Multiple groups

Suppose we deal with hierarchical data, that is, when observations emerge from $d > 1$ distinct groups (or levels) and the objective is to model heterogeneity both within and between these groups. As anticipated in Section 4.2.2, heterogeneity within groups is handled via mixture modeling to get group-specific clustering of observations, as well as density estimation. Concurrently, between-group heterogeneity can be addressed through two extreme modeling choices: (i) pooling all observations and (ii) conducting independent analyses for each of the d groups. However, both alternatives pose limitations. In the first case, differences across groups are not accounted for, while in the second one, groups are not linked, preventing sharing of statistical strength.

In Bayesian formalism, the sharing of information is naturally achieved through hierarchical modeling; parameters are shared among groups, and the randomness of the parameters induces dependencies among the groups. Mathematically, this is equivalent to assume that data are exchangeable within the same group and conditionally independent across different groups. This kind of dependence among multiple groups of observations was introduced in De Finetti (1938) and goes under the name of partial exchangeability. For the sake of clarity, we now consider the case of $d = 2$ ideally infinite sequences $\left(X_{j,i}\right)_{i\geq 1}$ and $j = 1, 2$ defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and both taking values in $(\mathbb{X}, \mathcal{X})$. Then, the array $\left\{\left(X_{j,i}\right)_{i\geq 1} : j = 1, 2\right\}$ is said to be partially exchangeable if for every $n_j > 0$ and for all permutations π_j of $\{1, \ldots, n_j\}$, with $j = 1, 2$ it holds that

$$\mathbb{P}\left(X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2}\right) = \mathbb{P}\left(X_{1,\pi_1(1)}, \ldots, X_{1,\pi_1(n_1)}, X_{2,\pi_2(1)}, \ldots, X_{1,\pi_2(n_2)}\right). \tag{4.20}$$

The definition in Equation (4.20) is easily generalizable to the case of $d > 2$.

The de Finetti's representation theorem (de Finetti, 1937) can be generalized to this setting. It establishes that the assumption of partial exchangeability is equivalent to the existence of a probability distribution $\mathcal{Q}$ on the space $\mathcal{P}_{\mathbb{X}}^d$ of all probability measures on $\mathbb{X}^d$, such that the following Bayesian hierarchical model holds:

$$\begin{aligned}(X_{1,i_1}, \ldots, X_{2,i_2}) \mid (P_1, \ldots, P_d) &\overset{\text{ind}}{\sim} P_1 \otimes \ldots \otimes P_d, \quad i_j \in \{1, \ldots, n_j\}, \quad j = 1, \ldots, d \\ (P_1, \ldots, P_d) &\sim \mathcal{Q}_d,\end{aligned}$$

A crucial problem consists in defining the Bayesian nonparametric prior $\mathcal{Q}_d$ which generates a vector of dependent probability measures $(P_1, \ldots, P_d)$. Specifically, the choice of the dependence structure between the components of the vector plays a pivotal role in applications, creating a borrowing of strength across the diverse, though related, groups of observations.

A substantial body of literature has developed around defining priors in this hierarchical framework, particularly following the pioneering introduction of the Hierarchical Dirichlet Process (HDP) by Teh et al. (2006). Since then, the HDP has been extended to encompass normalized random measures (Camerlenghi et al., 2019b; Argiento et al., 2020) and species sampling models (Bassetti et al., 2020). However, hierarchical constructions are not the only available approach. Many alternative strategies

have been proposed, including nested (Rodríguez et al., 2008; Camerlenghi et al., 2019a) and additive constructions (Müller et al., 2004; Lijoi et al., 2014), as well as normalized compound random measures (Griffin and Leisen, 2017), among others. Providing an exhaustive review of all these approaches is far beyond the scope of this thesis. Instead, we refer to Franzolini et al. (2025b), who introduced a general class of nonparametric priors, called multivariate species sampling models, which encompasses the vast majority of these proposals.

In model-based clustering, the dependence among the components of the vector of random probability measures $(P_1, \ldots, P_d)$ leads to group-specific clusters that are common among the groups, i.e., clusters that have the same interpretation across groups allow to define a global clustering. In general, the number of clusters within each group, as well as the number of global clusters, are unknown and need to be inferred from the data. In Chapter 5 we bridge the gap between infinite and finite mixtures in the hierarchical setting by introducing a novel family of Bayesian priors, named Vector of Finite Dirichlet Processes (Vec-FDR), to capture heterogeneity within and between groups. Interestingly, the novel prior is also part of multivariate species sampling models introduced in Franzolini et al. (2025b). In particular, we use the Vec-FDR prior to build a finite mixture model for each group, where the random number of mixture components, as well as the mixture parameters, are shared among groups. Rather, the mixture weights are assumed to be group-specific in order to accommodate differences between groups. This new class of hierarchical mixture models, named Hierarchical Mixture of Finite Mixtures, encompasses the popular Mixture of Finite Mixtures model (Miller and Harrison, 2018; Frühwirth-Schnatter et al., 2021) as a special case when the number of groups $d = 1$.

All previously cited approaches have been largely employed as Bayesian nonparametric priors in model-based clustering via mixture models while their application in species sampling problems, see 4.2.1, has been limited by the lack of closed-form estimators, particularly for predicting shared species across areas. To our knowledge, the only attempts we are aware of are Bacallado et al. (2015) and Camerlenghi et al. (2017), where posterior inference is mainly based on computational strategies. In Chapter 6, we take a step forward and develop a Bayesian nonparametric methodology to study the problem of unseen distinct and shared species when observations are collected in two different areas. Our approach answers the question: "How many species not yet observed in the two areas will be discovered in a future sample?" The proposed method is based on the Vector of Finite Dirichlet Process, introduced in Chapter 5 and provides closed-form expressions for estimating the quantity of interest. The Vec-FDP prior assumes a common and finite, yet random, total number of species in the two areas while allowing for variation in their proportions. Working with finite samples, not all species will necessarily be shared between areas, allowing the existence of area-specific species. Within this framework, a single underlying population is assumed, appearing to have varying characteristics across groups due to finite sampling.

Specifically, the main goals of this work are to derive a distributional theory for: (i) in-sample analysis for any observed samples of sizes n_1 and n_2, and (ii) out-of-sample predictions of the number of unseen distinct and shared species in additional unobserved samples of sizes m_1 and m_2. Hence, our results can be used to address the sample size determination in sampling problems designed to detect distinct and shared species.

Appendix of Chapter 4

C.1 Background on Point processes

Let $\mathbb{X}$ be a Polish space, and let $\mathcal{X}$ denote its associated σ-algebra. A measure $\tilde{\mu}$ on $(\mathbb{X}, \mathcal{X})$ is said to be locally finite if $\tilde{\mu}(B) < \infty$ for every relatively compact set $B \in \mathcal{X}$. We denote by $\mathcal{M}_{\mathbb{X}}$ the collection of all such locally finite measures and endow it with the corresponding Borel σ-algebra $\sigma(\mathcal{M}_{\mathbb{X}})$. Notice that the previously defined set $\mathcal{P}_{\mathbb{X}}$ of probability measures is a subset of $\mathcal{M}_{\mathbb{X}}$, consisting of those $\tilde{\mu} \in \mathcal{M}_{\mathbb{X}}$ satisfying $\tilde{\mu}(\mathbb{X}) = 1$.

Any measurable map

$$\Phi \,:\, (\Omega, \mathcal{A}, \mathbb{P}) \,\rightarrow\, (\mathcal{M}_{\mathbb{X}}, \sigma(\mathcal{M}_{\mathbb{X}}))\,,$$

is said to be a random measure on $(\mathbb{X}, \mathcal{X})$. Furthermore, if any realization of Φ is a counting measure, i.e., $\Phi(B)$ is a nonnegative integer almost surely for any relatively compact set $B \in \mathcal{X}$, then Φ is said to be a point process on $(\mathbb{X}, \mathcal{X})$. In this case, it admits the representation

$$\Phi(A) \,=\, \sum_{j \geq 1} \delta_{\tau_j}(A)\,,$$

for any set $A \in \mathcal{X}$, where $\left(\tau_j\right)_{j \geq 1}$ is a sequence of random variables taking values in $\mathbb{X}$. Thus, every point process is a random measure, but only those random measures whose realizations are almost surely counting measures qualify as point processes. The probability distribution of Φ, which is denoted as P_{Φ}, is the probability measure on $(\mathcal{M}_{\mathbb{X}}, \sigma(\mathcal{M}_{\mathbb{X}}))$ induced by Φ, that is $\mathbf{P}_{\Phi} = \mathbb{P} \circ \Phi^{-1}$. Namely,

$$\mathbf{P}_{\Phi} \,:\, \sigma(\mathcal{M}_{\mathbb{X}}) \,\rightarrow\, [0,1]\,; \quad \mathbf{P}_{\Phi}(L) \,=\, \mathbb{P}\left(\Phi \in L\right),$$

for any $L \in \sigma(\mathcal{M}_{\mathbb{X}})$. We also refer to $\mathbf{P}_{\Phi}$ as the law of the point process. It is in general a very complicated object and it is generally not possible to derive any representation of it. However, it is known that it is uniquely characterized by the Laplace functional,

$$\mathrm{L}_{\Phi}[f] \,=\, E\left[\exp\left\{-\Phi(f)\right\}\right],$$

where f is a measurable function $f : \mathbb{X} \rightarrow \mathbb{R}^{+}$ and we define the notation $\Phi(f) = \int_{\mathbb{X}} f(x)\Phi(\mathrm{d}x)$.

The mean measure of a point process Φ is a measure defined on $(\mathbb{X}, \mathcal{X})$ as

$$M_{\Phi}(B) \,=\, E\left(\Phi(B)\right)$$

for any set $B \in \mathcal{X}$. Moreover, let Φ^n be the n-power of Φ to be defined as the n-th measure theoretic

product, i.e., $\Phi^n(B_1 \times \cdots \times B_n) = \Phi(B_1) \dots \Phi(B_n)$. For any point process Φ, Φ^n can be expressed as

$$\Phi^n = \sum_{\boldsymbol{i} \in \Delta^n} \delta_{(\xi_{i_1}, \dots, \xi_{i_n})}, \tag{4.21}$$

where $\Delta^n = \left\{\boldsymbol{i} = (i_1, \dots, i_n) \ : \ i_k \in \{1, \dots, n\}, \ \forall k = 1, \dots, n\right\}$, hence elements of multi-index $\boldsymbol{i} \in \Delta^n$ may repeat and therefore atoms $(\xi_{i_1}, \dots, \xi_{i_n})$ do not need to be distinct. It can be proved that Φ^n is still a well defined point process, hence it is straightforward to define the n-th moment measure M_{Φ^n} of Φ as the mean measure of Φ^n. Finally, let $\Phi^{(n)}$ be the object obtained taking the summation in Equation (4.21) only over distinct elements,

$$\Phi^{(n)} = \sum_{\boldsymbol{i} \in \Delta^{(n)}} \delta_{(\xi_{i_1}, \dots, \xi_{i_n})},$$

where $\Delta^{(n)} = \left\{\boldsymbol{i} = (i_1, \dots, i_n) \ : \ i_k \in \{1, \dots, n\}, \text{ and } i_k \neq i_l \text{ for any } k \neq l\right\}$. $\Phi^{(n)}$ is called n-th factorial power of Φ, it is a point process itself and its mean measure, $M_{\Phi^{(n)}}$, is called n-th factorial mean measure of Φ. To conclude, it is crucial to observe that Φ^n and $\Phi^{(n)}$ coincide when evaluated over pairwise disjoint sets, which is a crucial assumption since it is often convenient to work in terms of factorial moment measure.

Before proceeding, we introduce two pivotal examples of point processes that play a central role in the construction of Bayesian nonparametric priors: the Poisson process and the Independent Finite Point Process.

Example (Poisson process)**.** Let Λ be a locally finite measure on $\mathbb{X}$. A point process Φ is said to be a Poisson process with intensity measure Λ if for all pairwise disjoint sets $B_1, \dots, B_k \in \mathcal{X}$, the random variables $\Phi(B_1), \dots, \Phi(B_k)$ are independent Poisson random variables with means $\Lambda(B_1), \dots, \Lambda(B_k)$, respectively. Namely,

$$\mathbb{P}\left(\Phi(B_1) = n_1, \dots, \Phi(B_k) = n_k\right) \ = \ \prod_{j=1}^{k} \frac{\left(\Lambda(B_j)\right)_j^n}{n_j!} \, e^{-\Lambda(B_j)} \, .$$

Throughout this dissertation, we assume that $\Lambda(A) = \int_A \rho(s)\mathrm{d}s$ and we refer to the density ρ as the Lévy intensity. Moreover, this implies that Λ is a diffuse measure which guarantees that the Poisson process is simple, i.e., the atoms of Φ are almost surely distinct. In this case, we write $\Phi \sim \mathrm{PP}(\rho)$. Equivalently, one can characterize the Poisson process as the point process whose Laplace functional is given by

$$\mathrm{L}_\Phi[f] \ = \ \exp\left\{-\int_{\mathbb{X}} \left(1 - e^{-f(s)}\right) \rho(s)\mathrm{d}s\right\} \, .$$

The mean measure of a Poisson process coincides with its intensity measure, that is,

$$M_\Phi(\mathrm{d}s) \ = \ \rho(s)\mathrm{d}s \, .$$

Moreover, the n-th factorial moment measure is given by

$$M_{\Phi^{(n)}}(\mathrm{d}s_1 \times \mathrm{d}s_n) \ = \ \prod_{j=1}^{n} \rho(s_j)\mathrm{d}s_j \, .$$

From Equation (C.1), the total mass $\Phi(\mathbb{X})$ of a Poisson process is distributed as a Poisson random variable with mean $\Lambda(\mathbb{X}) = \int_{\mathbb{X}} \rho(s)\mathrm{d}s$. This observation naturally leads to two important classes of models, depending on whether the Lévy intensity ρ is integrable. If ρ is not integrable, we obtain an infinitely active Poisson process, since the number of atoms in Φ is unbounded. Intuitively, this corresponds to a process that generates infinitely many small jumps. Conversely, if ρ is integrable, the process is referred to as finitely active, because the number of atoms is almost surely finite. In this case, we denote the number of atoms by M, and it follows that the atoms τ_j are independent and identically distributed according to $\Lambda(\cdot)/\Lambda(\mathbb{X})$.

Example (Independent Finite Point Process)**.** The IFPP is a generalization of the finitely active Poisson process where the number of points is not necessarily Poisson distributed. Specifically, we say that Φ distributes as a IFPP if it admits the following representation,

$$\Phi = \sum_{j=1}^{M} \delta_{\tau_j} ,$$

where M is an integer-valued random variable distributed as q_M and the atoms, conditionally to M, are i.i.d. distributed from some diffuse probability measure P_0. Equivalently, the IFPP is characterized by the following Laplace functional:

$$\mathrm{L}_{\Phi}[f] = \sum_{m=0}^{\infty} \left(\mathrm{L}_{P_0}[f]\right)^m q_M(m) ,$$

where $\mathrm{L}_{P_0}[f]$ is the Laplace transform of a random variable $\tau \sim P_0$, that is,

$$\mathrm{L}_{P_0}[f] = \int_{\mathbb{X}} e^{-f(x)} P_0(\mathrm{d}x) .$$

The mean measure and the n-th factorial moment measure are given by

$$\begin{aligned} M_{\Phi}(\mathrm{d}s) &= E\left(M\right) P_0(\mathrm{d}s) , \\ M_{\Phi^{(n)}}(\mathrm{d}s_1 \times \mathrm{d}s_n) &= E\left(M^{(n)}\right) \prod_{j=1}^{n} P_0(\mathrm{d}s_j) \end{aligned}$$

where $E\left(M^{(n)}\right)$ is the n-th factorial moment of the random variable M. Namely.

$$E\left(M^{(n)}\right) = E\left(M(M-1)\ldots(M-n+1)\right) = \sum_{m=0}^{\infty} q_m(m) \frac{(m+n)!}{m!} .$$

We conclude this section by introducing two fundamental operations on point processes that form the basis for the construction of Bayesian nonparametric priors: thinning and independent marking.

C.1.1 Thinning of points

The thinning of a point process consists in suppressing some subset of its points. Specifically, the only refer to the case of independent thinnings where the decision to suppress or keep each point is taken independently from the others. More precisely, let $p : \mathbb{G} \to [0, 1]$ be some measurable function called

the *retention function*. Then, given Φ, each point $\tau \in \Phi$ is erased with probability $1-p(X)$ independently from the other points. Moreover, let $\tilde{\Phi}$ be the thinning of Φ with retention function p. It is possible to show that $\tilde{\Phi}$ is a point process whose mean measure and Laplace functional are given by

$$M_{\tilde{\Phi}}(\mathrm{d}s) \;=\; p(s)M_{\Phi}(\mathrm{d}s)\,, \qquad \mathrm{L}_{\tilde{\Phi}}[f] \;=\; \mathrm{L}_{\Phi}[\tilde{f}]\,,$$

where $\tilde{f}(x) = -\log\left[1 - p(x)\left(1 - e^{-f(x)}\right)\right]$.

C.1.2 Independent marking of points

The independent marking of a point process refers to the operation by which each atom $\tau \in \Phi$ is associated with a random mark $Z \in \mathbb{Z}$, in such a way that the marks attached to different atoms are mutually independent. In the setting of interest, we restrict to the case of i.i.d. marks, sampled from some probability distribution f_Z, which we refer to as the mark distribution. The outcome of this construction is a new point process Ψ defined on the product space $\mathbb{X} \times \mathbb{Z}$:

$$\Psi \;=\; \sum_{j \geq 1} \delta_{(\xi_j, Z_j)}\,,$$

which is called an i.i.d. marked point process with ground process Φ and mark distribution f_Z.

We conclude this section introducing Palm calculus, a fundamental tool in the analysis of point processes. Palm calculus also allows to extend Fubini's theorem, and to enable the interchange of the expectation and integral operators when both pertain to a point process. However, there is a subtle distinction: the expectation is now taken with respect to the law of another point process, specifically, the reduced Palm version derived from the original process. In the case of Poisson processes, the reduced Palm version coincides with the law of the original Poisson process. This property characterizes Poisson processes and is known as the Slivnyak-Mecke theorem (Baccelli et al., 2020, Theorem 3.2.4). To effectively exploit this technique, it is essential to derive the reduced Palm distribution for any IFPP. This crucial result is presented in Theorem D.2.4.

C.1.3 Background on Palm calculus

The technique employed in the computations involves the disintegration of the Campbell measure of a point process Φ with respect to its mean measure M_{Φ}, usually called Palm kernel or family of Palm distributions of Φ. To be precise, we define the Campbell measure C_{Φ} on $\mathbb{X} \times \mathcal{M}_{\mathbb{X}}$ as follows

$$C_{\Phi}(B \times L) \;=\; E\left(\Phi(B)\,\mathbf{1}_L(\Phi)\right)\,, \quad B \in \mathcal{G}, L \in \mathbb{M}(\mathcal{G}).$$

for any measurable sets $B \in \mathcal{X}$ and $L \in \sigma(\mathcal{M}_{\mathbb{X}})$. By virtue of the Radon-Nikodym theorem, it follows that for any fixed L, there exists a unique disintegration probability kernel $\{\mathbf{P}^x_{\Phi}(\cdot)\}_{x \in \mathbb{X}}$ of C_{Φ} with respect to M_{Φ}, which satisfies the following integral equation

$$C_{\Phi}(B \times L) = \int_B \mathbf{P}^x_{\Phi}(L) M_{\Phi}(\mathrm{d}x),$$

for any $B \in \mathbb{X}, L \in \sigma(\mathcal{M}_{\mathbb{X}})$. Note that, for any $x \in \mathbb{X}$, $\mathbf{P}^x_\Phi$ is the distribution of a point process Φ_x on $\mathbb{X}$, such that $\mathbf{P}^x_\Phi(L) = \mathbb{P}(\Phi_x \in L)$. See (Kallenberg, 2021, Theorem 31.1). Notably, (Baccelli et al., 2020, Proposition 3.1.12) establishes that the point process Φ_x possesses an atom located at x, with probability one. This property enables us to interpret $\mathbf{P}^x_\Phi$ as the probability distribution of the point process Φ given that it contains an atom at x. Φ_x is commonly referred to as the *Palm version* of Φ at x. Finally, we define $\mathbf{P}^x_{\Phi^!}$ to be the distribution of the point process

$$\Phi^!_x = \Phi_x - \delta_x,$$

which is obtained by removing the point x from the Palm version Φ_x. The derived point process, $\Phi^!_x$, is known as *reduced Palm version* of Φ in x and its law, $\mathbf{P}^x_{\Phi^!}$, is called *reduced Palm kernel*. Hence, given a reduced Palm kernel, we can construct the non-reduced one by considering the distribution of $\Phi_x + \delta_x$.

The subsequent theorem is known as the Campbell-Little-Mecke (CLM) formula which extends Fubini's formula and allows the exchange of an integral and an expected value, taken with respect to the law of a point process.

Theorem C.1.1 (Campbell-Little-Mecke formula. Theorem 3.1.9 in Baccelli et al. (2020))**.** *Let Φ be a point process on $\mathbb{X}$ such that M_Φ is σ-finite, and denote by $\mathbf{P}_\Phi(\cdot)$ the distribution of Φ. Let $\left\{\mathbf{P}^x_\Phi(\cdot)\right\}_{x\in\mathbb{X}}$ be a family of Palm distributions of Φ. Then, for all measurable $g : \mathbb{X} \times \mathcal{M}_{\mathbb{X}} \to \mathbb{R}^+$, one has*

$$E\left(\int_{\mathbb{X}} g(x,\Phi)\Phi(\mathrm{d}x)\right) = \int_{\mathbb{X}} E\left(g(x,\Phi_x)\right) M_\Phi(\mathrm{d}x)\,, \tag{4.22}$$

where the expected value on the right-hand side is taken with respect to $\mathbf{P}^x_\Phi$, i.e., with respect to the law of the point process Φ_x.

Sometimes, it is useful to state Equation (4.22) in terms of the reduced Palm kernel $\mathbf{P}^x_{\Phi^!}$, i.e.,

$$E\left(\int_{\mathbb{X}} g\left(x,\Phi-\delta_x\right)\Phi(\mathrm{d}x)\right) = \int_{\mathbb{X}} E\left[g(x,\Phi^!_x)\right] M_\Phi(\mathrm{d}x)\,,$$

where the expected value on the right-hand side is taken with respect to $\mathbf{P}^x_{\Phi^!}$.

We conclude this section by introducing the multivariate extension of Equation (4.22), known as the higher order CLM formula. This requires the extension to the multivariate case of all previous definitions. Therefore, given a point process Φ define the n-th Campbell measure

$$C^n_\Phi(B\times L) = E\left(\int_B \mathbf{1}_L(\Phi)\Phi^n(\mathrm{d}\boldsymbol{x})\right),$$

for any $B \in \mathcal{X}^n$ and $L \in \sigma(\mathcal{M}_{\mathbb{X}})$ where $\mathrm{d}\boldsymbol{x} = (\mathrm{d}x_1 \ldots \mathrm{d}x_n)$ and $\Phi^n(\mathrm{d}\boldsymbol{x}) = \prod_{i=1}^n \Phi(\mathrm{d}x_i)$. Let M^n_Φ be the mean measure of Φ^n. Then, the n-th Palm distribution $\{\mathbf{P}^{\boldsymbol{x}}_\Phi\}_{\boldsymbol{x}\in\mathbb{X}^n}$ is defined as the disintegration kernel of C^n_Φ with respect to M_{Φ^n}, that is

$$C^n_\Phi(B\times L) = \int_B \mathbf{P}^x_\Phi(L) M_{\Phi^n}(\mathrm{d}\boldsymbol{x})\,.$$

The *higher order Palm version* $\Phi_{\boldsymbol{x}}$ is defined as the point process whose distribution is given by $\mathbf{P}^{\boldsymbol{x}}_\Phi$.

Consequently, the *higher order reduced Palm version* is the point process

$$\Phi^{!}_{\boldsymbol{x}} = \Phi_{\boldsymbol{x}} - \sum_{i=1}^{n} \delta_{x_i},$$

whose probability law, $\mathbf{P}^{\boldsymbol{x}}_{\Phi^{!}}$, is known as higher order reduced Palm kernel. We are finally ready to state the multivariate extension of Theorem C.1.1 and its reduced form formulation.

Theorem C.1.2 (Higher order CLM formula. Theorem 3.3.2 in Baccelli et al. (2020))**.** *Let Φ a point process on $\mathbb{X}$ such that M_{Φ^n} is σ-finite. Let $\left\{\mathbf{P}^{x}_{\Phi}(\cdot)\right\}_{\boldsymbol{x}\in\mathbb{X}^k}$ be a family of n-th Palm distributions of Φ. Then, for all measurable $g : \mathbb{R}^n \times \mathcal{M}_{\mathbb{X}} \to \mathbb{R}^+$*

$$E\left[\int_{\mathbb{X}^n} g(\boldsymbol{x}, \Phi)\Phi^n(\mathrm{d}\boldsymbol{x})\right] = \int_{\mathbb{X}^n} E\left(g(\boldsymbol{x}, \Phi_{\boldsymbol{x}})\right) M_{\Phi^n}(\mathrm{d}\boldsymbol{x}). \tag{4.23}$$

When $(x_1, \ldots, x_n)$ are distinct, we have that the n-th moment measure M_{Φ^n} coincides with the n-th factorial moment measure $M_{\Phi^{(n)}}$. In this setting, we state the CLM formula we are going to use most frequently in our computations (Baccelli et al., 2020, Theorem 3.3.6). This is when we consider the reduced Palm version of Φ in $\boldsymbol{x}$ and the integral is taken with respect to the n-th factorial power of Φ, namely,

$$E\left[\int_{\mathbb{X}^n} g\left(\boldsymbol{x}, \Phi - \sum_{i=1}^{n}\delta_{x_i}\right)\Phi^{(n)}(\mathrm{d}\boldsymbol{x})\right] = \int_{\mathbb{X}^n} E\left[g(\boldsymbol{x}, \Phi^{!}_{\boldsymbol{x}})\right] M_{\Phi^{(n)}}(\mathrm{d}\boldsymbol{x}). \tag{4.24}$$

Chapter 5

Hierarchical Mixture of Finite Mixtures

This chapter is based on Colombi et al. (2024a) and it is a joint work with Raffaele Argiento, Federico Camerlenghi and Lucia Paci.

5.1 Introduction

In the setting of statistical modeling of hierarchical data introduced in Section 4.3, we propose a Bayesian nonparametric mixture model we named Hierarchical Mixture of Finite Mixtures (HMFM). We point out that the name Hierarchical Mixture of Finite Mixtures was previously used by Miller (2014), whose model falls into a special case of the hierarchical species sampling models discussed by Bassetti et al. (2020). However, note that both Miller (2014) and Bassetti et al. (2020) use *hierarchical* to refer to the hierarchy among the random probability measures, while we refer to the hierarchy of the data.

The proposed model leverages the novel Bayesian nonparametric prior named Vector of Finite-Dirichlet Process (Vec-FDP) we here introduce, which relies on a random number of atoms shared among the different groups. Rather, the mixture weights are assumed to be group-specific in order to accommodate differences between groups. To this end, we follow the same approach as Argiento and De Iorio (2022) as we employ a more general construction for the mixing weights. This involves the normalization of positive unnormalized weights distributed according to any probability distribution over the positive real numbers. By doing so, we define a broader family of Bayesian priors, referred to as Vector of Normalized Independent Finite Point Processes (Vec-NIFPP), which encompasses the Vec-FDP as a special case when the unnormalized weights are Gamma distributed. Although we borrow Bayesian nonparametric tools to derive the distributional results and the clustering properties, the model leverages on a finite, yet random, number of mixture components, enhancing its accessibility to a broader audience, e.g., those interested in model-based clustering and mixture of experts models (Jacobs et al., 1991).

Posterior inference poses computational challenges in the context of Bayesian nonparametrics, particularly when dealing with hierarchical data. Rather, leveraging the posterior characterization of the Vec-FDP, we design an efficient MCMC sampling strategy, which significantly improves the existing methods based on infinite dimensional processes, such as the HDP. Notably, given a fixed number of clusters and groups, the HDP's computational time scales quadratically with the volume of data. In contrast, our approach achieves linear scaling.

The reason behind such improvement is evident from the restaurant franchise-like representation of

the prior. In contrast to the HDP's complex distinction between tables and menus, our representation is straightforward due to the absence of stacked layers of infinite dimensional objects in the model structure. This simplification enhances model interpretability, which is preserved when moving from a single group to multiple groups, unlike the HDP. The restaurant franchise metaphor also illuminates the flexibility of the sharing of information mechanism induced by the proposed method. In this regard, a comprehensive simulation study compares the proposed HMFM model with both the HDP mixture model and the MFM model assumed independently for each group. Experiments shed light on the advantages of employing joint modeling rather than independent analyses and demonstrate how the excessive sharing of information induced by the HDP may lead to misleading conclusions. Therefore, the HMFM strikes a balanced compromise between the HDP and the independent analyses, offering a tunable approach that combines the strengths of both methods.

The motivating example for this work comes from sports analytics. See Page et al. (2013) for Bayesian nonparametric methods in this domain. Our methodology finds application in the analysis of data from shot put, a track and field discipline that involves propelling a heavy spherical ball, or *shot*, over the greatest distance attainable. The dataset comprises measurements, specifically throw lengths or marks, recorded during professional shot put competitions from 1996 to 2016, for a total of 35, 637 marks on 403 athletes. The data are organized by aligning marks by season to ensure equitable comparison among athletes. Athletes have varying durations, with a maximum of 15 seasons, which correspond to the d different groups. We employ the proposed HMFM model to cluster athletes' performances within each season while preserving the interpretability of clusters across different seasons. This allows us to enrich the understanding of the evolution in athletes' performance trends. A remarkable finding is that the estimated clusters are gender-free, thanks to the inclusion of an additional season-specific regression parameter. Notably, the model identifies a special cluster consisting of six exceptional women's performances achieved by Olympic or world champions; no men have ever been able to outperform their competitors in such a neat way.

5.2 Hierarchical mixture of finite mixture

Given d groups (or levels), let $\boldsymbol{y}_j = \left(y_{j,1}, \ldots, y_{j,n_j}\right)$ denote the data collected over n_j individuals in group $j = 1, \ldots, d$, where $y_{j,i} \in \mathbb{Y}$ and $\mathbb{Y}$ is the sampling space. We assume that the data in each group j come from a finite mixture of M components, that is

$$y_{j,1}, \ldots, y_{j,n_j} \mid P_j \overset{\text{iid}}{\sim} \int_\Theta f(\cdot \mid \tau) P_j(\mathrm{d}\tau) \quad \text{for each } j = 1, \ldots, d, \tag{5.1}$$

where $\{f(\cdot \mid \tau),\ \tau \in \Theta\}$ is a parametric family of densities over $\mathbb{Y}$ and $(P_1, \ldots, P_d)$ is a vector of random probability measures over the parameter space $\Theta \subset \mathbb{R}^s$. We focus on random probability measures having almost surely discrete realizations. More specifically, we define a vector of finite dependent random probability measures $(P_1, \ldots, P_d)$ as follows,

$$P_j(\cdot) = \sum_{m=1}^{M} \frac{S_{j,m}}{T_j} \delta_{\tau_m}(\cdot), \tag{5.2}$$

where $S_{j,m}$ are the unnormalized weights, δ_{τ_m} stands for the delta-Dirac mass at τ_m, and $T_j = \sum_{m=1}^{M} S_{j,m}$ is referred to as the total mass.

As a prior for M, we place a 1-shifted Poisson distribution with parameter Λ, denoted by $\text{Pois}_1(\Lambda)$, so that we are sure that there always exists at least one mixture component. Then, conditionally to M, $(\tau_1, \dots, \tau_M)$ are common random atoms across the d random probability measures, which are assumed independent and identically distributed (i.i.d.) with common distribution P_0, that is a diffuse probability measure on Θ. Moreover, given M, the unnormalized weights $S_{j,m}$ are conditionally independent both within and between the groups. In particular, we assume the components of $\boldsymbol{S}_j = (S_{j,1}, \dots, S_{j,M})$ to be i.i.d. from Gamma $(\gamma_j, 1)$, independently with respect to $j = 1, \dots, d$. Throughout this work, we always refer to a shape-rate parametrization of the gamma distribution. The induced prior on the normalized mixture weights $(\pi_{j,1}, \dots, \pi_{j,M})$ is

$$(\pi_{j,1}, \dots, \pi_{j,M}) = \left(\frac{S_{j,1}}{T_j}, \dots, \frac{S_{j,M}}{T_j}\right) \sim \text{Dir}_M(\gamma_j, \dots, \gamma_j), \quad \text{for each } j = 1, \dots, d,$$

where $\text{Dir}_M(\gamma_j, \dots, \gamma_j)$ denotes the M-dimensional symmetric Dirichlet distribution with parameter γ_j. The mixing measure P_j is obtained by normalization:

$$P_j(\cdot) = \frac{\mu_j(\cdot)}{\mu_j(\Theta)}, \tag{5.3}$$

where $\mu_j(\cdot) = \sum_{m=1}^{M} S_{j,m}\delta_{\tau_m}(\cdot)$. The seminal contribution of Regazzini et al. (2003) has spurred the construction of random probability measures via the normalization approach, which turns out to be a convenient framework to face posterior inference; see, e.g., Lijoi et al. (2014) Camerlenghi et al. (2019b), Argiento et al. (2020) and Argiento and De Iorio (2022) for allied contributions. We point out that, marginally, each component P_j is a finite Dirichlet process as the one described in Argiento and De Iorio (2022). These discrete random measures are also known as Gibbs-type priors with a negative parameter (Gnedin and Pitman, 2006; De Blasi et al., 2015).

Our model construction generalizes the work of Argiento and De Iorio (2022) by allowing for the sharing of information across groups thanks to shared atoms and a shared number of components. Besides, the proposed model retains mathematical tractability thanks to the normalization approach and the conditional independence of the unnormalized weights. The prior specification for the mixture parameters $(M, \boldsymbol{S}, \boldsymbol{\tau})$, where $\boldsymbol{S} = (S_1, \dots, S_d)$ and $\boldsymbol{\tau} = (\tau_1, \dots, \tau_M)$, is equivalent to specify the joint law of the vector $(P_1, \dots, P_d)$, called *Vector of Finite Dirichlet Process* and denoted by

$$(P_1, \dots, P_d) \sim \text{Vec-FDP}(\Lambda, \boldsymbol{\gamma}, P_0), \tag{5.4}$$

where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)$. Summing up, the model can be formulated in the following hierarchical form,

$$\begin{aligned}
y_{j,1}, \dots, y_{j,n_j} \mid \boldsymbol{S}_j, \boldsymbol{\tau}, M &\overset{\text{iid}}{\sim} \sum_{m=1}^{M} w_{j,m} f(\cdot \mid \tau_m) \\
\tau_1, \dots, \tau_M \mid M &\overset{\text{iid}}{\sim} P_0(\cdot) \\
w_{j,1}, \dots, w_{j,M} \mid M, \gamma_j &\overset{\text{iid}}{\sim} \text{Dir}_M(\gamma_j, \dots, \gamma_j), \quad \text{for } j = 1, \dots, d \\
M \mid \Lambda &\sim \text{Pois}_1(\Lambda).
\end{aligned} \tag{5.5}$$

We notice that, when $d = 1$, the model in Equation (5.5) coincides with the MFM model (Miller and Harrison, 2018; Frühwirth-Schnatter et al., 2021). It follows that the proposed model is an extension of the MFM model to hierarchical data and so we refer to it as the *Hierarchical Mixture of Finite Mixture* model.

Remark The HMFM model can be framed into a more general and flexible class of Bayesian nonparametric models. In particular, the choice for the prior distribution of the number of components M and the unnormalized weights $S_{j,m}$, can be generalized with respect to the choices in Equation (5.5), while keeping the mathematical tractability. The theoretical properties of this broader family, named *Vector of Normalized Independent Finite Point Processes* are presented in Section D.2.

5.2.1 Clustering

It is worth noticing that the mixture model in Equation (5.1) can be equivalently written as

$$y_{j,i} \mid \theta_{j,i} \overset{\text{ind}}{\sim} f(\cdot \mid \theta_{j,i}), \qquad \theta_{j,i} \mid P_j \overset{\text{iid}}{\sim} P_j$$

with $i = 1, \dots, n_j$ and $j = 1, \dots, d$. Under this formulation, we get rid of the integral in Equation (5.1) by introducing a latent variable $\theta_{j,i}$ for each observation $y_{j,i}$. Conditionally on $(P_1, \dots, P_d)$, the latent variables $\theta_{j,i}$'s are i.i.d. within the same group and independent across groups. In other words, by virtue of the de Finetti representation theorem, $\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$, where $\boldsymbol{\theta}_j = (\theta_{j,1}, \dots, \theta_{j,n_j})$, is a sample from a partially exchangeable array of latent variables. This is tantamount to saying that the distribution of the $\theta_{j,i}$'s is invariant under a specific class of permutations; see (Kallenberg, 2005) and references therein. The distributional properties of $\boldsymbol{\theta}$ play a pivotal role in mixture models both in defining the clustering and devising efficient computational procedures. More precisely, since the P_j's in Equation (5.4) have discrete realizations, almost surely, the latent variables feature ties within and across groups; thus they naturally induce both a local and a global clustering of the observations.

Local clustering For each group $j = 1, \dots, d$, P_j is almost surely discrete, then ties are expected with positive probability among $\theta_{j,1}, \dots, \theta_{j,n_j}$. Let $K_j := K_{j,(n_j)}$ be the random number of distinct values in this sample. Such group-specific distinct values are collected in the set $\mathcal{T}_j = \left\{\theta^*_{j,1}, \dots, \theta^*_{j,K_j}\right\}$. Furthermore, let $\widetilde{\rho}_j = \left\{C_{j,1}, \dots, C_{j,K_j}\right\}$ be the random partition of $\{1, \dots, n_j\}$ induced by $\mathcal{T}_j$ through the following rule:

$$\theta_{j,i} \in C_{j,k} \iff \exists k \in \{1, \dots, K_j\} \text{ such that } \theta_{j,i} = \theta^*_{j,k},$$

for each $i = 1, \dots, n_j$. Note that the random partition $\widetilde{\rho}_j$ is exchangeable due to the exchangeability of $\theta_{j,1}, \dots, \theta_{j,n_j}$ and it is called *local clustering* of group j (or group-specific clustering).

Global clustering Since the P_j's share the same support, we also expect ties between groups, i.e., $\mathbb{P}\left(\theta^*_{j,k} = \theta^*_{j',k'}\right) > 0$, with $j \neq j'$. We define $\mathcal{T} = \left\{\theta^{**}_1, \dots, \theta^{**}_K\right\}$ the set of unique values among the $\theta^*_{j,k}$, $j = 1, \dots, d$ and $k = 1, \dots, K_j$; we also observe that $\mathcal{T} = \bigcup_{j=1}^{d} \mathcal{T}_j$. The corresponding random

partition $\rho = \{C_1, \ldots, C_K\}$ is induced by $\mathcal{T}$ as follows

$$\theta_{j,i} \in C_k \iff \exists k \in \{1, \ldots, K\} \text{ such that } \theta_{j,i} = \theta_k^{**},$$

for each $j = 1, \ldots, d$ and for each $i = 1, \ldots, n_j$. The random partition ρ is called *global clustering* and the number of global clusters has been denoted by $K := K_{(n_1,\ldots,n_d)}$. Since values are expected to be shared also across groups, then $K \leq \sum_{j=1}^{d} K_j$.

To shed light on the relationship between local and global clustering, we introduce $\rho_j = \{C_{j,1}, \ldots, C_{j,K}\}$ so that $C_k = \bigcup_{j=1}^{d} C_{j,k}$. Note that $C_{j,k}$ can be empty for some $k = 1, \ldots, K$. We define $n_{j,k} = |C_{j,k}|$ the number of observations of the j-th group in the k-th cluster. Then, we let $\boldsymbol{n}_j = (n_{j,1}, \ldots, n_{j,K})$ and the following hold true

$$\sum_{j=1}^{d} n_{j,k} > 0, \text{ and } \sum_{k=1}^{K} n_{j,k} = n_j, \tag{5.6}$$

for any $k = 1, \ldots, K$ and for any $j = 1, \ldots, d$, respectively.

The random partition induced by the whole sample $\boldsymbol{\theta}$ of size $n := n_1 + \cdots + n_d$ may be described through a probabilistic object called *partially Exchangeable Partition Probability Function* (pEPPF). The pEPPF, denoted here as $\Pi_K^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d)$ is the probability distribution of both the local and global clustering, where $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d$ satisfy the constraints given in Equation (5.6). The pEPPF is formally defined as follows:

$$\Pi_K^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d) := E\left[\int_{\Theta^K} \prod_{k=1}^{K} P_j^{n_{j,k}}\left(\mathrm{d}\theta_k^{**}\right)\right].$$

We refer to Camerlenghi et al. (2019b) for additional details.

Observe that, conditionally to M and $\boldsymbol{\tau}$, the unique values $\theta_1^{**}, \ldots, \theta_K^{**}$ are such that the following properties hold: (i) $K \leq M$; (ii) there exists $m \in \{1, \ldots, M\}$ such that $\theta_k^{**} = \tau_m$, for each $k = 1, \ldots, K$. Property (i) implies that a distinction between allocated mixture components (clusters) and non-allocated components is required (Nobile, 2004; Argiento and De Iorio, 2022). From property (ii), it follows that there are exactly K allocated components collected in the following set:

$$\mathcal{M}^{(a)} = \left\{m \in \{1, \ldots, M\} : \exists k \in \{1, \ldots, K\} \text{ such that } \theta_k^{**} = \tau_m\right\}.$$

5.3 Properties of the HMFM model

5.3.1 Distributional results

In this section, we derive all the theoretical properties for the latent variables $\theta_{j,i}$ modeled as follows

$$\theta_{j,i} \mid P_j \overset{\text{iid}}{\sim} P_j, \quad (P_1, \ldots, P_d) \sim \text{Vec-FDP}(\Lambda, \gamma, P_0), \tag{5.7}$$

where $j = 1, \ldots, d$ and $i = 1, \ldots, n_j$. In addition, we assume conditional independence across groups, i.e., $\boldsymbol{\theta}_j, \boldsymbol{\theta}_l \mid P_j, P_l$ are independent for $j \neq l$. The distributional results derived for the model in Equation (5.7) are the theoretical guidance to understand the clustering mechanism, to elicit the prior properly, and they play a pivotal role in devising efficient marginal and conditional algorithms to perform posterior inference. The proofs of the theoretical properties are obtained as a special case of the more general

theorems presented in Section D.3.

Before moving forward to the main results, we remind that for a Gamma$(\gamma_j, 1)$ distributed random variable, its Laplace transform $\psi_j(u_j)$ and the corresponding derivative $\kappa(u_j, n_{j,k})$ equal to

$$\psi_j\left(u_j\right) = \frac{1}{\left(1+u_j\right)^{\gamma_j}}, \quad \kappa_j\left(u_j, n_{j,k}\right) = \frac{1}{\left(1+u_j\right)^{n_{j,k}+\gamma_j}} \frac{\Gamma\left(n_{j,k}+\gamma_j\right)}{\Gamma\left(\gamma_j\right)}. \tag{5.8}$$

The almost sure discreteness of the P_j, coupled with their common supports, entails that the hierarchical sample $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ is equivalently characterized by $(\boldsymbol{\theta}^{**}, \rho)$, previously defined in Section 5.2.1. The following theorem specifies the distribution of the pEPPF for the HMFM model.

Theorem 5.3.1 (pEPPF)**.** *The probability to observe a sample $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ of size n from Equation (5.7) exhibiting K distinct values $\left(\theta_1^{**}, \ldots, \theta_K^{**}\right)$ with respective counts $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d$ is given by the following pEPPF*

$$\Pi_K^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d) = V(K; \gamma, \Lambda) \prod_{j=1}^{d} \prod_{k=1}^{K} \frac{\Gamma\left(n_{j,k}+\gamma_j\right)}{\Gamma\left(\gamma_j\right)}, \tag{5.9}$$

where $V(K; \gamma, \Lambda)$ equals

$$V(K; \gamma, \Lambda) = \int_{(\mathbb{R}^+)^d} \Psi(K, \boldsymbol{u}) \prod_{j=1}^{d} \frac{u_j^{n_j-1}}{\Gamma\left(n_j\right)} \frac{1}{\left(1+u_j\right)^{n_j+K\gamma_j}} \mathrm{d}u_1 \ldots \mathrm{d}u_d,$$

and, setting $\psi(\boldsymbol{u}) = \prod_{j=1}^{d} \psi_j\left(u_j\right)$, $\Psi(K, \boldsymbol{u})$ is defined as follows

$$\Psi(K, \boldsymbol{u}) \;=\; \Lambda^{K-1}\left(K + \Lambda\psi(\boldsymbol{u})\right) e^{-\Lambda(1-\psi(\boldsymbol{u}))}. \tag{5.10}$$

See Section D.3.1 for an extended discussion and derivation of $\Psi(K, \boldsymbol{u})$.

The predictive distributions, i.e., the distribution of θ_{j,n_j+1} given $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ for each possible group j, easily follow from Theorem 5.3.1. These unveil important intuitions about the clustering mechanism, and they are the building block for a marginal posterior sampler; we defer their detailed presentation to Section 5.3.3. Recall that in Section 5.2.1 we defined the number of global clusters, denoted as $K_{(n_1,\ldots,n_d)}$. This is a random quantity and in the case of $d = 2$ groups whose cardinalities are n_1 and n_2, respectively, we can compute an explicit expression for the prior distribution of $K_{(n_1,n_2)}$.

Theorem 5.3.2 (Prior distribution of global number of clusters)**.** *Consider $d = 2$ groups of observations with size n_1 and n_2, respectively. Then, the prior distribution for the global number of clusters $K_{(n_1,n_2)}$ is*

$$\begin{aligned} &\mathbb{P}\left(K_{(n_1,n_2)} = K\right) \\ &\quad = V(K; \gamma, \Lambda) \sum_{r_1=0}^{K} \sum_{r_2=0}^{K-r_1} \binom{K-r_1}{r_2} \frac{(K-r_2)!}{r_1!} \prod_{j=1}^{2} \left|C(n_j, K-r_j; -\gamma_j)\right| \end{aligned} \tag{5.11}$$

where for any non-negative integers $n \geq 0$ and $0 \leq K \leq n$, $C(n, K; -\gamma_j)$ denotes the central generalized factorial coefficients. See Charalambides (2002).

The following theorem provides functionals of $(P_1, \ldots, P_d)$, which are useful for prior elicitation and to shed light on the dependence structure introduced by our prior.

Theorem 5.3.3 (Mixed moments). *Let* $(P_1, \dots, P_d) \sim$ Vec-FDP (Λ, γ, P_0) *be a vector of normalized random probability measures defined through normalization as in Equation* (5.3)*. Then the following hold:*

(i) for any measurable sets A, B *and for any* $j, l \in \{1, \dots, d\}$*,*

$$E\left[P_j(A)P_l(B)\right] = \mathbb{P}\left(K_{(1,1)} = 1\right)\left(P_0(A \cap B) - P_0(A)P_0(B)\right); \tag{5.12}$$

(ii) for any measurable set A *and for any* $j, l \in \{1, \dots, d\}$*,*

$$E\left[P_j(A)^{n_j}P_l(A)^{n_l}\right] = E\left[P_0(A)^{K_{(n_j,n_l)}}\right] = \sum_{k=1}^{n_j+n_l} P_0(A)^k \mathbb{P}\left(K_{(n_j,n_l)} = k\right), \tag{5.13}$$

where $K_{(n_j,n_l)}$ *is the global number of clusters across two groups with size* n_j *and* n_l *and it can be evaluated via Equation* (5.11)*.*

Furthermore, both Equations (5.12) and (5.13) can be extended to the case of more than two groups. As a byproduct of Theorem 5.3.3, we obtain a closed form expression for pairwise correlation between the components of $(P_1, \dots, P_d)$ evaluated on specific sets. Let A be a measurable set, then, for any $j, l \in \{1, \dots, d\}$:

$$\text{corr}\left(P_j(A), P_l(A)\right) = \frac{1 - e^{-\Lambda}}{\Lambda\left(\gamma_j + 1\right)\left(\gamma_l + 1\right) I\left(\gamma_j, \Lambda\right) I\left(\gamma_l, \Lambda\right)}, \tag{5.14}$$

where $I\left(\gamma_j, \Lambda\right) = \int_0^1 (1 + \Lambda x)\, e^{-\Lambda(1-x)}(1 - x^{1/\gamma_j})\mathrm{d}x$. The expression in Equation (5.14) does not depend on the choice of the set A. Thus, it may be considered an overall measure of dependence between the two random probability measures. The limits of Equation (5.14) when both γ_j and γ_l goes to 0 and $+\infty$ equal to

$$\lim_{\gamma_j, \gamma_l \to 0} \text{corr}\left(P_j(A), P_l(A)\right) = \frac{1 - e^{-\Lambda}}{\Lambda}, \quad \lim_{\gamma_j, \gamma_l \to \infty} \text{corr}\left(P_j(A), P_l(A)\right) = 1. \tag{5.15}$$

These limits are interesting because we see that, given Λ, decreasing γ_j and γ_l, the correlation does not go to 0 but reaches a lower bound that depends on Λ, which, in turn, goes to 0 if Λ increases. On the other hand, increasing values of γ_j and γ_l lead correlation equal to 1, regardless of the value of Λ. See the left panel in Figure 5.1.

5.3.2 Graphical representation of the correlation function

We now provide a graphical visualization of the correlation function. Equation (5.14) depends on three parameters, γ_1, γ_2 and Λ. For graphical convenience, we always refer to the case when γ_1 and γ_2 are set to a common value, namely γ. In the left panel of Figure 5.1, we show the correlation function in Equation (5.14) evaluated on a grid of values of γ and fixed values of Λ. Then, in the right panel of the same figure, we do the opposite, evaluating the correlation over a grid of values of Λ and some fixed values of γ. In the first case, curves are monotonically increasing functions of γ, whose lowest value is given in the first row of Equation (5.15). They rapidly increase for small values of γ and then slowly

tend to their limiting value, that is 1, as shown in the second row of Equation (5.15). The right panel of Figure 5.1 depicts a function that monotonically decrease as Λ increases. In particular, we note that the curve representing $\gamma = 1$ is much higher than the other ones, showing how crucial is this parameter in determining the value of the correlation. To conclude, the right panel of Figure 5.1 shows the case when both Λ and γ are free to change. We refer to Section D.3.6 for the proofs of Equations (5.14) and (5.15) and a generalization of Equation (5.14) .

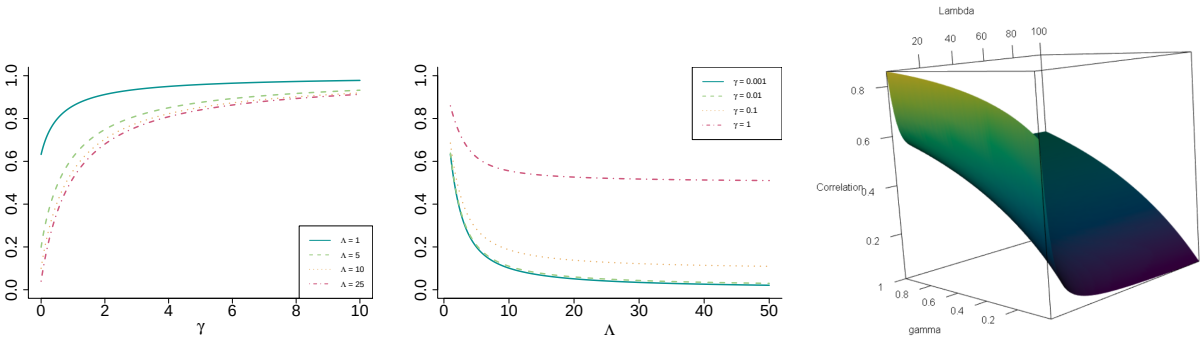

Figure 5.1: Left panel: correlation for $\gamma_1 = \gamma_2$ varying over a grid of values. Each curve is obtained by fixing Λ to the values reported in the legend. Middle panel: correlation for Λ varying over a grid of values. Each curve is obtained by fixing $\gamma_1 = \gamma_2$ to the values reported in the legend. Right panel: correlation function as a function of $\gamma_1 = \gamma_2$ and Λ.

We now aim at giving a posterior characterization for a vector $(P_1, \ldots, P_d)$ distributed as in Equation (5.4). Since $(P_1, \ldots, P_d)$ is obtained via normalization of $(\mu_1, \ldots, \mu_d)$, it is sufficient to provide a posterior characterization for the latter vector. In order to do this, we follow the same approach as James et al. (2009), Camerlenghi et al. (2019b) and Argiento and De Iorio (2022). Thus, we introduce a vector of auxiliary variables $\boldsymbol{U}_n = (U_1, \ldots, U_d)$ such that $U_j \mid T_j \overset{\text{ind}}{\sim} \text{Gamma}\left(n_j, T_j\right)$, where $T_j = \mu_j(\mathbb{X})$. This is possible since the marginal distribution of $\boldsymbol{U}_n$ does exist, see Section D.3.7. Hence, conditionally to $\boldsymbol{U}_n$ and to $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$, $(\mu_1, \ldots, \mu_d)$ is a superposition of two independent processes, one driving the non-allocated components and the other one driving the allocated components.

Theorem 5.3.4 (Posterior representation)**.** *Let $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ be a sample from the statistical model in Equation (5.7). Then, the posterior distribution of $(\mu_1, \ldots, \mu_d)$ is characterized as the superposition of two independent processes on $(\mathbb{R}^+)^d \times \Theta$:*

$$(\mu_1, \ldots, \mu_d) \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \boldsymbol{U}_n \overset{\text{d}}{=} \left(\mu_1^{(a)}, \ldots, \mu_d^{(a)}\right) + \left(\mu_1^{(na)}, \ldots, \mu_d^{(na)}\right), \text{ where:}$$

(i) the process of allocated components $\left(\mu_1^{(a)}, \ldots, \mu_d^{(a)}\right)$ equals

$$\mu_j^{(a)} = \sum_{k=1}^{K} S_{j,k}^{(a)} \delta_{\theta_k^{**}}, \text{ as } j = 1, \ldots, d,$$

where the random variables $S_{j,k}^{(a)} \mid \boldsymbol{U}_n \overset{ind}{\sim} \text{Gamma}(n_{j,k} + \gamma_j, u_j + 1)$, for each $j \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, K\}$;

(ii) the process of non-allocated components $\left(\mu_1^{(na)}, \ldots, \mu_d^{(na)}\right)$ *equals*

$$\mu_j^{(na)} = \sum_{m^*=0}^{M^*} S_{j,k}^{(na)} \delta_{\tau_{m^*}}, \text{ as } j = 1, \ldots, d.$$

In particular, M^* *is a random variable distributed as a mixture of Poisson distributions, namely*

$$M^* \sim (1 - w_K(\boldsymbol{u})) \operatorname{Pois}_1\left(\Lambda \prod_{j=1}^{d} \psi_j\left(u_j\right)\right) + w_K(\boldsymbol{u}) \operatorname{Pois}\left(\Lambda \prod_{j=1}^{d} \psi_j\left(u_j\right)\right),$$

where we have set $w_K(\boldsymbol{u}) := K/(\Lambda \prod_{j=1}^{d} \psi_j\left(u_j\right) + K)$*, and the random variables* $S_{j,1}^{(na)}, \ldots, S_{j,M^*}^{(na)} \mid \boldsymbol{U}_n, M^* \overset{iid}{\sim} \operatorname{Gamma}(\gamma_j, u_j + 1)$*, for* $j \in \{1, \ldots, d\}$.

Note that for the process of non-allocated components $\mathbb{P}(M^* = 0) > 0$ even if $\mathbb{P}(M = 0) = 0$. Hence, it is possible to have zero non-allocated components.

5.3.3 Predictive distribution and franchise metaphor

Further intuitions on the cluster mechanism described in Section 5.2.1 are available when considering the predictive distributions. Consider a realization $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ with K distinct values $\left(\theta_1^{**}, \ldots, \theta_K^{**}\right)$ and a partition $\rho = \{C_1, \ldots, C_K\}$ with counts $(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d)$ satisfying the constraints in Equation (5.6). Following the approach of James et al. (2009), Favaro and Teh (2013) and Argiento and De Iorio (2022) we work conditionally to $\boldsymbol{U}_n = \boldsymbol{u}$. Then, for each group, say j, we have

$$\begin{aligned} &\mathbb{P}\left(\theta_{j,n_j+1} \in \cdot \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \boldsymbol{u}, \boldsymbol{\gamma}, \Lambda\right) \\ &\qquad \propto \sum_{k=1}^{K} \left(n_{j,k} + \gamma_j\right) \delta_{\theta_k^{**}}(\cdot) + \psi(\boldsymbol{u}) \gamma_j \Lambda \frac{K + 1 + \Lambda\psi(\boldsymbol{u})}{K + \Lambda\psi(\boldsymbol{u})} P_0(\cdot). \end{aligned} \tag{5.16}$$

Such a predictive distribution can be interpreted in terms of a restaurant franchise metaphor. Consider a franchise of d Chinese restaurants each with possibly infinitely many tables. Here $\theta_{j,i}$ represents the dish served to customer i in restaurant j, and each θ_k^{**} represents a dish. All customers sitting at the same table must eat the same dish. The same dish can not be served at different tables in the same restaurant, but it can be served across different restaurants. According to the predictive law, the first customer entering the first restaurant sits at the first table eating dish $\theta_{1,1} = \theta_1^{**}$, which is drawn from P_0. At the same time, an empty table serving dish θ_1^{**} must be prepared in all the other restaurants: this step corresponds to the first cluster allocation, i.e., C_1. Then, the second customer of the first restaurant arrives and can either: (i) sit at the same table as the first customer, with probability proportional to $1 + \gamma_1$ or (ii) sit at a new table with probability proportional to $\psi(\boldsymbol{u})\gamma_1\Lambda\frac{2+\Lambda\psi(\boldsymbol{u})}{1+\Lambda\psi(\boldsymbol{u})}$. In the latter case, the customer chooses a new dish θ_2^{**}, drawn from P_0, and the number of clusters K is increased by 1; moreover, an empty table serving dish θ_2^{**} must be prepared in all the other restaurants of the franchise. Then, the process evolves according to Equation (5.16). Figure 5.2 displays a graphical representation of the process.

Interestingly, our model is more parsimonious than the HDP by Teh et al. (2006) in sharing information across restaurants. While the HDP relies on the popularity of a dish throughout the entire franchise to influence a new customer's choice, in our model, such probability depends on the sample only through

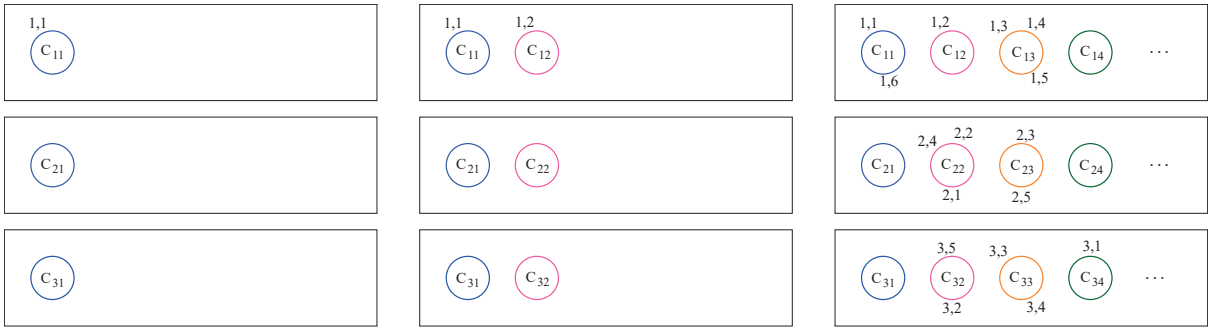

Figure 5.2: Chinese restaurant franchise process representation for Vec-FDP based on a sample of six, five and five observations in $d = 3$ groups, respectively.

the dish's popularity within the specific restaurant the customer enters. This distinctive feature proves to be appealing as it mitigates the excessive borrowing of information across groups that is induced by hierarchical processes. Subsequent numerical experiments highlight the advantage of our model by showing that the HDP can lead to misleading results in posterior inference. An additional advantage of our model with respect to the HDP is that different tables within the same restaurant cannot serve the same dish. This simplifies the local clustering structure, and it improves the computational efficiency in posterior sampling, as discussed in Section 5.4.2. We conclude by reminding that there are situations where the stronger capability of sharing information of the HDP can be a preferable feature. Experiment 1 in Section 5.5.2 is one such case, although the HMFM still remains a competitive alternative.

5.4 Fitting details

5.4.1 Hyperpriors

We consider the following hyperpriors for the process hyperparameters (Λ, γ),

$$\pi(\Lambda, \gamma) = \pi(\gamma \mid \Lambda)\pi(\Lambda) = \prod_{j=1}^{d} \text{Gamma}(a_\gamma, \Lambda b_\gamma) \times \text{Gamma}(a_\Lambda, b_\Lambda). \tag{5.17}$$

The prior distribution in Equation (5.17) extends, to our setting, the prior choice introduced by Frühwirth-Schnatter and Malsiner-Walli (2019) to encourage sparsity in the mixture, whose advantages have been studied both theoretically (Rousseau and Mengersen, 2011; Van Havre et al., 2015) and empirically (Malsiner-Walli et al., 2016, 2017). Furthermore, this prior formulation assumes the γ_js to be conditionally independent given Λ, so tuning the sharing of information between groups, see also Figure 5.1. In particular, note that $\Lambda \mid \gamma$ is still Gamma distributed, i.e., $\Lambda \mid \gamma \sim \text{Gamma}\left(a_\Lambda + da_\gamma, b_\Lambda + b_\gamma \sum_{j=1}^{d} \gamma_j\right)$, which yields tractable posterior inference.

We rely on the sample equivalence principle (Diaconis and Ylvisaker, 1979) to design a suitable reparametrization of the prior

$$\Lambda \mid \gamma \sim \text{Gamma}\left(a_\Lambda + da_\gamma, b_\Lambda + b_\gamma \sum_{j=1}^{d} \gamma_j\right),$$

with an easier interpretation. Indeed, it is straightforward to show that

$$E[\Lambda \mid \gamma] = \left(\frac{1}{E[\Lambda]} \frac{a_\Lambda / a_\gamma}{a_\Lambda / a_\gamma + d} + \frac{1}{\widehat{\Lambda}} \frac{d}{a_\Lambda / a_\gamma + d} \right)^{-1},$$

where $E[\Lambda]$ is the prior mean of Λ, i.e., a_Λ / b_Λ and $\widehat{\Lambda} = \frac{a_\gamma}{b_\gamma} \frac{d}{\sum_{j=1}^{d} \gamma_j}$ is the maximum likelihood estimator for Λ based on a sample $\gamma_1, \ldots, \gamma_d$. Hence, a_Λ / a_γ represents the size of a prior sample and it can be compared to the size d. Let Λ_0 and V_Λ be the prior expected value and variance of Λ, respectively, and γ_0 be a common prior guess for the γ_j's, so that $\frac{1}{d} \sum_{j=1}^{d} \gamma_j = \gamma_0$. Then, assuming $E[\Lambda] = \widehat{\Lambda}$, we obtain

$$a_\gamma = \frac{1}{d} \frac{\Lambda_0^2}{\mathrm{V}_\Lambda}, \quad b_\gamma = \frac{a_\gamma}{\gamma_0 \Lambda_0}, \quad a_\Lambda = \frac{\Lambda_0^2}{\mathrm{V}_\Lambda}, \quad b_\Lambda = \frac{\Lambda_0}{\mathrm{V}_\Lambda}.$$

We now provide an intuition for choosing Λ_0, V_Λ and γ_0. We would like our mixture to be sparse, which, in practice, means that the expected number of clusters is (much) smaller than the expected number of mixture components. Under the HMFM model, this translates to small values of γ_j's. Looking at each season separately, we note that each element of the is a finite DP. Hence, it is easy to compute the prior distribution for each local number of cluster K_j, using Equation (5.11) with $n_2 = 0$, and compare it with the law of M. Setting a small number of clusters for each season implies being a priori noninformative about the heterogeneity across different seasons.

5.4.2 Computational methods

As customary in hierarchical modeling, we introduce latent allocation vectors $\boldsymbol{c}_j = \left(c_{j,1}, \ldots, c_{j,n_j}\right)$ whose element $c_{j,i} \in \{1, \ldots, M\}$ denotes to which component observation $y_{j,i}$ is assigned, for each $j = 1, \ldots, d$. Setting $\theta_{j,i} = \tau_{c_{j,i}}$, we are able to link the mixture parameters and the observation-specific parameters. We suggest two MCMC strategies to carry out posterior inference for mixture modeling.

Conditional algorithm The first strategy we present is a conditional algorithm that provides full Bayesian inference on both the mixing parameters $(P_1, \ldots, P_d)$ and the clustering structure ρ. Namely, we draw a sample of the vector of random probability measures $(P_1, \ldots, P_d)$ from its posterior distribution given in Theorem 5.3.4 by sampling from the joint posterior distribution of $(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_d, \boldsymbol{\tau}, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_d, M)$. Note that the global number of clusters K is automatically deduced from the cluster allocation vectors $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_d)$. To do so, we resort to auxiliary variables $\boldsymbol{U}_n$ and the hyperparameters (Λ, γ). For the sake of brevity, we denote $\boldsymbol{S} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_d)$ and $\boldsymbol{c} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_d)$. We adopt a blocked Gibbs sampling strategy. In particular, let $\Delta = (\boldsymbol{S}, \boldsymbol{\tau}, \boldsymbol{c}, M, \boldsymbol{U}, \gamma, \Lambda)$ be a vector collecting all the parameters and let $\boldsymbol{y}$ be the collection of all variables $y_{j,i}$, for each group j and for each individual i. We partition Δ in two blocks $\Delta_1 = (\boldsymbol{S}, \boldsymbol{\tau}, \boldsymbol{c}, \Lambda)$ and $\Delta_2 = (\boldsymbol{U}, M, \gamma)$. We use the following notation, $\mathcal{L}(X \mid \text{rest})$, to indicate the distribution of a random variable X conditionally to everything but itself. The algorithm is a blocked Gibbs sampler whose state space is $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_d, \boldsymbol{\tau}, \boldsymbol{S}, M, \boldsymbol{U}, \Lambda, \gamma)$. The algorithm is composed of two main steps.

Step 1: Metropolis-within-Gibbs algorithm to update one parameter at a time.

1.1 Update cluster allocations $\boldsymbol{c}$. This step is done independently for each group j and for each

individual i by sampling $c_{j,i}$ from a discrete distribution such that

$$\mathbb{P}(c_{j,i} = m \mid \text{rest}) \propto S_{j,m} f(y_{j,i} \mid \tau_m), \quad \text{for } m = 1, \ldots, M.$$

After sampling the whole vector, compute the number of allocated components K. This operation takes $O(Mn)$ time.

1.2 Update mixture parameters $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)$. From Theorem 5.3.4, the full conditional distribution further factorizes in two independent parts, one for the first K allocated components and one for the remaining M^* non-allocated components. In particular, for $k = 1, \ldots, K$, sample τ_k independently from

$$\mathbb{P}\left(\tau_k \in d\tau_k \mid \text{rest}\right) \propto \left\{ \prod_{(j,i):c_{j,i}=k} f(y_{j,i} \mid \tau_k) \right\} P_0(d\tau_k)$$

while for $m = K+1, \ldots, M$, sample τ_m independently from the prior P_0. In general, to find elements belonging to some cluster k requires performing a search a vector of length n, which takes $O(n)$ time. Hence, overall, this operation would take $O(Kn)$ time. Nevertheless, a smart implementation of the code can avoid such a search. Therefore, in practice, the computational time of this update scales sublinearly with respect to n.

1.3 Update the unnormalized mixture weights $\boldsymbol{S}$ differently for the allocated and non-allocated parts of the process. In particular, for $k = 1, \ldots, K$, sample $S_{j,k}$ from $\text{Gamma}(\gamma_j + n_{j,k}, U_j + 1)$, independently for each $j = 1, \ldots, d$, while for $m = K+1, \ldots, M$, sample $S_{j,m}$ from $\text{Gamma}(\gamma_j, U_j + 1)$. This operation takes $O(Md)$ time.

1.4 Update Λ from the mixture of gamma densities

$$\begin{aligned} &\frac{K\,(b^* + 1 - \psi(\boldsymbol{u}))}{(a^* - 1)(\psi(\boldsymbol{u}) + K(b^* + 1)} \,\text{Gamma}\,(\Lambda; a^* + K - 1, b^* + 1 - \psi(\boldsymbol{u})) \\ &\quad + \frac{(a^* + K - 1)(\psi(\boldsymbol{u})}{(a^* - 1)\psi(\boldsymbol{u}) + K(b^* + 1)} \,\text{Gamma}\,(\Lambda; a^* + K, b^* + 1 - \psi(\boldsymbol{u})) \end{aligned}$$

where $a^* = a_\Lambda + da_\gamma$; $b^* = b_\Lambda + b_\gamma \sum_{j=1}^d \gamma_j$. This update is straightforward as it only takes $O(1)$ time.

Step 2: To avoid conditioning to $\boldsymbol{U}$ when updating $\boldsymbol{\gamma}$ and M, we write the full conditional distribution of Δ_2 as $\mathcal{L}\,(\boldsymbol{U} \mid \text{rest})\,\mathcal{L}\,(M, \boldsymbol{\gamma} \mid \Delta_1, \boldsymbol{y})$.

2.1 Update U_j from $\text{Gamma}(n_j, T_j)$, independently for each $j = 1, \ldots, d$. This operation takes $O(d)$ time.

2.2 Update $M = K + M^*$, where K has been updated in Step 1.1 and M^* is sampled from

$$\mathbb{P}\left(M^* = m^* \mid \boldsymbol{\gamma}, \Delta_2, \boldsymbol{y}\right) \propto \frac{m^* + K}{m^*!} \Lambda^{m^*} \prod_{j=1}^d \frac{\Gamma\left(\gamma_j(m^* + K)\right)}{\Gamma\left(\gamma_j(m^* + K) + n_j\right)}.$$

This step is accomplished using the an adaptive Metropolis-Hastings step. This update takes $O(1)$ time.

2.3 Update γ_j from a density proportional to

$$\frac{\Gamma\left(\gamma_j(M^*+K)\right)}{\Gamma\left(\gamma_j(M^*+K)+n_j\right)}\prod_{k=1}^{K}\left(\frac{\Gamma\left(n_{j,k}+\gamma_j\right)}{\Gamma\left(\gamma_j\right)}\right)\text{Gamma}(\gamma_j;a_\gamma,\Lambda b_\gamma),$$

independently for each $j = 1,\dots,d$. This update is performed via the adaptive Metropolis-Hastings step. This operation takes $O(d)$ time.

Overall, the computational time is driven by the update of cluster allocation variables and the unique values. As discussed above, in practice the latter can be easily reduced. In general, each iteration takes $O((M+K)n)$ even though in practice the computational time is much closer to $O(Mn)$.

Marginal algorithm The second algorithm is a marginal sampler that simplifies the computation by integrating the mixture parameters while providing inference on the sole clustering structure. The algorithm is derived from the predictive distributions detailed in Section 5.3.3, and the full conditional distribution of $\boldsymbol{U}_n$ given in Theorem D.2.3. The algorithm is based on the Chinese restaurant franchise characterization of the Vector of Finite Dirichlet process. The state space of the Markov Chain is identified by the parameters $(\boldsymbol{c}_1,\dots,\boldsymbol{c}_d,\boldsymbol{U},\Lambda,\boldsymbol{\gamma})$. The algorithm is composed of four steps:

Step 1: Update cluster allocations $\boldsymbol{c}$. This is done in $n=\sum_{j=1}^{d}n_j$ sub-steps. For each restaurant $j=1,\dots,d$ and for each customers $i=1,\dots,n_j$, we remove observation i from the cluster and denote by $\rho_j^{-i}=\left\{C_{j,1}^{-i},\dots,C_{j,K^{-i}}^{-i}\right\}$ the resulting partition in K^{-i} clusters. Then, the i-th observation is assigned to a new cluster $C_{j,K^{-i}+1}^{-i}$ with probability proportional to

$$\mathbb{P}\left(i\in C_{j,K^{-i}+1}^{-i}\mid \boldsymbol{u},\rho_j^{-i},\boldsymbol{y}\right)\propto\psi(\boldsymbol{u})\gamma_j\Lambda\frac{K^{-i}+1+\Lambda\psi(\boldsymbol{u})}{K^{-i}+\Lambda\psi(\boldsymbol{u})}\mathcal{M}\left(y_{j,i}\right)$$

while it is assigned to each of the existing clusters $C_{j,l}^{-i}$, for $l=1,\dots,K^{-i}$, with probability proportional to

$$\mathbb{P}\left(i\in C_{j,l}^{-i}\mid \boldsymbol{u},\rho_j^{-i},\boldsymbol{y}\right)\propto\left(n_{j,k}^{-i}+\gamma_j\right)\frac{\mathcal{M}\left(\boldsymbol{y}_{C_l^{-i}\cup i}\right)}{\mathcal{M}\left(\boldsymbol{y}_{C_l^{-i}}\right)}$$

where $\psi(\boldsymbol{u})=\prod_{j=1}^{d}\psi_j\left(u_j\right)$ and, for each global cluster $C_l\in\rho$, $\boldsymbol{y}_{C_l}$ is the vector of the $y_{h,g}$, $h\in\{1,\dots,d\}$ and $g\in\{1,\dots,n_h\}$, such that $h,g\in C_l$ and

$$\mathcal{M}\left(\boldsymbol{y}_{C_l}\right)=\int_{\mathbb{X}}\prod_{(h,g)\in C_l}f\left(y_{h,g}\mid\theta\right)P_0(d\theta)$$

is the marginal distribution of the data within the cluster C_l with sampling model $f\left(y_{h,g}\mid\theta\right)$ and prior $P_0(d\theta)$. The computational time is proportional to $(K+1)n$ times the evaluation time for the marginal distribution $\mathcal{M}(\cdot)$). This latter cost may be problem specific. In our experiments and application, $\mathcal{M}(\cdot)$ is available in closed analytical form and its parameters are easily computed by updating the sufficient statistics of the likelihood f. Hence, for our purposes, this update takes $O((K+1)n)$.

Step 2: Update the whole vector $\boldsymbol{U}$ by sampling from a density proportional to

$$\prod_{j=1}^{d}\left\{U_j^{n_j-1}\frac{1}{(1+U_j)^{n_j+K\gamma_j}}\right\}\left(K+\Lambda\prod_{j=1}^{d}\frac{1}{(1+U_j)^{\gamma_j}}\right)\exp\left\{\Lambda\prod_{j=1}^{d}\frac{1}{(1+U_j)^{\gamma_j}}\right\}.$$

This operation takes $O(d)$ time.

Step 3: Update the whole vector $\boldsymbol{\gamma}$ by sampling from a density proportional to

$$\prod_{j=1}^{d}\left\{\text{Gamma}\left(\gamma_j;a_\gamma,\Lambda b_\gamma\right)\frac{1}{(1+u_j)^{K\gamma_j}}\prod_{m=1}^{K}\frac{\Gamma(\gamma_j+n_{j,m})}{\Gamma(\gamma_j)}\right\}$$
$$\times\left(K+\Lambda\prod_{j=1}^{d}\frac{1}{(1+U_j)^{\gamma_j}}\right)\exp\left\{\Lambda\prod_{j=1}^{d}\frac{1}{(1+U_j)^{\gamma_j}}\right\}$$

This operation takes $O(d)$ time.

Step 4: Update Λ from the mixture of gamma densities

$$\frac{K\,(b^*+1-\psi(\boldsymbol{u}))}{(a^*-1)(\psi(\boldsymbol{u})+K(b^*+1)}\,\text{Gamma}\,(\Lambda;a^*+K-1,b^*+1-\psi(\boldsymbol{u}))$$
$$+\frac{(a^*+K-1)\psi(\boldsymbol{u})}{(a^*-1)(\psi(\boldsymbol{u})+K(b^*+1)}\,\text{Gamma}\,(\Lambda;a^*+K,b^*+1-\psi(\boldsymbol{u}))$$

where $a^*=a_\Lambda+da_\gamma$; $b^*=b_\Lambda+b_\gamma\sum_{j=1}^{d}\gamma_j$. This update is straightforward as it only takes $O(1)$ time. Note that both Step 2 and 3 require sampling from d-dimensional distributions which are not available in closed form and which can not be factorized in the product of d independent components. Joint updates must be performed. We suggest to use adaptive Metropolis-Hastings techniques or gradient-based methods. In particular, we use here the Metropolis Adjusted Langevin Algorithm (MALA) introduced by Roberts and Rosenthal (1998). Overall, each iteration takes $O((K+1)n)$.

Notably, the HMFM achieves linear scaling, taking $O(n(M+K))$ and $O(n(K+1))$ time for one iteration of the conditional and marginal sampler, respectively. In contrast, the HDP faces scalability issues that may eventually limit its practical feasibility. A naive implementation of HDP, based on the traditional Chinese restaurant franchise process, takes $O(n^2)$ time. To overcome this computational burden, Teh et al. (2006) also propose a direct assignment scheme whose computational bottleneck is the computation of the unsigned Stirling numbers $s(n_{j,k},m)$ for each $j=1,\dots,d$, $k=1,\dots,K$ and for all positive integers $m\leq n_{j,k}$, where $n_{j,k}$ is the number of observations in group j assigned to cluster k. The computational time to compute $s(n_{j,k},m)$ is $O((n_{j,k})^2)$. By doing so, the quadratic cost of the algorithm is deferred to the calculation of the Stirling numbers which, however, can be precomputed and saved. Once they are available, the cost per iteration is also linear for the HDP. Thus, while the precomputation of Stirling numbers makes HDP competitive with HMFM for moderate values of n the linear complexity of our proposed method makes it more scalable and appealing for large datasets. The R code implementing both MCMC algorithms is available at `https://github.com/alessandrocolombi/HMFM`, along with the simulation study.

5.5 Simulation study

We carried out an extensive simulation study comparing the HMFM in Equation (5.5) with: (i) the HDP (Teh et al., 2006) mixture model; (ii) the MFM model assumed independently for each group (MFM-indep); (iii) the MFM model assumed for the pooled data, i.e., ignoring the groups (MFM-pooled). The MFM model is fitted using the algorithm by Argiento and De Iorio (2022). Regarding the HMFM, we investigate the performance of both the conditional (HMFM-cond) and the marginal (HMFM-marg) sampler.

In summary, the simulation study consists of three experiments: the first one considers $d = 2$ groups that share a component and clearly shows the advantage of the joint modeling approach against both independent group-specific analyses and pooled analyses. The second experiment is an illustrative example with again $d = 2$ groups, but without any components shared across the groups. The study highlights the limitations of the HDP in situations where borrowing information from other groups can lead to misleading conclusions. Rather, this issue is mitigated by the HMFM that borrows less information across the groups relative to the HDP. The third experiment generates data from $d = 15$ groups and further evidences how the HMFM outperforms the HDP

5.5.1 Experimental setting and performance evaluation

For a fair comparison, we fit all competing models by setting the same base measure P_0. In particular, we follow the strategy explained in Section 5.2 and set a Normal-InvGamma prior with $\mu_0 = \text{mean}(\mathbf{y})$, $k_0 = 1/(\text{range}(\mathbf{y}))^2$, $\nu_0 = 4$ and $\sigma_0^2 = 1/2$. Regarding the hyperparameters, for the HDP we set default hyperprior $\alpha \sim \text{Gamma}(1, 1)$ and $\gamma \sim \text{Gamma}(1, 0.1)$ (Teh et al., 2006). Also, we use default hyperpriors for both the independent and pooled analyses, as suggested in the `AntMAN` package, i.e., $\gamma \sim \text{Gamma}(1, 1)$ and $\Lambda \sim \text{Gamma}(1, 1)$. Finally, for the HMFM, we follow the strategy described in Section 5.4.1 to set the values of Λ_0, V_Λ and γ_0; the values are detailed for each experiment below.

To assess the performance of recovering the true clustering, we compute the Co-clustering error (CCE; Dahl 2006; Bassetti et al. 2020) and the Adjusted Rand Index (ARI; Hubert and Arabie 1985). The CCE is based on the posterior probabilities $\hat{\pi}_{lk} \in [0, 1]$ of two data points l and k to belong to the same cluster, i.e., the proportion of times that two observations have been assigned to the same mixing component over the MCMC iterations. Specifically, the error is defined as the average L_1 distance between the true pairwise co-clustering matrix with elements $\pi_{lk} \in \{0, 1\}$ and the estimated co-clustering probability (similarity), namely

$$\text{CCE} = \frac{1}{n^2} \sum_{l=1}^{n} \sum_{k=1}^{n} | \, \pi_{lk} - \hat{\pi}_{lk} \, |,$$

where $n = \sum_{j=1}^{d} n_j$. Being a distance, lower values of CCE indicate better fit, attaining zero in the absence of co-clustering error.

On the other hand, the ARI is computed using the estimated partition which is obtained by minimizing the variation of information function (Wade and Ghahramani, 2018; Dahl et al., 2022). The index ranges between zero (the true and the estimated clustering do not agree on any pair of points) and one (the two clusterings are the same).

5.5.2 Experiment 1

This experiment considers $d = 2$ groups, both having two local clusters ($K_1 = 2$, $K_2 = 2$) one of which is shared; hence, the global number of clusters is $K = 3$. The mixing probabilities are set so that the shared component has a lower value in the second group. In particular, 50 independent datasets are generated from

$$y_{1,i} \overset{\text{iid}}{\sim} 0.5\text{N}(-3, 0.1) + 0.5\text{N}(0, 0.5) \text{ and } y_{2,i} \overset{\text{iid}}{\sim} 0.2\text{N}(0, 0.5) + 0.8\text{N}(1.75, 1.5), \text{ for } i = 1, \dots, 300.$$

Note that the second group is defined so that the two components strongly overlap. As a result, the shared component is completely masked in this group. Nevertheless, the masked component can be spotted by exploiting the sharing of information with the other group. Figure 5.3 shows the empirical distribution of a simulated dataset, as well as the pooled data, and the underlying densities.

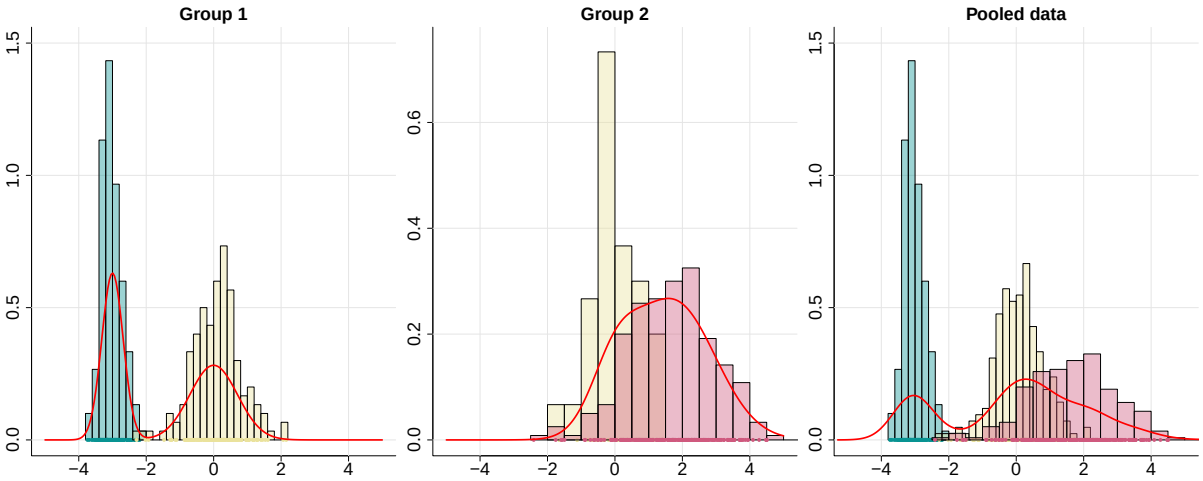

Figure 5.3: Empirical distributions of a dataset simulated under Experiment 1. Dots represent the observations while lines represent the underlying densities. Colors relate to the mixing components.

For each simulated dataset, we fit the HDP, the independent group-specific MFM, and the proposed HMFM by setting $\Lambda_0 = 5$, $\text{V}_\Lambda = 5$, $\gamma_0 = 0.5$, following the guidelines in Section 5.4.1. A clear advantage of joint modeling is the possibility of deriving model-based clustering also across different groups, which is not possible when we run independent analyses.

Table 5.1 presents the results of model comparison based on the mean and standard deviation (in brackets) of the ARI over the simulated datasets. The results show that all methods, with negligible differences, are able to perfectly recover the true clustering in the first group, where the two components are well separated. Instead, the advantage of the sharing of information allowed by hierarchical modeling is evident in the second group, where components overlap. Indeed, the MFM-indep fails to identify the presence of two different clusters in the second group, gathering all observations together and obtaining an ARI value close to zero. On the other hand, the HDP and the HMFM are able to borrow information from the first group to recognize the presence of two clusters in the second group, leading to higher values of ARI, with HMFM (both marginal and conditional samplers) outperforming the HDP. The MFM-pooled method represents an intermediate situation between independent analyses and hierarchical approaches. By pooling all the data, it retains information from the first group, enabling it to identify the presence of two clusters even in the second group with performance comparable to the HMFM and the HDP. To clarify, we compute the percentage of simulated datasets for which each method obtains a partition of

Table 5.1: The first two columns report the mean and the standard deviation (in brackets) of the ARI for the two groups over the 50 simulated datasets under Experiment 1. The third column shows the percentage of times that each method gathers all observations of the second group in a single cluster. The final column reports mean and standard deviation of the ARI relative to the final partition of the data.

	Group 1	Group 2	% 1 cluster	Global
HMFM-cond	0.991 (0.012)	0.133 (0.185)	62%	0.720 (0.024)
HMFM-marg	0.991 (0.011)	0.130 (0.181)	62%	0.719 (0.024)
HDP	0.993 (0.010)	0.097 (0.169)	74%	0.721 (0.018)
MFM-indep	0.993 (0.011)	0.027 (0.100)	90%	-
MFM-pooled	0.971 (0.033)	0.100 (0.126)	28%	0.540 (0.071)

the second group consisting only of a single cluster. Note that lower values are better as we know that the true number of clusters is two. The results are reported in the third column of Table 5.1 and show how the MFM-indep consistently fails to identify at least two clusters for the second group, showing the limitations of the independent analyses against the hierarchical approaches. This limitation is even more pronounced with respect to the MFM-pooled, which represents the extreme case of information sharing. By pooling all data together, the presence of two clusters in the second group becomes evident.

However, we point out that the competitiveness of the MFM-pooled compared to the HMFM and the HDP holds only when looking at the group-specific metrics. Rather, when we evaluate the global ARI in the final column of Table 5.1, i.e., the ARI relative to the final partition of all the data, it becomes clear that the pooled solution is still sub-optimal compared to the joint modeling; recall that the ARI for the global partition is not defined for MFM-indep.

To conclude the clustering comparison, Figure 5.4 provides a picture of the distribution of the CCE within each group over the simulated datasets. Unlike the ARI, which considers only the final clustering estimate, the CCE accounts for all MCMC iterations, thus better reflecting the posterior variance. However, since it relies on the group-specific posterior similarity matrix, it is not well-defined for MFM-pooled. Therefore, the latter case is discarded from the results. In the first group, differences between the HMFM (conditional and marginal) and the MFM-indep are negligible. On the other hand, the MFM-indep commits a higher error in the second group, with also higher variability among different datasets, which confirms the findings reported in Table 5.1 about the sub-optimality of using independent analyses relative to joint modeling of the groups. Finally, we point out that the HDP performs better than the HMFM in the first and second groups. Indeed, in this example, the high sharing of information of the HDP enhances the inference as the two clusters strongly overlap, i.e., sharing as much information as possible from the two groups becomes beneficial. Hence, Figure 5.4 indicates that the HDP is preferable overall, considering all the MCMC iterations. However, such improvement becomes negligible, if not completely reversed, when considering only the final estimate obtained by minimizing the variation of information criterion, see Table 5.1.

5.5.3 Experiment 2

This experiment considers $d = 2$ groups coming from two components, not shared across the groups, namely $K_1 = 2$, $K_2 = 2$, so that $K = 4$. In particular, 50 datasets are generated from the model $y_{1,i} \overset{\text{iid}}{\sim} 0.5\text{N}(-3, 1) + 0.5\text{N}(1, 1)$ and $y_{2,i} \overset{\text{iid}}{\sim} 0.5\text{N}(-4, 1) + 0.5\text{N}(0, 1)$, each for $i = 1, \dots, n_j$. We repeat the experiment for an increasing number of observations, $n = 50, 100, 200$, which we equally

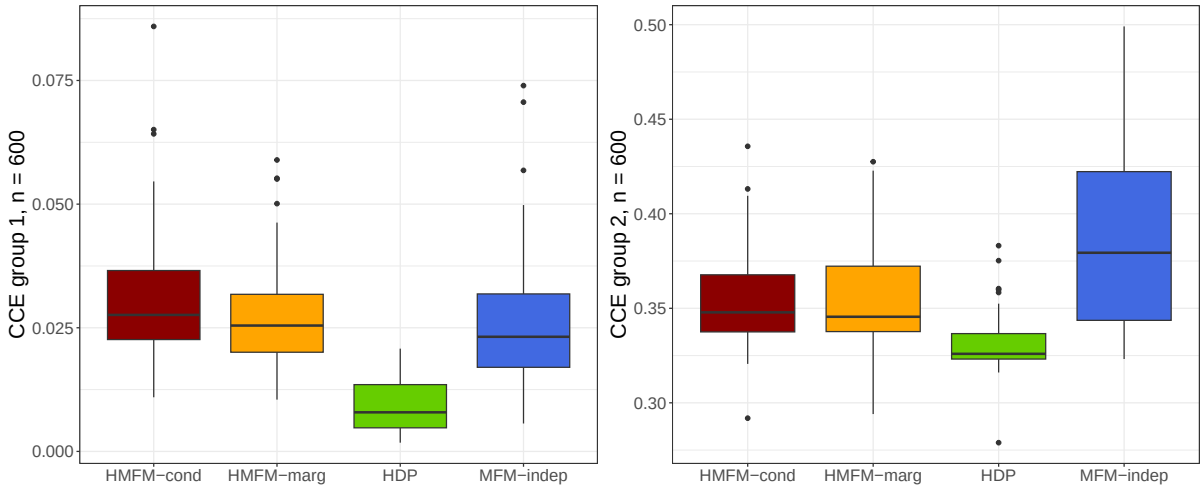

Figure 5.4: Co-Clustering Error in the first group (left) and second group (right). Boxplots are obtained over 50 datasets simulated under Experiment 1.

divided into the two groups. See Figure 5.5 for an example of the empirical distribution of a simulated dataset with $n = 100$ and the underlying densities. This experiment is designed to assess whether the borrowing of information may lead to misleading results in situations where groups do not share any features. Hence, for this scenario, we confine the comparison in terms of global clustering among the HMFM, the HDP, and the MFM-pooled. Here, the HMFM is fitted by setting $\Lambda_0 = 10$, $V_\Lambda = 2$, $\gamma_0 = 0.01$ while the HDP and MFM are fitted employing default hyperpriors.

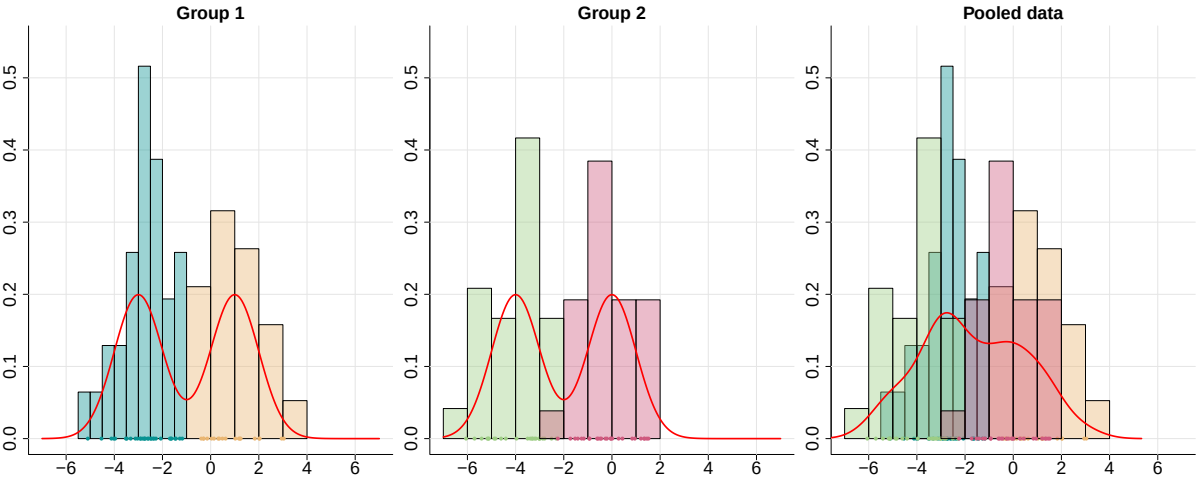

Figure 5.5: Empirical distributions of a dataset simulated under Experiment 2. Dots represent the observations, while lines represent the underlying densities. Colors relate to the mixing components.

Table 5.2 presents the results of model comparison based on the mean and standard deviation (in brackets) of the ARI over the simulated datasets. Although the scenario is simple, the table shows that the HDP struggles to find the underlying global clustering, and it is outperformed by the HMFM, both marginal and conditional. For what concerns MFM-pooled, Figure 5.5 clearly shows that discarding the group-membership information makes the clustering task much harder as all clusters strongly overlap. This explains the poor performances of the MFM-pooled reported in Table 5.2, which do not significantly improve increasing the sample size. The same conclusions can be drawn from the CCE relative to

Table 5.2: Mean and standard deviation (in brackets) of the ARI for the two groups over the 50 simulated datasets under Experiment 2.

	$n = 50$	$n = 100$	$n = 200$
HMFM-cond	0.60 (0.35)	0.74 (0.24)	0.90 (0.10)
HMFM-marg	0.59 (0.33)	0.74 (0.24)	0.89 (0.08)
HDP	0.30 (0.27)	0.48 (0.28)	0.73 (0.18)
MFM-pooled	0.36 (0.13)	0.40 (0.72)	0.41 (0.04)

the global partition (Figure 5.6) which takes into account all the posterior pairwise probabilities of observations to be clustered together. Clearly, the HMFM outperforms both HDP and the MFM-pooled in terms of clustering estimation.

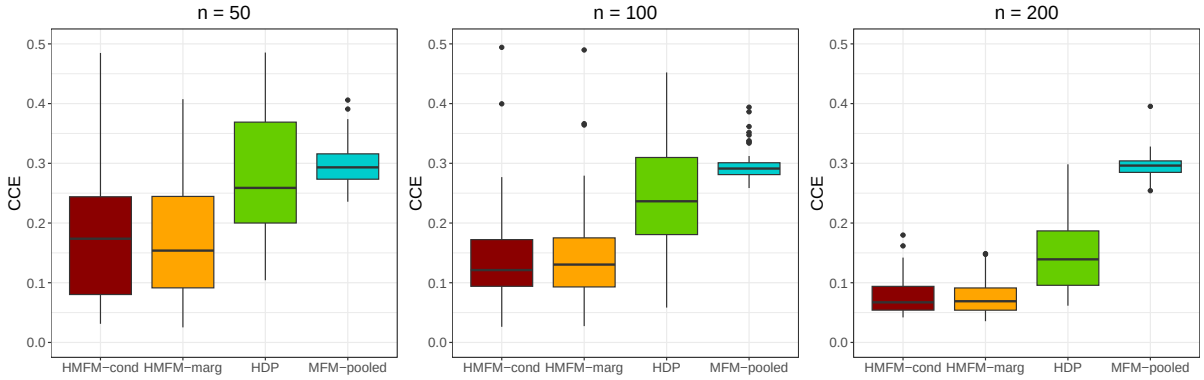

Figure 5.6: Co-Clustering Error for different sample sizes. Boxplots are obtained over 50 datasets simulated under Experiment 2.

Finally, Figure 5.7 shows the frequencies of the estimated number of clusters over the 50 different datasets. The HMFM always prefers a number of clusters greater than two and, as the sample size increases, it selects four clusters, which is the true value. In contrast, the HDP tends to identify only two clusters when n is small, discarding this preference when more data are added. This example showcases that the poorer clustering performance of the HDP is due to the oversharing of information which can compromise the recovery of the true underlying partition. Indeed, in the extreme situation (i.e., ignoring the groups), the MFM-pooled remains fixed at the two-cluster solution, even when the sample size increases.

To better clarify the clustering mechanism of the different models, we resort to the restaurant franchise metaphor discussed in Section 5.3.3. According to the HDP, when a new customer arrives, the probability of consuming a dish not yet served in that specific restaurant, but available in other franchise restaurants, hinges on the cumulative number of clients eating that dish across all restaurants. In contrast, the HMFM uses information only about whether a dish is being consumed or not in other restaurants. This experiment unveils a potential issue in the HDP's clustering mechanism; the concentration of all clients within the first restaurant (group) eating the same dish increases the probability of that dish being offered in the second restaurant (group), even though no components are shared between them. Rather, this confusion is circumvented by the HMFM's clustering mechanism. We notice that this phenomenon is directly tied to the choice of the prior, and its impact diminishes as the sample size increases, as shown in Table 5.2. Nevertheless, the HMFM model still outperforms the competing approaches even when $n = 200$.

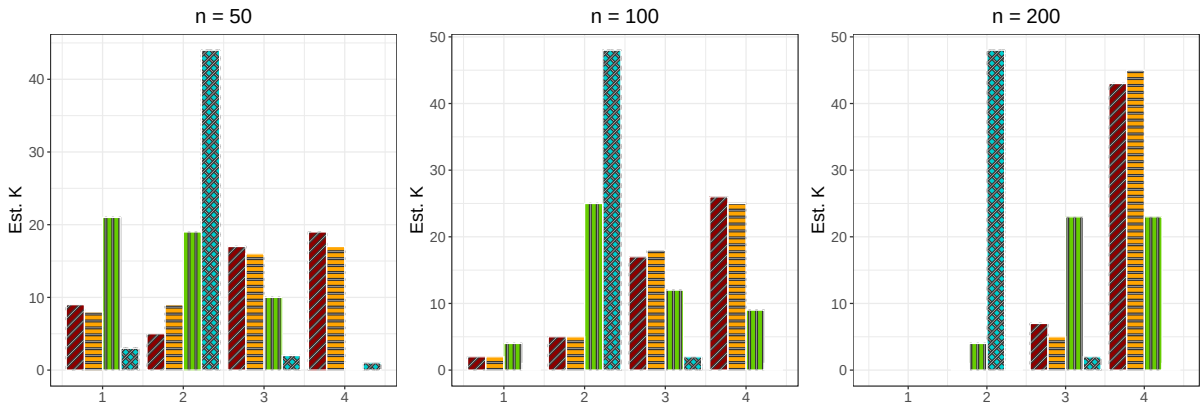

Figure 5.7: Estimated number of global clusters. Frequencies are obtained over 50 datasets simulated under Experiment 2. Red and orange bars represent the HMFM, conditional and marginal algorithms, respectively. Green bars are for the HDP, while the light blue bars are for the MFM-pooled.

5.5.4 Experiment 3

The final experiment considers $d = 15$ groups, each consisting of $n_j = 30$ observations. Data have been generated so that for the first 12 groups, we consider the following component specific parameters $\mu_1 = (-3, 0, 1)$, $\sigma_1 = (\sqrt{0.5}, \sqrt{0.5}, \sqrt{0.5})$. To increase the variability and differences between the groups, for each $j = 1, \ldots, 12$ we first sample the number of local components $K_j \in \{2, 3\}$ and then, conditionally to K_j, we choose uniformly the components to be selected. Regarding the weights, for each of the 12 groups, we assign most of the mass to the second component, if included, namely, $\pi_{j,2} = 0.5$ if $K_j = 3$ or $\pi_{j,2} = 2/3$ if $K_j = 2$. The remaining components receive the remaining mass split in equal parts. Finally, the last three groups all share two components whose parameters are $\mu_2 = (-1.5, 1.5)$, $\sigma_2 = (\sqrt{0.5}, \sqrt{0.5})$ with equal mixing weights. Then, the global number of components is $K = 5$. Figure 5.8 shows an example of data simulated under this scenario, while Figure 5.9 reports the corresponding pooled data.

For each simulated dataset, we fit the HDP, the MFM-pooled and the HMFM; in particular, the latter, is fitted by setting the hyperparameters $\Lambda_0 = 15$, $V_\Lambda = 3$, $\gamma_0 = 0.05$. Figure 5.10 shows the boxplot of the ARI (left) and CCE (middle) of the global clustering evaluated over the simulated datasets. The right panel of Figure 5.10 displays the distribution of the estimated number of clusters according to the different approaches over the simulated datasets. Both the ARI and the CCE show that the HMFM significantly outperforms the HDP. The same conclusions can be drawn by inspecting the ARI and the CCE within each group (not shown here for brevity). The MFM-pooled is also outperformed by the HMFM, although its performance is better or at least comparable with the HDP. One plausible explanation is that the MFM-pooled consistently reaches a solution with three clusters, as seen in the rightmost panel of Figure 5.10, which aligns with the empirical density reported in Figure 5.9. This solution is preferable, according to ARI and CCE, to those proposed by the HDP, which tends to overestimate the number of clusters. Indeed, the HDP only rarely estimates four or five clusters (five being the true value), and in these cases, the estimated ARI value is very close to one, similar to the HMFM. Finally, the barplot of the estimated number of clusters, see Figure 5.10, shows that the HMFMs consistently recover the correct number of clusters.

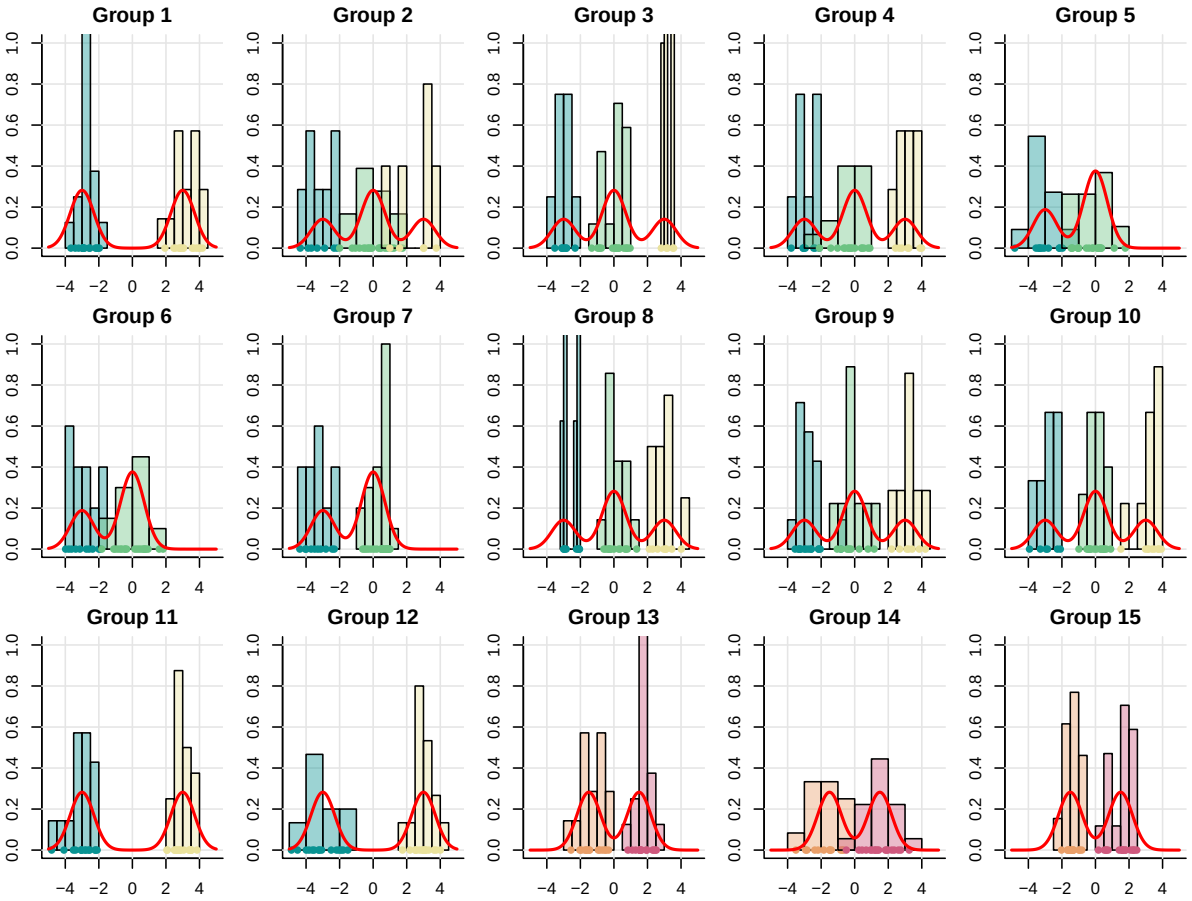

Figure 5.8: Empirical distributions of a dataset simulated under Experiment 3. Dots represent the observations while lines represent the underlying densities. Colours relate to the mixing components.

5.5.5 Density estimation

Regarding density estimation, we compute a Predictive Score (PS; Bassetti et al. 2020) for each group $j = 1, \dots, d$, that is the L_1 distance between the group-specific density $f(y_{j,n_j+1})$ and the corresponding predictive density $\hat{f}(y_{j,n_j+1} \mid \mathbf{Y})$, namely

$$\text{PS}_j = \int_{-\infty}^{\infty} |f(y_{j,n_j+1}) - \hat{f}(y_{j,n_j+1} \mid \mathbf{Y})| \, dy_{j,n_j+1}.$$

The PS is a distance so lower values indicate better fit, with zero being the lowest possible value. Moreover, note that the PS is meaningless in the case of MFM-pooled as the estimated densities in the different groups would coincide and would significantly differ from the true ones.

Figure 5.11 displays the boxplots for the group-specific PS under Experiment 1. The figure shows that differences are negligible in the first group, for which all methods provide a good estimate of the density. However, the density for the second group estimated using the MFM-indep is worse than the one obtained by the HMFM and the HDP. Group-specific density estimation is not performed in the pooled case since it is not clear how to fairly define the group-specific densities.

Similarly, the HMFM models provide better estimates of the densities over the HDP under Experiment 2 and 3, as shown in Figures 5.12 and 5.13, respectively. Nevertheless, such differences are less evident than for the clustering.

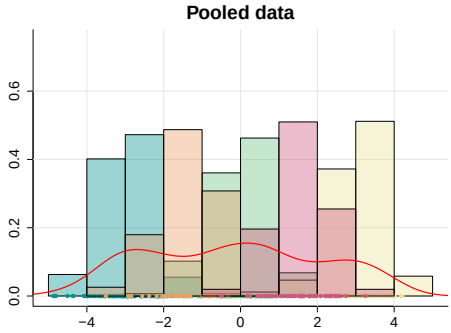

Figure 5.9: Empirical distributions of pooled dataset simulated under Experiment 3. Dots represent the observations while lines represent the underlying densities. Colours relate to the mixing components.

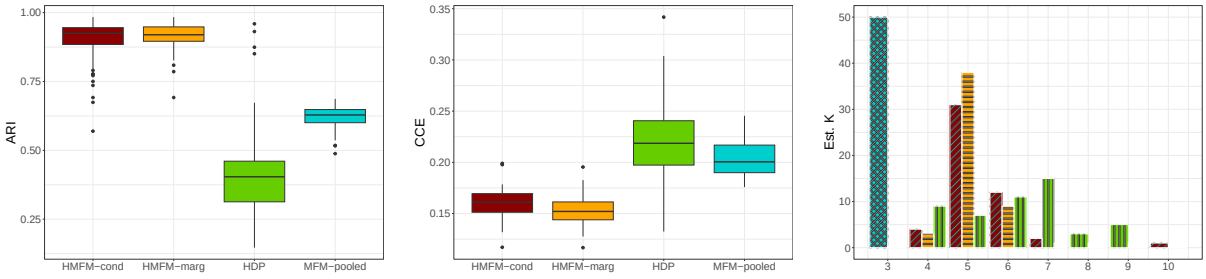

Figure 5.10: ARI (left) and CCE (middle) of the global clustering and estimated number of clusters (right) evaluated over the datasets simulated under Experiment 3.

5.6 Analysis of shot put data

Shot put is a track and field event in which athletes throw a heavy spherical ball, known as the shot, as far as possible. Our dataset comprises measurements, specifically the throw lengths or marks, recorded during professional shot put competitions from 1996 to 2016, for a total of 35, 637 measurements on 403 athletes. Each athlete's record includes the mark achieved, competition details, and personal information, namely, age, gender, and whether the event took place indoors or outdoors. The analyzed data are publicly available (`www.tilastopaja.eu`). Our objective is to model the seasonal performance for each shot putter, interpreted as the mean and variance of his/her seasonal marks. In particular, the season number assigned to each observation corresponds to the number of seasons the athlete has participated in, excluding seasons where he/she did not compete. For example, season 1 represents the athlete's first active season. This grouping of observations into seasons reflects the athletes' years of experience. Figure 5.18 visually illustrates the performance evolution throughout the career of four randomly selected shot putters from the dataset. Each athlete has different participation in competitions, and the length and trajectory of their performance careers vary. While performance is expected to vary over the athlete's career, the figure evidences that the performance remains relatively consistent within each season. We characterized the seasonal performances as arising primarily from random fluctuations around a mean value. The values of this mean and the associated variability are unknown and are inferred from the data. Although it is a simplified representation, this captures the essential characteristics of athletes' careers.

In a previous study, Dolmeta et al. (2023) employed a GARCH model to account for the volatility clustering of athletes' results over time. Rather, in this work, we frame the data into a hierarchical structure where each season represents a different group. Hence, we assume the HMFM model described

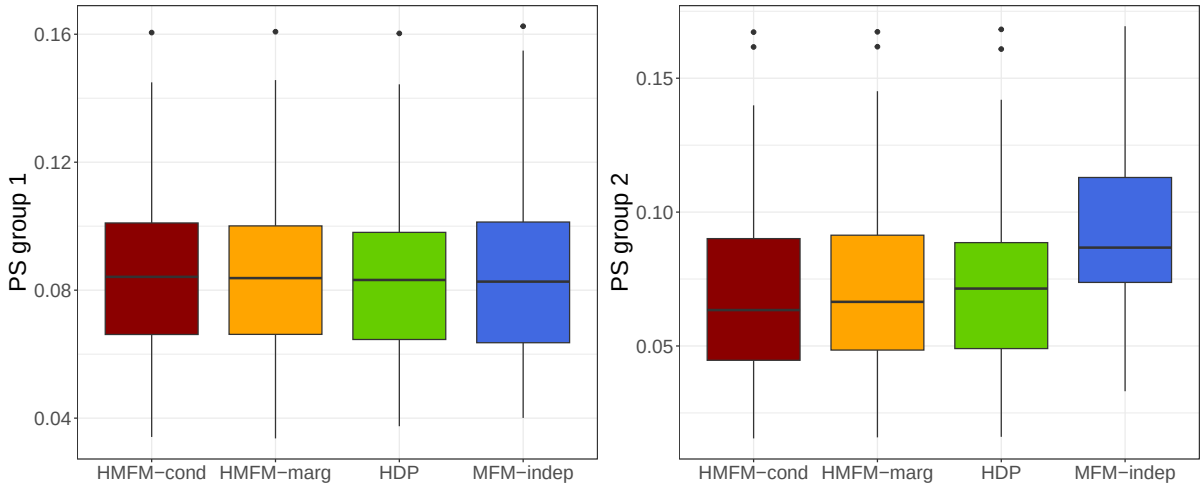

Figure 5.11: Predictive Score in the first (left) and second (right) group. Boxplots are obtained over 50 datasets simulated under Experiment 1.

in Section 5.2 for analyzing the athletes' performance; the proposed model allows us to capture the variability among different seasons and clustering the performances both within the seasons and across them.

Let n_j be the number of athletes competing in season j, with $j = 1, \dots, d$. The longest career consists of 15 seasons, which is then the total number of groups $d = 15$. Each active athlete i in season j, with $i = 1, \dots, n_j$, takes part in $N_{j,i}$ events. At each event, indexed by $h = 1, \dots, N_{j,i}$, the athlete's mark $y_{j,i,h}$ is measured. Moreover, r event-specific covariates are available, $\boldsymbol{x}_{j,i,h} \in \mathbb{R}^r$, and collected in the design matrices $X_{j,i} \in \mathbb{R}^{N_{j,i} \times r}$.

Assuming that observations are noisy measurements of an underlying athlete-specific function, the model we employ for these data is $y_{j,i,h} = \mu_{j,i} + X_{j,i}\boldsymbol{\beta}_j + \varepsilon_{j,i,h}$, with $\varepsilon_{j,i,h} \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma^2_{j,i}\right)$, where $\mu_{j,i}$ is a season-specific random intercept, $\boldsymbol{\beta}_j$ is a r-dimensional vector of regression parameters, shared among all the athletes in season j, and $\sigma^2_{j,i}$ denotes the error variance. Therefore, within each season j, the athlete's observations $\boldsymbol{y}_{j,i} = (y_{j,i,1}, \dots, y_{j,i,N_{j,i}})$ are distributed as

$$\boldsymbol{y}_{j,i} \mid \mu_{j,i}, \sigma^2_{j,i}, \boldsymbol{\beta}_j, X_{i,j} \overset{\text{ind}}{\sim} \mathrm{N}_{N_{j,i}}\left(\mu_{j,i}\mathbf{1}_{N_{j,i}} + X_{j,i}\boldsymbol{\beta}_j, \sigma^2_{j,i}\mathbf{I}_{N_{j,i}}\right), \tag{5.18}$$

where $\mathrm{N}_{N_{j,i}}$ denotes the $N_{j,i}$-dimensional normal distribution, $\mathbf{1}_{N_{j,i}}$ is a vector of length $N_{j,i}$ with all entries equal to 1 and $\mathbf{I}_{N_{j,i}}$ is the identity matrix of size $N_{j,i}$. To ensure identifiability, observations $\boldsymbol{y}_{j,i}$ have been centered within each season, i.e., $\sum_{i=1}^{n_j} \sum_{h=1}^{N_{j,i}} y_{j,i,h} = 0$ for each j.

Letting $\theta_{j,i} = \left(\mu_{j,i}, \sigma^2_{j,i}\right)$, we place a Vec-FDP prior for $\theta_{j,i}$ so that a clustering of athletes' performances both within and across different seasons is achieved. We emphasize that the model is designed to cluster performances rather than athletes themselves, which would not be feasible since the observational units are not consistent across seasons. We assume a multivariate normal prior distribution for the regression coefficients, whose prior mean is denoted by $\boldsymbol{\beta}_0$ and covariance matrix Σ_0. We define $\boldsymbol{y}$ as the collection of all observations across seasons j and athletes i. Based on evidence from a previous analysis (Dolmeta et al., 2023) and for ease of interpretation, we use only gender as a covariate in our analysis. In particular, we use male athletes as reference baseline and set $\boldsymbol{\beta}_0 = -2\mathbf{1}_d$ and $\Sigma_0 = \mathbf{I}_d$

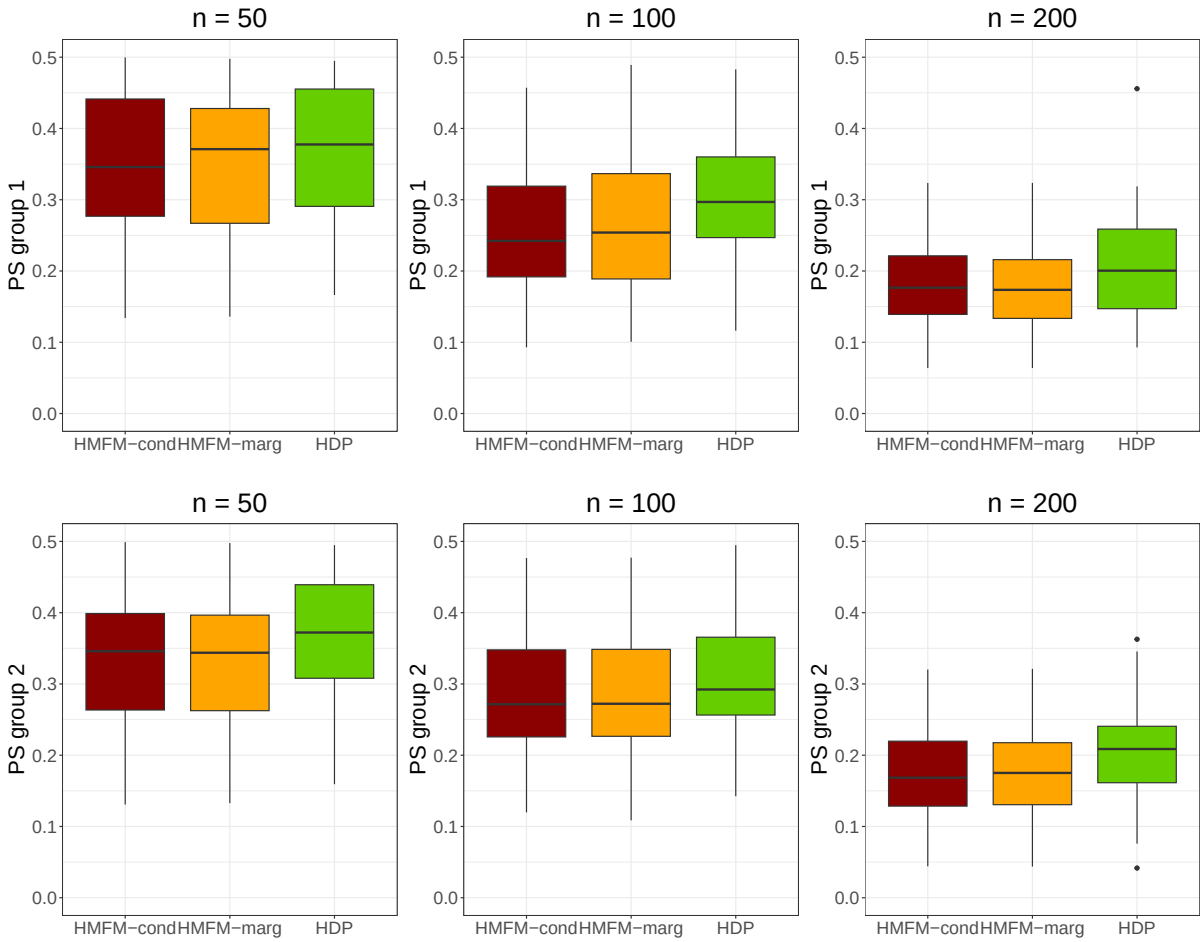

Figure 5.12: Predictive score for the first group (top panels) and the second group (bottom panels) for different sample sizes. Boxplot are obtained over 50 datasets simulated under Experiment 2.

expecting male athletes to throw longer than females.

We set the base probability measure P_0 for $\theta = (\mu, \sigma^2)$ to be a Normal-Inverse Gamma, exploiting the conjugacy with the likelihood in Equation (5.18). In particular, the Normal-Inverse Gamma distribution is parametrized as in Hoff (2009), i.e.,

$$\left(\mu, \sigma^2\right) \sim \text{InvGamma}(\mu_0, k_0, \nu_0, \sigma_0^2) = \text{N}\left(\mu_0, \frac{\sigma^2}{k_0}\right) \times \text{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right).$$

Following the approach of Richardson and Green (1997) and Lijoi et al. (2007b), we set $\mu_0 = 0$ and $k_0 = \frac{1}{\text{range}(\boldsymbol{y})^2}$. Then, we set $\nu_0 = 4$ and $\sigma_0^2 = 10$ to have a vague InvGamma with infinite variance. For the process hyperparameters Λ and γ, we set the hyperprior in Equation (5.17). To achieve sparsity in the mixture, we follow the approach in Section 5.4.1, and set $\Lambda_0 = 25$, $\text{V}_\Lambda = 3$ and $\gamma_0 = 1/\sum_{j=1}^{d} n_j = 0.00027$, leading to $a_\gamma = 13.89$, $b_\gamma = 2007.78$, $a_\Lambda = 208.33$, $b_\Lambda = 8.33$. The complete formulation of the hierarchical model can be found in Section D.5. The burn-in period has been set equal to $50,000$, then $200,000$ additional iterations were run with a thinning of 10. The initial partition has been set using the k-means on the pooled dataset with 20 centers. Posterior analysis is not sensitive for such a choice.

Figure 5.14 shows the posterior 95% credible intervals of the regression coefficients. The posterior

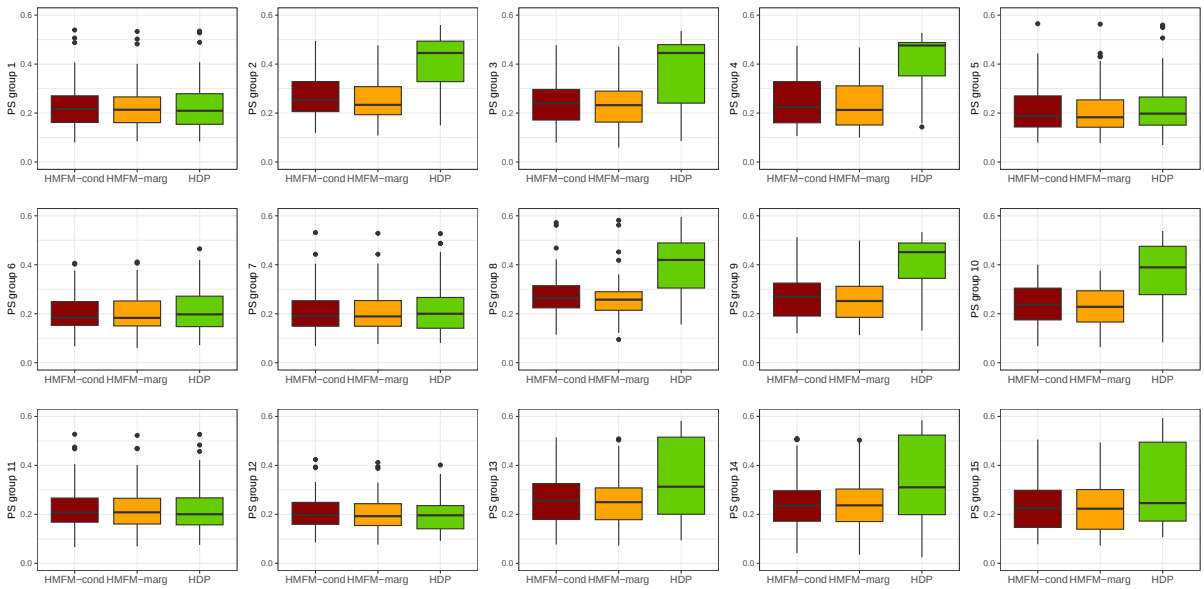

Figure 5.13: Predictive Score in all the groups. Boxplot are obtained over 50 datasets simulated under Experiment 3.

distribution is concentrated on negative values, meaning that the athletes' marks are, on average, higher for males than females. Also, the effect of gender on athletes' performance is significantly different across seasons, e.g., it is more evident in the first years and it reduces over career years, with the exception of the final season.

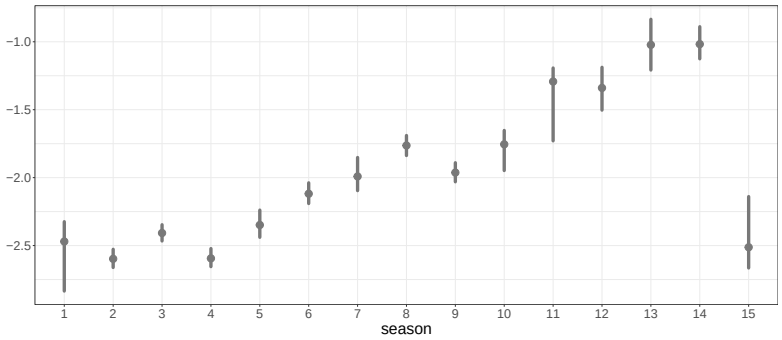

Figure 5.14: 95% posterior credible intervals of season-specific coefficients β's.

The final clustering has been obtained through minimization of the variation of information (Wade and Ghahramani, 2018; Dahl et al., 2022) loss function, and it consists of 13 clusters, with the posterior mode being equal to 15. Among these, we identify 11 main clusters since two of them capture noisy observations with high variance. The estimated clusters have been relabeled according to decreasing means.

A remarkable finding is that the cluster interpretation does not depend on gender, whose effect has been filtered out by the season-specific parameter β . In other words, our clustering does not trivially distinguish between males and females, but it models the athletes' performance regardless of their gender. This claim is supported by the fact that, when ordering the clusters according to their means, both males' and females' average performance are ordered too, with the exception of one cluster which is made of female athletes only. Moreover, Figure 5.15 reports the athletes' marks, colored according to their cluster

membership, for male and female players, respectively. The two plots are similar, highlighting that the cluster interpretation is gender-free.

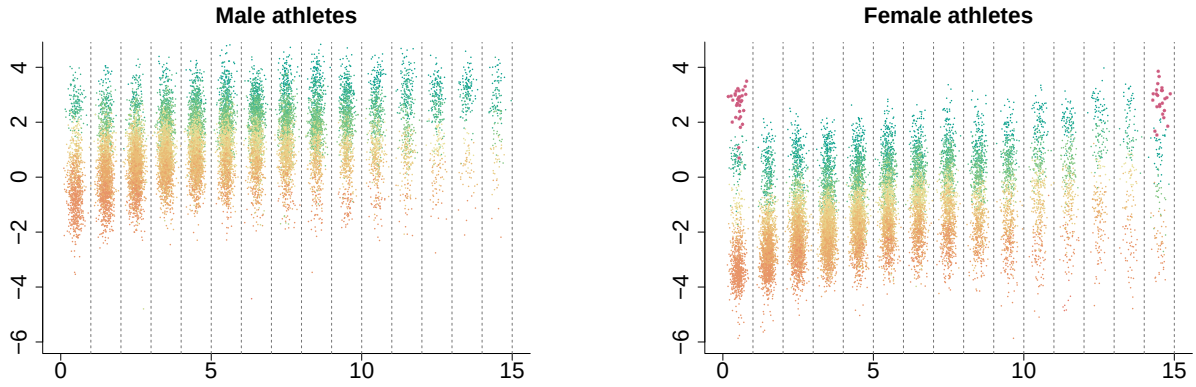

Figure 5.15: Shot put marks for male athletes (left panel) and female athletes (right panel). Vertical dotted lines delimit seasons. Dots are colored according to their cluster membership.

Nevertheless, we are able to identify the presence of a particular cluster, whose points are highlighted in the right panel of Figure 5.15, including three exceptional women performances, which are much above the average mean throw for female athletes. No man belongs to such a cluster, meaning that no one has ever been able to outperform competitors in such a neat way. In particular, in this cluster, we find Astrid Kumbernuss, who is a three-time World champion and one-time Olympic champion; Valerie Adams, who, during her outstanding career, won two Olympic Games and four World Championships and Nadzeya Ostapchuk, who won a bronze medal at the Olympic Games.

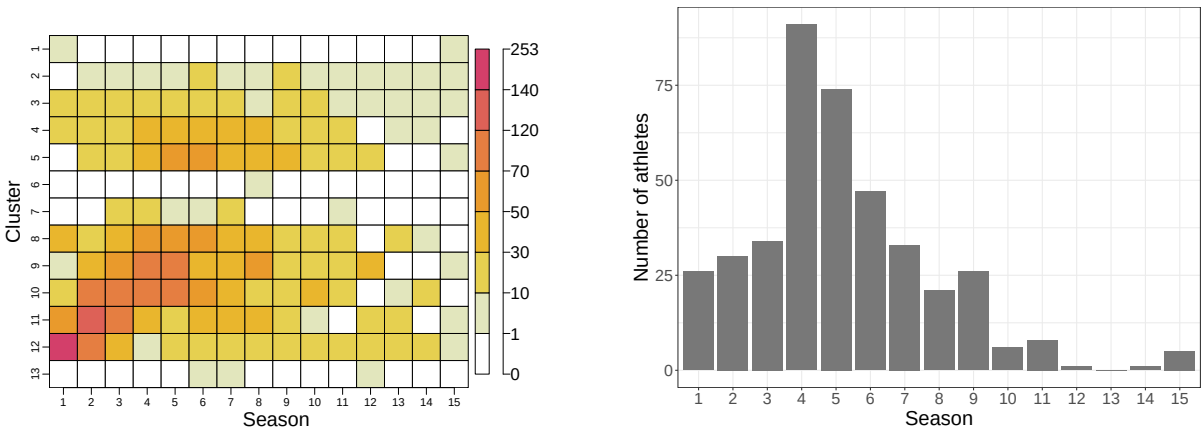

Figure 5.16: Left panel: local cluster sizes. Right panel: absolute frequencies of the seasons in which each athlete reaches their peak cluster for the first time, i.e., the one with the highest average.

According to the model (5.18), clusters are interpreted according to their corresponding mean and variances, but many insights can be gained from the clustering results. Notably, our analysis focuses on global considerations about the average evolution of the athletes' careers, as well as on the individual-specific sequence of clusters that characterizes every athlete. We refer to Section D.5 for tables reporting all season-specific cluster sizes and cluster summaries and to the left panel of Figure 5.16 for a representation of the local cluster sizes. Finally, see Section D.5.1 for a comparison with a naive modeling, which consists of fitting a mixture model to the pooled data.

From the global interpretation of the clustering results, we observe that in the first season of their careers, most of the athletes are grouped into the lowest-ranked cluster. This finding aligns with the expectation that rookies tend to exhibit similar performances. Then, the remaining athletes are divided among other low-ranked clusters, with the exception of some athletes who already belong to high-ranked clusters. The intermediate-level clusters are notably empty, highlighting a highly polarized situation. An interesting feature of our model is the ability to evaluate how the larger cluster changes across seasons. Notice in the left panel of Figure 5.16 the darkest squares progressively shift from the lowest-ranked cluster toward the intermediate-ranked clusters up to season 7. This trend likely reflects both the athletes' increasing experience and ongoing physical development during the early stages of their careers. However, from season 7 onwards, this progression is eventually tempered by external factors not captured by the model, such as training consistency and injuries. Furthermore, the right panel of Figure 5.16 shows the absolute frequencies of the seasons in which each athlete first reached the highest-ranked cluster of their career. This depicts a skewed distribution, with a mode at the season 4, and subsequent seasons being more frequent compared to the first three seasons. The frequencies decrease significantly from season 10 onwards.

The analysis of cluster summaries provides additional insights. From Table 5.4, we note that male athletes in the highest-ranked clusters have higher average ages than female athletes. This indicates that women tend to reach their peak performance at a younger age than men. A possible explanation for this disparity is that female bodies develop earlier than male bodies; this is supported by the fact that women begin their careers at an average age of 18.2, while men start at an average age of 19.4. Cluster 1 stands out from this trend as it comprises female athletes with a mean age of 28, which is considerably older than the average peak age. Lastly, it is worth noting that top-level clusters, specifically clusters 2 and 3, have a higher proportion of female athletes than male athletes. This is despite the overall number of observations for women being smaller than for men. This suggests that male competitions may be more balanced, making it more challenging for athletes to distinguish themselves from the average level.

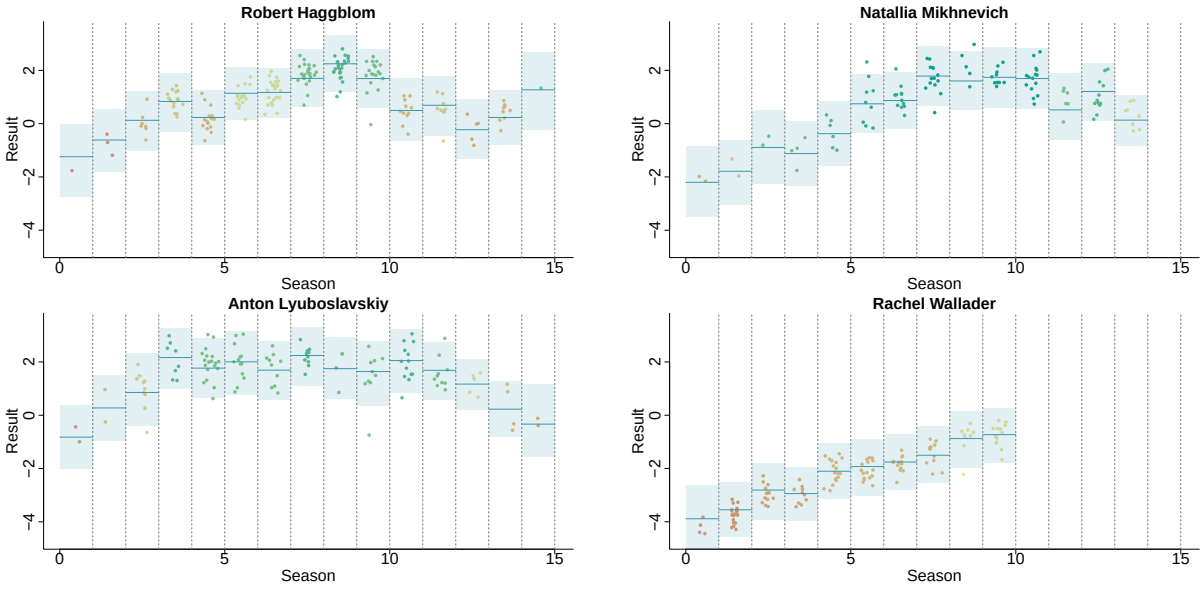

Figure 5.17: Shot put measurement for four randomly selected athletes. Points are colored according to the cluster membership of the corresponding performance. Solid means represent the estimated cluster means. Shaded areas represent the 95% credible bands.

A key feature of our analysis is the possibility of studying the evolution of season-specific cluster membership of each player. This is exemplified in Figure 5.17, which showcases the trajectories of four players (two men and two women). In the plot, marks representing each season are colour-coded according to their cluster memberships, while solid lines represent the estimated seasonal mean performances, and the shaded areas represent 95% credible bands. One notable pattern emerges from the trajectory of Robert Häggblom, where a drop occurred immediately after the peak of his career, where he even participated in the 2008 Olympic games. A back injury conditioned the final part of its career. This sudden change would have been challenging to capture by a time-smoothing model. In contrast, Rachel Wallader demonstrates significant improvements throughout her career, starting from cluster 12, which we recall being the predominant cluster among rookies, and eventually reaching the higher-performing cluster before retiring, also managing to win the British title. Anton Lyuboslavskiy and Natallia Mikhnevich share remarkable career paths, showcasing exceptional performance not only for a single season but for extended periods of time, both around 9 years. Indeed, both are Olympic level players. Unlike Rachel Wallader, Anton Lyuboslavskiy continued to compete beyond his prime, maintaining high levels of performance but eventually transitioning to intermediate cluster levels. Lastly, we highlight that Robert Häggblom and Natallia Mikhnevich achieved comparable marks, but the performances of the second athlete, a woman, are assigned to higher-ranked clusters. This demonstrates our model's ability to recognize top players, regardless of their gender. Indeed, Natallia Mikhnevich's career has been richer in success, as she won both gold and silver at the European Championships. Finally, we point out that we identify athlete-specific sequences of clusters, which could themselves be clustered to capture similarities and differences in the development trajectories of the athletes. A model-based solution to this task would necessitate moving beyond the framework of partial exchangeability across the seasons and adopting temporal modeling; see Page et al. (2022) for a possible alternative.

Appendix of Chapter 5

D.1 Independent Finite Point Processes

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let us denote by $\mathbb{G}$ a Polish space endowed with its corresponding Borel σ-algebra $\mathcal{G}$. Moreover, let $\mathbb{M}(\mathbb{G})$ be the space of locally finite measures on $\mathbb{G}$, and $\mathcal{M}(\mathbb{G})$ represents the corresponding Borel σ-algebra. A point process Φ on $\mathbb{G}$ is a measurable map $\Phi : \mathbb{G} \to \mathbb{M}(\mathbb{G})$, defined as

$$\Phi(B) = \sum_{m \geq 1} \delta_{\xi_m}(B), \quad B \in \mathcal{G}.$$

having denoted by δ_{ξ_m} the delta-Dirac mass at ξ_m. The sequence $(\xi_m)_{m \geq 1}$ is a random countable subset of $\mathbb{G}$ and its elements are called atoms.

We recall that Independent Finite Point Process are random measures, whose definition is provided in Section 4.1.2 Here, we consider *Vectors of Independent Finite Point Processes* (Vec-IFPPs), thus we work on $\mathbb{G} = (\mathbb{R}^+)^d \times \mathbb{X}$. A Vec-IFPP is a point process whose unnormalized weights are elements of $(\mathbb{R}^+)^d$, namely $\boldsymbol{S}_m = (S_{1,m}, \ldots, S_{d,m})$. The atoms τ_m are still i.i.d. distributed according to P_0, leading to the following definition of the point process Φ,

$$\Phi(A \times B) = \sum_{m=1}^{M} \delta_{(\boldsymbol{S}_m, \tau_m)}(A \times B), \tag{5.19}$$

where A is a Borel set of $(\mathbb{R}^+)^d$ and $B \in \mathcal{X}$. The unnormalized weights $\boldsymbol{S}_m$ are still i.i.d. distributed according to H, that is now a probability distribution on $(\mathbb{R}^+)^d$. For the sake of simplicity, we only resort to the case when this is defined as the measure theoretic product of probability distributions H_j defined on $\mathbb{R}^+$, i.e., $H(\mathrm{d}\boldsymbol{s}) = H_1(\mathrm{d}s_1) \times \cdots \times H_d(\mathrm{d}s_d)$. In addition, we also assume all H_j to belong to the same parametric family having density $h(\cdot \mid \gamma_j)$ with respect to the Lebesgue measure on $\mathbb{R}^+$, i.e $H_j(\mathrm{d}s) = h(s \mid \gamma_j)\mathrm{d}s$, where $\gamma_j \in \mathbb{R}^+$ are component-specific parameters. We write $\Phi \sim \text{Vec-IFPP}(q_M, H, P_0)$.

D.1.1 Normalization

Given $\Phi \sim \text{Vec-IFPP}(q_M, H, P_0)$ on $\mathbb{G} = (\mathbb{R}^+)^d \times \mathbb{X}$, we define a vector of length d of unnormalized random measures $(\mu_1, \ldots, \mu_d)$ on $\mathbb{X}$ as follows,

$$\mu_j(B) = \int_0^\infty \int_B s_j \Phi(\mathrm{d}\boldsymbol{s}, \mathrm{d}\tau) = \sum_{m=1}^{M} S_{j,m} \delta_{\tau_m}(B), \quad B \in \mathcal{X}. \tag{5.20}$$

Then, the vector of random probability measures $(P_1, \dots, P_d)$ is defined through normalization as

$$P_j(\cdot) = \frac{\mu_j(\,\cdot\,)}{\mu_j(\mathbb{X})}. \tag{5.21}$$

If $q_M(0) = 0$, then (5.21) is well defined since $0 < \mu_j(\mathbb{X}) < \infty$ almost surely. We write $(\mu_1, \dots, \mu_d) \sim$ Vec-IFPP(q_M, H, P_0) and $(P_1, \dots, P_d) \sim$ Vec-NIFPP(q_M, H, P_0), which stands for vector of Normalized Independent Finite Point Process.

D.2 Vector of Normalized Independent Finite Point Processes

In this section, we define the general class of Bayesian nonparametric priors for the vector $(P_1, \dots, P_d)$ called *Vector of Normalized Independent Finite Point Process* (Vec-NIFPP) introduced in Section 5.2 and denoted as

$$(P_1, \dots, P_d) \;\sim\; \text{Vec-NIFPP}\,(q_M, H, P_0)\,. \tag{5.22}$$

We recall that the Vec-NIFPP encompasses the HMFM model in Equation (5.5) as a special case.

Here, q_M is a probability mass function on the positive integers $\{1, 2, 3, \dots\}$ and it acts as a prior for the random number of components M. Then, given M, we still assume the unnormalized weights $S_{j,m}$ to be conditionally independent both within and between the groups. However, we now assume the components of $\boldsymbol{S}_j = \left(S_{j,1}, \dots, S_{j,M}\right)$ to be, given M, i.i.d. from H_j, a probability distribution over $\mathbb{R}^+$. For the sake of simplicity, we assume all H_j to belong to the same parametric family having density $h(\cdot \mid \gamma_j)$, i.e., $H_j(\mathrm{d}s) = h(s \mid \gamma_j)\mathrm{d}s$, where $\gamma_j \in \mathbb{R}^+$ are group-specific parameters. We set $H := H_1 \times \cdots \times H_d$. Finally, conditionally to M, the common random $(\tau_1, \dots, \tau_M)$ atoms are i.i.d. from P_0, as in Section 5.2. We recall that the P_j's are defined through normalization, see Equation (5.3), hence we also define the joint law of the $(\mu_1, \dots, \mu_d)$, called Vector of Independent Finite Point Process (Vec-IFPP).

D.2.1 Distributional results of the Vec-NIFPP

In this section, we derive all theoretical properties for the latent variables $\theta_{j,i}$ modeled as follows

$$\theta_{j,i} \mid P_j \overset{\text{iid}}{\sim} P_j, \quad (P_1, \dots, P_d) \sim \text{Vec-NIFPP}\,(q_M, H, P_0)\,, \tag{5.23}$$

where $j = 1, \dots, d$ and $i = 1, \dots, n_j$. In addition, we assume conditional independence across groups, i.e., $\boldsymbol{\theta}_j, \boldsymbol{\theta}_l \mid P_j, P_l$ are independent for $j \neq l$. Such properties are used to prove the theorems and propositions presented in Section 5.3 as particular cases.

The almost sure discreteness of the P_j, coupled with their common supports, entails that the hierarchical sample $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$ is equivalently characterized by $(\boldsymbol{\theta}^{**}, \rho)$, previously defined in Section 5.2.1. The following theorem specifies the distribution of the pEPPF for the model under investigation.

Theorem D.2.1 (pEPPF)**.** *The probability to observe a sample $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$ of size n from Equation (5.23) exhibiting K distinct values $\left(\theta_1^{**}, \dots, \theta_K^{**}\right)$ with respective counts $\boldsymbol{n}_1, \dots, \boldsymbol{n}_d$ is given by the*

following pEPPF

$$\Pi_K^{(n)}(\boldsymbol{n}_1,\ldots,\boldsymbol{n}_d)=\int_{(\mathbb{R}^+)^d}\prod_{j=1}^{d}\left\{\frac{u_j^{n_j-1}}{\Gamma(n_j)}\prod_{k=1}^{K}\kappa_j(u_j,n_{j,k})\right\}\Psi(K,\boldsymbol{u})\mathrm{d}u_1,\ldots,\mathrm{d}u_d$$

where

$$\Psi(K,\boldsymbol{u})=\sum_{m=0}^{\infty}q_M(m+K)\frac{(m+K)!}{m!}\prod_{j=1}^{d}\left(\psi_j\left(u_j\right)\right)^m,$$

$\psi_j\left(u_j\right)$ is the Laplace transform of a random variable $S_j\sim H_j$ and $\kappa_j\left(u_j,n_{j,k}\right)$ is its derivative. Namely,

$$\psi_j\left(u_j\right)=E\left[e^{-u_jS_j}\right]=\int_0^{\infty}e^{-u_js}H_j(\mathrm{d}s),\qquad \kappa_j\left(u_j,n_{j,k}\right)=\int_0^{\infty}e^{-u_js}s^{n_{j,k}}H_j(\mathrm{d}s).$$

The predictive distributions, i.e., the distribution of θ_{j,n_j+1} given $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_d)$ for each possible group j, follow from Theorem D.2.1 and their presentation is deferred to Section D.4.

The following theorem provides functionals of $(\mu_1,\ldots,\mu_d)$.

Theorem D.2.2 (Mixed moments)**.** *Let $(P_1,\ldots,P_d)\sim$ Vec-NIFPP (q_M,H,P_0) be a vector of normalized random probability measures defined through normalization as in Equation (5.3). The following holds:*

(i) For any measurable set A and for any $j\in\{1,\ldots,d\}$,

$$E\left[P_j(A)\right]=P_0(A).$$

(ii) for any measurable sets A,B and for any $j,l\in\{1,\ldots,d\}$,

$$E\left[P_j(A)P_l(B)\right]=\mathbb{P}\left(K_{(1,1)}=1\right)\left(P_0(A\cap B)-P_0(A)P_0(B)\right);\tag{5.24}$$

(iii) for any measurable set A and for any $j,l\in\{1,\ldots,d\}$,

$$E\left[P_j(A)^{n_j}P_l(A)^{n_l}\right]=E\left[P_0(A)^{K_{(n_j,n_l)}}\right]=\sum_{k=1}^{n_j+n_l}P_0(A)^k\mathbb{P}\left(K_{(n_j,n_l)}=k\right),\tag{5.25}$$

where $K_{(n_j,n_l)}$ is the global number of clusters across two groups with size n_j and n_l.

The prior probabilities for the number of clusters can be written in terms of the pEPPF, e.g., $\mathbb{P}\left(K_{(1,1)}=1\right)=\Pi_1^{(2)}\left((1),(1)\right)$. Furthermore, both Equation (5.24) and (5.25) can be extended to the case of more than two groups. As a byproduct of Theorem D.2.2 we obtain a closed form expression for pairwise correlation between the components of $(P_1,\ldots,P_d)$ evaluated on specific sets. Let A be a measurable set, then, for any $j,l\in\{1,\ldots,d\}$:

$$\operatorname{corr}\left(P_j(A),P_l(A)\right)=\frac{\mathbb{P}\left(K_{(1,1)}=1\right)}{\sqrt{\mathbb{P}\left(K_{j,(2)}=1\right)\mathbb{P}\left(K_{l,(2)}=1\right)}},\tag{5.26}$$

where $\mathbb{P}\left(K_{j,(2)}=1\right)$ is the prior probability of having a single local cluster out of two observations in group j.

Equation (5.26) evaluates both P_j and P_l over the same Borel set A. To prove the previous result, we go through the more general case regarding the correlation between $P_j(A)$ and $P_l(B)$, where A and B are measurable sets in $\mathcal{X}$. This equals to

$$\operatorname{corr}(P_j(A), P_l(B)) = \frac{E\Big[P_j(A)P_l(B)\Big] - E\Big[P_j(A)\Big]E\Big[P_l(B)\Big]}{\sqrt{\operatorname{Var}(P_j(A))\operatorname{Var}(P_l(B))}}. \tag{5.27}$$

The numerator in Equation (5.27) easily follows from Theorem D.2.2. As for the denominator, we use the following result from Argiento et al. (2020),

$$\operatorname{Var}(P_j(A)) = \mathbb{P}\left(K_{j,(2)} = 1\right) P_0(A)(1 - P_0(A)).$$

If we plug the previous expressions in (5.27), we obtain

$$\begin{aligned}&\operatorname{corr}\left(P_j(A), P_l(B)\right)\\ &\quad= \frac{\mathbb{P}\left(K_{(1,1)} = 1\right)\left(\, P_0(A \cap B) - P_0(A)P_0(B)\,\right)}{\sqrt{\mathbb{P}\left(K_{j,(2)} = 1\right)\mathbb{P}\left(K_{l,(2)} = 1\right) P_0(A)(1 - P_0(A))P_0(B)(1 - P_0(B))}}.\end{aligned} \tag{5.28}$$

In particular, Equation (5.26) from Equation (5.28) when $A = B$.

Since Theorem D.2.2 also involves higher-moments results, we compute the coskewness between pairs of the random vector $(P_1, \ldots, P_d)$ to further study their dependence. Consider two random variables, X and Y, both having finite third moments. Let μ_X denote the mean of X, σ_X^2 its variance, and define the skewness of X as $\operatorname{Sk}(X) = \frac{\mathbb{E}\left[(X-\mu_X)^3\right]}{\sigma_X^3}$. The same definitions and notation can be applied to Y. The coskewness of X over Y is defined as

$$\begin{aligned}\operatorname{CoSk}(X, Y) &= \frac{\mathbb{E}\left((X - \mu_X)^2 (Y - \mu_Y)\right)}{\sigma_X^2 \sigma_Y} = \frac{\mathbb{E}\left(X^2 Y\right) - 2\mu_X \mathbb{E}(XY) - \mathbb{E}\left(X^2\right)\mu_Y + 2\mu_X^2 \mu_Y}{\sigma_X^2 \sigma_Y}\\ &= \frac{\operatorname{Cov}\left(X^2, Y\right) - 2\mu_X \operatorname{Cov}(X, Y)}{\sigma_X^2 \sigma_Y}\end{aligned}$$

We note that the coskewness operator CoSk is not symmetric, and the coskewness of Y over X is defined analogously. Additionally, the coskewness can deviate from zero even when X and Y are symmetric, as it depends on the random variable $X + Y$. This mixed third moment serves as an indicator of the joint asymmetry of X and Y (see, for instance, Friend and Westerfield (1980); Fang and Lai (1997)). In the field of Econometrics, the coskewness is frequently employed to study the risk associated with financial portfolios. In our setting, if we let $X = P_j(A)$ and $Y = P_l(A)$. From Theorem D.2.2 we have that both μ_X and μ_Y equal $P_0(A)$. Then, using Equations (5.24) and (5.25) we get

$$\begin{aligned}\operatorname{CoSk}\left(P_j(A), P_l(A)\right) &= \frac{1}{\left(\mathbb{P}(K_{(2)} = 1)P_0(A)(1 - P_0(A))\right)^{3/2}}\\ &\times \left(E[P_0(A)^{K_{(2,1)}}] - P_0(A)\left(E[P_0(A)^{K_{(2)}}] - 2P_0(A)\left(\mathbb{P}(K_{(2)} = 1) + \mathbb{P}(K_{(1,1)} = 2)\right)\right)\right).\end{aligned}$$

A more general formulation of the previous equation for $\operatorname{CoSk}\left(P_j(A), P_l(B)\right)$ could also be computed using Equation (5.45).

We now aim at giving a posterior characterization for a vector $(P_1, \dots, P_d)$ distributed as in Equation (5.22). Since $(P_1, \dots, P_d)$ is obtained via normalization of $(\mu_1, \dots, \mu_d)$, it is sufficient to provide a posterior characterization for the latter vector. In order to do this, we follow the same approach of James et al. (2009), Camerlenghi et al. (2019b) and Argiento and De Iorio (2022). Thus, we introduce a vector of auxiliary variables $\boldsymbol{U}_n = (U_1, \dots, U_d)$ such that $U_j \mid T_j \overset{\text{ind}}{\sim} \text{Gamma}\left(n_j, T_j\right)$, where $T_j = \mu_j(\mathbb{X})$. This is possible since the marginal distribution of $\boldsymbol{U}_n$ does exist, see Section D.3.7.

Hence, conditionally to $\boldsymbol{U}_n$ and to $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$, $(\mu_1, \dots, \mu_d)$ is a superposition of two independent processes, one driving the non-allocated components and the other one driving the allocated components.

Theorem D.2.3 (Posterior representation)**.** *Let $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$ be a sample from the statistical model in Equation (5.23). Then, the posterior distribution of $(\mu_1, \dots, \mu_d)$ conditionally to $\boldsymbol{U}_n = \boldsymbol{u}$ and a realization of $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$ in K distinct values $\left(\theta_1^{**}, \dots, \theta_K^{**}\right)$ with counts $(\boldsymbol{n}_1, \dots, \boldsymbol{n}_d)$ is characterized as the superposition of two independent processes on $(\mathbb{R}^+)^d \times \Theta$:*

$$(\mu_1, \dots, \mu_d) \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d, \boldsymbol{U}_n \overset{\text{d}}{=} \left(\mu_1^{(a)}, \dots, \mu_d^{(a)}\right) + \left(\mu_1^{(na)}, \dots, \mu_d^{(na)}\right), \text{ where:}$$

(i) the process of allocated components $\left(\mu_1^{(a)}, \dots, \mu_d^{(a)}\right)$ equals

$$\mu_j^{(a)} = \sum_{k=1}^{K} S_{j,k}^{(a)} \delta_{\theta_k^{**}}, \text{ as } j = 1, \dots, d,$$

where the random variables $S_{j,k}^{(a)}$, for $j \in \{1, \dots, d\}$ and $k \in \{1, \dots, K\}$, are independent with densities on $\mathbb{R}^+$ given by

$$f_{S_{j,k}^{(a)}}(s) \propto e^{-u_j s} s^{n_{j,k}} h_j(s);$$

(ii) the process of non-allocated components $\left(\mu_1^{(na)}, \dots, \mu_d^{(na)}\right)$ is a Vec-IFPP $\left(q_M^*, H^*, P_0\right)$*, with $H^*(\mathrm{d}\boldsymbol{s}) = H_1^*(\mathrm{d}s_1) \times \cdots \times H_d^*(\mathrm{d}s_d)$ and*

$$H_j^*(\mathrm{d}s) \propto e^{-u_j s} h_j(s) \mathrm{d}s, \qquad q_M^*(m) \propto q_M(m+K) \frac{(m+K)!}{m!} \prod_{j=1}^{d} \left(\psi_j\left(u_j\right)\right)^m.$$

(iii) Moreover, the conditional distribution of $\boldsymbol{U}_n$ depends on the distinct values only through their counts and it has density on $(\mathbb{R}^+)^d$ given by

$$f_{U_n}(\boldsymbol{u} \mid \boldsymbol{n}_1, \dots, \boldsymbol{n}_d) \propto \Psi(K, \boldsymbol{u}) \prod_{j=1}^{d} \left\{ \frac{u_j^{n_j - 1}}{\Gamma(n_j)} \prod_{k=1}^{K} \kappa(u_j, n_{j,k}) \right\}.$$

The proofs of the results stated in the current section are given in Section D.3 and rely on a point process construction of the Vec-NIFPP prior detailed in Section D.1.

D.2.2 Properties of Vec-IFPP

In this section we derive the main properties of Vec-IFPP, in particular we characterize the higher order reduced Palm distribution associated to any IFPP defined on some Polish space $\mathbb{G}$ as well as its Laplace

functional.

Theorem D.2.4. *Let $\Phi \sim \text{IFPP}(q_M, \nu)$ defined on a Polish space $\mathbb{G}$, where q_M is a discrete distribution over $\{0, 1, 2, \dots\}$ and ν is probability distribution on $\mathbb{G}$, independent of q_M. The higher order reduced Palm version $\Phi^!_{\boldsymbol{\xi}}$ associated to Φ at $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \dots \boldsymbol{\xi}_k)$ is a point process*

$$\Phi^!_{\boldsymbol{\xi}} = \sum_{m=1}^{M} \delta_{\boldsymbol{\xi}_m}$$

such that $\Phi^!_{\boldsymbol{\xi}} \sim \text{IFPP}(q^!_M, H)$. In particular, $q^!_M$ is defined as

$$q^!_M(m) = \frac{1}{E[M^{(k)}]} q_M(m+k) \frac{(m+k)!}{m!}, \tag{5.29}$$

where

$$E[M^{(k)}] := E[M(M-1)\dots(M-k+1)] = \sum_{m=0}^{\infty} q_M(m+k) \frac{(m+k)!}{m!}$$

is the normalizing constant of $q^!_M$ and the atoms ξ_m are still i.i.d. according to H.

Proof. We write the higher order reduced Campbell-Little-Mecke (CLM) formula (4.24) in our case:

$$E\left[\int_{\mathbb{G}^k} f\left(\xi_1, \dots, \xi_k, \Phi - \sum_{j=1}^{k} \delta_{\xi_k}\right) \Phi^{(k)}(\mathrm{d}\boldsymbol{\xi})\right] = \int_{\mathbb{G}^k} E\left[f\left(\xi_1, \dots, \xi_k, \Phi^!_{\boldsymbol{\xi}}\right)\right] \mathrm{M}_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{\xi}), \tag{5.30}$$

where $\mathrm{d}\boldsymbol{\xi} = (\mathrm{d}\xi_1, \dots, \mathrm{d}\xi_k)$ and $\mathrm{M}_{\Phi^{(k)}}$ is the k-th factorial moment measure of Φ. We now focus on the left-hand side of (5.30). Using the definition of k-th factorial process, the integral inside the expected value is

$$\begin{aligned}\int_{\mathbb{G}^k} f\left(\xi_1, \dots, \xi_k, \Phi - \sum_{j=1}^{k} \delta_{\xi_k}\right) \Phi^{(k)}(\mathrm{d}\boldsymbol{\xi}) &= \sum_{\boldsymbol{i} \in \Delta^{(k)}} f\left(\xi_{i_1}, \dots, \xi_{i_k}, \sum_{m=1}^{M} \delta_{\xi_m} - \sum_{j=1}^{k} \delta_{\xi_{i_j}}\right) \\ &= \sum_{\boldsymbol{i} \in \Delta^{(k)}} f\left(\xi_{i_1}, \dots, \xi_{i_k}, \sum_{j=1}^{M-k} \delta_{\xi_j}\right),\end{aligned}$$

where the set of k-tuple $\Delta^{(k)}$ is defined as $\Delta^{(k)} := \{\boldsymbol{i} = (i_1, \dots, i_k) \ : \ i_j \in \{1, \dots, k\} \text{ and } i_h \neq i_j \text{ for all } h \neq j\}$. In the equation above, the equality from the first to the second line holds, given that $k \leq M$, since k out of M are subtracted from the process. The order of the atoms does not matter since they are exchangeable, hence it is fine to remove the first k ones. Then, using the result above for the inner integral and by definition of IFPP, the expected value on the left-hand side of (5.30) equals

$$\begin{aligned}&E\left[\int_{\mathbb{G}^k} f\left(\xi_1, \dots, \xi_k, \Phi - \sum_{j=1}^{k} \delta_{\xi_k}\right) \Phi^{(k)}(\mathrm{d}\boldsymbol{\xi})\right] \\ &\qquad = \sum_{m=k}^{\infty} q_M(m) \sum_{\boldsymbol{i} \in \Delta^{(k)}} E\left[f\left(\xi_{i_1}, \dots, \xi_{i_k}, \sum_{j=1}^{m-k} \delta_{\xi_j}\right)\right]\end{aligned} \tag{5.31}$$

where the outer sum is constrained to values $m \geq k$; otherwise, the inner sum is not even defined. Now,

we focus on the expected value on the right-hand side. Note that this is an integral on $\mathbb{G}^m$ with respect to $(\xi_1, \dots, \xi_m)$. Using Fubini's theorem, such integral can be written in two iterative integrals, one on the first k atoms and one on the last $m-k$ ones:

$$\begin{aligned}\sum_{\boldsymbol{i}\in\Delta^{(k)}} E\left[f\left(\xi_{i_1},\dots,\xi_{i_k},\sum_{j=1}^{m-k}\delta_{\xi_j}\right)\right] &= \sum_{\boldsymbol{i}\in\Delta^{(k)}} E\left[\int_{\mathbb{G}^k} f\left(x_1,\dots,x_k,\sum_{j=1}^{m-k}\delta_{\xi_j}\right)P_0^k(\mathrm{d}\boldsymbol{x})\right] \\ &= \frac{m!}{(m-k)!}\int_{\mathbb{G}^k} E\left[f\left(x_1,\dots,x_k,\sum_{j=1}^{m-k}\delta_{\xi_j}\right)\right]\prod_{j=1}^{k}P_0(\mathrm{d}x_j)\end{aligned} \tag{5.32}$$

where $\mathrm{d}\boldsymbol{x} = (\mathrm{d}x_1, \dots, \mathrm{d}x_k)$ and the expected value is now taken with respect to the law of the final $m-k$ atoms. Then, the final equality follows using Fubini's theorem once again and by noticing that the function in the sum does not depend on the index $\boldsymbol{i}$; in addition we have observed that the cardinality of set $\Delta^{(k)}$ equals $\binom{m}{k}/k! = m!/(m-k)!$. Then, we complete our manipulation of the left-hand side of (5.30) by plugging (5.32) into (5.31), and we exploit Fubini's theorem to exchange the integral and the series. Therefore, renaming $\xi = x$, (5.30) is equivalent to

$$\begin{aligned}\sum_{m=k}^{\infty} q_M(m)\frac{m!}{(m-k)!}\int_{\mathbb{G}^k} E\left[f\left(\xi_1,\dots,\xi_k,\sum_{j=1}^{m-k}\delta_{\xi_j}\right)\right]\prod_{j=1}^{k}P_0(\mathrm{d}\xi_j) \\ = \int_{\mathbb{G}^k} E\left[f\left(\xi_1,\dots,\xi_k,\Phi^!_{\boldsymbol{\xi}}\right)\right]\mathrm{M}_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{\xi}).\end{aligned} \tag{5.33}$$

We conclude using identity (5.33) to identify $\Phi^!_{\boldsymbol{\xi}}$. To this aim, we recall that $\mathrm{M}_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{\xi}) = E[M^{(k)}]\prod_{j=1}^{k}P_0(\xi_j)$ (Baccelli et al., 2020) and, by changing the index of the summation, we get

$$\begin{aligned}\sum_{m=0}^{\infty} q_M(m+k)\frac{(m+k)!}{m!}\frac{1}{E[M^{(k)}]}\int_{\mathbb{G}^k} E\left[f\left(\xi_1,\dots,\xi_k,\sum_{h=1}^{m}\delta_{\xi_h}\right)\right]\mathrm{M}_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{\xi}) \\ = \int_{\mathbb{G}^k} E\left[f\left(\xi_1,\dots,\xi_k,\Phi^!_{\boldsymbol{\xi}}\right)\right]\mathrm{M}_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{\xi}).\end{aligned}$$

Then, the statement of the theorem simply follows by identification. □

We now state the main results we use in subsequent proofs, these just trivially follow from Theorem D.2.4.

Corollary 1. *Let $\Phi \sim$ Vec-IFPP(q_M, H, P_0) defined on $(\mathbb{R}^+)^d \times \mathbb{X}$, where q_M is a discrete distribution over $\{1, 2, 3, \dots\}$, H and P_0 are probability distributions over $(\mathbb{R}^+)^d$ and $\mathbb{X}$, respectively, such that q_M, H and P_0 are independent. Then, the Palm distribution $\{P_\Phi^{\boldsymbol{s},x}\}_{(\boldsymbol{s},x)\in(\mathbb{R}^+)^d\times\mathbb{X}}$ is the distribution of the point process $\delta_{(\boldsymbol{s},x)} + \Phi^!_{\boldsymbol{s},x}$, where $\Phi^!_{\boldsymbol{s},x} \sim$ Vec-IFPP$(q^!_M, H, P_0)$ and $q^!_M$ is given in (5.29) by setting $k = 1$.*

Similarly, let $(\boldsymbol{s}, \boldsymbol{x}) = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_k, x_1, \dots, x_k)$, then the higher order Palm distribution $\{P_\Phi^{\boldsymbol{s},\boldsymbol{x}}\}_{(\boldsymbol{s},x)\in(\mathbb{R}^+)^{dk}\times\mathbb{X}^k}$ is the distribution of the point process $\sum_{i=1}^{k}\delta_{(\boldsymbol{s}_i,x_i)} + \Phi^!_{\boldsymbol{s},\boldsymbol{x}}$, where $\Phi^!_{\boldsymbol{s},\boldsymbol{x}} \sim$ Vec-IFPP$(q^!_M, H, P_0)$ and $q^!_M$ is given in (5.29).

In particular, note that $\Phi^!_{\boldsymbol{s},\boldsymbol{x}}$ in Corollary 1 depends on $(\boldsymbol{s}, \boldsymbol{x})$ only through the length k. Moreover, see that $q^!_M(0) > 0$ even if $q_M(0) = 0$.

The following result provides the Laplace functional of a Vec-IFPP(q_M, H, P_0) distributed point process.

Lemma D.2.5. *Let* $\Phi \sim$ Vec-IFPP(q_M, H, P_0) *on* $\mathbb{G} = (\mathbb{R}^+)^d \times \mathbb{X}$*, where* q_M *is a discrete distribution over* $\{0, 1, 2, \dots\}$*,* H *is probability distribution on* $(\mathbb{R}^+)^d$*, independent of* q_M *and* P_0 *is a probability measure on* $\mathbb{X}$*, independent of* H *and* q_M*. Then, for any* $f : (\mathbb{R}^+)^d \times \mathbb{X} \to \mathbb{R}^+$ *the Laplace functional of* Φ *is*

$$\begin{aligned} \mathrm{L}_\Phi[f] :&= E\left[\exp\left\{-\int_{(\mathbb{R}^+)^d \times \mathbb{X}} f(\boldsymbol{s}, x)\Phi(\mathrm{d}\boldsymbol{s}, \mathrm{d}x)\right\}\right] \\ &= \sum_{m=0}^{\infty} q_M(m)\left(\int_{\mathbb{G}} e^{-f(\boldsymbol{s},x)} H(\mathrm{d}\boldsymbol{s})P_0(\mathrm{d}x)\right)^m. \end{aligned}$$

Proof. First recall the identity $\mathrm{L}_\Phi[f] = \mathrm{G}_\Phi(e^{-f})$ between the Laplace functional and the generating functional. Then, we use (Baccelli et al., 2020, Lemma 4.3.14) to write

$$\begin{aligned} \mathrm{L}_\Phi[f] &= E\left[\prod_{(\boldsymbol{s},x)\in\Phi} \exp\{-f(\boldsymbol{s}, x)\}\right] = \sum_{m=0}^{\infty} q_M(m) E\left[\prod_{(\boldsymbol{s},x)\in\Phi} \exp\{-f(\boldsymbol{s}, x)\} \mid \Phi(\mathbb{G}) = m\right] \\ &= q_M(0) + \sum_{m=1}^{\infty} q_M(m) \int_{\mathbb{G}^m} \left(\prod_{i=1}^{m} \exp\{-f(\boldsymbol{s}_i, x_i)\}\right) H^m(d\boldsymbol{s}_1, \dots, \boldsymbol{s}_m) P_0^m(\mathrm{d}x_1, \dots, \mathrm{d}x_m). \end{aligned}$$

We now leverage the mutual independence between unnormalized weights and the fact that atoms are i.i.d. to obtain

$$\mathrm{L}_\Phi[f] \;=\; \sum_{m=0}^{\infty} q_M(m)\left(\int_{\mathbb{G}} \exp\{-f(\boldsymbol{s}, x)\} H(\mathrm{d}\boldsymbol{s})P_0(\mathrm{d}x)\right)^m,$$

which concludes the proof. □

In particular, Lemma D.2.5 can be used to compute the Laplace functional of the random measures $(\mu_1, \dots, \mu_d)$ defined in (5.20).

Lemma D.2.6. *Let* $\Phi \sim$ Vec-IFPP(q_M, H, P_0) *be defined as in Lemma D.2.5. Let* $(\mu_1, \dots, \mu_d)$ *be random measures on* $\mathbb{X}$ *defined through* Φ *as in* (5.20)*. Then, for any set of measurable functions* $f_1, \dots, f_d$ *such that each* $f_j : \mathbb{X} \to \mathbb{R}^+$*, we have*

$$E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{X}} f_j(x)\mu_j(\mathrm{d}x)\right\}\right] = \sum_{m=0}^{\infty} q_M(m)\left(\int_{\mathbb{G}} \exp\left\{-\sum_{j=1}^{d} f_j(x)s_j\right\} H(\mathrm{d}\boldsymbol{s})P_0(\mathrm{d}x)\right)^m.$$

Proof. To prove the result, it is enough to show that this is a special case of Lemma D.2.5. Indeed,

relying on Equation (5.20), we can express $\mu_j(\mathrm{d}x)$ as $\mu_j(\mathrm{d}x) = \int_{\mathbb{G}} \delta_{\tau_m}(\mathrm{d}x) s_j \phi(\mathrm{d}\boldsymbol{s}, \mathrm{d}\tau)$, then

$$E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{X}} f_j(x)\mu_j(\mathrm{d}x)\right\}\right] = E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{X}} f_j(x)\int_{\mathbb{G}} \delta_y(\mathrm{d}x) s_j \Phi(\mathrm{d}\boldsymbol{s}, \mathrm{d}y)\right\}\right]$$
$$= E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}} f_j(x) s_j \Phi(\mathrm{d}\boldsymbol{s}, \mathrm{d}x)\right\}\right] = \mathrm{L}_{\Phi}[g],$$

where $g(\boldsymbol{s}, x) = \sum_{j=1}^{d} f_j(x) s_j$. The statement now follows by an application of Lemma D.2.5 with $f(\boldsymbol{s}, x) = g(\boldsymbol{s}, x)$. □

The following lemma is extensively used throughout our proofs. It consists of the Laplace transform of the reduced Palm version of Φ at $(\boldsymbol{s}, \boldsymbol{x})$, evaluated at a special choice of f function.

Lemma D.2.7. *Let Φ be defined as in Lemma D.2.5. Let $(\boldsymbol{s}, \boldsymbol{x}) = (s_1, \ldots, s_k, x_1, \ldots, x_k)$ and let $\Phi^{!}_{\boldsymbol{s},\boldsymbol{x}}$ be the higher order reduced Palm version of Φ at $(\boldsymbol{s}, \boldsymbol{x})$. Then, for any $(u_1, \ldots, u_d) \in (\mathbb{R}^+)^d$, the following holds*

$$E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}} u_j w_j \Phi^{!}_{\boldsymbol{s},\boldsymbol{x}}(\mathrm{d}\boldsymbol{w}, \mathrm{d}y)\right\}\right] = \frac{1}{E[M^{(k)}]}\Psi(k, \boldsymbol{u})$$

where $\Psi(k, \boldsymbol{u})$ is given by

$$\Psi(k, \boldsymbol{u}) = \sum_{m=0}^{\infty} q_M(m+k)\frac{(m+k)!}{m!}\prod_{j=1}^{d}\left(\psi_j(u_j)\right)^m, \tag{5.34}$$

and $\psi_j(u_j)$ is the Laplace transform of a random variable $S_j \sim H_j$, namely

$$\psi_j(u_j) = E[e^{-u_j S_j}] = \int_0^{\infty} e^{-u_j s} H_j(\mathrm{d}s).$$

Proof. The proof follows by combining the reduced Palm characterization given in Corollary 1 and Lemma D.2.5 with $f(\boldsymbol{w}, y) = \sum_{j=1}^{d} u_j w_j$. In particular, note that $f(\boldsymbol{w}, y)$ does not depend on y. □

D.3 Proofs of the main results

D.3.1 Proof of Equation (5.10)

Lemma D.2.7 is extensively used throughout this section and it introduces the key quantity $\Psi(k, \boldsymbol{u})$ defined in Equation (5.10). Hence, we start this section proving such result in the special case of HMFM, that is when q_M is chosen to be a 1−shifted Poisson distribution with parameter Λ. Namely,

$$\mathbb{P}(M = m \mid \Lambda) = q_M(m) = \frac{e^{-\Lambda}\Lambda^{m-1}}{(m-1)!}, \quad \text{for } m = 1, 2, \ldots \tag{5.35}$$

In such a case, the $\Psi(K, \boldsymbol{u})$ function, defined in Equation (5.34) for $K \geq 0$ and $\boldsymbol{u} \in (\mathbb{R})^d$ equals

$$\begin{aligned}\Psi(K, \boldsymbol{u}) &= \sum_{m=0}^{\infty} \frac{(m+K)!}{m!} \frac{e^{-\Lambda}\Lambda^{m+K-1}}{(m+K-1)!} \prod_{j=1}^{d} \left(\psi_j(u_j)\right)^m \\ &= e^{-\Lambda}\Lambda^{K-1} \sum_{m=0}^{\infty} \frac{(m+K)\Lambda^m}{m!} \psi(\boldsymbol{u})^m,\end{aligned}$$

where $\psi(\boldsymbol{u}) = \prod_{j=1}^{d}(\psi_j(u_j)$. We note that, in the case under consideration, $\psi_j(u_j) = 1/(u_j+1)^{\gamma_j} < 1$, hence, the series does converge. As a consequence, we can split the series in two parts

$$\Psi(K, \boldsymbol{u}) = e^{-\Lambda}\Lambda^{K-1} \left(\sum_{m=1}^{\infty} \frac{\Lambda^m}{(m-1)!} \psi(\boldsymbol{u})^m + K \sum_{m=0}^{\infty} \frac{\Lambda^m}{(m)!} \psi(\boldsymbol{u})^m \right),$$

where we point out that the first sum starts from $m = 1$ because we simplified $m/m!$, which is exactly equal to 0 when $m = 0$. Then, we change of the index in the first sum ($h = m - 1$)

$$\Psi(K, \boldsymbol{u}) = e^{-\Lambda}\Lambda^{K-1} \left(\Lambda\psi(\boldsymbol{u}) \sum_{h=0}^{\infty} \frac{1}{h!} \left(\Lambda\psi(\boldsymbol{u})\right)^h + K \sum_{m=0}^{\infty} \frac{\Lambda^m}{(m)!} \prod_{j=1}^{d} \left(\psi_j(u_j)\right)^m \right).$$

The result follows after recognizing the exponential series

$$\Psi(K, \boldsymbol{u}) = e^{-\Lambda}\Lambda^{K-1} \left(\Lambda\psi(\boldsymbol{u}) e^{\Lambda\psi(\boldsymbol{u})} + K e^{\Lambda\psi(\boldsymbol{u})} \right) = \Lambda^{K-1} \left(K + \Lambda\psi(\boldsymbol{u})\right) e^{-\Lambda(1-\psi(\boldsymbol{u}))}.$$

D.3.2 Proof of Theorem D.2.1

We compute the pEPPF by resorting to the definition given by Camerlenghi et al. (2019b):

$$\Pi_K^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d) = E\left[\int_{\mathbb{X}^K} \prod_{j=1}^{d} \prod_{k=1}^{K} P_j(\mathrm{d}\theta_k^{**})^{n_{j,k}} \right]. \tag{5.36}$$

By an application of Fubini's theorem, we can exchange the integral and the expected value in (5.36). Thus, we focus on the evaluation of the following expected value:

$$\begin{aligned} E\left[\prod_{j=1}^{d} \prod_{k=1}^{K} P_j(\mathrm{d}\theta_k^{**})^{n_{j,k}} \right] &= E\left[\prod_{j=1}^{d} \prod_{k=1}^{K} \frac{1}{\mu_j(\mathbb{X})^{n_{j,k}}} \mu_j(\mathrm{d}\theta_k^{**})^{n_{j,k}} \right] \\ &= E\left[\prod_{j=1}^{d} \prod_{k=1}^{K} \left\{ \int_0^{\infty} \frac{u_j^{n_{j,k}-1}}{\Gamma(n_{j,k})} e^{-u_j\mu_j(\mathbb{X})} \mathrm{d}u_j \right\} \mu_j(\mathrm{d}\theta_k^{**})^{n_{j,k}} \right] \\ &= \int_{[0,\infty]^d} E\left[\prod_{j=1}^{d} \frac{u_j^{n_j-1}}{\Gamma(n_j)} e^{-u_j\mu_j(\mathbb{X})} \prod_{k=1}^{K} \mu_j(\mathrm{d}\theta_k^{**})^{n_{j,k}} \right] \mathrm{d}\boldsymbol{u} \\ &= \int_{[0,\infty]^d} \prod_{j=1}^{d} \frac{u_j^{n_j-1}}{\Gamma(n_j)} E\left[\exp\left\{ -\sum_{j=1}^{d} u_j\mu_j(\mathbb{X}) \right\} \prod_{k=1}^{K} \prod_{j=1}^{d} \mu_j(\mathrm{d}\theta_k^{**})^{n_{j,k}} \right] \mathrm{d}\boldsymbol{u}, \end{aligned} \tag{5.37}$$

where $\mathrm{d}\boldsymbol{u} = (\mathrm{d}u_1, \ldots, \mathrm{d}u_d)$. The first equality in (5.37) follows by definition of P_j, the second one by the identity $\frac{1}{x^n} = \int_0^\infty \frac{u^{n-1}}{\Gamma(n)} e^{-ux} dx$, the third one follows by Fubini's theorem and recalling that $\sum_{k=1}^K n_{j,k} = n_j$ while the fourth one is just a convenient arrangement of terms. Observe that the product of the random measures μ_j's in the last term of (5.37) equals

$$\begin{aligned}\prod_{k=1}^K \prod_{j=1}^d \mu_j(\mathrm{d}\theta_k^{**})^{n_{j,k}} &= \prod_{k=1}^K \int_{\mathbb{G}} \delta_{x_k}(\mathrm{d}\theta_k^{**}) \prod_{j=1}^d s_{j,k}^{n_{j,k}} \Phi(\mathrm{d}\boldsymbol{s}_k, \mathrm{d}x_k) \\ &= \int_{\mathbb{G}^K} \prod_{k=1}^K \left(\delta_{x_k}(\mathrm{d}\theta_k^{**}) \prod_{j=1}^d s_{j,k}^{n_{j,k}} \right) \Phi^K(\mathrm{d}\boldsymbol{s}, \mathrm{d}\boldsymbol{x}),\end{aligned} \tag{5.38}$$

where $\boldsymbol{s}_k = (s_{1,k}, \ldots, s_{d,k})$ are $(\mathbb{R}^+)^d$-dimensional integration variables representing the unnormalized weights of the Vec-IFPP with atoms x_k. Then, $\boldsymbol{x} = (x_1, \ldots, x_k)$ is a k-dimensional vector while $\boldsymbol{s} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ is a k-dimensional vector of $(\mathbb{R}^+)^d$-dimensional components. Finally, $\Phi^K(\mathrm{d}\boldsymbol{s}, \mathrm{d}\boldsymbol{x})$ is the K-th power point process of Φ, as defined in Section C.1.3. See Baccelli et al. (2020, Ch. 1.3.1) for further details.

Plugging (5.38) into the last member of (5.37) and exploiting the definition of $\mu_j(\mathbb{X}) = \int_{\mathbb{G}} w_j \Phi(\mathrm{d}\boldsymbol{w}, \mathrm{d}z)$, we get

$$\begin{aligned}E\left[\prod_{j=1}^d \prod_{k=1}^K P_j(\mathrm{d}\theta_k^{**})^{n_{j,k}} \right] = \int_{[0,\infty]^d} \prod_{j=1}^d \frac{u_j^{n_j - 1}}{\Gamma(n_j)} E\left[\int_{\mathbb{G}^K} \prod_{k=1}^K \left(\delta_{x_k}(\mathrm{d}\theta_k^{**}) \prod_{j=1}^d s_{j,k}^{n_{j,k}} \right) \right. \\ \left. \times \exp\left\{ -\sum_{j=1}^d \int_{\mathbb{G}} u_j w_j \Phi(\mathrm{d}\boldsymbol{w}, \mathrm{d}z) \right\} \Phi^K(\mathrm{d}\boldsymbol{s}, \mathrm{d}\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{u}.\end{aligned} \tag{5.39}$$

Now, we only focus on the expected value on the right-hand side of (5.39). In the sequel, we assume that $d\theta_k^{**}$ are infinitesimally disjoint sets which do not overlap, say balls centrered at θ_k^{**} with a sufficiently small radius $\varepsilon \to 0$. We now apply the higher order CLM formula, stated in Equation (4.23), with

$$f(\boldsymbol{s}, \boldsymbol{x}, \Phi) = \prod_{k=1}^K \left(\delta_{x_k}(\mathrm{d}\theta_k^{**}) \prod_{j=1}^d s_{j,k}^{n_{j,k}} \right) \exp\left\{ -\sum_{j=1}^d \int_{\mathbb{G}} u_j w_j \Phi(\mathrm{d}\boldsymbol{w}, \mathrm{d}z) \right\}.$$

As a consequence, the expected value under study boils down to

$$
\begin{aligned}
&E\left[\int_{\mathbb{G}^K}\prod_{k=1}^{K}\left(\delta_{x_k}(\mathrm{d}\theta_k^{**})\prod_{j=1}^{d}s_{j,k}^{n_{j,k}}\right)\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}u_j w_j\Phi(\mathrm{d}\boldsymbol{w},\mathrm{d}z)\right\}\Phi^K(\mathrm{d}\boldsymbol{s},\mathrm{d}\boldsymbol{x})\right]\\
&=\int_{\mathbb{G}^K}\prod_{k=1}^{K}\left(\delta_{x_k}(\mathrm{d}\theta_k^{**})\prod_{j=1}^{d}s_{j,k}^{n_{j,k}}\right)\\
&\qquad\times E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}u_j w_j\left(\Phi^!_{\boldsymbol{s},\boldsymbol{x}}+\sum_{k=1}^{K}\delta_{(s_k,x_k)}\right)(\mathrm{d}\boldsymbol{w},\mathrm{d}z)\right\}\right]\mathrm{M}_{\Phi^K}(\mathrm{d}\boldsymbol{s},\mathrm{d}\boldsymbol{x})\\
&=\int_{\mathbb{G}^K}\prod_{k=1}^{K}\left(\delta_{x_k}(d\theta_k^{**})\prod_{j=1}^{d}s_{j,k}^{n_{j,k}}\right)\exp\left\{-\sum_{j=1}^{d}\sum_{k=1}^{K}u_j s_{j,k}\right\}\\
&\qquad\times E\left[\exp\{-\sum_{j=1}^{d}\int_{\mathbb{G}}u_j w_j\Phi^!_{\boldsymbol{s},\boldsymbol{x}}(\mathrm{d}\boldsymbol{w},\mathrm{d}z)\}\right]E\left[M^{(K)}\right]H^K(\mathrm{d}\boldsymbol{s})P_0^K(\mathrm{d}\boldsymbol{x}).
\end{aligned}
\tag{5.40}
$$

Note that in the second equality of Equation (5.40) we replaced $\mathrm{M}_{\Phi^K}(\mathrm{d}\boldsymbol{s},\mathrm{d}\boldsymbol{x})$ with

$$
\mathrm{M}_{\Phi^{(K)}}(\mathrm{d}\boldsymbol{s},\mathrm{d}\boldsymbol{x})=E\left[M^{(K)}\right]H^K(\mathrm{d}\boldsymbol{s})P_0^K(\mathrm{d}\boldsymbol{s}),
$$

because the sets $\mathrm{d}\theta_k^{**}$ are infinitesimally pairwise disjoint. The higher order reduced Palm version $\Phi^!_{\boldsymbol{s},\boldsymbol{x}}$ of Φ at $(\boldsymbol{s},\boldsymbol{x})$ has been defined in Corollary 1. The integral with respect to $\mathbb{G}^K$ is now simple to compute as it factorizes in the product of the base measures evaluated in small sets centrered over the distinct values, namely,

$$
\begin{aligned}
&E\left[\int_{\mathbb{G}^K}\prod_{k=1}^{K}\left(\delta_{x_k}(\mathrm{d}\theta_k^{**})\prod_{j=1}^{d}s_{j,k}^{n_{j,k}}\right)\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}u_j w_j\Phi(\mathrm{d}\boldsymbol{w},\mathrm{d}z)\right\}\Phi^K(\mathrm{d}\boldsymbol{s},\mathrm{d}\boldsymbol{x})\right]\\
&=\prod_{k=1}^{K}P_0(\mathrm{d}\theta_k^{**})E\left[M^{(K)}\right]\int_{(\mathbb{R}^+)^{dK}}\prod_{j=1}^{d}\prod_{k=1}^{K}e^{-u_j s_{j,k}}s_{j,k}^{n_{j,k}}\\
&\qquad\times E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}u_j w_j\Phi^!_{\boldsymbol{s},\boldsymbol{x}}(\mathrm{d}\boldsymbol{w},\mathrm{d}z)\right\}\right]\prod_{k=1}^{K}\prod_{j=1}^{d}H_j(\mathrm{d}s_{j,k}).
\end{aligned}
\tag{5.41}
$$

Then, we use Lemma D.2.7 to compute the expected value appearing in the final line of Equation (5.41). This allows us to eliminate the normalizing constant $E\left[M^{(K)}\right]$. Moreover, we note that such a result does not depend on $\boldsymbol{s}$. Hence, plugging it into (5.39), we get

$$
\begin{aligned}
&E\left[\prod_{j=1}^{d}\prod_{k=1}^{K}P_j(\mathrm{d}\theta_k^{**})^{n_{j,k}}\right]\\
&\quad=\int_{[0,\infty]^d}\prod_{j=1}^{d}\left\{\frac{u_j^{n_j-1}}{\Gamma(n_j)}\prod_{k=1}^{K}\kappa_j\left(u_j,n_{j,k}\right)\right\}\Psi(K,\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\prod_{k=1}^{K}P_0(\mathrm{d}\theta_k^{**})
\end{aligned}
\tag{5.42}
$$

where $\kappa_j\left(u_j, n_{j,k}\right) = \int_0^\infty e^{-u_j s_j} s_j^{n_{j,k}} H_j(\mathrm{d}s_j)$. We can now exploit (5.42) to evaluate the pEPPF in (5.36), and the result easily follows by integrating out the θ_k^{**}.

D.3.3 Proof of Theorem 5.3.1

The pEPPF induced by the Vec-FDP prior follows plugging Equations (8) and (10) into Theorem D.2.1.

D.3.4 Proof of Theorem D.2.2

We prove the three parts of the theorem separately.

(i) For any Borel set $A \in \mathcal{X}$, we have

$$\begin{aligned} E[P_j(A)] &= E\left[\frac{\mu_j(A)}{\mu_j(\mathbb{X})}\right] = E\left[\int_0^\infty e^{-u\mu_j(\mathbb{X})}\mathrm{d}u \int_{\mathbb{G}} s_j \delta_x(A)\Phi(\mathrm{d}\boldsymbol{s}, \mathrm{d}x)\right] \\ &= \int_0^\infty E\left[\int_{\mathbb{G}} \exp\left\{-\int_{\mathbb{G}} u w_j \Phi(\mathrm{d}\boldsymbol{w}, \mathrm{d}y)\right\} s_j \delta_x(A)\Phi(\mathrm{d}\boldsymbol{s}, \mathrm{d}x)\right]\mathrm{d}u. \end{aligned}$$

The first equality follows by definition of $P_j(A)$, the second one exploits the definition of μ_j as well as the identity $1/x = \int_0^\infty e^{-ux}\mathrm{d}x$. For the third equality we use Fubini's theorem. Finally use the CLM formula, i.e., Equation (4.22), with

$$f(\boldsymbol{s}, x, \Phi) = \exp\left\{-\int_{\mathbb{G}} u w_j \Phi(\mathrm{d}\boldsymbol{w}, \mathrm{d}y)\right\} s_j \delta_x(A).$$

As a consequence, the previous expression becomes

$$\begin{aligned} E[P_j(A)] &= \int_0^\infty \int_{\mathbb{G}} s_j \delta_x(A) E\left[\exp\left\{-\int_{\mathbb{G}} u w_j \left(\Phi^!_{\boldsymbol{s},x} + \delta_{(\boldsymbol{s},x)}\right)(\mathrm{d}\boldsymbol{w}, \mathrm{d}y)\right\}\right] \mathrm{M}_\Phi(\mathrm{d}\boldsymbol{s}, \mathrm{d}x)\mathrm{d}u \\ &= \int_0^\infty \int_{\mathbb{G}} e^{-us_j} s_j \delta_x(A) E\left[\exp\left\{-\int_{\mathbb{G}} u w_j \Phi^!_{\boldsymbol{s},x}(\mathrm{d}\boldsymbol{w}, \mathrm{d}y)\right\}\right] E[M] H(\mathrm{d}\boldsymbol{s}) P_0(\mathrm{d}x)\mathrm{d}u. \end{aligned} \tag{5.43}$$

To conclude, note that $H(\mathrm{d}\boldsymbol{s}) = \prod_{l=1}^d H_l(\mathrm{d}s_l)$ with only the j-th component that does not integrate to 1. Moreover the expected value in the final line of Equation (5.43) can be computed using Lemma D.2.5 with $f(\boldsymbol{w}, y) = u w_j$. Hence,

$$\begin{aligned} E[P_j(A)] &= P_0(A) \int_0^\infty \left(\int_0^\infty e^{-us_j} s_j H_j(\mathrm{d}s_j)\right) \prod_{l\neq j}\left(\int_0^\infty H_l(\mathrm{d}s_l)\right) \\ &\quad \times \left(\sum_{m=0}^\infty q_M(m+1)(m+1)(\psi_j(u))^m\right)\mathrm{d}u \\ &= P_0(A)\int_0^\infty \kappa_j(u,1)\Psi(1,u)\mathrm{d}u = P_0(A). \end{aligned}$$

The last equality holds since the final integral is equal to the pEPPF given in Theorem 5.3.1 when $d = 1$, $n = 1$ and $k = 1$. Hence it is trivially equal to 1.

(iii) Without loss of generality, we set $j = 1$ and $l = 2$. Leveraging on the integral identity $\frac{1}{x^n} =$

$\int_0^\infty \frac{u^{n-1}}{\Gamma(n)} e^{-ux} \mathrm{d}x$, for any Borel sets $A_1, A_2 \in \mathcal{X}$, we get

$$\begin{aligned} &E\left[P_1(A_1)^{n_1} P_2(A_2)^{n_2}\right] \\ &\quad = E\left[\frac{1}{\mu_1(\mathbb{X})^{n_1}\mu_2(\mathbb{X})^{n_2}}\mu_1(A_1)^{n_1}\mu_2(A_2)^{n_2}\right] \\ &\quad = E\left[\int_{[0,\infty]^2} \frac{u_1^{n_1-1}u_2^{n_2-1}}{\Gamma(n_1)\Gamma(n_2)} e^{-(u_1\mu_1(\mathbb{X})+u_2\mu_2(\mathbb{X}))}\mu_1(A_1)^{n_1}\mu_2(A_2)^{n_2}\right]. \end{aligned} \tag{5.44}$$

As done in previous calculations, we now want to expand the definitions of $\mu_j(A_j)^{n_j}$, for $j = 1, 2$, using Equation (5.20). More specifically, we need to express the $\mu_j(A_j)$s as $n = n_1 + n_2$ different integrals, which are explicitly introduced through these expressions

$$\begin{aligned} \mu_1(A_1)^{n_1} &= \left(\int_{\mathbb{G}} \delta_{v_1}(A_1) r_{1,1} \Phi(\mathrm{d}\boldsymbol{r}_1, \mathrm{d}v_1)\right) \cdots \left(\int_{\mathbb{G}} \delta_{v_{n_1}}(A_1) r_{n_1,1} \Phi(\mathrm{d}\boldsymbol{r}_{n_1}, \mathrm{d}v_{n_1})\right), \\ \mu_2(A_2)^{n_2} &= \left(\int_{\mathbb{G}} \delta_{z_1}(A_2) t_{1,2} \Phi(\mathrm{d}\boldsymbol{t}_1, \mathrm{d}z_1)\right) \cdots \left(\int_{\mathbb{G}} \delta_{z_{n_2}}(A_2) t_{n_2,2} \Phi(\mathrm{d}\boldsymbol{t}_{n_2}, \mathrm{d}z_{n_2})\right), \end{aligned}$$

where $\boldsymbol{r}_i = (r_{i,1}, \ldots, r_{i,d})$ for $i = 1, \ldots, n_1$ and $\boldsymbol{t}_i = (t_{i,1}, \ldots, t_{i,d})$ for $i = 1, \ldots, n_2$. As a consequence, Equation (5.44) becomes

$$\begin{aligned} E\left[P_1(A_1)^{n_1} P_2(A_2)^{n_2}\right] = E\Bigg[&\int_{[0,\infty]^2} \frac{u_1^{n_1-1}u_2^{n_2-1}}{\Gamma(n_1)\Gamma(n_2)} e^{-(u_1\mu_1(\mathbb{X})+u_2\mu_2(\mathbb{X}))} \\ &\times \prod_{i=1}^{n_1} \int_{\mathbb{G}} r_{i,1}\delta_{v_i}(A_1)\Phi(\mathrm{d}\boldsymbol{r}_i, \mathrm{d}v_i) \prod_{l=1}^{n_2} \int_{\mathbb{G}} t_{l,2}\delta_{z_l}(A_2)\Phi(\mathrm{d}\boldsymbol{t}_l, \mathrm{d}z_l)\ \mathrm{d}\boldsymbol{u}, \Bigg] \end{aligned}$$

where $\mathrm{d}\boldsymbol{u} = (\mathrm{d}u_1, \mathrm{d}u_2)$. We then collect all n integration variables by defining a vector $\boldsymbol{x} = (v_1, \ldots, v_{n_1}, z_1, \ldots, z_{n_2})$ and a vector of vectors $\boldsymbol{s} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n_1}, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_{n_2})$. Then, the Fubini's theorem implies

$$\begin{aligned} E[P_1(A_1)^{n_1} P_2(A_2)^{n_2}] = &\int_{[0,\infty]^2} \frac{u_1^{n_1-1}u_2^{n_2-1}}{\Gamma(n_1)\Gamma(n_2)} \\ &\times E\left[\int_{\mathbb{G}^n} e^{-(u_1\mu_1(\mathbb{X})+u_2\mu_2(\mathbb{X}))}\delta_{\boldsymbol{x}}(A_1^{n_1}\times A_2^{n_2})\left(\prod_{i=1}^{n_1} s_{i,1}\right)\left(\prod_{i=n_1+1}^{n_1+n_2} s_{i,2}\right)\Phi^n(\mathrm{d}\boldsymbol{s}, \mathrm{d}\boldsymbol{x})\right]\mathrm{d}\boldsymbol{u}. \end{aligned}$$

where $A^n := A \times \cdots \times A$ for n times and $s_{i,j}$ represents the j-th component of the vector in the i-th position of $\boldsymbol{s}$.

Differently from previous computations, it is not convenient to immediately apply the higher order CLM formula. Indeed, integrals are not defined over infinitely small sets but over possibly overlapping sets $A_1^{n_1} \times A_2^{n_2}$. As a consequence, we can not replace the n-power Φ^n with the n-factorial power $\Phi^{(n)}$ point process. Instead, we must compute the expected value using (Baccelli et al., 2020, Lemma 14.E.4)

$$\begin{aligned} E[P_1(A_1)^{n_1} P_2(A_2)^{n_2}] = &\int_{[0,\infty]^2} \frac{u_1^{n_1-1}u_2^{n_2-1}}{\Gamma(n_1)\Gamma(n_2)} \sum_{k=1}^{n} \sum_{(*)_k} \frac{1}{k!} \prod_{j=1}^{2} \binom{n_j}{n_{j,1}, \ldots, n_{j,k}} \\ &\times E\left[\int_{\mathbb{G}^k} e^{-(u_1\mu_1(\mathbb{X})+u_2\mu_2(\mathbb{X}))} \prod_{m=1}^{k}\left(\delta_{y_m}(B_m)\prod_{j=1}^{2} w_{j,m}^{n_{j,m}}\right)\Phi^{(k)}(\mathrm{d}\boldsymbol{w}, \mathrm{d}\boldsymbol{y})\right]\mathrm{d}\boldsymbol{u}, \end{aligned}$$

where we defined the set $B_m := \bigcap_{i=1}^{n_{1,m}} A_1 \cap \bigcap_{i=1}^{n_{2,m}} A_2$ and we are summing over all possible partitions of k elements but, since the order of the elements does not matter, we only consider their counts. Namely, the set $(*)_k$ is defined as

$$(*)_k := \Big\{(n_{1,1} \dots, n_{1,k}, n_{2,1} \dots, n_{2,k}) : n_{j,m} \geq 0 \text{ for } j = 1, 2 \; m = 1, \dots, k, \text{ and } \sum_{m=1}^{k} n_{j,m} = n_j \text{ for } j = 1, 2, \text{ and } n_{1,m} + n_{2,m} \geq 1 \text{ for } m = 1, \dots, k\Big\}.$$

Consequently, note that the integration set B_m is never empty but it is equal to $A_1 \cap A_2$ if $n_{j,m} > 0$ for both $j = 1, 2$ or it is equal to A_1 when $n_{2,m} = 0$ or A_2 if $n_{1,m} = 0$. We now apply the higher order CLM formula (4.24), and we use the expression of the k-th factorial moment measure to get

$$\begin{aligned}
&E[P_1(A_1)^{n_1} P_2(A_2)^{n_2}] \\
&\quad = \int_{[0,\infty]^2} \frac{u_1^{n_1-1} u_2^{n_2-1}}{\Gamma(n_1)\Gamma(n_2)} \sum_{k=1}^{n} \sum_{(*)_k} \frac{1}{k!} \prod_{j=1}^{2} \binom{n_j}{n_{j,1}, \dots, n_{j,k}} \\
&\qquad \times E\left[M^{(k)}\right] \int_{\mathbb{G}^k} \prod_{m=1}^{k} \delta_{y_m}(B_m) \prod_{j=1}^{2} \prod_{m=1}^{k} e^{u_j w_{j,m}} w_{j,m}^{n_{j,m}} \\
&\qquad \times E\left[\exp\left\{-\sum_{j=1}^{2} \int_{\mathbb{G}} u_j s_j \Phi^{!}_{\boldsymbol{w},\boldsymbol{y}}(\mathrm{d}\boldsymbol{s}, \mathrm{d}z)\right\}\right] P_0^k(\mathrm{d}\boldsymbol{y}) H^k(\mathrm{d}\boldsymbol{w}) \mathrm{d}\boldsymbol{u} \qquad (5.45) \\
&\quad = \sum_{k=1}^{n} \Bigg\{ \prod_{m=1}^{k} P_0(B_m) \sum_{(*)_k} \frac{1}{k!} \prod_{j=1}^{2} \binom{n_j}{n_{j,1}, \dots, n_{j,k}} \\
&\qquad \times \int_{[0,\infty]^2} \frac{u_1^{n_1-1} u_2^{n_2-1}}{\Gamma(n_1)\Gamma(n_2)} \prod_{j=1}^{2} \prod_{m=1}^{k} \kappa_j(u_j, n_{j,m}) \Psi(2, \boldsymbol{u}) \mathrm{d}\boldsymbol{u} \Bigg\}.
\end{aligned}$$

To conclude, we recognize that the final integral is the pEPPF when $d = 2$, $K_{(n_1,n_2)} = k$ and counts $\boldsymbol{n}_j = (n_{j,1}, \dots, n_{j,k})$, for each $j = 1, 2$. Moreover, note that in the special case $A_1 = A_2 = A \in \mathcal{X}$, we have $B_m = A$, in particular B_m does not depend on m. In such a case, we have

$$\sum_{k=1}^{n} P_0(A)^k \sum_{(*)_k} \frac{1}{k!} \prod_{j=1}^{2} \binom{n_j}{n_{j,1}, \dots, n_{j,k}} \Pi_k^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \sum_{k=1}^{n_1+n_2} P_0(A)^k \mathbb{P}(K_{(n_1,n_2)} = k), \qquad (5.46)$$

thus, (ii) easily follows from (5.46).

D.3.5 Proof of Theorem 5.3.3

The proof follows the same steps presented in Section D.3.4.

D.3.6 Proofs of Equations (5.14) and (5.15)

Firstly, we prove Equation (5.14). Without loss of generality, we set $j = 1$ and $l = 2$. Then, we recall that when $A = B$, Equation (5.28) boils down to

$$\text{corr}\left(P_1(A), P_2(A)\right) = \frac{\mathbb{P}\left(K_{(1,1)} = 1\right)}{\sqrt{\mathbb{P}\left(K_{1,(2)} = 1\right)\mathbb{P}\left(K_{2,(2)} = 1\right)}}, \tag{5.47}$$

where both the numerator and the denominator admit an explicit representation in terms of pEPPF and EPPF, respectively. We first focus on the numerator in (5.47). Since we are dealing with Vec-FDP, we can exploit the results of Section 5.3 to obtain

$$\begin{aligned}\mathbb{P}\left(K_{(1,1)} = 1\right) &= \Pi_1^{(2)}(1,1) \\ &= \int_{[0,\infty]^2}\left(1 + \frac{\Lambda}{(1+u_1)^{\gamma_1}(1+u_2)^{\gamma_2}}\right) \\ &\times \exp\left\{-\Lambda\left(1 - \frac{1}{(1+u_1)^{\gamma_1}(1+u_2)^{\gamma_2}}\right)\right\}\frac{\gamma_1\gamma_2}{(1+u_1)^{1+\gamma_1}(1+u_2)^{1+\gamma_2}}\mathrm{d}u_1\mathrm{d}u_2.\end{aligned} \tag{5.48}$$

Applying the following change of variables $x_j = \frac{1}{(1+u_j)^{\gamma_j}}$, for $j = 1, 2$, Equation (5.48) equals

$$\begin{aligned}\mathbb{P}\left(K_{(1,1)} = 1\right) &= \int_{[0,1]^2}(1 + \Lambda x_1 x_2)e^{-\Lambda(1-x_1x_2)}\mathrm{d}x_1\mathrm{d}x_2 \\ &= e^{-\Lambda}\left(\int_{[0,1]^2} e^{\Lambda x_1 x_2}\mathrm{d}x_1\mathrm{d}x_2 + \Lambda\int_{[0,1]^2} x_1x_2e^{\Lambda x_1x_2}\mathrm{d}x_1\mathrm{d}x_2\right).\end{aligned}$$

Integrand functions are positive, hence we use Fubini's theorem to compute the bi-dimensional integral as two iterative integrals. In particular, we first integrate by parts with respect to x_2 to get,

$$\begin{aligned}\mathbb{P}\left(K_{(1,1)} = 1\right) &= e^{-\Lambda}\left(\int_0^1 \frac{1}{\Lambda x_1}(e^{\Lambda x_1} - 1)\mathrm{d}x_1 + \int_0^1 e^{\Lambda x_1}\mathrm{d}x_1 - \int_0^1 \frac{1}{\Lambda x_1}(e^{\Lambda x_1} - 1)\mathrm{d}x_1\right) \\ &= e^{-\Lambda}\int_0^1 e^{\Lambda x_1}\mathrm{d}x_1 = \frac{1 - e^{-\Lambda}}{\Lambda}.\end{aligned} \tag{5.49}$$

One can exploit similar arguments to compute the denominator $\mathbb{P}\left(K_{1,(2)} = 1\right)$ in (5.47). Indeed, we note that, marginally, P_1 and P_2 are Finite Dirichlet Processes of Argiento and De Iorio (2022). Then,

$$\mathbb{P}\left(K_{j,(2)} = 1\right) = \Pi_1^{(2)}(2) = \int_0^\infty\left(1 + \frac{\Lambda}{(1+u)^{\gamma_j}}\right)e^{-\Lambda\left(1 - \frac{1}{(1+u)^{\gamma_j}}\right)}\frac{u\gamma_j(1+\gamma_j)}{(1+u)^{2+\gamma_j}}\mathrm{d}u, \tag{5.50}$$

for $j = 1, 2$. The change of variable, $x_j = \frac{1}{(1+u_j)^{\gamma_j}}$, for $j = 1, 2$, is applied to Equation (5.50) which boils down to

$$\mathbb{P}\left(K_{j,(2)} = 1\right) = (\gamma_j + 1)e^{-\Lambda}\int_0^1 (1 + \Lambda x)e^{\Lambda x}(1 - x^{1/\gamma_j})\mathrm{d}x. \tag{5.51}$$

A similar formula holds true for the probability $\mathbb{P}\left(K_{j,(1)} = 1\right)$. Substituting the expressions (5.49) and (5.51) in (5.47), we get Equation (5.14).

We now focus on the proof of Equation (5.15) to obtain the limit of the correlation in Equation (5.14)

as γ_j and γ_l go to either 0 or ∞, simultaneously. We note that the integral $I(\gamma_j, \Lambda)$ in Equation (5.14) does not admit an analytical solution, hence we need to study the limiting cases for small and large values of γ_j.
Firstly, when $\gamma_j \to 0$, we use the dominated convergence theorem to show that

$$\begin{aligned}\lim_{\gamma_j \to 0} I(\gamma_j, \Lambda) &= \lim_{\gamma_j \to 0} \int_0^1 (1+\Lambda x) e^{-\Lambda(1-x)} (1 - x^{1/\gamma_j}) \mathrm{d}x \\ &= e^{-\Lambda} \int_0^1 (1+\Lambda x) e^{\Lambda x} \mathrm{d}x = 1.\end{aligned}$$

The final equality has been obtained integrating by parts after noting that $\int (1+\Lambda x) e^{\Lambda x} dx = x e^{\Lambda x} + c$. Hence, we can evaluate the limit in Equation (5.14) to obtain:

$$\lim_{\gamma_j, \gamma_l \to 0} \operatorname{corr}(P_j(A), P_l(A)) = \frac{1 - e^{-\Lambda}}{\Lambda}.$$

The limiting case for $\gamma_j, \gamma_l \to \infty$ is obtained similarly. Once again, we first exchange the limit and the integral leveraging on the dominated convergence theorem and then we integrate by parts as before to obtain,

$$\begin{aligned}\lim_{\gamma_j \to \infty} (\gamma_j + 1) I(\gamma_j, \Lambda) &= \lim_{\gamma_j \to \infty} (\gamma_j + 1) e^{-\Lambda} \int_0^1 (1+\Lambda x) e^{\Lambda x} (1 - x^{1/\gamma_j}) \mathrm{d}x \\ &= -e^{-\Lambda} \int_0^1 (1+\Lambda x) e^{\Lambda x} \log(x) \mathrm{d}x = \frac{1 - e^{-\Lambda}}{\Lambda}.\end{aligned} \tag{5.52}$$

Hence, the limit of Equation (5.14) can be evaluate on the basis of (5.52), to get

$$\lim_{\gamma_j, \gamma_l \to \infty} \operatorname{corr}(P_j(A), P_l(A)) = 1.$$

D.3.7 Proof of Theorem D.2.3

We now introduce useful notation and a vector of additional variables $\boldsymbol{U}_n = (U_1, \ldots, U_d)$ which helps us to describe Bayesian inference for our model. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ be the collection of all random variables $\theta_{j,i}$ across all possible groups, for $j = 1, \ldots, d$ and $i = 1, \ldots, n_j$.
The conditional distribution of $(\boldsymbol{\theta} \mid \mu_1, \ldots, \mu_d)$ is given by

$$\mathbb{P}(\boldsymbol{\theta} \in d\boldsymbol{\theta} \mid \mu_1, \ldots, \mu_d) = \prod_{j=1}^{d} \frac{1}{T_j^{n_j}} \prod_{i=1}^{n_j} \mu_j \left(\mathrm{d}\theta_{j,i}\right), \tag{5.53}$$

where $T_j = \mu_j(\mathbb{X})$. Similarly to (James et al., 2009; Argiento and De Iorio, 2022), we introduce a vector of auxiliary variables $\boldsymbol{U}_n = (U_1, \ldots, U_d)$ such that $U_j \mid T \overset{\text{ind}}{\sim} \text{Gamma}(n_j, T_j)$. Leveraging on the integral identity $\frac{1}{x^n} = \int_0^\infty \frac{u^{n-1}}{\Gamma(n)} e^{-ux} \mathrm{d}x$ we introduce a suitable augmentation of the underlying probability space and consider the joint conditional distribution of $(\boldsymbol{\theta}, \boldsymbol{U}_n \mid \mu_1, \ldots, \mu_d)$

$$\mathbb{P}\left(\boldsymbol{\theta} \in \mathrm{d}\boldsymbol{\theta}, \boldsymbol{U}_n \in \mathrm{d}\boldsymbol{u} \mid \mu_1, \ldots, \mu_d\right) = \prod_{j=1}^{d} \frac{u_j^{n_j - 1}}{\Gamma(n_j)} e^{-u_j T_j} \mathrm{d}\boldsymbol{u} \prod_{i=1}^{n_j} \mu_j \left(\mathrm{d}\theta_{j,i}\right).$$

As explained in Section 5.2.1, since $(\mu_1, \ldots, \mu_d)$ are almost surely discrete, with positive probability there will be ties within the sample $\boldsymbol{\theta}$. For this reason, $\boldsymbol{\theta}$ is equivalently characterized by the couple $(\boldsymbol{\theta}^{**}, \rho)$, introduced in Section 5.2.1. Moreover, since the order is not relevant, we recall that ρ is further characterized given its number of clusters $K_{(n_1,\ldots,n_d)} = k$ and their counts $(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k)$. As a consequence we may write

$$\mathbb{P}\left(\boldsymbol{\theta} \in \mathrm{d}\boldsymbol{\theta}, \boldsymbol{U}_n \in \mathrm{d}\boldsymbol{u} \mid \mu_1, \ldots, \mu_d\right) = \prod_{j=1}^{d} \frac{u_j^{n_j-1}}{\Gamma(n_j)} e^{-u_j T_j} \mathrm{d}\boldsymbol{u} \prod_{k=1}^{K} \mu_j\left(\mathrm{d}\theta_k^{**}\right)^{n_{j,k}}. \tag{5.54}$$

The augmentation of (5.53) through $\boldsymbol{U}_n$ is made possible since the marginal distribution of $\boldsymbol{U}_n$ exists. Indeed, for each component, say j, it holds that

$$\begin{aligned}\mathbb{P}\left(U_j \in \mathrm{d}u_j\right) &= E\left[\mathbb{P}\left(U_j \in \mathrm{d}u_j \mid T_j\right)\right] = E\left[\frac{u_j^{n_j-1}}{\Gamma(n_j)} T_j^{n_j} e^{-T_j u_j} \mathrm{d}u_j\right] \\ &= \frac{u_j^{n_j-1}}{\Gamma(n_j)} \mathrm{d}u_j \int_0^{\infty} t^{n_j} e^{-t u_j} f_{T_j}(t) \mathrm{d}t,\end{aligned}$$

which is well defined since T_j is, almost surely, a sum of a finite number of positive random variables hence it is finite and admits density f_{T_j}. Finally, leveraging on the independence of the components of $\boldsymbol{U}_n$, it follows that it has a density, with respect to the Lebesgue measure, that is

$$f_{\boldsymbol{U}_n}(\boldsymbol{u}) = \prod_{j=1}^{d} \frac{u_j^{n_j-1}}{\Gamma(n_j)} \int_0^{+\infty} t^{n_j} e^{-t_j u_j} f_{T_j}(t_j) \mathrm{d}t_j. \tag{5.55}$$

To simplify the notation, we drop subscript n for random variables $\boldsymbol{U}_n$ and simply write $\boldsymbol{U} = (U_1, \ldots, U_d)$. Nevertheless, we recall that such a random vector depends on the group sizes $(n_1, \ldots, n_d)$.

Proof. The strategy to prove Theorem 5.3.4 consists in describing the Laplace functional of $(\mu_1, \ldots, \mu_d)$ given $\boldsymbol{\theta}$ and $\boldsymbol{U}$. Let $f_j : \mathbb{X} \to \mathbb{R}^+$ be $\mathcal{X}$ measurable for each $j \in \{1, \ldots, d\}$. Then, following James et al. (2009); Beraha et al. (2025a), we have

$$\begin{aligned}&E\left[\exp\left\{-\sum_{j=1}^{d} \int_{\mathbb{X}} f_j(x)\mu_j(\mathrm{d}x)\right\} \mid \boldsymbol{\theta} \in \mathrm{d}\theta, \boldsymbol{U} \in \mathrm{d}\boldsymbol{u}\right] \\ &= \frac{E\left[\exp\left\{-\sum_{j=1}^{d} \int_{\mathbb{X}} f_j(x)\mu_j(\mathrm{d}x)\right\} \mathbb{P}\left(\boldsymbol{\theta} \in \mathrm{d}\theta, \boldsymbol{U} \in \mathrm{d}\boldsymbol{u} \mid \mu_1, \ldots, \mu_d\right)\right]}{E\left[\mathbb{P}\left(\boldsymbol{\theta} \in \mathrm{d}\theta, \boldsymbol{U} \in \mathrm{d}\boldsymbol{u} \mid \mu_1, \ldots, \mu_d\right)\right]}.\end{aligned} \tag{5.56}$$

We now focus on the numerator of (5.56) since the denominator will follow by setting $f_j = 0$ for each j. We note that Equation (5.54) equals the probability appearing in the expected value. Then, the numerator

under consideration equals to

$$
\begin{aligned}
&E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{X}} f_j(x)\mu_j(\mathrm{d}x)\right\}\mathbb{P}\left(\boldsymbol{\theta}\in\mathrm{d}\theta,\boldsymbol{U}\in\mathrm{d}\boldsymbol{u}\mid\mu_1,\ldots,\mu_d\right)\right]\\
&\quad= E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{X}} f_j(x)\mu_j(\mathrm{d}x)\right\}\prod_{j=1}^{d}\left\{\frac{u_j^{n_j-1}}{\Gamma(n_j)}e^{-u_jT_j}\mathrm{d}\boldsymbol{u}\prod_{k=1}^{K}\mu_j\left(\mathrm{d}\theta_k^{**}\right)^{n_{j,k}}\right\}\right]\\
&\quad= \prod_{j=1}^{d}\left(\frac{u_j^{n_j-1}}{\Gamma(n_j)}\right)\mathrm{d}\boldsymbol{u}\,E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{X}}\left(u_j+f_j(x)\right)\mu_j(\mathrm{d}x)\right\}\prod_{j=1}^{d}\prod_{k=1}^{K}\mu_j\left(\mathrm{d}\theta_k^{**}\right)^{n_{j,k}}\right].
\end{aligned}
$$

Reasoning as for (5.38), the previous expression equals

$$
\begin{aligned}
&E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{X}} f_j(x)\mu_j(\mathrm{d}x)\right\}\mathbb{P}\left(\boldsymbol{\theta}\in\mathrm{d}\theta,\boldsymbol{U}\in\mathrm{d}\boldsymbol{u}\mid\mu_1,\ldots,\mu_d\right)\right]=\prod_{j=1}^{d}\left(\frac{u_j^{n_j-1}}{\Gamma(n_j)}\right)\mathrm{d}\boldsymbol{u}\times\\
&E\left[\int_{\mathbb{G}^k}\prod_{k=1}^{K}\left(\delta_{y_k}(\mathrm{d}\theta_k^{**})\prod_{j=1}^{d}s_{j,k}^{n_{j,k}}\right)\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}w_j\left(u_j+f_j(x)\right)\Phi(\mathrm{d}\boldsymbol{w},\mathrm{d}x)\right\}\Phi^K(\mathrm{d}\boldsymbol{s},\mathrm{d}\boldsymbol{y})\right].
\end{aligned}
\tag{5.57}
$$

Note that (5.57) has the same form of (5.39). Therefore, applying the higher order CLM formula we get

$$
\begin{aligned}
&E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{X}} f_j(x)\mu_j(\mathrm{d}x)\right\}\mathbb{P}\left(\boldsymbol{\theta}\in\mathrm{d}\theta,\boldsymbol{U}\in\mathrm{d}\boldsymbol{u}\mid\mu_1,\ldots,\mu_d\right)\right]\\
&\quad=\prod_{j=1}^{d}\left\{\left(\frac{u_j^{n_j-1}}{\Gamma(n_j)}\right)\mathrm{d}\boldsymbol{u}\prod_{k=1}^{K}\int_0^{\infty}e^{-s_j\left(f_j(\theta_k^{**})+u_j\right)}s^{n_{j,k}}H_j(\mathrm{d}s_j)\right\}\\
&\quad\times E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}w_j\left(u_j+f_j(\theta_k^{**})\right)\Phi^{!}_{\boldsymbol{s},\boldsymbol{y}}(\mathrm{d}\boldsymbol{w},\mathrm{d}x)\right\}\right]E\left[M^{(K)}\right]\prod_{k=1}^{K}P_0(\mathrm{d}\theta_k^{**}).
\end{aligned}
\tag{5.58}
$$

As a consequence, the denominator of (5.56) equals

$$
\begin{aligned}
E\left[\mathbb{P}(\boldsymbol{\theta}\in\mathrm{d}\theta,\boldsymbol{U}\in\mathrm{d}\boldsymbol{u}\mid\mu_1,\ldots,\mu_d)\right]&=\prod_{j=1}^{d}\left\{\frac{u_j^{n_j-1}}{\Gamma(n_j)}\prod_{k=1}^{K}\kappa_j(u_j,n_{j,k})\right\}\\
\times\quad E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}w_ju_j\Phi^{!}_{\boldsymbol{s},\boldsymbol{y}}(\mathrm{d}\boldsymbol{w},\mathrm{d}x)\right\}\right]&E\left[M^{(K)}\right]\prod_{k=1}^{K}P_0(\mathrm{d}\theta_k^{**}).
\end{aligned}
\tag{5.59}
$$

Taking the ratio between (5.58) and (5.59), we obtain

$$
\prod_{j=1}^{d}\prod_{k=1}^{K}\int_0^{\infty}\frac{e^{-s_jf_j(\theta_k^{**})}e^{-s_ju_j}s_j^{n_{j,k}}}{\kappa_j(u_j,n_{j,k})}H_j(\mathrm{d}s_j)\frac{E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}w_j\left(u_j+f_j(x)\right)\Phi^{!}_{\boldsymbol{s},\boldsymbol{y}}(\mathrm{d}\boldsymbol{w},\mathrm{d}x)\right\}\right]}{E\left[\exp\left\{-\sum_{j=1}^{d}\int_{\mathbb{G}}w_ju_j\Phi^{!}_{\boldsymbol{s},\boldsymbol{y}}(\mathrm{d}\boldsymbol{w},\mathrm{d}x)\right\}\right]},
\tag{5.60}
$$

which is the Laplace functional of two independent processes, conditionally to $\boldsymbol{U}$, $K_{(n_1,\ldots,n_d)}$ and $\boldsymbol{\theta}^{**}$. To conclude the proof of the theorem, it is enough to show that the Laplace functionals of the random

measures described in the statement coincide with those in (5.60).

(i) We first show that the first term in Equation (5.60) is the Laplace functional of the vector of random measures $\left(\mu_1^{(a)}, \dots, \mu_d^{(a)}\right)$, where $\mu_j^{(a)} = \sum_{k=1}^K S_{j,k}^{(a)} \delta_{\theta_k^{**}}$. We refer to Chapter 5 for the definition of random variables $S_{j,k}^{(a)}$, for $j = 1, \dots, d$ and $k = 1, \dots, K$. The Laplace transform of $(\mu_1^{(a)}, \dots, \mu_d^{(a)})$ equals

$$
\begin{aligned}
&E\left[\exp\left\{-\sum_{j=1}^d \int_{\mathbb{X}} f_j(x)\mu_j^{(a)}(dx)\right\} \mid \boldsymbol{\theta} \in \mathrm{d}\theta, \boldsymbol{U} \in \mathrm{d}\boldsymbol{u}\right] \\
&\quad = E\left[\prod_{j=1}^d \exp\left\{-\sum_{k=1}^K S_{j,k}^{(a)} f_j(\theta_k^{**})\right\} \mid \boldsymbol{\theta} \in \mathrm{d}\theta, \boldsymbol{U} \in \mathrm{d}\boldsymbol{u}\right] \\
&\quad = E\left[\prod_{j=1}^d \prod_{k=1}^K \exp\left\{-S_{j,k}^{(a)} f_j(\theta_k^{**})\right\} \mid \boldsymbol{\theta} \in \mathrm{d}\theta, \boldsymbol{U} \in \mathrm{d}\boldsymbol{u}\right] \\
&\quad = \prod_{j=1}^d \prod_{k=1}^K \int_0^\infty \frac{e^{s_j\left(u_j + f_j(\theta_k^{**})\right)} s_j^{n_{j,k}}}{\kappa_j(u_j, n_{j,k})} H_j(\mathrm{d}s_j).
\end{aligned}
\tag{5.61}
$$

The first equality exploits the fact that the atoms of each of the $\mu_j^{(a)}$ are non-random. The second equality holds thanks to the independence between the unnormalized jumps and finally for the third equality we computed the Laplace transform of of random variables $S_{j,m}^{(a)}$. The final term of Equation (5.61) is exactly the same term appearing in Equation (5.60), which concludes the proof.

(ii) We now show that the second term in Equation (5.60) is the Laplace functional of the vector of random measures $\left(\mu_1^{(na)}, \dots, \mu_d^{(na)}\right) \sim$ Vec-IFPP(q_M^*, H^*, P_0). An explicit formulation of q_M^* and H^* is here provided. They equal

$$
\begin{aligned}
H_j^*(\mathrm{d}s_j) &= \frac{e^{-u_j s_j} H_j(\mathrm{d}s_j)}{\psi_j(u_j)}, \text{ for } j = 1, \dots, d \\
q_M^*(m) &= \frac{q_M(m+k)\frac{(m+k)!}{m!}\prod_{j=1}^d \left(\psi_j(u_j)\right)^m}{\sum_{h=0}^\infty q_M(h+k)\frac{(h+k)!}{h!}\prod_{j=1}^d \left(\psi_j(u_j)\right)^h},
\end{aligned}
$$

where $\psi_j(u_j) = \int_0^\infty e^{-u_j s_j} H_j(\mathrm{d}s_j)$. To compute the Laplace transform of $\left(\mu_1^{(na)}, \dots, \mu_d^{(na)}\right)$, we use Lemma D.2.6 and we get

$$
\begin{aligned}
&E\left[\exp\left\{-\sum_{j=1}^d \int_{\mathbb{X}} f_j(x)\mu_j^{(na)}(\mathrm{d}x)\right\} \mid \boldsymbol{\theta} \in \mathrm{d}\theta, \boldsymbol{U} \in \mathrm{d}\boldsymbol{u}\right] = \\
&\quad = \sum_{m=0}^\infty q_M^*(m)\left(\int_{\mathbb{G}} \exp\left\{\sum_{j=1}^d -s_j f_j(x)\right\} P_0(\mathrm{d}x)\prod_{j=1}^d H_j^*(\mathrm{d}s_j)\right)^m \\
&\quad = \sum_{m=0}^\infty q_M^*(m)\left(\prod_{j=1}^d \int_0^\infty \int_{\mathbb{X}} \frac{\exp\left\{-s_j\left(f_j(x) + u_j\right)\right\}}{\psi_j(u_j)} H_j(\mathrm{d}s_j) P_0(\mathrm{d}x)\right)^m,
\end{aligned}
\tag{5.62}
$$

where the second equality follows by the definition of H_j^*.

The Laplace transform derived in Equation (5.62) must coincide with the second term of Equation

(5.60). To show this, we focus on the expected value in the numerator of the second term in Equation (5.60), since the denominator will follow by setting $f_j = 0$. We use Lemma D.2.5 for $\Phi^!_{\boldsymbol{s},\boldsymbol{y}}$ with $f(\boldsymbol{w}, x) = \sum_{j=1}^d w_j(u_j + f_j(x))$ to get

$$\begin{aligned} &E\left[\exp\left\{-\sum_{j=1}^d \int_{\mathbb{G}} w_j\left(u_j + f_j(x)\right)\Phi^!_{\boldsymbol{s},\boldsymbol{y}}(\mathrm{d}\boldsymbol{w}, \mathrm{d}x)\right\}\right] \\ &= \sum_{m=0}^{\infty} q_M(m+k)\frac{(m+k)!}{m!}\frac{1}{E\left[M^{(K)}\right]}\left(\prod_{j=1}^d \int_0^{\infty}\int_{\mathbb{X}} e^{-w_j(u_j+f_j(x))}H_j(\mathrm{d}w_j)P_0(\mathrm{d}x)\right)^m. \end{aligned} \tag{5.63}$$

Therefore, the denominator equals to

$$E\left[\exp\left\{-\sum_{j=1}^d \int_{\mathbb{G}} w_j u_j \Phi^!_{\boldsymbol{s},\boldsymbol{y}}(\mathrm{d}\boldsymbol{w}, \mathrm{d}x)\right\}\right] = \sum_{m=0}^{\infty} q_M(m+k)\frac{(m+k)!}{m!}\frac{1}{E[M^{(K)}]}\left(\prod_{j=1}^d \psi_j(u_j)\right)^m. \tag{5.64}$$

To conclude, we multiply and divide the numerator by $\prod_{j=1}^d \left(\psi_j(u_j)\right)^m$ and then take the ratio between equations (5.63) and (5.64), simplifying common terms. The identity with Equation (5.62) follows after recognizing the definition of $q^*_M(m)$, which has been stated above.
(iii) The statement follows from Equation (5.59), conditioning on $\boldsymbol{\theta}$ and computing the expected value as done in Equation (5.64). □

D.3.8 Proof of Theorem 5.3.2

Proof. Consider a sequence of n observations divided into $d = 2$ groups of length n_1 and n_2, respectively. We want to compute the probability of having $K_{(n_1,n_2)} = K$ global clusters under a Vec-FDP(q_M, γ, Λ) prior.
Firstly, we remind the following formula (Charalambides, 2002, Theorem 8.16)

$$|C(n, K; -\gamma, -\rho)| = \frac{1}{K!}\sum \binom{n}{r_1, \ldots, r_K}\prod_{j=1}^K (\gamma)_{r_j} \tag{5.65}$$

where $(x)_n$ denotes the Pochhammer symbol, i.e., $(x)_n = \frac{\Gamma(x+n)}{\Gamma(x)}$. In general, this is defined for any $n \in \mathbb{R}^+$ but, for positive and integer values of n, it reduces to the rising factorial of x of order n. Moreover, in Equation (5.65) the sum is extended over all the vectors $(r_1, \ldots, r_K)$ of positive integers such that $\sum_{j=1}^K r_j = n$.
The probability under study $\mathbb{P}(K_{(n_1,n_2)} = K)$ can be written as follows in terms of the pEPPF function,

$$\mathbb{P}(K_{n_1,n_2} = K) = \sum_{(\star)} \frac{1}{K!}\prod_{j=1}^2 \binom{n_j}{n_{j,1}, \ldots, n_{j,K}} \cdot \Pi_K^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)$$

where the sum $(\star)$ is extended over all the vectors $(n_{1,1}, \ldots, n_{1,K})$ and $(n_{2,1}, \ldots, n_{2,K})$ of non-negative integers satisfying the following constraints

$$n_{j,k} \geq 0,\ k \in 1, \ldots, K,\ j \in \{1, 2\},\ \sum_{k=1}^K n_{j,k} = n_j\ j \in \{1, 2\} \text{ and } n_{1,k} + n_{2,k} \geq 1 \text{ as } k = 1, \ldots, K.$$

By exploiting the expression of the pEPPF we get

$$\begin{aligned}\mathbb{P}(K_{(n_1,n_2)}=K)=&\sum_{\bar{M}=0}^{\infty}\frac{(\bar{M}+K)!}{\bar{M}!}q_M(\bar{M}+K)\prod_{j=1}^{2}\frac{1}{(\gamma_j(\bar{M}+K))_{n_j}}\\&\times\sum_{(\star)}\frac{1}{K!}\prod_{j=1}^{2}\binom{n_j}{n_{j,1},\ldots,n_{j,K}}\cdot\prod_{j=1}^{2}\prod_{k=1}^{K}(\gamma_j)_{n_{j,k}}.\end{aligned}\tag{5.66}$$

We now focus on the evaluation of the sums over the partitions in (5.66). This may be calculated by counting how many zeros we have in the vectors $(n_{j,1},\ldots,n_{j,K})$. We denote by r_1 (resp. r_2) the number of possible zeros in the first vector (resp. second vector). Thanks to the validity of the condition $n_{1,k}+n_{2,k}\geq 1$, we know that if $n_{1,k}$ then $n_{2,k}\geq 1$ and vice versa. Moreover we observe that one has $\binom{K}{r_1}$ ways to choose the zeros in the first vector and $\binom{K-r_1}{r_2}$ possibilities to choose the zeros in the second one. Thus, we get

$$\begin{aligned}&\sum_{(\star)}\frac{1}{K!}\prod_{j=1}^{2}\binom{n_j}{n_{j,1},\ldots,n_{j,K}}\cdot\prod_{j=1}^{2}\prod_{k=1}^{K}(\gamma_j)_{n_{j,k}}\\&\quad=\sum_{r_1=0}^{K}\sum_{r_2=0}^{K-r_1}\binom{K}{r_1}\binom{K-r_1}{r_2}\sum_{(\star_1)}\sum_{(\star_2)}\frac{1}{K!}\prod_{j=1}^{2}\binom{n_j}{n_{j,1},\ldots,n_{j,K-r_j}}\cdot\prod_{j=1}^{2}\prod_{k=1}^{K-r_j}(\gamma_j)_{n_{j,k}}\end{aligned}$$

where

$$(\star_j)=\left\{(n_{j,1},\ldots,n_{j,K-r_j}):\ n_{j,k}\geq 1,k=1,\ldots,K-r_j,\ \sum_{k=1}^{K-r_j}n_{j,k}=n_j\right\}$$

as $j=1,2$. We now exploit (5.65) to obtain

$$\begin{aligned}&\sum_{(\star)}\frac{1}{K!}\prod_{j=1}^{2}\binom{n_j}{n_{j,1},\ldots,n_{j,K}}\cdot\prod_{j=1}^{2}\prod_{k=1}^{K}(\gamma_j)_{n_{j,k}}\\&\quad\times\sum_{r_1=0}^{K}\sum_{r_2=0}^{K-r_1}\prod_{j=1}^{2}|C(n_j,K-r_j;-\gamma_j)|\end{aligned}\tag{5.67}$$

If we now replace (5.67) in (5.66), the statement of the theorem follows. □

D.3.9 Proof of Theorem 5.3.4

Here, we derive the results shown in Theorem 5.3.4 as a special case of Theorem D.2.3.

Proof. The distributions of $S_{j,k}^{(a)}$ and $S_{j,k}^{(na)}$, for each $j=1,\ldots,d$ and $k=1,\ldots,K$ follow from Theorem D.2.3 setting $h_j(s)\mathrm{d}=1/\Gamma(\gamma_j)\ s^{\gamma_j-1}\exp^s$. While, to prove that $M^*\sim q_M^*$ where q_M^* is a mixture of

Poisson distributions, we plug Equation (5.35) into Theorem D.2.3 to get, for each integer $m \geq 0$,

$$
\begin{aligned}
q_M^*(m) \;&\propto\; \frac{(m+K)!}{m!}\frac{\Lambda^m}{(m+K-1)!}\left(\prod_{j=1}^{d}\psi_j\left(u_j\right)\right)^m \\
&\propto\; \frac{m+K}{m!}\left(\Lambda\prod_{j=1}^{d}\psi_j\left(u_j\right)\right)^m \;=\; \frac{m}{m!}\left(\Lambda\prod_{j=1}^{d}\psi_j\left(u_j\right)\right)^m + \frac{K}{m!}\left(\Lambda\prod_{j=1}^{d}\psi_j\left(u_j\right)\right)^m \\
&\propto\; \left(\Lambda\prod_{j=1}^{d}\psi_j\left(u_j\right)\right)\frac{1}{(m-1)!}\left(\Lambda\prod_{j=1}^{d}\psi_j\left(u_j\right)\right)^{m-1}\delta_{m\geq 1} + \frac{K}{m!}\left(\Lambda\prod_{j=1}^{d}\psi_j\left(u_j\right)\right)^m .
\end{aligned}
$$

By recognizing that the first term becomes zero when $m = 0$, the final equality holds. This completes the proof as we identify the kernel of a mixture of Poisson distributions, where the first is shifted by 1. □

D.4 Predictive distributions under the Vec-NIFPP

In this section, we provide and prove the formula for the predictive distribution for a new client entering a restaurant j. This has already been stated in Equation (5.16) for the special case of a Vec-FDP prior. Instead, in this section we refer to the more general Vec-NIFPP. More precisely, consider a realization $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ from $(P_1, \ldots, P_d) \sim \text{Vec-NIFPP}(q_M, H, P_0)$ with K distinct values $(\theta_1^{**}, \ldots, \theta_K^{**})$ and a partition $\rho = \{C_1, \ldots, C_K\}$ with counts $(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d)$ satisfying the constraints in Equation (5.6). Following the approach of James et al. (2009), Favaro and Teh (2013) and Argiento and De Iorio (2022) we work conditionally to $\boldsymbol{U}_n = \boldsymbol{u}$. Then, for each group, say j, we have

$$
\begin{aligned}
&\mathbb{P}(\theta_{j,n_j+1} \in \cdot \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \boldsymbol{u}) \\
&\quad\propto \sum_{k=1}^{K}\frac{\kappa_j(u_j, n_{j,k}+1)}{\kappa_j(u_j, n_{j,k})}\delta_{\theta_k^{**}}(\cdot) \;+\; \kappa_j(u_j, 1)\left(\prod_{h\neq j}\kappa_h(u_h, 0)\right)\frac{\Psi(K+1, \boldsymbol{u})}{\Psi(K, \boldsymbol{u})}P_0(\cdot)
\end{aligned}
\tag{5.68}
$$

Without loss of generality, consider a new client entering the first restaurant, where n_1 customers are currently seated. This (n_1+1)-th client can (i) choose an already existing table, or (ii) sit to a non-existing table. Note that, as detailed in Section 5.3.3, we distinguish between empty and non-existing tables. The latter refers to tables that do not appear in any restaurant of the franchise, while we admit the possibility of having tables with no clients in some restaurants as long as at least one customer in the whole franchise is seated at the corresponding table, see Figure 5.2.

As far as the proof of Equation (5.68) is concerned, considering case (i): we compute the probability of sitting at table k, $k = 1, \ldots, K$, where dish θ_k^{**} is served. This probability equals

$$
\mathbb{P}\left(\theta_{1,n_1+1} = \theta_k^{**} \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\right) \;=\; \Pi_K^{(n+1)}(\widetilde{\boldsymbol{n}}_1, \ldots, \boldsymbol{n}_d),
$$

where $\widetilde{\boldsymbol{n}}_1 = (n_{1,1}, \ldots, n_{1,k}+1, \ldots, n_{1,K})$. If we further condition to $\boldsymbol{U}_n = u$, we obtain

$$\begin{aligned}
&\mathbb{P}\big(\theta_{1,n_1+1} = \theta_k^{**} \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \boldsymbol{u}\big) \\
&\quad \propto \Psi(\boldsymbol{u}, K)\left(\prod_{j=2}^{K} \frac{(u_j)^{n_j-1}}{\Gamma(n_j)} \prod_{l=1}^{K} \kappa_j(u_j, n_{j,l})\right) \frac{(u_1)^{n_1}}{\Gamma(n_1+1)} \kappa_1(u_1, n_{1,k}+1) \prod_{l\neq k} \kappa_1(u_1, n_{1,l}) \\
&\quad = \Psi(\boldsymbol{u}, K)\left(\prod_{j=2}^{K} \frac{(u_j)^{n_j-1}}{\Gamma(n_j)} \prod_{l=1}^{K} \kappa_j(u_j, n_{j,l})\right) \frac{(u_1)^{n_1}}{\Gamma(n_1+1)} \frac{\kappa_1(u_1, n_{1,k}+1)}{\kappa_1(u_1, n_{1,k})} \prod_{l=1}^{K} \kappa_1(u_1, n_{1,l}),
\end{aligned}$$

where the second equality follows since we just multiplied and divided by $\kappa_1(u_1, n_{1,k})$.

Now, considering case (ii): we are interested in the probability of sitting at a new, non-existing, table where a new dish θ_{K+1}^{**} drawn from P_0 is served. This equals to

$$\mathbb{P}\big(\theta_{1,n_1+1} \in \cdot \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\big) \;=\; \Pi_{K+1}^{(n+1)}(\boldsymbol{n}_1^{\dagger}, \ldots, \boldsymbol{n}_d^{\dagger}),$$

where $\boldsymbol{n}_1^{\dagger} = (n_{1,1}, \ldots, n_{1,K}, 1)$ and $\boldsymbol{n}_j^{\dagger} = (n_{j,1}, \ldots, n_{j,K}, 0)$ for each $j = 2, \ldots, d$. As before, we further condition to $\boldsymbol{U}_n = u$ and we obtain

$$\begin{aligned}
\mathbb{P}\left(\theta_{1,n_1+1} \in \cdot \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \boldsymbol{u}\right) \propto\; & \Psi(\boldsymbol{u}, K+1)\left(\prod_{j=2}^{K} \frac{(u_j)^{n_j-1}}{\Gamma(n_j)} \kappa_j(u_j, 0) \prod_{l=1}^{K} \kappa_j(u_j, n_{j,l})\right) \\
& \times \frac{(u_1)^{n_1}}{\Gamma(n_1+1)} \kappa_1(u_1, 1) \prod_{l=1}^{K} \kappa_1(u_1, n_{1,l}) P_0(\cdot).
\end{aligned}$$

We can collect the previous expression in the following compact form

$$\begin{aligned}
&\mathbb{P}\left(\theta_{1,n_1+1} \in \cdot \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \boldsymbol{u}\right) \propto \\
&\Psi(\boldsymbol{u}, K) \sum_{k=1}^{K} \left\{\left(\prod_{j=2}^{K} \frac{(u_j)^{n_j-1}}{\Gamma(n_j)} \prod_{l=1}^{K} \kappa_j(u_j, n_{j,l})\right) \frac{(u_1)^{n_1}}{\Gamma(n_1+1)} \frac{\kappa_1(u_1, n_{1,k}+1)}{\kappa_1(u_1, n_{1,k})} \prod_{l=1}^{K} \kappa_1(u_1, n_{1,l}) \delta_{\theta^{**}}(\cdot)\right\} \\
&+ \Psi(\boldsymbol{u}, K+1)\left(\prod_{j=2}^{K} \frac{(u_j)^{n_j-1}}{\Gamma(n_j)} \kappa_j(u_j, 0) \prod_{l=1}^{K} \kappa_j(u_j, n_{j,l})\right) \frac{(u_1)^{n_1}}{\Gamma(n_1+1)} \kappa_1(u_1, 1) \prod_{l=1}^{K} \kappa_1(u_1, n_{1,l}) P_0(\cdot).
\end{aligned}$$

The latter is a mixture (with unnormalized weights) of Dirac point masses centered in the distinct values θ_k^{**} and the prior P_0. Finally, we cancel out the common term in the unnormalized weights and we get

$$\begin{aligned}
&\mathbb{P}\big(\theta_{1,n_1+1} \in \cdot \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d, \boldsymbol{u}\big) \\
&\quad \propto \Psi(\boldsymbol{u}, K) \sum_{k=1}^{K} \left\{\frac{\kappa_1(u_1, n_{1,k}+1)}{\kappa_1(u_1, n_{1,k})} \delta_{\theta^{**}}(\cdot)\right\} + \Psi(\boldsymbol{u}, K+1)\left(\prod_{j=2}^{K} \kappa_j(u_j, 0)\right) \kappa_1(u_1, 1) P_0(\cdot).
\end{aligned}$$

Dividing by $\Psi(\boldsymbol{u}, K)$, Equation (5.68) is obtained.

Figure 5.2 provides a graphical representation of the Chinese restaurant franchise process for Vec-NIFPP based on a sample of six, five and five observations in three groups, respectively. The first (left panel) and the second client (middle panel) enter the first restaurant and sit at different tables. Tables having the same color must be prepared in all other restaurants. Finally, (right panel) one of the possible

configurations when all clients are seated. Global clustering is obtained by merging tables having the same color.

D.5 Additional results on shot put data analysis

Here we detail the complete specification of the HMFM model for the shot put data:

$$
\begin{aligned}
\boldsymbol{y}_{j,1},\ldots,\boldsymbol{y}_{j,n_j} \mid \boldsymbol{S}_j,\theta,M,\boldsymbol{\beta}_j &\overset{\text{ind}}{\sim} \sum_{m=1}^{M} \frac{S_{j,m}}{T_j} d\mathrm{N}_{N_{j,i}}(\boldsymbol{y}_{j,i} \mid \mu_m \mathbf{1}_{N_{j,i}} + X_{j,i}\boldsymbol{\beta}_j, \sigma_m^2 \mathbf{I}_{N_{j,i}}), \quad \text{for } j = 1,\ldots,d \\
\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d &\overset{\text{iid}}{\sim} \mathrm{N}_d(\boldsymbol{\beta}_0, \Sigma_0) \\
\tau_1,\ldots,\tau_M \mid M &\overset{\text{iid}}{\sim} \text{Normal-InvGamma}(\mu_0, k_0, \nu_0, \sigma_0) \\
S_{j,1},\ldots,S_{j,M} \mid M,\gamma_j &\overset{\text{iid}}{\sim} \text{Gamma}(\gamma_j, 1), \quad \text{for } j = 1,\ldots,d \\
M \mid \Lambda &\sim \text{Pois}_1(\Lambda) \\
\gamma_1,\ldots,\gamma_d \mid \Lambda &\sim \text{Gamma}(a_\gamma, \Lambda b_\gamma) \\
\Lambda &\sim \text{Gamma}(a_\Lambda, b_\Lambda),
\end{aligned}
$$

where $d\mathrm{N}_{N_{j,i}}(\boldsymbol{y} \mid \boldsymbol{\mu}, \Sigma)$ is the density of a multivariate normal distribution evaluated in $\boldsymbol{y}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

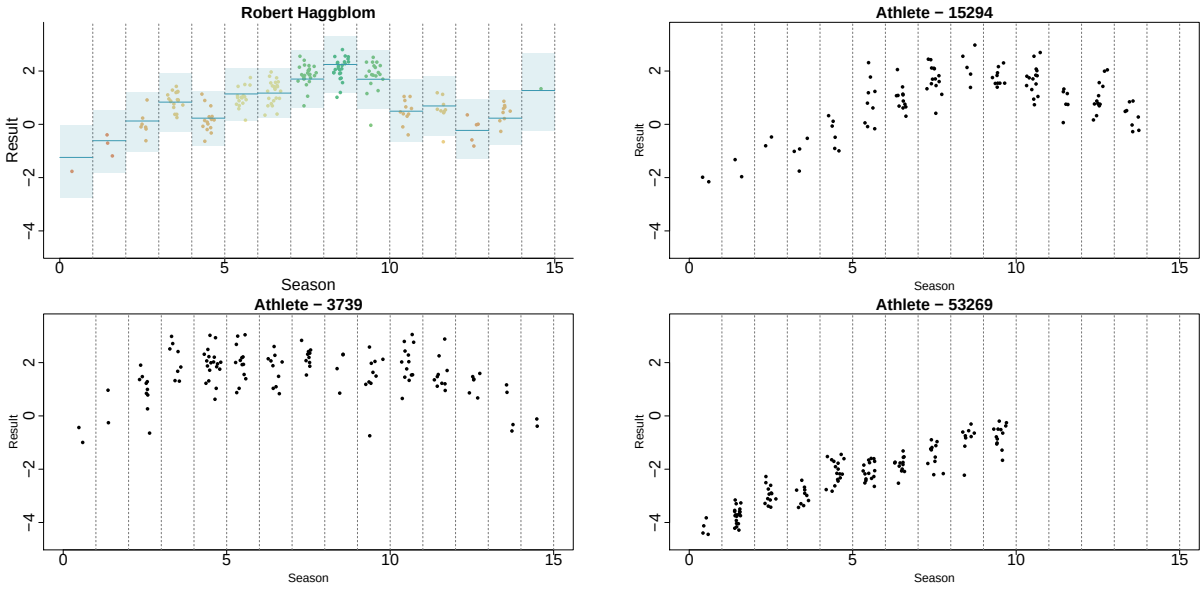

Figure 5.18: Shot put measurements collected throughout an athlete's career for four randomly selected athletes. Vertical dotted lines delimit seasons.

Table 5.3 presents the local clustering sizes, for each season and for each cluster. Finally, Table 5.4 shows some summary statistics for the global clusters and compares them with the overall dataset.

D.5.1 Comparison with pooled data modeling

A natural competitor for the proposed model is a naive modeling, which consists of fitting a mixture model to the pooled data. Discarding the group membership, the statistical units are the vectors $\boldsymbol{y}_{j,i}$

Table 5.3: Cluster cardinalities for each cluster (rows) and for each season (columns).

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
Cluster 1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Cluster 2	0	4	5	9	7	12	6	9	11	9	9	5	6	5	3
Cluster 3	14	11	14	13	18	16	21	9	11	12	3	9	3	6	4
Cluster 4	11	14	10	33	36	31	34	31	24	18	16	0	8	2	0
Cluster 5	0	13	20	43	58	51	41	37	49	23	11	20	0	0	9
Cluster 6	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
Cluster 7	0	0	12	11	6	9	17	0	0	0	4	0	0	0	0
Cluster 8	32	28	38	56	64	61	46	37	17	11	10	0	10	6	0
Cluster 9	9	40	69	114	74	49	33	52	29	20	20	30	0	0	4
Cluster 10	28	82	112	75	79	56	45	23	29	36	21	0	9	12	0
Cluster 11	55	139	78	43	29	44	34	30	22	5	0	12	14	0	5
Cluster 12	253	72	45	6	12	12	17	27	19	29	27	13	12	11	6
Cluster 13	0	0	0	0	0	1	1	0	0	0	0	3	0	0	0

(representing the shots of athlete i in season j). Namely,

$$\boldsymbol{y}_{j,i} \mid \boldsymbol{S}, \theta, M, \boldsymbol{\beta} \overset{\text{ind}}{\sim} \sum_{m=1}^{M} \frac{S_m}{T} d\mathrm{N}_{N_{j,i}}(\boldsymbol{y}_{j,i} \mid \mu_m \mathbf{1}_{N_{j,i}} + X_{j,i}\boldsymbol{\beta}, \sigma_m^2 \mathbf{I}_{N_{j,i}}),$$

for $j = 1, \ldots, d,\ i = 1, \ldots, n_j$ and where $T = \sum_{m=1}^{M} S_m$. After fitting the model using the same hyper-parameters setting and number of iterations reported in Section 5.6, we consider the group membership of the observations so that we can compare both the global, the season-specific, and the athlete-specific findings to the fit obtained using the proposed HMFM model. Similarly to Section 5.5, we refer to the pooled data analysis as MFM-pooled.

There are many similarities regarding global clustering with respect to the HMFM analysis. The MFM-pooled identifies 15 clusters that we ordered according to their decreasing means. Similarly to the HMFM, the first cluster comprises only exceptional performances by women. Just as clusters 6 and 13 in the HMFM results, the MFM-pooled also reveals two clusters (6 and 15) composed of noisy observations characterized by high variance and a very low mean for cluster 15. Another similarity concerns the evolution of the largest cluster in the early seasons. In the first season, clusters 13 and 14 have roughly the same number of athletes. Then, from season 2 to 5, the largest clusters are, respectively, 13, 12, 11, and 9. Finally, cluster 9 remains the largest one from seasons 6 to 8. As evidenced in the HMFM analysis, the MFM-pooled also indicates an almost linear improvement in the early years of an athlete's career, likely due to both acquired experience and ongoing physical development. Furthermore, the MFM-pooled reaches similar conclusions to HMFM about when athletes first reach their peak cluster as the distribution of the absolute frequencies is very similar to the one shown in the right panel of Figure 5.16.

If the global clustering has many points in common between the two models, the differences are more evident in the local clustering. Indeed, comparing the left panel of Figure 5.16 for the corresponding visualization of the local cluster sizes we get with the pooled analysis reported in the left panel of Figure 5.19 we notice that the latter shows fewer empty clusters and many more clusters whose cardinalities range between 1 and 10. Furthermore, the right panel of Figure 5.19 compares the number of season-specific clusters obtained with the two models, HMFM (red bars) and MFM-pooled (light blue bars), with

Table 5.4: Cluster statistics for each global cluster. The first row represents the case where all data are pooled in a single cluster. Each subsequent row corresponds to a specific global cluster. The columns display the sample mean of various cluster statistics, differentiated by M (male) or F (female) athletes in the cluster.

Cl	Mean	SizeM	SizeF	MeanM	MeanF	VarM	VarF	AgeM	AgeF
Pooled	0	1925	1689	1.13	-1.29	1.51	2.16	24.5	22.8
1	2.69	0	3	0	2.69	-	0.36	-	28.0
2	2.49	42	58	3.47	1.59	0.28	0.51	27.7	26.0
3	1.87	78	86	2.86	0.67	0.28	0.35	27.5	24.7
4	1.38	147	121	2.26	0.20	0.28	0.47	25.8	25.1
5	1.01	245	130	1.69	-0.38	0.29	0.43	25.7	23.6
6	0.81	2	1	1.08	0.00	13.05	1.83	27.0	28.0
7	0.39	36	23	1.20	-0.95	0.86	0.97	24.2	21.8
8	0.20	238	178	1.17	-1.02	0.20	0.34	24.4	22.2
9	-0.26	313	230	0.74	-1.54	0.29	0.39	23.8	22.0
10	-0.87	327	280	0.26	-2.02	0.26	0.37	22.8	22.0
11	-1.68	241	269	-0.31	-2.68	0.26	0.34	22.2	21.6
12	-2.38	254	307	-1.04	-3.24	0.34	0.43	22.0	21.6
13	-4.31	2	3	-4.76	-4.31	18.10	2.60	36.3	25.4

the latter overestimating the number of clusters with respect to our proposed model. We claim this is a clear advantage of the hierarchical modeling, achieving better shrinkage. The practical and interpretative differences are evident. For example, in the first season, a remarkable finding of our model was its ability to discriminate between the large group of rookies in low-ranked clusters and the ones debuting directly in high-ranked clusters, leaving the intermediate clusters empty. In the pooled analysis, this sharp separation is not observable. Similar considerations can be made for the final three seasons, where the HMFM can effectively discriminate between high- and low-level athletes, capturing differences that the MFM-pooled cannot.

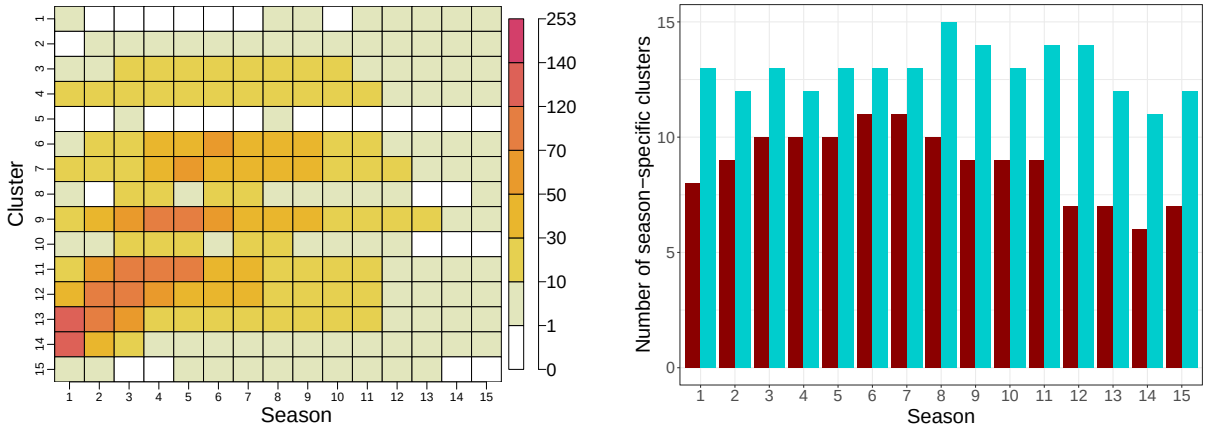

Figure 5.19: Local cluster sizes (left panel) and barplot of season-specific number of clusters obtained under HMFM (red bars) and MFM-pooled (light blue bars).

Table 5.5: Cluster cardinalities for each cluster (rows) and for each season (columns). Findings from pooled data analysis.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**
Cluster 1	1	0	0	0	0	0	0	1	1	0	3	1	3	2	2
Cluster 2	0	2	1	1	5	8	4	6	8	7	4	6	2	4	2
Cluster 3	4	2	10	12	11	14	19	16	12	13	8	6	6	4	2
Cluster 4	13	12	11	16	21	15	28	19	16	14	12	8	4	4	3
Cluster 5	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0
Cluster 6	7	17	14	34	45	54	40	41	36	21	14	8	5	3	2
Cluster 7	12	14	26	36	54	49	43	33	32	15	13	11	6	3	4
Cluster 8	2	0	10	13	9	12	13	6	4	3	4	2	0	0	3
Cluster 9	28	43	64	79	85	68	46	46	30	24	18	18	11	6	2
Cluster 10	3	7	21	26	13	7	13	10	4	6	1	3	0	0	0
Cluster 11	26	61	74	93	74	44	39	26	26	21	14	6	7	6	4
Cluster 12	45	86	89	60	42	49	32	26	22	14	10	7	7	3	3
Cluster 13	127	111	68	29	19	20	15	20	13	16	14	7	7	2	4
Cluster 14	126	45	14	4	4	1	2	5	6	8	5	8	3	5	2
Cluster 15	9	3	0	0	1	1	1	1	1	1	1	1	1	0	0

Table 5.6: Cluster statistics for each global cluster. The columns display the sample mean of various cluster statistics, differentiated by M (male) or F (female) athletes in the cluster. Findings from pooled data analysis.

Cl	**Mean**	**SizeM**	**SizeF**	**MeanM**	**MeanF**	**VarM**	**VarF**	**AgeM**	**AgeF**
Pooled	0	1925	1689	1.13	-1.29	1.51	2.16	24.5	22.8
1	2.55	0	14	0.00	2.55	-	0.32	-	27.1
2	2.50	22	38	3.61	1.63	0.24	0.28	27.8	26.2
3	2.05	68	71	3.07	0.90	0.30	0.24	27.3	26.1
4	1.46	95	101	2.49	0.35	0.23	0.25	26.5	25.5
5	1.31	2	1	1.59	0.00	5.85	1.83	22.8	27.9
6	1.24	215	126	1.95	-0.22	0.31	0.30	25.9	24.4
7	0.59	210	141	1.41	-0.75	0.22	0.22	24.8	23.5
8	0.54	50	31	1.34	-0.83	0.66	0.69	24.8	22.3
9	0.06	355	213	0.89	-1.30	0.23	0.23	24.2	22.2
10	-0.44	67	47	0.55	-1.67	0.63	0.60	22.9	22.2
11	-0.71	288	233	0.33	-1.84	0.22	0.23	22.9	22.0
12	-1.45	244	251	-0.18	-2.36	0.27	0.24	22.5	21.8
13	-2.19	204	268	-0.75	-2.95	0.21	0.21	21.9	20.4
14	-3.04	96	142	-1.57	-3.67	0.19	0.19	22.2	18.7
15	-3.52	9	12	-2.17	-4.39	6.16	2.21	27.7	20.2

Chapter 6

How Many Unseen Species are in Multiple Areas?

This chapter, based on Colombi et al. (2025a), is a joint work with Raffaele Argiento, Federico Camerlenghi and Lucia Paci. We develop a Bayesian nonparametric framework for studying unseen distinct and shared species across two sampling areas. Building on the Vector of Finite Dirichlet Process introduced earlier, we provide closed-form results for both in-sample analysis and out-of-sample prediction. Our novel methodology is demonstrated on both synthetic data, highlighting similarities and differences with frequentist approaches, and a real-world dataset. Specifically, we analyze an ant population in Trieste, sampled from two parks: one just outside the city and the other in its center. We start by reviewing some basic concepts and estimators developed in the frequentist literature that will be needed throughout this chapter. We refer to Section 4.2.1 for a review of the Bayesian nonparametric approach in the case of a single area.

6.1 Frequentist estimators

The case in which the investigation focuses on a single area is a long-established and extensively discussed problem in statistics, dating back to Fisher et al. (1943). Here, the $n \geq 1$ observed individuals represent a random sample $(X_1, \ldots, X_n)$ drawn from an unknown discrete distribution P. In this setting, the celebrated Good-Turing estimator (Good, 1953) provides an estimate of the probability that the $n+1$-th observation coincides with a species whose frequency in the original sample is exactly equal to f, with $f \geq 0$. In particular, for $f = 0$, this estimates the probability of observing a new, previously undetected species as the relative frequency of the species observed exactly once in the sample, commonly known as singletons. The Good-Toulmin estimator (Good and Toulmin, 1956) represents a m-steps ahead generalization for the probability of discovering a new species. These estimators are nonparametric since they do not rely on any assumption for the unknown generating distribution P, making them flexible and easy to use both for deriving stopping rules in sampling strategies Rasmussen and Starr (1979) or to estimate species richness (Chao, 1984). See Orlitsky et al. (2016) for an exhaustive review of generalizations and improvements of the Good-type estimators.

The study of estimators for multiple areas is much less extensive than for a single population. Even with just $d = 2$ different areas, the mathematical complexity increases significantly because we need to

estimate the number of shared species, i.e., those occurring in both areas. Indeed, this statistic is crucial to evaluate the similarity or dissimilarity between the two areas. To get an idea of the increased difficulty, we point out that the total number of distinct species can be expressed as the sum of two quantities: the number of observed distinct species (a known quantity) and the number of unobserved species (an unknown quantity). Rather, for shared species, there are three distinct unknowns: species that are shared but undetected in both areas and species that are shared but observed in only one of the two areas. The exponential increase in the number of unknowns (which is $2^d - 1$) explains why the case of two areas is the most studied in the literature. For scenarios with $d > 2$ areas, comparisons are made between all possible pairs. In fact, commonly used multi-area indices are designed for the two-area case, since summarizing comparisons across three or more areas into a single index is challenging, see Pan et al. (2009).

In the frequentist framework, the model assumes two discrete and unknown probability distributions, P_1 and P_2, with possible different lengths, but ordered in such a way that the first S species are the shared ones. Then, two random samples $\boldsymbol{X}_j = \left(X_{j,1}, \ldots, X_{j,n_j}\right)$, $j = 1, 2$, of sizes n_1 and n_2 are taken from P_1 and P_2, respectively. In this setting, Yue and Clayton (2012), Chuang et al. (2015) and Chao et al. (2017) proposed Turing-type estimators for the probability of discovering a new shared species in the next pair of observations. However, the three studies addressed only the case of the one-step ahead prediction, despite Yue and Clayton (2012) highlighting the need for sampling strategies where $m > 1$ pairs of observations are considered. Consequently, a generalization of the Good-Toulmin estimator for the shared species problem remains open. On the other hand, various methods have been proposed to estimate shared species richness. The seminal work by Chao et al. (2000) introduced a method based on sample coverage, which has been improved in two ways. The first improvement relies on Laplace's approximations (Chao et al., 2006), while the second version presents a lower bound of the initial method (Pan et al., 2009), which proves to be more useful in practice. Finally, Chuang et al. (2015) present a jackknife estimator for the richness of shared species.

6.2 Vector of Finite Dirichlet Processes

Consider a sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ from $d = 2$ partially exchangeable sequences. Specifically, let $\boldsymbol{X}_j = \left(X_{j,1}, \ldots, X_{j,n_j}\right)$, for $j = 1, 2$, and assume that each $X_{j,i}$ takes values in a Polish space $\mathbb{X}$. As anticipated in Section 4.3.2, this is equivalent to assuming that the distribution of $\boldsymbol{X}$ is invariant under permutations occurring within elements of vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, but it is not invariant under permutations among them. Hereafter, we refer to observations coming from two areas or groups, keeping in mind that the definition of a 'group' is problem-specific and can be interpreted in a broader sense.

As we only refer to the case of $d = 2$ groups, the de Finetti's representation theorem introduced in Equation (4.3.2) reduces to assuming the following Bayesian nonparametric model:

$$\begin{aligned} (X_{1,i_1}, X_{2,i_2}) \mid (P_1, P_2) &\overset{\text{ind}}{\sim} P_1 \otimes P_2, \quad (i_1, i_2) \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\} \\ (P_1, P_2) &\sim \mathcal{Q}_2 \,. \end{aligned} \tag{6.1}$$

In this chapter, we study the previous model under a Vec-FDP prior, as introduced in Chapter 5. We recall that the random probability measures P_1 and P_2 are defined as in Equation (5.2). In what follows, we let $w_{m,j} = S_{j,m}/T_j$ be the normalized weights and we let $\boldsymbol{w}_j = \left(w_{j,1}, \ldots, w_{j,M}\right)$ be the vector

of species' proportions specific to population j, $j = 1, 2$. From Section 5.2, we have that defining a prior for (P_1, P_2) is equivalent to place a joint prior over $(M, \tau_1, \ldots, \tau_M, \boldsymbol{w}_1, \boldsymbol{w}_2)$. Here, conditionally to M, $(\tau_1, \ldots, \tau_M)$ are common random atoms across the two random probability measures, which are assumed independent and identically distributed with common distribution P_0, that is a diffuse probability measure on $\mathbb{X}$. Moreover, $\boldsymbol{w}_1$ is independent of $\boldsymbol{w}_2$ and such that $\boldsymbol{w}_j \mid M \sim \text{Dir}_M(\gamma_j, \ldots, \gamma_j)$, for $j = 1, 2$, where $\text{Dir}_M(\gamma_j, \ldots, \gamma_j)$ denotes the M-dimensional symmetric Dirichlet distribution with group-specific parameter γ_j.

Finally, M is supposed to be a positive integer-valued random variable whose probability mass function is denoted with q_M. Following Chapter 5, we choose a 1-shifted Poisson distribution. Nevertheless, we point out that all our results hold for any distribution on positive integers. We summarize this prior construction by writing

$$(P_1, P_2) \sim \text{Vec-FDP}(\Lambda, \boldsymbol{\gamma}, P_0), \tag{6.2}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. Placing a Vec-FDP prior on (P_1, P_2) we assume that the two groups share the same finite, yet random, number of species M that appear with different frequencies $w_{j,m}$ in the two groups, i.e., in the two areas. Since a Dirichlet prior is chosen for these probabilities, the same M species would eventually be observed in the two groups if we were able to get an infinite sample from P_1 and P_2. However, in practice, we always work with finite samples of sizes $n_1 \geq 1$ and $n_2 \geq 1$.

Let $n = n_1 + n_2$ be the total number of observations. Within each group j, a random number of K_{j,n_j} distinct species would be observed, which we label as $\boldsymbol{X}_j^* = \left\{X_{j,1}^*, \ldots, X_{j,K_{j,n_j}}^*\right\}$. Let $K_{j,n_j} = r_j$ be a realization of this random variable. Since P_1 and P_2 share the same support, ties between the two groups are expected, i.e., $\mathbb{P}\left(X_{1,k}^* = X_{2,k'}^*\right) > 0$. Thus, the set of labels for the distinct species in the whole sample is obtained as $\boldsymbol{X}^{**} = \boldsymbol{X}_1^* \cup \boldsymbol{X}_2^* = \left\{X_1^{**}, \ldots, X_{K_{n_1,n_2}}^{**}\right\}$, where K_{n_1,n_2} denotes the overall number of distinct species which is, in general, smaller than the sum of K_{1,n_1} and K_{2,n_2}. Such difference is due to the random number of species shared between the two groups, namely

$$S_{n_1,n_2} = K_{1,n_1} + K_{2,n_2} - K_{n_1,n_2}. \tag{6.3}$$

In the following, we let $K_{n_1,n_2} = r$ and $S_{n_1,n_2} = t$ denote the realizations of the distinct and shared number of species, respectively. Moreover, we refer to group-specific quantities as *local* quantities while we call *global* quantities those related to the joint sequence $\boldsymbol{X}$. Therefore, K_{j,n_j} is also named the local number of distinct species in group j while K_{n_1,n_2} is the global number of distinct species. We refer to Section 6.3 for a more detailed description of these quantities.

6.2.1 pEPPF and predictive distribution

The sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ from model (6.1) with $\mathcal{Q}_2$ chosen as in Equation (6.2) is uniquely determined by the species labels $\boldsymbol{X}^{**}$ and their abundances. The latter are defined by vectors $\boldsymbol{n}_j = (n_{j,1}, \ldots, n_{j,r})$, such that $n_{j,l}$ represents the abundance of the l-th species in the j-th group, for $j = 1, 2$ and $l = 1, \ldots, r$. These counts must satisfy the following constraints,

$$n_{j,l} \geq 0, \quad n_{1,l} + n_{2,l} > 0, \quad \sum_{l=1}^{r} n_{j,l} = n_j \quad l = 1, \ldots, r; \; j = 1, 2. \tag{6.4}$$

Consistent with the description in Section 6.2, some of the counts $n_{j,l}$ may also be zero.

Pitman (1996) shows that the marginal likelihood $\mathscr{L}(\boldsymbol{X})$, which is obtained by integrating (P_1, P_2) out of model (6.1), admits a factorization in terms of the species abundances and the corresponding species labels. Namely, $\mathscr{L}(\boldsymbol{X}) = \mathscr{L}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{X}^{**}) = \mathscr{L}(\boldsymbol{n}_1, \boldsymbol{n}_2) \prod_{l=1}^{r} P_0(\mathrm{d}X_l^{**})$. In what follows, we drop the value of the labels $\boldsymbol{X}^{**}$ since it is not relevant for our purposes. Consequently, the main object of interest is the law of abundances $\mathscr{L}(\boldsymbol{n}_1, \boldsymbol{n}_2)$. Camerlenghi et al. (2019b) shows that the law of the random partition induced by a partially exchangeable sequence $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ may be described through the pEPPF, introduced in Section 5.2.1. Under a Vec-FDP prior, the pEPPF equals

$$\Pi_r^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = V_{n_1,n_2}^r \prod_{j=1}^{d} \prod_{l=1}^{r} (\gamma_j)_{n_{j,l}}, \tag{6.5}$$

where $d = 2$, $(\boldsymbol{n}_1, \boldsymbol{n}_2)$ satisfy the constraints given in Equation (6.4) and

$$V_{n_1,n_2}^r = \sum_{m=1}^{\infty} (m)_{r\downarrow}\, q_M(m) \prod_{j=1}^{d} \frac{1}{(\gamma_j m)_{n_j}}. \tag{6.6}$$

See Section E.2.1 and Section 5.3 for an alternative expression. In this work, we let $(m)_{r\downarrow} = m(m-1)\dots(m-r+1)$ denote the falling factorial of order r and $(x)_n = \Gamma(x+n)/\Gamma(x)$ is the Pochhammer symbol, also known as the rising factorial when n is a natural number. The sum in Equation (6.6) can be started from $m = r$ since $(m)_{r\downarrow} = 0$ for all $m < r$. See Section 6.2.2 for further analysis and properties of the V coefficients such as convergence, asymptotics, and a recurrence relationship. The pEPPF in Equation (6.5) represents the sampling model we assume generates the data. However, since its form is rather involved and hard to interpret, it is common to describe the data-generating mechanism by inspecting the predictive distributions, which are a generalization of the well-known Chinese restaurant franchise process introduced in Teh et al. (2006). Confining our attention to the first group, in Section 5.3.3 we showed that, taken $(\boldsymbol{n}_1, \boldsymbol{n}_2)$ observations as in Equation (6.4), the $n_1 + 1$-th observation can either be equal to one of the 'old' (already observed) species, with probability proportional to $V_{n_1+1,n_2}^r (n_{1,l} + \gamma_1)$, or to a 'new' (never observed before) species, whose label is drawn from P_0, with probability proportional to $V_{n_1+1,n_2}^{r+1} \gamma_1$. Namely,

$$\mathbb{P}\left(X_{1,n_1+1} \in A \mid \boldsymbol{X}\right) = \frac{V_{n_1+1,n_2}^r}{V_{n_1,n_2}^r} \sum_{l=1}^{r} (n_{1,l} + \gamma_1)\, \delta_{X_l^{**}}(A) + \frac{V_{n_1+1,n_2}^{r+1}}{V_{n_1,n_2}^r} \gamma_1 P_0(A),$$

The case of a new observation in group 2 trivially follows. Instead, we report here the general case when a new pair of clients arrives, one for each group, i.e., the $n_1 + 1$-th and $n_2 + 1$-th clients in the first and

Table 6.1: Unnormalized probabilities of observing an old species and a new species in each group when a new pair of observations is considered.

		Group 1	
		Old	New
Group 2	Old	$V^r_{n_1+1,n_2+1} q_1^{\text{old}} q_2^{\text{old}}$	$V^{r+1}_{n_1+1,n_2+1} q_1^{\text{new}} q_2^{\text{old}}$
	New	$V^{r+1}_{n_1+1,n_2+1} q_1^{\text{old}} q_2^{\text{new}}$	$\left(V^{r+1}_{n_1+1,n_2+1} + V^{r+2}_{n_1+1,n_2+1}\right) q_1^{\text{new}} q_2^{\text{new}}$

second restaurants, respectively. The predictive distribution is

$$\begin{aligned}
&\mathbb{P}\left(X_{1,n_1+1} \in A\,, X_{2,n_2+1} \in B \mid \boldsymbol{X}\right) \\
&= \frac{V^r_{n_1+1,n_2+1}}{V^r_{n_1,n_2}} \sum_{l_1=1}^{r}\sum_{l_2=1}^{r} \left(n_{1,l_1}+\gamma_1\right)\left(n_{1,l_2}+\gamma_2\right)\delta_{X^{**}_{l_1}}(A)\delta_{X^{**}_{l_2}}(B) \\
&\quad + \frac{V^{r+1}_{n_1+1,n_2+1}}{V^r_{n_1,n_2}} \left\{ \sum_{l_1=1}^{r} \left(n_{1,l_1}+\gamma_1\right)\delta_{X^{**}_{l_1}}(A)\gamma_2 P_0(B) + \gamma_1 P_0(A) \sum_{l_2=1}^{r}\left(n_{1,l_2}+\gamma_2\right)\delta_{X^{**}_{l_2}}(B) \right\} \\
&\quad + \frac{V^{r+1}_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\gamma_1\gamma_2 P_0(A\cap B) + \frac{V^{r+2}_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\gamma_1\gamma_2 P_0(A)P_0(B),
\end{aligned} \tag{6.7}$$

for any measurable sets A and B. The one-step ahead predictive distribution in Equation (6.7) follows from Equation (6.27) after noticing that $|C(1,0;-\gamma_j,-(\gamma_j r_j+n_j))| = \gamma_j r_j + n_j$ and $|C(1,1;-\gamma_j,-(\gamma_j r_j + n_j))| = \gamma_j$. Finally, Equation (6.7) can also be summarized in a concise version that highlights the distinction between old and new species, thereby neglecting which of the r species is selected in the case of an old species. The unnormalized probabilities are summarized in Table 6.1, where we have defined, for each group j, $q_j^{\text{old}} = \sum_{l=1}^{r}(n_{j,l}+\gamma_j)$ the weight of generating an observed species, and $q_j^{\text{new}} = \gamma_j$ as the weight associated with a new species. The normalizing constant is $V^r_{n_1+1,n_2+1}$. Table 6.1 includes all four possible cases, which involve either generating or not generating a new observation in each of the two groups. The apex of each coefficient V indicates the total number of distinct species in the enlarged sample of size (n_1+1, n_2+1). Specifically, we highlight that r can be increased by a single unit even in the scenario where a new species is observed in both groups. This is because the new species could be the same in both groups, i.e., a previously unobserved shared species.

6.2.2 Analysis of the $V^r_{n_1,n_2}$ coefficients

We discuss here some properties of the coefficients $V^r_{n_1,n_2}$, defined in Equation (6.6). In the exchangeable case, i.e., when all observations are drawn from the same area, Gnedin and Pitman (2006) shows that the EPPF admits the product form

$$\Pi_r^{(n)}(n_1,\ldots,n_r) \;=\; V_n^r \prod_{l=1}^{r}(1-\sigma)_{n_l-1}\,, \tag{6.8}$$

for any $\sigma < 1$, $n \geq 1$, $r \leq n$ and positive integers $n_1,\ldots,n_r$ that sum up to n if and only if the set of non-negative weights $\{V_r^n \;:\; n \geq 1\,, 1 \leq r \leq n\}$ satisfies the recurrence relationship $V_n^r = V_{n+1}^{r+1} + (n-\sigma r)V_{n+1}^r$.

The sampling models in Equation (6.8) are known as Gibbs-type models, see De Blasi et al. (2015) for further discussion. Within this class, it is important to distinguish between two cases, namely when $\sigma \in [0, 1)$ and when $\sigma < 0$. The former corresponds to models with a potentially infinite number of species, whereas the latter assumes a finite -though random- number of species. This assumption aligns with the hypothesis of our model, and we therefore confine our attention to this case. Finally, Miller and Harrison (2018) reparametrize the model in Equation (6.8) so that the recurrence relationship above takes the form

$$V_n^r = \gamma V_{n+1}^{r+1} + (\gamma r + n) V_{n+1}^r \,, \tag{6.9}$$

and the coefficients V_n^r admits the infinite sum representation,

$$V_n^r \;=\; \sum_{m=1}^{\infty} (m)_{r\downarrow}\, q_M(m)\, \frac{1}{(\gamma m)_n}. \tag{6.10}$$

The latter expression is also recovered when assuming Dirichlet distributed weights in the Normalized Independent Finite Point Process model by Argiento and De Iorio (2022).

Since Equation (6.10) is recovered from Equation (6.6) by setting either n_1 or n_2 equal to zero, we say that the coefficients V_{n_1,n_2}^r are a multi-group extension of the V_n^r coefficients in the case of finitely many species. Indeed, we show that they share similar properties to V_n^r. In particular, V_{n_1,n_2}^r multiplies the general term of the series in (6.10) by a factor that rapidly decreases to zero, both with respect to the series index and the number of observations. As a consequence, V_{n_1,n_2}^r converges even faster than V_n^r. Hence, the next proposition states that the V_{n_1,n_2}^r coefficients are well defined and, for sufficiently large sample sizes, can be accurately approximated by the r-th term of the series.

Proposition 1. *The V_{n_1,n_2}^r coefficients introduced in Equation (6.6) are well defined for every choice of probability mass function q_M. Moreover, for any integer $r \geq 1$ such that $q_M(r) > 0$, the following approximation holds:*

$$\begin{aligned} V_{n_1,n_2}^r \;=\; & \frac{r!\, q_M(r)}{(\gamma_1 r)_{n_1} (\gamma_2 r)_{n_2}} \\ & \times \left\{ 1 + n_1^{-\gamma_1} n_2^{-\gamma_2} (r+1)(\gamma_1 r)_{\gamma_1} (\gamma_2 r)_{\gamma_2} \frac{q_M(r+1)}{q_M(r)} + o(n_1^{-\gamma_1} n_2^{-\gamma_2}) \right\} . \end{aligned} \tag{6.11}$$

The latter result extends Gnedin and Pitman (2006); Miller and Harrison (2018) to the multi-group setting. The proof is given in Section E.3.1. Finally, we conclude deriving the recurrence relationship satisfied by V_{n_1,n_2}^r.

Proposition 2. *Let V_{n_1,n_2}^r be the coefficients defined in Equation (6.6). Then, the following 1-step recurrence relationships holds*

$$V_{n_1,n_2}^r \;=\; \gamma_2 V_{n_1,n_2+1}^{r+1} + (\gamma_2 r + n_2) V_{n_1,n_2+1}^r \,. \tag{6.12}$$

or equivalently,

$$V_{n_1,n_2}^r \;=\; \gamma_1 V_{n_1+1,n_2}^{r+1} + (\gamma_1 r + n_1) V_{n_1+1,n_2}^r \,. \tag{6.13}$$

Moreover, the following 2*-steps recurrence relationships holds,*

$$\begin{aligned} V^r_{n_1,n_2} &= \gamma_1\gamma_2\left\{r^2 V^r_{n_1+1,n_2+1} + (2r+1)V^{r+1}_{n_1+1,n_2+1} + V^{r+2}_{n_1+1,n_2+1}\right\} \\ &\quad + n_1 V^r_{n_1+1,n_2} + n_2 V^r_{n_1,n_2+1} - n_1 n_2 V^r_{n_1+1,n_2+1}\,. \end{aligned} \tag{6.14}$$

The proof is given in Section E.3.2.

6.3 In-sample analysis

6.3.1 Correlation

A natural question that arises when moving from analyzing a single group to multiple groups is quantifying the interaction between these two. Here, we aim to provide a quantitative answer by examining the statistical dependence between the data generating models for the two groups, namely between P_1 and P_2. Expanding the result of Chapter 5, we obtain a closed-form expression for the correlation between P_1 and P_2 when evaluated on the same measurable set A, namely

$$\operatorname{cor}\left(P_1(A), P_2(A)\right) \;=\; \frac{E\left(1/M\right)}{\sqrt{(1+\gamma_1)\,(1+\gamma_2)}\sqrt{E\left(\frac{1}{1+\gamma_1 M}\right)E\left(\frac{1}{1+\gamma_2 M}\right)}}\,. \tag{6.15}$$

The expression (6.15) does not depend on the choice of the set A. Thus, it may be considered an overall measure of dependence between the two random probability measures. Furthermore, if $M \sim \text{Pois}_1\left(\Lambda\right)$, then the numerator in Equation (6.15) is available in closed-form and it equals

$$E\left(1/M\right) = \Lambda^{-1}\left(1 - \mathrm{e}^{-\Lambda}\right)\,. \tag{6.16}$$

Additionally, the following limiting values hold:

$$\lim_{\gamma_1,\gamma_2\to 0} \operatorname{cor}\left(P_1(A), P_2(A)\right) \;=\; E\left(1/M\right), \qquad \lim_{\gamma_1,\gamma_2\to +\infty} \operatorname{cor}\left(P_1(A), P_2(A)\right) \;=\; 1\,.$$

The above limits suggest an interpretation of the γ_j's as homogeneity parameters: large values of γ_j's indicate similar groups that share most of the distinct species. Conversely, small values of γ_j's result in the minimum value of Equation (6.15). Interestingly, this value is not zero but depends on the choice of the prior distribution of M. Intuitively, the larger expected values of M drive the correlation between the two populations toward zero.

6.3.2 In-sample statistics

In this section, we investigate the distributions of the most relevant in-sample statistics for a sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ of sizes n_1 and n_2 from model (6.1) under the Vec-FDP prior given in Equation (6.2). Mathematically, this is equivalent to studying the properties of the Bayesian nonparametric prior.

We start presenting the main quantities of interest by means of an example, reported in Figure 6.1, which shows the observed samples in two different areas. Moreover, Table 6.2 summarizes the notation and the meaning of each random variable.

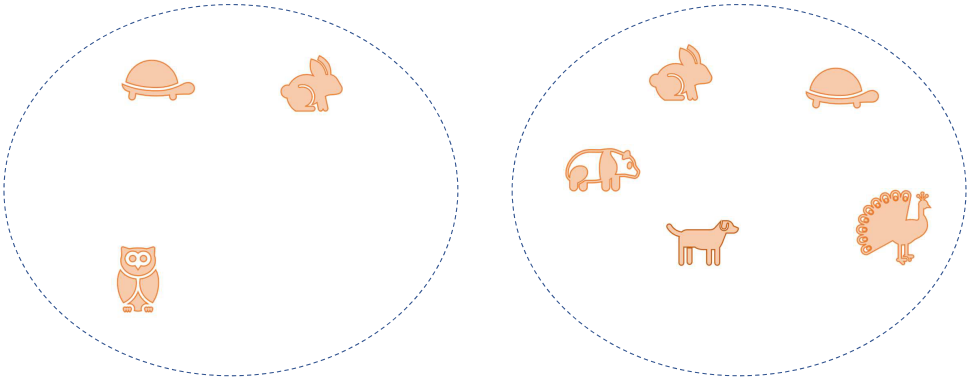

Figure 6.1: Observed species in the observed sample. Each dotted circle delimits an area.

The first area is composed of $K_{1,n_1} = 3$ distinct species. Following the notation of Section 6.2, this is a local quantity because it does not require any knowledge of the second area to be computed. Similarly, the local number of distinct species in the second area, K_{2,n_2}, equals five. In general, we denote the observed values of K_{j,n_j} as r_j, for $j = 1, 2$. Then, moving to global quantities, i.e., those that require both areas to be computed, let the global number of distinct species be K_{n_1,n_2}. This can be computed pooling all species and discarding those repeated twice. In our running example, $K_{n_1,n_2} = 6$ and, in general, this is denoted as r. We highlight that $r \geq r_j$, $j = 1, 2$, because each local distinct species is also a global distinct species but we have $r \leq r_1 + r_2$, because some species may appear in both areas. These are called shared species, the associated random variable is denoted as S_{n_1,n_2}, and its realization is t. In Figure 6.1, $t = 2$ (the rabbit and the turtle).

We also introduce two additional quantities, $K^*_{1,n}$ and $K^*_{2,n}$, which are useful in our proofs and for the understanding of our framework. Consider $j = 1$, $K^*_{1,n}$ represents the number of species observed in the first group but not in the second one. For example, the owl is only seen in area 1, hence $K^*_{1,n} = 1$. On the other hand, $K^*_{2,n} = 3$ (the panda, the turkey, and the dog). In general, we let $K^*_{j,n} = r^*_j$, $j = 1, 2$. Specifically, even though r^*_j refers to the number of species of a single area, this is also a global quantity as it is not well defined if a second area is not observed.

Summing up, we introduced six random variables which are dependent. In the following, we report their main relationships

$$t = r_1 + r_2 - r, \quad r = t + r^*_1 + r^*_2 \quad r_j = t + r^*_j, \quad j = 1, 2. \tag{6.17}$$

Among these four linear equations, only three of them are linearly independent. As a consequence, the number of linearly independent random variables is three. We only need three out of six quantities to properly characterize the observed sample, the remaining three are deduced using the system of Equations (6.17).

Moving toward the distribution of such random quantities, we recall that previous works derived the marginal distribution of both K_{j,n_j} (Lijoi et al., 2007a; Argiento and De Iorio, 2022) (see Equation (6.21)) and K_{n_1,n_2} (see Equation (5.11) in Section 5.3). Conversely, the distribution of S_{n_1,n_2} has not yet been derived, although it is linearly related to the number of local and global distinct species by Equation (6.3). This is because also the joint distribution of K_{1,n_1}, K_{2,n_2} and K_{n_1,n_2} is required to derive the distribution of S_{n_1,n_2}. Theorem 6.3.1 overcomes this limitation.

Table 6.2: In-sample statistics

Random variable	Realizations	Description
K_{n_1,n_2}	r	# of global distinct species
S_{n_1,n_2}	t	# of shared species
K_{j,n_j}	r_j	# of local distinct species in group j
$K^*_{j,n}$	r^*_j	# of distinct species observed in group j but which are missing in group j'

Theorem 6.3.1. *Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ be a sample of sizes n_1 and n_2 from model (6.1) under the* Vec-FDP *prior given in Equation (6.2). Then, the joint distribution of the local number of distinct species K_{1,n_1} and K_{2,n_2} and the global number of distinct species K_{n_1,n_2} equals*

$$\mathbb{P}\left(K_{n_1,n_2} = r,\ K_{1,n_1} = r_1,\ K_{2,n_2} = r_2\right) = V^r_{n_1,n_2}\frac{r_1!r_2!}{r^*_1!r^*_2!t!}\prod_{j=1}^{2}|C(n_j, r_j; -\gamma_j)|, \tag{6.18}$$

for $r \in \{1, \ldots, r_1 + r_2\}$ and $r_j \in \left\{\{1, \ldots, \min\{r, n_j\}\right\}$ $(j = 1, 2)$ and where we defined $r^*_1 = r - r_2$, $r^*_2 = r - r_1$ and $t = r_1 + r_2 - r$. The coefficient $C(\cdot, \cdot; \cdot)$ in Equation (6.18) denotes the generalized factorial coefficient, as defined in Charalambides (2002).

The proof of Theorem 6.3.1 is provided in Section E.4. The generalized factorial coefficients in Equation (6.18) can be computed via the triangular recurrence relationships described in Charalambides (2002). See Section E.1 for further details on generalized factorial coefficients. Equation (6.18) is a joint distribution and enables the evaluation of all other linearly dependent in-sample statistics. For instance, the distribution of the shared species, $\mathbb{P}\left(S_{n_1,n_2} = t\right)$, is obtained summing the expression in Equation (6.18) for all r, r_1, r_2 such that $t = r_1 + r_2 - r$, $t = 0, \ldots, \min\{n_1, n_2\}$. Alternatively, one may draw Monte Carlo samples from Equation (6.18) to obtain a Monte Carlo estimation of the distribution of interest. Furthermore, since Equation (6.3) is a linear relationship, the choice of S_{n_1,n_2} as a dependent variable is arbitrary.

It is possible to derive the joint prior $\mathbb{P}\left(K_{n_1,n_2} = r, S_{n_1,n_2} = t\right)$, integrating out one of the local quantities. By doing so, we obtain the following joint prior distribution

$$\begin{aligned}\mathbb{P}\left(K_{n_1,n_2} = r,\ S_{n_1,n_2} = t\right) &= V^r_{n_1,n_2} \\ &\times \sum_{k^*_1=0}^{r-t}\binom{r-k^*_1}{t}\frac{(t+k^*_1)!}{k^*_1!}|C(n_1, t+k^*_1; -\gamma_1)||C(n_2, r-k^*_1; -\gamma_2)|\,,\end{aligned} \tag{6.19}$$

for $r \in \{1, \ldots, n_1 + n_2\}$ and $t \in \{0, \ldots, \min\{r, n_1, n_2\}$. Additionally, the marginal distribution of S_{n_1,n_2} can be obtained by marginalizing K_{n_1,n_2} out of Equation (6.19). We have that,

$$\begin{aligned}\mathbb{P}\left(S_{n_1,n_2} = t\right) = \sum_{r=1}^{n_1+n_2}\sum_{k^*_1=0}^{r-t} V^r_{n_1,n_2}\binom{r-k^*_1}{t}\frac{(t+k^*_1)!}{k^*_1!} \\ \times |C(n_1, t+k^*_1; -\gamma_1)||C(n_2, r-k^*_1; -\gamma_2)|\,.\end{aligned} \tag{6.20}$$

Finally, we consider the marginal distributions of the local quantities, hence fixing $j \in \{1, 2\}$. We have the following expression for the local number of prior distinct species:

$$\mathbb{P}\left(K_{j,n_j} = r_j\right) = V_{n_j}^{r_j} |C\left(n_j, r_j; -\gamma_j\right)|, \tag{6.21}$$

for $r_j \in \{1, \ldots, n_j\}$ and where $V_{n_j}^{r_j}$ in obtained from V_{n_1,n_2}^{r}, defined in Equation (6.6), setting $n_{j'} = 0$, for $j' \neq j$. In particular, this implies $r = r_j$. We also notice that $V_{n_j}^{r_j}$ coincides with the definition of the V coefficients in Gnedin and Pitman (2006) and Miller and Harrison (2018). The proof of Equation (6.21) is given in Argiento and De Iorio (2022).

6.4 Out-of-sample prediction

The present section addresses the task of out-of-sample prediction of new distinct and shared species. Given $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, an additional sample comprising m_1 and m_2 individuals is considered, resulting in an enlarged sample of sizes $n_1 + m_1$ and $n_2 + m_2$, namely $\left(X_{j,n_j+1}, \ldots, X_{j,n_j+m_j}\right)$, for $j = 1, 2$. Similarly to the previous section, we start presenting the posterior quantities of interest by means of an updated example, reported in Figure 6.2. In red, we show the same observed species as in Figure 6.1 while we add the additional species in green. These are only present in a future sample, and therefore, they are still unobserved. We say that Figure 6.2 represents the enlarged sample, made of both the observed and the future ones. Finally, Table 6.3 summarizes the notation and the meaning of each random variable.

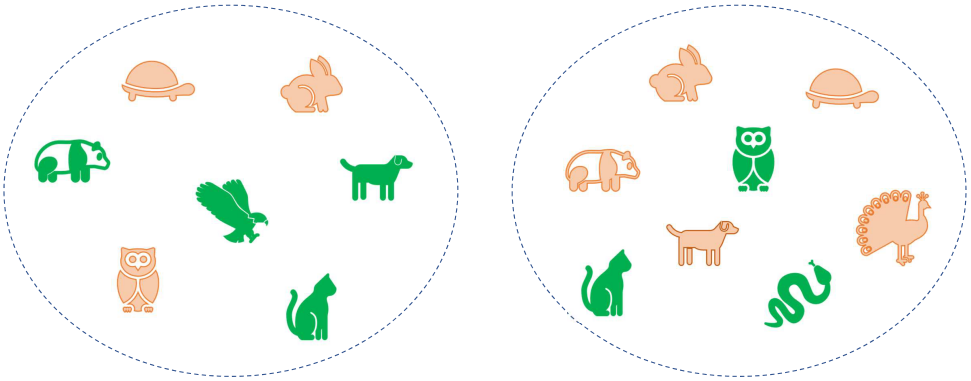

Figure 6.2: Observed and future species in the enlarged sample, green species represent those belonging to the future, additional, sample, hence they are unobserved.

Once again, we start from the local number of new distinct species, $K_{j,m_j}^{(n_j)}$, for $j = 1, 2$, and generally denoted as k_j. Let $j = 1$, this is defined as the number of distinct species in the enlarged sample minus the one that were already observed in the observed sample. In our running example, this is simply computed as the number of green species in area 1, that is four, $k_1 = 4$. This does not imply that a future sample can not contain turtles and rabbits, but these would not count as new species because they were already present in the previous sample. Similarly, $k_2 = 3$. Let us move to the global quantities, starting from the global number of distinct species $K_{m_1,m_2}^{(n_1,n_2)}$, whose realization is denoted as k. This is also defined as the number of global species in the enlarged sample minus the same quantity in the observed sample, namely, $K_{m_1,m_2}^{(n_1,n_2)} = K_{n_1+m_1,n_2+m_2} - K_{n_1,n_2}$. Some care is required when computing this quantity from Figure 6.2. Indeed, in Figure 6.1 it was enough to pool the red species and discard the repeated ones. Hence here one may tempted to pool the green species and discard the repeated ones, but

this would be wrong. Indeed, note that the owl and the dog have already been observed in area 1 and 2, respectively, and must not count as new global species. Hence, $k = 3$ (the eagle, the cat and the snake). Although trivial, this important example shows that we can not compute k from k_1 and k_2 only looking at the future sample (green species) but we must also take into account the past sample (red species). Let us also have a look at the other main global quantity, that is the number of new shared species, $S_{m_1,m_2}^{(n_1,n_2)}$, also defined as $S_{m_1,m_2}^{(n_1,n_2)} = S_{n_1+m_1,n_2+m_2} - S_{n_1,n_2}$ and denoted as s. Once again, if we repeat what we did in Section 6.3, i.e., pooling the red species and counting the number of the repeated ones, with the green species in Section 6.2, we end up committing a mistake. Indeed, the only repeated green species is the cat, while we claim that $s = 4$ in our running example. This follows since the correct way to count is to follow the definition and not to miss the mixed cases, i.e., those species that are first observed only in one group and then also in the future sample of the other one. In this case, the owl, the panda and the dog are shared species in the enlarged sample $S_{n_1+m_1,n_2+m_2}$, as well as the previously mentioned cat and the turtle, but the latter has already been counted as a shared species in the observed sample S_{n_1,n_2}, hence s equals four. Let us then characterize and name those quantities we implicitly computed to derive k and s. Let S_m^* be the number of those species that do not belong to the r observed global species and that are shared among the two areas and the s^* be its realization. In our example, $s^* = 1$, that is the cat, that is the only green species appearing in both areas. Moreover, let $S_{j',j}$ be the number of species that only appears in area j' for what concerns the observed sample but that are then also seen in area j once the future sample is considered. Their realizations are denoted as $s_{j',j}$, for $j', j = 1, 2$ and $j' \neq j$. For example, if $j = 1$, then $j' = 2$ and $s_{2,1} = 2$, i.e., the panda and the dog. The computation of k and s can be summarized as $k = k_1 + k_2 - s^* - s_{1,2} - s_{2,1}$ and $s = s^* + s_{1,2} + s_{2,1}$. Hence, $k = k_1 + k_2 - s$.

Finally, we follow Section 6.3 and introduce two additional auxiliary quantities, $K_{j,m}^{*(n)}$, for $j = 1, 2$, and denoted as k_j^*. Let $j = 1$, extending the prior quantity $K_{1,n}^*$, this counts the number of new distinct species that were not part of the observed r global species and that are present in the first area. In our example, $k_1^* = 1$ (the eagle) and $k_2^* = 1$ (the snake). As for $K_{j,n}^*$, these are also global quantities that are not defined in the case of a single area.

One further consideration is that we can divide the global posterior random variables in two categories, (i) those involving some of the r observed species ($S_{1,2}$ and $S_{2,1}$) and (ii) those considering only new species, never observed neither in area 1 nor in area 2 (S_m^*, $K_{1,m}^{*(n)}$ and $K_{2,m}^{*(n)}$). What about the two main quantities of interest $K_{m_1,m_2}^{(n_1,n_2)}$ and $S_{m_1,m_2}^{(n_1,n_2)}$? For what concerns the number of new shared species, this is directly related to $S_{1,2}$ and $S_{2,1}$, hence it clearly belongs to (i). On the other hand, the number of new distinct species, by definition, must belong to (ii). Indeed, for the sake of clarity, this has been derived as a function of $K_{j,m_j}^{(n_j)}$ and $S_{j',j}$ but it can also be computed ad $k = s^* + k_1^* + k_2^*$, which are all quantities belonging to (ii). This difference has a major impact when considering the computation of joint and marginal distributions of the posterior quantities of interest.

The main linear relationships among the nine introduced posterior random variables are the following

$$\begin{aligned} &k = s^* + k_1^* + k_2^*, \quad s = s^* + s_{1,2} + s_{2,1}, \quad s = k_1 + k_2 - k, \\ &k_j = s^* + k_j^* + s_{j',j}, \quad j', j = 1, 2; j' \neq j \end{aligned} \tag{6.22}$$

Only four of these equations are linearly independent, which means that the posterior set of random variables needs five linearly independent quantities to be fully characterized. In particular, we highlight

the following crucial linear relationship

$$S_{m_1,m_2}^{(n_1,n_2)} = K_{1,m_1}^{(n_1)} + K_{2,m_2}^{(n_2)} - K_{m_1,m_2}^{(n_1,n_2)}. \tag{6.23}$$

In Section 6.3.2, we pointed out that $r \leq r_1 + r_2$ and $r \geq r_j$, for $j = 1, 2$. In this posterior case, the analogous condition $k \leq k_1 + k_2$ still holds but in the main manuscript we notice that $k \geq k_j$ does not hold any longer. For example, in Section 6.2 we have $k = 3$ which is smaller that $k_1 = 4$. Why is so? Intuitively, it is possible that k_j is growing because of the discovery of many species that were only observed in area j', hence the presence of these in area j is not increasing k. More precisely, the set of equations in (6.17), it can be shown that the condition for $k_j > k$ to happen is $s_{j',j} > k_{j'}^*$.

Table 6.3: Out-of-sample statistics.

Random Variable	Realization	Description
$K_{m_1,m_2}^{(n_1,n_2)}$	k	# of new global distinct species
$S_{m_1,m_2}^{(n_1,n_2)}$	s	# of new shared species
$K_{j,m_j}^{(n_j)}$	k_j	# of new local distinct species in group j
$K_{j,m}^{*(n)}$	k_j^*	# of new distinct species in group j but missing in group j'
S_m^*	s^*	# of new shared species among the k new distinct species
$S_{j',j}$	$s_{j',j}$	# of species which were first *only* observed in group j' and that are then observed in group j

6.4.1 Posterior of the total number of species

The model introduced in Section 6.2 assumes a finite number of species M, which is usually unknown and is assumed to be random. Recall that q_M denotes its prior distribution. Following the posterior representation theorem of a Vec-FDP prior presented in Theorem 5.3.4, our belief about M is updated once data have been collected since the total number of species equals, in distribution, $r + M^\star$, where $M^\star$ is a random variable distributed according to $q_{M|\boldsymbol{X}}^\star$, whose probability mass function is

$$q_{M|\boldsymbol{X}}^\star(m^\star) = \frac{1}{V_{n_1,n_2}^r}(m^\star + r)_{r\downarrow} q_M(m^\star + r) \prod_{j=1}^{d} \frac{1}{(\gamma_j(m^\star + r))_{n_j}}, \tag{6.24}$$

for $m^\star \in \mathrm{N} \cup \{0\}$. Specifically, $M^\star$ may equal zero, since $q_{M|\boldsymbol{X}}^\star(0) > 0$. See Section E.5.1 for the equivalence between Equation (6.24) and the corresponding formulation presented in Theorem 5.3.4. Furthermore, in Section E.5.2, we show that posterior expected value of $M^\star$ equals

$$E\left(M^\star \mid \boldsymbol{X}\right) = \frac{V_{n_1,n_2}^{r+1}}{V_{n_1,n_2}^r}. \tag{6.25}$$

In particular, it admits the following asymptotic approximation for large sample sizes:

$$E\left(M^\star \mid \boldsymbol{X}\right) = (r+1)\frac{q_M(r+1)}{q_M(r)}(\gamma_1 r)_{\gamma_1}(\gamma_2 r)_{\gamma_2} n_1^{-\gamma_1} n_2^{-\gamma_2}\,(1 + o(1)). \tag{6.26}$$

Equation (6.26) shows that $E\left(M^{\star} \mid \boldsymbol{X}\right)$ goes to zero when $n_1, n_2 \to \infty$. This aligns with our modelling assumption, i.e., with an infinite amount of data, all possible species would have already been observed, leaving no room for further discoveries. Moreover, this also highlights the crucial role of the parameters γ_1 and γ_2 in governing the discovery rate of new species. If the values are greater than one, the expression quickly approaches zero. This means that after only a few observations, the expectation of discovering new species rapidly decreases. Conversely, values much smaller than one enable the discovery of new species even when a large number of observations is considered.

6.4.2 Joint predictive distribution

Following the same approach of Section 6.3.2, it is of primary interest to derive both the marginal and the joint distributions of the random variables on the right side of Equation (6.23). These are stated in the following theoretical results.

Theorem 6.4.1. *Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ be a sample of sizes n_1 and n_2 from model (6.1) under the* Vec-FDP *prior given in Equation (6.2). Let $K_{n_1,n_2} = r$ and $S_{n_1,n_2} = t$ be the observed number of global distinct and shared species, and let $K_{j,n_j} = r_j$, for $j = 1, 2$ be the observed local distinct species. Then,*

$$\begin{aligned}&\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2} = k,\ K^{(n_1)}_{1,m_1} = k_1,\ K^{(n_2)}_{2,m_2} = k_2 \mid \boldsymbol{X}\right) = \\ &\frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^{r}_{n_1,n_2}} \prod_{j=1}^{2} |C(m_j, k_j; -\gamma_j, -(\gamma_j r_j + n_j))| \sum_{s^*=0}^{k} \sum_{k_1^*=0}^{k-s^*} \frac{k_1! k_2!}{s^*! k_1^*! k_2^*!} \binom{r_1^*}{s_{12}} \binom{r_2^*}{s_{21}}\end{aligned} \tag{6.27}$$

for non-negative integers k, k_1, k_2 such that $0 \le k \le k_1 + k_2$ and $0 \le k_j \le m_j$ for $j = 1, 2$. Specifically, in Equation (6.27) we set $k_2^* = k - s^* - k_1^*$, $s_{12} = k_2 + k_1^* - k$ and $s_{21} = k_1 - k_1^* - s^*$. Finally, the coefficient $C(\,\cdot\,, \cdot\,; \cdot\,, \cdot\,)$ in Equation (6.27) denotes the non-central generalized factorial coefficient, as defined in Charalambides (2002).

The proof of Theorem 6.4.1 is provided in Section E.5.4. The non-central generalized factorial coefficient satisfies a specific recursive relation that helps its evaluation. See Charalambides (2002) for details. As discussed at the beginning of this section, all auxiliary quantities involved in Equation (6.27) are interpretable, see Table 6.3. The marginal distributions for the local and global number of distinct species are reported in Equations (6.28) and (6.29). These are the posterior counterparts of Equations (5.11) and (6.21).

Proposition 3. *Under the same hypothesis of Theorem 6.4.1, the marginal distribution of the global number of new distinct species $K^{(n_1,n_2)}_{m_1,m_2}$ is*

$$\begin{aligned}\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2} = k \mid \boldsymbol{X}\right) &= \frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^{r}_{n_1,n_2}} \\ &\times \sum_{k_1^*=0}^{k} \sum_{k_2^*=k-k_1^*}^{k} \frac{(k_1^*+s^*)!(k_2^*+s^*)!}{k_1^*! k_2^*! s^*!} \prod_{j=1}^{2} |C(m_j, k_j^*+s^*; -\gamma_j, -(\gamma_j r + n_j))|,\end{aligned} \tag{6.28}$$

for $k \in \{0, \ldots, m_1 + m_2\}$ and where $s^* = k - k_1^* - k_2^*$.

The proof of Proposition 3 is provided in Section E.5.5. The marginal distributions of the local

number of new distinct species in group j, $K^{(n_j)}_{j,m_j}$, follows from Equation (6.28) and it equals

$$\mathbb{P}\left(K^{(n_j)}_{j,m_j} = k_j \mid \boldsymbol{X}\right) = \frac{V^{r_j+k_j}_{n_j+m_j}}{V^{r_j}_{n_j}} |C(m_j, k_j; -\gamma_j, -(\gamma_j r_j + n_j))|, \tag{6.29}$$

for $k_j \in \{0, \ldots, m_j\}$. The latter coincides with the findings in De Blasi et al. (2015) about Gibbs-type priors with negative parameters. Furthermore, we highlight that Theorem 6.3.1 requires $K_{j,n_j} \leq K_{n_1,n_2}$. This condition, however, is not necessary in Theorem 6.4.1, as it is possible for the number of global discoveries to exceed the number of local discoveries, i.e., $K^{(n_j)}_{j,m_j} > K^{(n_1,n_2)}_{m_1,m_2}$.

6.4.3 Discovering shared species

Similarly to Section 6.3.2, Equations (6.29) and (6.28) allow us to compute the posterior expected values of the number of new distinct species and, by means of Equation (6.23), to derive the Bayesian estimator of the number of new shared species, that is

$$E\left[S^{(n_1,n_2)}_{m_1,m_2} \mid \boldsymbol{X}\right] = E\left[K^{(n_1)}_{1,m_1} \mid \boldsymbol{X}\right] + E\left[K^{(n_2)}_{2,m_2} \mid \boldsymbol{X}\right] - E\left[K^{(n_1,n_2)}_{m_1,m_2} \mid \boldsymbol{X}\right]. \tag{6.30}$$

The associated uncertainty is quantified through the posterior distribution, namely, $\mathbb{P}\left(S^{(n_1,n_2)}_{m_1,m_2} = s \mid \boldsymbol{X}\right)$ that is obtained summing the expression in Equation (6.27) for all k, k_1, k_2 such that $s = k_1 + k_2 - k$ and for all $s = \{0, \ldots, m_1 + m_2\}$. Alternatively, it can be estimated via Monte Carlo sampling, drawing samples from Equation (6.27).

The shared species sample coverage is defined as the proportion of shared species that are observed in the sample. Being able to estimate it allows us to assess the number of shared species that have been observed, and, therefore, to decide whether it is worth continuing the experiment and sampling more and more observations. Moreover, the shared species sample coverage on m-steps ahead facilitates determining the size of the additional sample that ensures the coverage exceeds a specified threshold. In our setting, the shared coverage probability for m-steps ahead, i.e., the probability of not discovering new shared species in the additional sample, follows from Equation (6.27) and it equals

$$\mathbb{P}\left(S^{(n_1,n_2)}_{m_1,m_2} = 0 \mid \boldsymbol{X}\right) = \sum_{k_1=0}^{m_1} \sum_{k_2=0}^{m_2} \frac{V^{r+k_1+k_2}_{n_1+m_1,n_2+m_2}}{V^{r}_{n_1,n_2}} \prod_{j=1}^{d} |C(m_j, k_j; -\gamma_j, -(\gamma_j r_j + n_j))|. \tag{6.31}$$

In particular, the one-step ahead coverage probability, is obtained from Equation (6.31) setting $m_1 = 1$ and $m_2 = 1$. In this case, we can explicitly write the whole distribution $\mathbb{P}\left(S^{(n_1,n_2)}_{1,1} = s \mid \boldsymbol{X}\right)$ for each

$s \in \{0, 1, 2\}$ and not just for $s = 0$ as it happens for $m_j > 1$. Specifically, we have that

$$\begin{aligned}
\mathbb{P}\left(S_{1,1}^{(n_1,n_2)} = 0 \mid \boldsymbol{X}\right) &= \frac{V^r_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}(\gamma_1 r_1 + n_1)(\gamma_2 r_2 + n_2) \\
&\quad + \frac{V^{r+1}_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\{\gamma_1(\gamma_2 r_2 + n_2) + \gamma_2(\gamma_1 r_1 + n_1)\} + \frac{V^{r+2}_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\gamma_1\gamma_2\,, \\
\mathbb{P}\left(S_{1,1}^{(n_1,n_2)} = 1 \mid \boldsymbol{X}\right) &= \frac{V^r_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\left\{r_2^*\gamma_1(\gamma_2 r_2 + n_2) + r_1^*\gamma_2(\gamma_1 r_1 + n_1)\right\} + \frac{V^{r+1}_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\gamma_1\gamma_2(r_1^* + r_2^*)\,, \\
\mathbb{P}\left(S_{1,1}^{(n_1,n_2)} = 2 \mid \boldsymbol{X}\right) &= \frac{V^r_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\gamma_1\gamma_2 r_1^* r_2^*\,.
\end{aligned} \tag{6.32}$$

In particular, from Equation (6.32) we derive the probability of discovering at least one new shared species, that is

$$\begin{aligned}
\mathbb{P}\left(S_{1,1}^{(n_1,n_2)} > 0 \mid \boldsymbol{X}\right) &= 1 - \frac{V^r_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}(\gamma_1 r_1 + n_1)(\gamma_2 r_2 + n_2) \\
&\quad - \frac{V^{r+1}_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\{\gamma_1(\gamma_2 r_2 + n_2) + \gamma_2(\gamma_1 r_1 + n_1)\} - \frac{V^{r+2}_{n_1+1,n_2+1}}{V^r_{n_1,n_2}}\gamma_1\gamma_2.
\end{aligned} \tag{6.33}$$

The ratios of V coefficients in Equation (6.33) represent the three different scenarios where the new pair of observations yield none, one, or two new distinct global species. It is worth noting that increasing values of the probability in Equation (6.33) indicates that the observed sample is sufficiently exhaustive, suggesting that further data collection may not be necessary. Equation (6.33) can be compared with the Good-Turing estimators proposed in the frequentist literature. These are presented in Section 6.6.2 where a simulation study is carried out to compare our model with the competing estimators. Nevertheless, we note that we are not aware of any available result, nor in the frequentist or Bayesian literature, about the m-step ahead discovery probability to which we can compare Equation (6.31).

6.5 Diversity-based estimation strategy

6.5.1 Diversity and similarity indices

In ecology, the concept of diversity is tied not only to the number of distinct species present in an area but also to their heterogeneity. For instance, having ten equally represented species is intuitively very different from having one highly abundant species and the remaining nine extremely rare (Colwell et al., 2009). In the literature, there is no unique quantitative definition of diversity; instead, a variety of indices have been proposed to measure it. Among these, we focus on Simpson's diversity index (Simpson, 1949), which captures both richness and evenness in species distributions. Assuming that the unknown discrete distribution P_j that generates the population is made up of M distinct species with species proportions $w_{j,m}$, Simpson's diversity index is

$$\rho_j = \sum_{m=1}^{M} w_{j,m}^2. \tag{6.34}$$

Useful alternatives are suitable transformations of ρ_j, such as the Gini-Simpson index, defined as $1-\rho_j$, or the inverse-Simpson index, $1/\rho_j$ (Colwell et al., 2009). The index ρ_j ranges between $1/M$ (all species are uniformly distributed) and one (one species is abundant and the remaining are negligible). These extreme cases correspond to the cases of maximum and minimum heterogeneity (minimum and maximum homogeneity), respectively.

By definition, the Simpson's diversity index quantifies diversity within a single area, not across two areas. Instead, when we assume the population is generated from a vector (P_1, P_2) of unknown discrete distribution, each having M different species whose proportions are $w_{j,m}$, for $j = 1, 2$ and $m = 1, \ldots, M$, the Morisita index (Morisita, 1959) can be used to quantify the similarity between the two areas, namely, $2\rho_{12}/(\rho_1 + \rho_2)$, where

$$\rho_{12} = \sum_{m=1}^{M} w_{1,m} w_{2,m}. \tag{6.35}$$

Increasing values of the Morisita index show evidence for identical communities, i.e., with the same species proportions. Chao et al. (2017) highlighted an important probabilistic interpretation of the Morisita index. The numerator represents the probability of selecting the same shared species when two observations, one for each group, are randomly drawn from this population. The denominator of the Morisita index is instead the sum of the Simpson's diversity indices in the two groups. The original unbiased estimator proposed by Simpson Simpson (1949) is

$$\hat{\rho}_{j,\text{unb}} = \sum_{l=1}^{r} \frac{n_{j,l}}{n_j} \frac{(n_{j,l}-1)}{(n_j-1)}. \tag{6.36}$$

Since ρ_j is interpreted as the probability that two randomly and independently chosen individuals belong to the same species, the estimator in Equation (6.36) is obtained by dividing the total number of within-species pairs, $\sum_{l=1}^{r} n_{j,l}(n_{j,l}-1)/2$, by the total number of possible pairs, $n_j(n_j-1)/2$.

6.5.2 Parameter interpretation and estimation

The Bayesian nonparametric model in Equation (6.1) under the Vec-FDP prior in Equation (6.2) is governed by the parameters $\gamma = (\gamma_1, \gamma_2)$ and Λ. These are not known and must be estimated from the data. To do so, we first provide a diversity-based interpretation of such parameters. Equation (6.1) does not assume the existence of a unique, unknown vector of probability distributions (P_1, P_2) but rather assumes that (P_1, P_2) is random. However, we can integrate out this source of randomness by taking the expected values, obtaining

$$E(\rho_j) = E\left(\sum_{m=1}^{M} w_{j,m}^2\right) = (1+\gamma_j)\, E\left(\frac{1}{1+\gamma_j M}\right), \tag{6.37}$$

$$E(\rho_{12}) = E\left(\sum_{m=1}^{M} w_{1,m} w_{2,m}\right) = E(1/M). \tag{6.38}$$

Proof of Equations (6.37) and (6.38) are deferred to Section E.6.1. The expected value of ρ_{12} in Equation (6.38) solely depends on q_M. This is due to the fact that we model the dependence in the vector of random probability measures (P_1, P_2) through the random number of species M and by imposing common atoms, which are irrelevant at this stage. Note that our choice of q_M as $\text{Pois}_1(\Lambda)$ yields explicit

expressions for Equation (6.38) depending on Λ; see Equation (6.16). Then, Λ regulates the amount of similarity between the two areas. Furthermore, the limits of the expected Simpson index in Equation (6.37) further illuminate the interpretation of γ_1 and γ_2 as homogeneity parameters, as described in Section 6.3.1. In fact, the limits of Equation (6.37) for $\gamma_j \to 0$ and $\gamma_j \to \infty$ are equal to one and $E\,(1/M)$, respectively, which represent the case of minimum and maximum heterogeneity. Therefore, γ_1 and γ_2 are homogeneity parameters since heterogeneity decreases (increases) as γ_j increases (decreases).

Our strategy for estimating γ_1, γ_2 and Λ consists of two steps. Firstly, we use the observed data to get estimates of ρ_j $j = 1, 2$ in Equation (6.34) and ρ_{12} in Equation (6.35). This step can be achieved with standard routines, such as using the estimator in Equation (6.36). Then, we plug such estimates in the left-hand sides of Equations (6.37) and (6.38) and solve with respect to γ_1, γ_2 and Λ. In particular, we first solve Equation (6.38), which involves only Λ. Since $E\,(1/M) = \Lambda^{-1}\left(1 - \mathrm{e}^{-\Lambda}\right)$ is monotonically decreasing towards zero, this step is trivial. Then, given Λ, Equation (6.37) decouples in two independent conditions, one for each group, which can be solved in parallel.

Alternative estimation procedures can also be considered. For instance, in the case of a single population, Camerlenghi et al. (2024) proposed to maximise the marginal likelihood, while Lijoi et al. (2007a) and Favaro et al. (2009) suggested maximising the prior distribution of the number of distinct species evaluated at their observed values. Alternatively, Balocchi et al. (2024a) employed a fully Bayesian approach by specifying suitable hyperpriors and estimating the parameters via Markov Chain Monte Carlo. However, among these possibilities, we recommend our diversity-based approach for its straightforward interpretability and ease of implementation. In fact, the computational effort required by the proposed strategy is negligible compared to the mentioned alternatives.

6.5.3 Posterior quantities

In this section, we present the Bayesian estimators of ρ_j, $j = 1, 2$, and ρ_{12} given an observed sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ of sizes n_1 and n_2, with r distinct species and t shared species. In particular, the posterior expected value of the Simpson diversity index in area j equals

$$E\left(\rho_j \mid \boldsymbol{X}\right) = E_{q^{\star}_{M|\boldsymbol{X}}}\left(\frac{\sum_{l=1}^{r} n_{j,l}(n_{j,l}+1) + \gamma_j\left(\gamma_j(r+M^{\star}) + r + M^{\star} + 2n_j\right)}{n_j(n_j+1) + \gamma_j^2(r+M^{\star})^2 + \gamma_j(r+M^{\star})(2n_j+1)}\right). \tag{6.39}$$

Note that the posterior is computed with respect to the multilevel sample $\boldsymbol{X}$ since $M^{\star}$ depends on both areas. The proof is given in Section E.6.3.

The limiting behavior for γ_j going to zero or infinity further explains how to interpret such a parameter and sheds light on the properties of the model. Firstly, consider the limit for γ_j going to zero while $\gamma_{j'}$ is kept fixed, with $j' \neq j$, that is

$$\lim_{\gamma_j \to 0} E(\rho_j \mid \boldsymbol{X}) = \sum_{l=1}^{r} \frac{n_{j,l}}{n_j}\frac{(n_{j,l}+1)}{(n_j+1)}. \tag{6.40}$$

Equation (6.40) is derived in Section E.6.4. In the model formulation, the limiting case $\gamma_1 \to 0$ represents a prior belief in which all the mass is concentrated on a single species. In this case, the Simpson diversity index is one, which corresponds to the minimal heterogeneity, i.e., no diversity. A posteriori, this belief is updated only based on the observed sample, which results in an increase in the estimated diversity.

Finally, we note the similarity between the limit in Equation (6.40) and the Simpson estimator in Equation (6.36). The intuition behind the result in Equation (6.40) is that the prior effectively contributes to the estimation through an additional observation with unknown group assignment. Thus, the total sample size increases by one, and the number of within-species pairs is augmented to include n_j additional pairs formed between each observation in group j and the prior-induced pseudo-observation. Since the Bayesian estimator in Equation (6.40) is larger than the estimator in Equation (6.36), small values of γ_j shrink the Simpson estimator towards one, coherently with the interpretation of the corresponding prior belief.

We now turn our attention to the limit of $\gamma_j \to +\infty$, which results in

$$\lim_{\gamma_j \to +\infty} E(\rho_j \mid \boldsymbol{X}) = E_{q^{\star}_{M|\boldsymbol{X},\infty}} \left(\frac{1}{r + M^{\star}} \mid \boldsymbol{X} \right), \tag{6.41}$$

where $q^{\star}_{M|\boldsymbol{X},\infty}$ is the limiting probability distribution of $q^{\star}_{M|\boldsymbol{X}}$ when $\gamma_j \to +\infty$. Further details are provided in Section E.6.2 while the explicit expression is reported in Equation (6.86). The proof of Equation (6.41) is provided in Section E.6.4. We recall that large values of γ_1 and γ_2 correspond to setting a uniform prior over all species, i.e., the situation of maximum heterogeneity. A posteriori, after observing a sample containing r distinct species, the expected heterogeneity remains the largest possible one: it is equal to the average of the inverse of $r + M^{\star}$, where $M^{\star}$ is the (random) number of unobserved species and is distributed according to the limiting distribution $q^{\star}_{M|\boldsymbol{X},\infty}$.

We now focus on the Bayesian posterior estimator of ρ_{12}, given by

$$E\left(\rho_{12} \mid \boldsymbol{X}\right) = E_{q^{\star}_{M|\boldsymbol{X}}} \left(\frac{\sum_{l=1}^{t} n_{1,l} n_{2,l} + \gamma_1 \gamma_2 (r + M^{\star}) + n_1 \gamma_2 + n_2 \gamma_1}{(\gamma_1 (r + M^{\star}) + n_1)(\gamma_2 (r + M^{\star}) + n_2)} \right). \tag{6.42}$$

Again, we turn our attention to the limiting case for the γ_j's parameters. The limit of $E\left(\rho_{12} \mid \boldsymbol{X}\right)$ for γ_1, γ_2 both going towards zero equals the plug-in estimator, i.e.,

$$\lim_{\gamma_1, \gamma_2 \to 0} E\left(\rho_{12} \mid \boldsymbol{X}\right) = \sum_{l=1}^{t} \frac{n_{1,l}}{n_1} \frac{n_{2,l}}{n_2}. \tag{6.43}$$

This result supports the interpretation of γ_1 and γ_2 as homogeneity parameters which can be employed to impose sparsity within each area when pushed towards zero. Indeed, we recall that values of γ_j close to zero represent the prior belief of minimum correlation between P_1 and P_2, that is $E\left(1/M\right)$ (see Section 6.3.1). Then, the posterior expectation $E\left(\rho_{12} \mid \boldsymbol{X}\right)$ achieves its minimum equal to the plug-in estimator, obtained by replacing the species probabilities by normalized observed counts, see e.g., Archer et al. (2014). This means that the amount of similarity among the two areas must be at least equal to the observed ones. In the undersampled regime, the plug-in estimator is negatively biased, while the prior offers a correction for this bias.

Finally, when we assume a priori that all species are equally present in the population, i.e., the case of $\gamma_1, \gamma_2 \to +\infty$, we obtain

$$\lim_{\gamma_1, \gamma_2 \to +\infty} E(\rho_{12} \mid \boldsymbol{X}) = E_{q^{\star}_{M|\boldsymbol{X},\infty}} \left(\frac{1}{r + M^{\star}} \mid \boldsymbol{X} \right). \tag{6.44}$$

Specifically, Equation (6.44) coincides with the corresponding limit of the Simpson index in Equation

(6.41). This implies that the estimated Morisita index reaches its maximum value of one, which is, once again, consistent with our interpretation of the γ_j's parameters. In fact, pushing the γ_j's towards infinity reflects a strong prior belief that the two populations are perfectly correlated, i.e., that they are identical.

6.6 Simulation study

We carried out a simulation study to assess our methodology and compare it with the existing estimators. The R code performing the analysis is available at `https://github.com/alessandrocolombi/HSSM`. The simulation study consists of two different experiments and we follow Yue and Clayton (2012) for generating the data.

6.6.1 Data generation

As a data generating mechanism, we follow Yue and Clayton (2012) and assume (P_1, P_2) to be two discrete probability distributions, each made of $M_{\text{true}} = 60$ species. The group-specific species proportions are denoted by $p_{j,m}$, $j = 1, 2$; $m = 1, \dots, M_{true}$ and set according to a geometric decay, i.e., $p_{j,m} \propto \alpha_j^m$, with $\alpha_j \in \{0.8, 0.85, 0.9\}$. We consider all possible combinations of these three values, for a total of six scenarios. Figure 6.3 shows the species probabilities for each value of α_j.

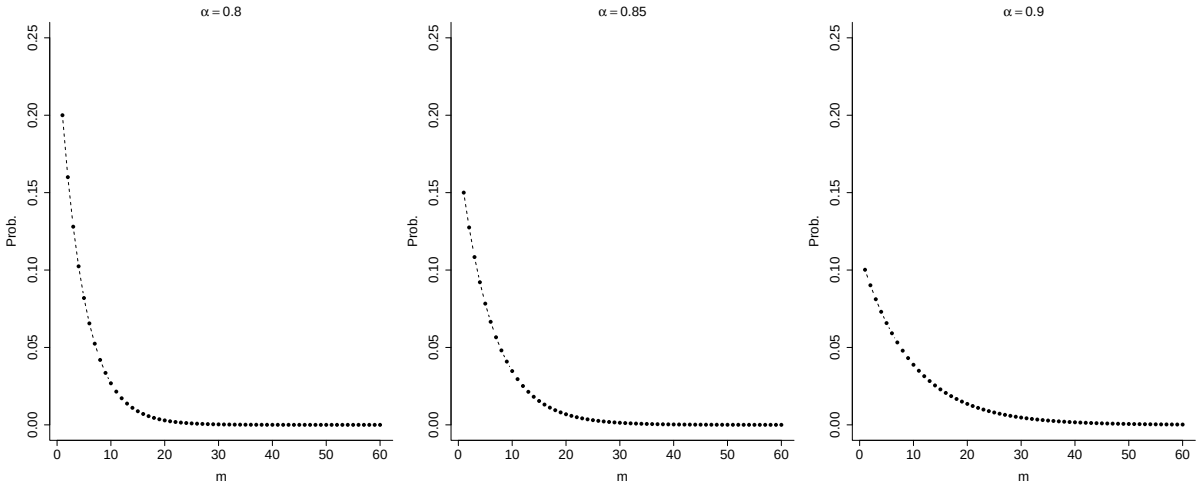

Figure 6.3: Graphical representations of the $M_{\text{true}} = 60$ species proportions used to generate data. These probabilities are plotted before shuffling the data.

For each couple (α_1, α_2), we also define the corresponding couple (P_1, P_2) from which the data are generated by random sampling. However, since the probabilities $p_{j,m}$ are monotonically decreasing in m, the species having a high probability of being observed in one group are likely to appear in the other group. Consequently, only these few species will have a high probability of being observed as shared, leaving a negligible probability for all others. To mitigate this effect, we randomly shuffle the species (P_1, P_2). In particular, this approach prevents the Morisita index described in Section 6.5.1 from being exactly equal to one when $\alpha_1 = \alpha_2$. Finally, in each scenario, we let $n_1 = n_2$ be the number of observed data in each group and $n = n_1 + n_2$ the total number of observations.

6.6.2 Experiment 1

The first experiment compares the ability of our Bayesian model to estimate the one-step ahead probability of discovering a new shared species against two nonparametric frequentist competitors proposed by Yue and Clayton (2012) and Chao et al. (2017).

Specifically, let f_{ν_1,ν_2} be the number of species that appeared exactly ν_1 and ν_2 times in the first and second groups, respectively. Moreover, let $f_{\nu_1,+}$ (f_{+,ν_2}) be the number of species that appeared exactly ν_1 (ν_2) times in the first (second) group and at least once in the second (first) group. Then, the estimators proposed by Yue and Clayton (2012) (YueDP) and Chao et al. (2017) (ChaoDP) are the following:

$$\mathbb{P}_{\text{YueDP}}\left(S_{1,1}^{(n_1,n_2)} > 0 \mid \boldsymbol{X}\right) = \frac{1}{n_1}\left(f_{1+} + f_{+1} + f_{11}\right), \tag{6.45}$$

$$\mathbb{P}_{\text{ChaoDP}}\left(S_{1,1}^{(n_1,n_2)} > 0 \mid \boldsymbol{X}\right) = \frac{f_{1+}}{n_1} + \frac{f_{+1}}{n_2} + \frac{f_{11}}{n_1 n_2}. \tag{6.46}$$

Equation (6.45) is defined only for $n_1 = n_2$. Estimators in Equations (6.45) and (6.46) will be compared with the discovery probability in Equation (6.33) (BayesDP). A third alternative, proposed by Yue et al. (2022), is also available in the literature. However, we did not include this option because, in the experiments we conducted, it produced results almost identical to those of ChaoDP.

Given (P_1, P_2) generated as described in Section 6.6.1, we evaluate the competing methods over a grid of sample sizes, $n_1 = n_2 = \{50, 100, \ldots, 400\}$. The experiment is then repeated for 140 datasets. Results are compared with the true probability of discovering a new species, as reported in Yue and Clayton (2012), which equals

$$\begin{aligned}\mathbb{P}_{\text{true}}\left(S_{1,1}^{(n_1,n_2)} > 0 \mid \boldsymbol{X}\right) = \sum_{m=1}^{M_{\text{true}}} & p_{1,m}p_{2,m}\mathbf{1}(n_{1,m} = 0,\, n_{2,m} = 0) + p_{1,m}\mathbf{1}(n_{1,m} = 0,\, n_{2,m} > 0) \\ & + p_{2,m}\mathbf{1}(n_{1,m} > 0,\, n_{2,m} = 0),\end{aligned}$$

where $n_{j,m}$ is the observed absolute frequency of the m-th species in the j-th group.

Figure 6.4 displays the results for the different scenarios. Dots represent the median value across the 140 replications, while the vertical bars correspond to the associated 95% confidence intervals. The uncertainty is considerably high when n is small, mainly due to the data generation mechanism. Indeed, even the true values (in black) show significant uncertainty. This arises because, with few observations, the number of shared species detected is small, meaning that even slight changes, such as observing one or two additional shared species, can significantly impact the estimates across replications. Regarding the performance of our proposed method (in green), overall, our estimates closely align with the true values. In particular, the results show a trend comparable to the competitors, with some differences in point estimates for small datasets and convergence to the true values as n increases. For instance, in Scenario I with $n_1 = 50$, the proposed method underestimates the true probability, although the median value still falls within the true interval. As n increases, the proposed estimates align with the competitors, even though the confidence intervals remain somehow wider and asymmetric. Scenario III is the most challenging for our method since it is characterized by high species heterogeneity. In this scenario, the proposed method slightly overestimates the true values for small n and underestimates the true probabilities for larger n. Among the competitors, YueDP (blue) shows the highest error, particularly for small sample sizes, as already by Yue et al. (2022). Conversely, ChaoDP (yellow) is highly precise,

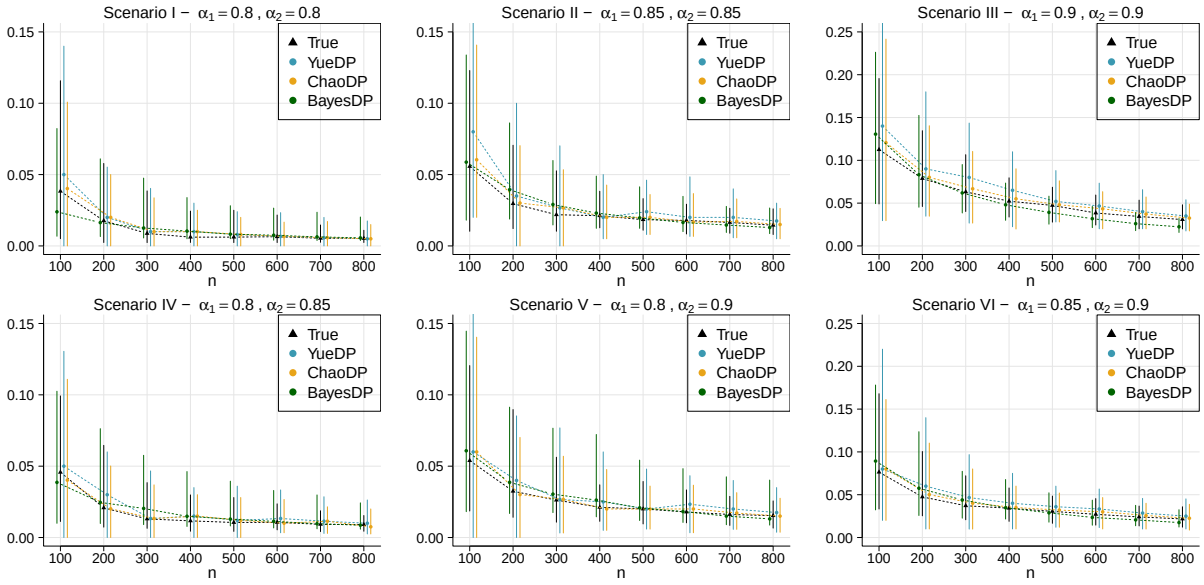

Figure 6.4: True and estimated probabilities of discovering a new shared species in Experiment 1, evaluated for different sample sizes.

with median estimates close to true values across all scenarios. However, the lower tails of the confidence intervals for YueDP and ChaoDP are often longer than the true and proposed ones. In some scenarios (I, II, IV, V), these tails extend to include zero, which is particularly inconvenient in practical applications.

Summing up, the estimator in Equation (6.33) for the shared discovery probability is comparable to the state-of-the-art estimators. However, our model offers the additional benefit of providing estimates for the m-steps ahead coverage probability given in Equation (6.31) and the m-steps ahead estimates of the number of shared species, $E\left(S_{m_1,m_2}^{(n_1,n_2)} \mid \boldsymbol{X}\right)$, which is the focus of the second experiment.

6.6.3 Experiment 2

The second experiment showcases the ability of estimator in Equation (6.30) (BayesSH) to predict the number of additional shared species in a future, unobserved test set. Unlike the first experiment in Section 6.6.2, we are not aware of any frequentist estimator specifically designed for the number of unseen shared species in a finite, future sample. However, we compare our approach to the state-of-the-art estimator for shared species richness introduced by Chao et al. (2000), as implemented in the `R` package `SpadeR` (Chao et al., 2016). This estimator, referred to as ChaoSH, focuses on estimating the total number of shared species across two groups. However, there are significant differences between BayesSH and ChaoSH. Our BayesSH estimator is based on the assumption that the two groups are generated from a model that, under infinite sampling, produces the same species in the two groups but with different proportions. On the other hand, Chao et al. (2000) follow a different approach. For each group j, they allow some of the probabilities $w_{j,m}$ to be zero for some m. This implies that they assume the possibility that, even with infinite sampling, some species that are present in one group can never be found in the other group. This is equivalent to assuming that the total number of species differs between areas, which, in the end, allows to find the number of shared species in the two groups. Consequently, the purpose of this simulation experiment is not to compare the performance of the two methods but to highlight their different assumptions and interpretation of the results. Although later works have extended the

ChaoSH estimator (see among others Pan et al. (2009); Chuang et al. (2015); Chao et al. (2017)), we chose to compare our results against Chao et al. (2000) because of its relevance in the literature as the first proposal of its kind.

The experiment proceeds as follows: firstly, we generate data as described in Section 6.6.1 with $n_1 = n_2 = 400$. Then, we set a grid of percentages ranging from 0.1 to 0.9, and for each value, we defined the training set to consist of that corresponding percentage of the entire dataset. Finally, we computed the main statistic of interest, S_{true}, which serves as the reference value for our comparisons.The training set is used to estimate the model parameters as described in Section 6.5 and compute the observed statistics, S_{obs}. Given the estimated parameters, we predict the expected number of new shared species in the test set, S_{est}, as explained in Section 6.4. To conclude, we compare S_{true} with $S_{\text{obs}} + S_{\text{est}}$. We replicate the experiment generating for each percentage value, 140 different training and test sets.

The results in Figure 6.5 show that our BayesSH estimator converges to the true value S_{true} as the training set percentage increases, for all scenarios. A slight underestimation is registered only in scenarios II and III, consistent with the findings of the experiment in Section 6.6.2. On the other hand, ChaoSH tends to stabilize at slightly higher values than S_{true}, showing a clear discrepancy between the dashed line in Figure 6.5 and the final prediction of ChaoSH, i.e., when the training set comprises nearly the entire dataset. This gap corresponds to the estimate of the number of shared species present in the population but unobserved in the initial dataset. Exceptions are scenarios I and IV, where ChaoSH also converges to S_{true}. This behavior is again consistent with the first simulation experiment. In scenarios I and IV, the probability of discovering new shared species after observing 800 individuals is negligible, indicating that the observed dataset is already saturated with all possible shared species.

As the results show, the ChaoSH estimator is inferential in nature, aiming to estimate an unknown quantity of the model by relying on asymptotic assumptions, such as sufficiently large sample sizes. In contrast, we propose a predictive approach, applicable to any observed sample sizes n_1 and n_2, as well as any future sample sizes m_1 and m_2, although we expect some bias for large values of m_1 and m_2 when this assumption is not matched in practice. In other words, the BayesSH answers the question, "How many species that we have not yet observed will we find in the future?" while ChaoSH answers the question, "How many species that we have not yet observed actually exist?", though in this second case, some species may never be encountered.

6.7 Analysis of ants data

In this section, we illustrate the methodology outlined in the previous sections to analyse real-world data. Zara et al. (2021) conducted a case study aimed at evaluating the impact of urbanization on biodiversity, with a specific focus on the beneficial effects of urban green areas. Although urbanization is often cited as one of the main drivers of species extinction, due to factors such as pollution, changes in land use, and the introduction and spread of alien species, the authors argue that urban green spaces can serve as important refuges, not only promoting human well-being but also supporting high levels of species diversity. This is because resources such as water and food are readily available in these areas, fostering biodiversity in potentially complex communities. To explore this topic, the authors conducted their study in the urban setting of Trieste, a city in north-eastern Italy. To account for areas with varying levels of urbanization, Zara et al. (2021) identified a few distinct sites in the city. For our analysis, we focus on

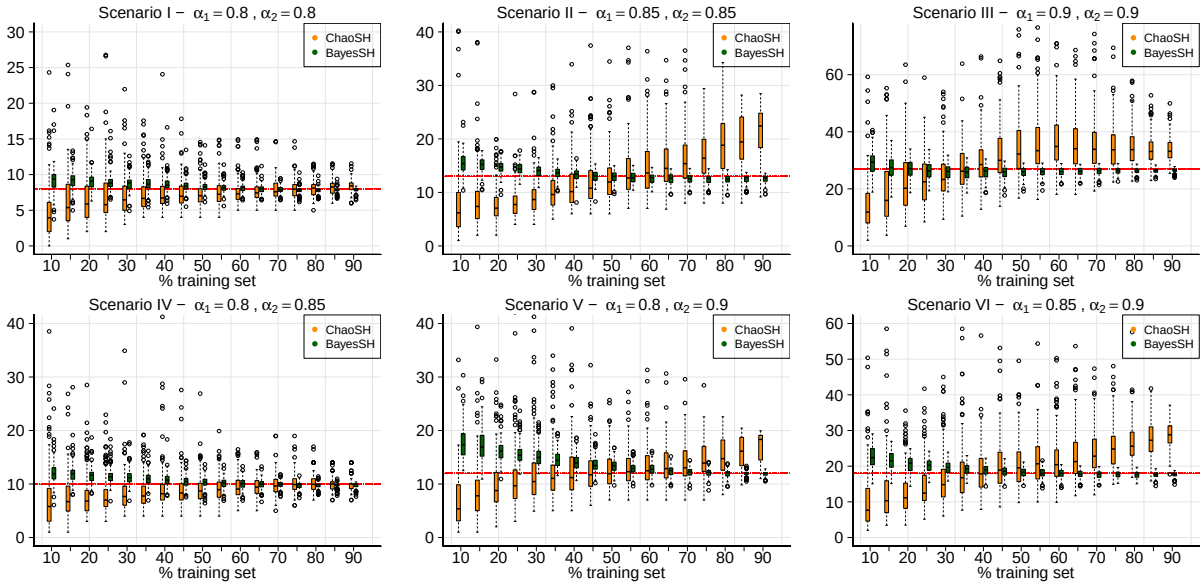

Figure 6.5: Predicted number of shared species for different training set percentages. The red line represents S_{true}.

two of these areas because of their ecological distinction: Bosco Bovedo (BB) and Orto Lapidario (OL). The former is a semi-natural urban forest located just outside residential areas, acting as a transition zone between urban and natural environments. The latter is a city park within a local museum complex. The two selected sites are sufficiently far from each other (about 6 kilometres) to rule out migratory contamination between the areas. Hence, since spatiotemporal effects are negligible, this experiment is well suited to be modelled using the framework introduced in Section 6.2. Ground-dwelling arthropods were sampled using pitfall traps. In particular, each site was sampled three times during the spring and summer months, with ten pitfall traps randomly placed in each area, maintaining a minimum distance of 20 meters between traps. Formicidae were identified at species level on the basis of their morphology and their abundance was recorded. Overall, a total number of $2,971$ and $3,489$ ants were taken in the first area (BB) and in the second area (OL), respectively. However, 2037 out of 2971 (around 68%) observations in the first group belong to the same species (crematogaster schmidti), which we excluded from the analysis. Similarly, in the second area (OL), 1229 out of 3489 (around 35%) also belong to the same species (pheidole pallidula), which we also excluded. This procedure is common in ecological studies, as highly abundant species are known to contribute little to understanding species diversity. Hence, the dataset analyzed consists of $n_1 = 934$ and $n_2 = 2235$ observations. The observed numbers of local distinct species are $r_1 = 17$ and $r_2 = 23$ while the global number of distinct species and the number of shared species are $r = 30$ and $t = 10$, respectively. The observed species proportions, sorted in decreasing order, are displayed in Figure 6.6.

For the dataset analysis, we follow the same approach used in the simulation study, see Section 6.6. Firstly, we compute the discovery probability in Equation (6.33) over a grid of sample sizes. Since this experiment is based on a real dataset, we cannot compare the results with any true curve. However, we compare the results obtained with our model with those obtained using the estimators by Yue and Clayton (2012) (YueDP) and Chao et al. (2017) (ChaoDP); see Equations (6.45) and (6.46). This comparison is displayed in the left panel of Figure 6.7. The analysis was repeated on 140 datasets obtained by randomly reordering the data. The uncertainty across the various replications is represented by vertical

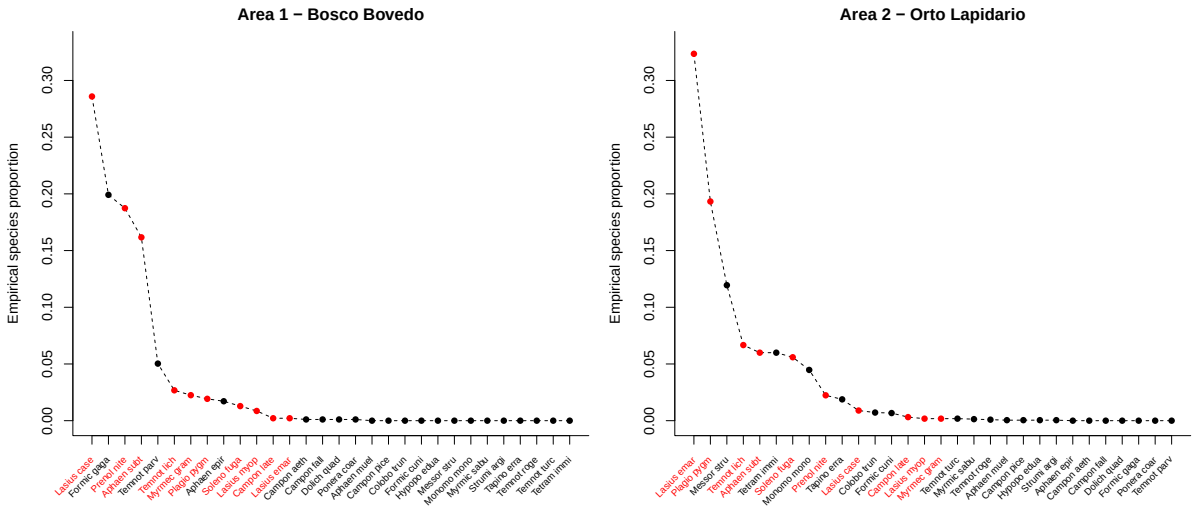

Figure 6.6: Graphical representations of observed species proportions in the dataset. Species have been sorted, within each group, in decreasing order. Red points and names represent shared species.

bars depicting the first and third quartiles of the estimates. We observe that our approach deviates from competing methods when the sample size is small but converges to the same values as the sample size increases. This initial difference aligns with the results observed in the simulation study and should be interpreted in light of the high variability that characterizes the results. Regarding this variability, the lower tails of the confidence intervals for both YueDP and ChaoDP always include zero, a behavior also identified in Section 6.6.2. On the other hand, our model does not suffer from this drawback, ensuring a discovery probability for new shared species that is strictly greater than zero.

We now illustrate the predictive capability of our model. We divide the observed dataset into training and test sets: the former is used as input to estimate the model parameters, which are then plugged into the equations presented in Section 6.2.1 to predict the number of shared species in the test set. The results are shown in the right panel of Figure 6.7, alongside those obtained from the estimator by Chao et al. (2000) (ChaoSH). We observe that the proposed estimator and ChaoSH converge to the same value. This pair of results suggests that the experiment was conducted exhaustively and that all shared species have been effectively observed. Indeed, evidence indicates that the sampling effort was even excessive relative to the number of shared species discovered. This is where a potential advantage of our analysis comes in. Indeed, if the sampling procedure had been informed by the discovery probability in Equation (6.33), in combination with an optimal stopping rule as presented in Yue and Clayton (2012), the experiment might have been concluded after gathering about half of the observations. This consideration applies only to the study of the number of shared species, which is the primary focus of this work but was not necessarily the ultimate goal of the original analysis of Zara et al. (2021).

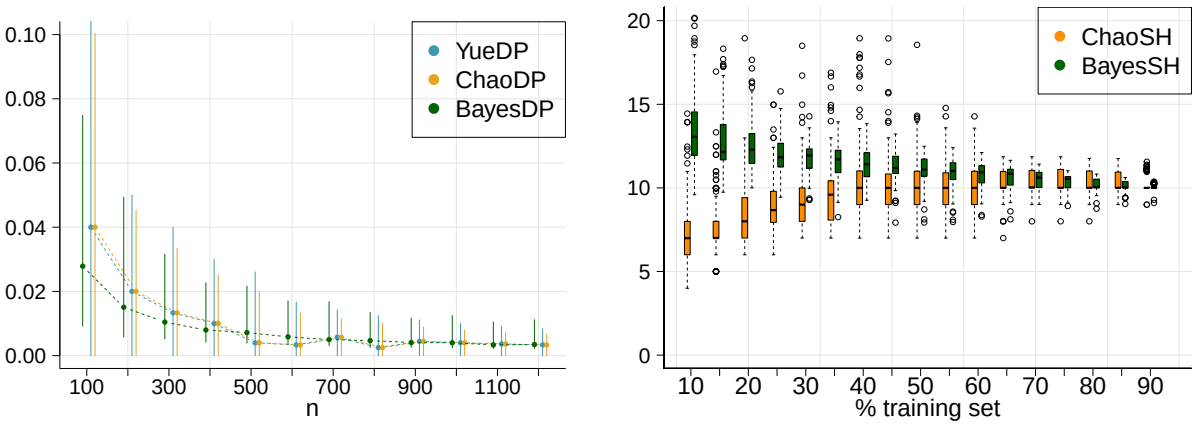

Figure 6.7: Left panel: estimated probability of discovering a new shared species, evaluated for different sample sizes. Right panel: predicted shared species for different training set percentages.

Appendix of Chapter 6

E.1 Review of generalized factorial coefficients

The results presented in Section 6.3 and Section 6.4 rely on both central and non-central generalized factorial coefficients. In this section, we provide some background about these combinatorial objects and report the most relevant formulae which are extensively used in subsequent sections. We refer to (Charalambides, 2002, Ch. 8) for a detailed discussion on this topic.

For any positive integers n and k, with $k \leq n$, the generalized factorial coefficient $C(n, k; \gamma)$ is the coefficient of the k-th order falling factorial of t in the expansion of the n-th order generalized factorial of t with scale parameter γ, namely

$$(\gamma t)_{n\downarrow} = \sum_{k=0}^{n} C(n, k; \gamma)(t)_{k\downarrow}; \tag{6.47}$$

Sometimes, we refer to $C(n, k; \gamma)$ as the central generalized factorial coefficient to distinguish it to its non-central generalisation, which is

$$(\gamma t - \rho)_{n\downarrow} = \sum_{k=0}^{n} C(n, k; \gamma, \rho)(t)_{k\downarrow}; \tag{6.48}$$

In particular, we have that $C(n, k; \gamma, 0) = C(n, k; \gamma)$. We wish to highlight the use of the falling factorial in Equations (6.47) and (6.48). If one were to replace it with the rising factorial, this would lead to a different definition of the generalized factorial coefficient, as used, for instance, in Lijoi et al. (2007a). We denote this alternative form as $\mathscr{C}(n, k; \gamma, \rho)$. The two definitions are connected by the identity $\mathscr{C}(n, k; \gamma, \rho) = (-1)^{n-k} C(n, k; \gamma, \rho)$.

An important formula which relates the central and the non-central generalized factorial coefficients is the following one

$$C(n, k; \gamma, \rho) = \sum_{j=k}^{n} \binom{n}{j} (\rho)_{(n-j)\downarrow} C(j, k; \gamma). \tag{6.49}$$

Moreover, let $(x)_n$ denote the n-th order rising factorial of x and recall that $(x)_{n\downarrow} = (-1)^n (-x)_n$. From Equation (6.49) we also derive the following generalisation of Equation (6.49) involving the absolute values of the generalized factorial coefficients,

$$|C(n, k; \gamma, \rho)| = \sum_{j=k}^{n} \binom{n}{j} (-\rho)_{n-j} |C(j, k; \gamma)|; \tag{6.50}$$

where in Equation (6.50) we also exploited the fact that $|C(n,k;\gamma,\rho)| = (-1)^n C(n,k;\gamma,\rho)$. Finally, for $\gamma > 0$, we also remind the following formula

$$|C(n,k;-\gamma)| = \frac{1}{k!}\sum_{(\star)}\binom{n}{n_1,\ldots,n_k}\prod_{l=1}^{k}(\gamma)_{r_l} \tag{6.51}$$

where the sum is taken over the following set

$$(\star) = \{(n_1,\ldots,n_k):\ n_l \geq 1\,,\ n_1+\cdots+n_k = n\}\,.$$

It will be also useful to remind an important generalisation of Vandermonde's identity:

$$\sum_{(\star\star)}\binom{n}{n_1,\ldots,n_k}\prod_{l=1}^{k}(\gamma_l)_{r_l} = (\gamma_1+\ldots+\gamma_k)_n \tag{6.52}$$

where the sum is taken over the following set

$$(\star\star) = \{(n_1,\ldots,n_k):\ n_l \geq 0,\ n_1+\cdots+n_k = n\}$$

and $\gamma_l > 0$ for $l = 1,\ldots,k$.

E.2 Proofs of results in Section 6.2.1

E.2.1 Proof of Equation (6.5)

Colombi et al. (2024a) derived the following equivalent form for the pEPPF,

$$\begin{aligned}&\Pi_r^{(n)}(\boldsymbol{n}_1,\boldsymbol{n}_2)\\ &= \int_{[0,\infty]\times[0,\infty]}\Psi(r,u_1,u_2)\prod_{j=1}^{2}\frac{u_j^{n_j-1}}{\Gamma(n_j)\,(1+u_j)^{n_j+\gamma_j r}}\mathrm{d}u_1\mathrm{d}u_2\times\prod_{j=1}^{2}\prod_{l=1}^{r}(\gamma_j)_{n_{j,l}}\,,\end{aligned}$$

and, as we are only interested in the case of symmetric Dirichlet distributed random weights, $\Psi(r,u_1,u_2)$ takes the following form

$$\Psi(r,u_1,u_2) \;=\; \sum_{m^\star=0}^{\infty}\frac{(m^\star+r)!}{m^\star!}q_M(m^\star+r)\prod_{j=1}^{2}\left(1+u_j\right)^{\gamma_j m^\star}\,.$$

In summary, we want to exchange the integral and the infinite sum and solve the remaining integrals with respect to u_1 and u_2. By doing do, we have that

$$\begin{aligned}&\Pi_r^{(n)}(\boldsymbol{n}_1,\boldsymbol{n}_2)\\ &= \prod_{j=1}^{2}\prod_{l=1}^{r}(\gamma_j)_{n_{j,l}}\sum_{m^\star=0}^{\infty}\left\{\frac{(m^\star+r)!}{m^\star!}q_M(m^\star+r)\prod_{j=1}^{2}\int_0^\infty\frac{1}{\Gamma(n_j)}\frac{u_j^{n_j-1}}{(1+u_j)^{n_j+\gamma_j(r+m^\star)}}\mathrm{d}u_j\right\}\,.\end{aligned}$$

The integral equals a Beta function (Abramowitz and Stegun, 1964, p.258), hence we have that

$$\begin{aligned}
&\Pi_r^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) \\
&= \sum_{m^\star=0}^{\infty} \left\{ \frac{(m^\star + r)!}{m^\star!} q_M(m^\star + r) \prod_{j=1}^{2} \frac{B(n_j, \gamma_j(m^\star + r))}{\Gamma(n_j)} \right\} \prod_{j=1}^{2} \prod_{l=1}^{r} (\gamma_j)_{n_{j,l}} \\
&= \sum_{m^\star=0}^{\infty} \left\{ \frac{(m^\star + r)!}{m^\star!} q_M(m^\star + r) \prod_{j=1}^{2} \frac{1}{(\gamma_j(m^\star + r))_{n_j}} \right\} \prod_{j=1}^{2} \prod_{l=1}^{r} (\gamma_j)_{n_{j,l}} .
\end{aligned} \tag{6.53}$$

To complete the proof, it is enough to change variables in the infinite sum and define $V^r_{n_1,n_2}$ as

$$V^r_{n_1,n_2} = \sum_{m^\star=0}^{\infty} (m^\star)_{r\downarrow} q_M(m^\star) \prod_{j=1}^{2} \frac{1}{(\gamma_j(m^\star))_{n_j}} .$$

The latter coincides with the definition given in Equation (6.6).

E.3 Proofs of the results in Section 6.2.2

E.3.1 Proof of Proposition 1

We start proving the first statement about the convergence of the infinite sum. Firstly, we use the change of variables $m^\star = m - r$ and we rewrite $V^r_{n_1,n_2}$ as

$$V^r_{n_1,n_2} = \sum_{m^\star=0}^{\infty} \frac{(m + r)!}{m!} \prod_{j=1}^{2} \frac{1}{(\gamma_j(m^\star + r))_{n_j}} q_M(m^\star + r) . \tag{6.54}$$

Then, to prove the statement we consider the series of the asymptotic expansion of the general term. To do so, we use the following Stirling approximation for large values of $m^\star$:

$$\frac{(m^\star + r)!}{m^\star!} \sim e^{-r} \sqrt{\frac{m^\star + r}{m^\star}} \left(\frac{m^\star + r}{m^\star} \right)^{m^\star} (m^\star + r)^r , \tag{6.55}$$

where we use the notation $f(x) \sim g(x)$ as a shortform for $f(x) = o_{x_0}(g(x))$ for $x \to x_0$. We also recall the following approximation for the ratio of Gamma function $\Gamma(a + cm)/\Gamma(b + cm) \sim (cm)^{a-b}$, when $m \to \infty$. Hence, we have that

$$\frac{\Gamma(\gamma_j m^\star + \gamma_j r)}{\Gamma(\gamma_j m^\star + \gamma_j r + n_j)} \sim (\gamma_j m^\star)^{-n_j} . \tag{6.56}$$

Using Equations (6.55) and (6.56), the following is the asymptotic expression of the general term in Equation (6.54) for large values of $m^\star$

$$\begin{aligned}
&\frac{(m^\star+r)!}{m^\star!}\prod_{j=1}^{2}\frac{1}{(\gamma_j(m^\star+r))_{n_j}}q_M(m^\star+r)\\
&\sim e^{-r}\sqrt{\frac{m^\star+r}{m^\star}}\left(\frac{m^\star+r}{m^\star}\right)^{m^\star}(m^\star+r)^r(\gamma_1)^{-n_1}(\gamma_2)^{-n_2}(m^\star)^{-n}q_M(m^\star+r)\\
&\sim(\gamma_1)^{-n_1}(\gamma_2)^{-n_2}\left(\frac{m^\star+r}{m^\star}\right)^{r}(m^\star)^{-n+r}q_M(m^\star+r)\,,\\
&\sim(\gamma_1)^{-n_1}(\gamma_2)^{-n_2}\frac{1}{(m^\star)^{n-r}}q_M(m^\star+r)\,,
\end{aligned}$$

where we defined $n = n_1 + n_2$, we used $\left(\frac{m^\star+r}{m^\star}\right)^{m^\star} \sim e^r$ and we wrote $(m^\star)^{-n}$ as $(m^\star)^{-(n-r)-r}$. As a consequence, the convergence of $V^r_{n_1,n_2}$ can be assessed studying the convergence of the following series,

$$\sum_{m^\star=0}^{\infty}(\gamma_1)^{-n_1}(\gamma_2)^{-n_2}\frac{1}{(m^\star)^{n-r}}q_M(m^\star+r)\leq(\gamma_1)^{-n_1}(\gamma_2)^{-n_2}\sum_{m^\star=0}^{\infty}q_M(m^\star+r)<\infty\,.$$

The latter follows since $r \leq n$ implies that $\frac{1}{(m^\star)^{n-r}} \leq 1$. Additionally, being q_M a probability mass function, the final sum is less or equal to one.

We now continue proving the statement about the asymptotic expansion of $V^r_{n_1,n_2}$. For sake of notation simplicity, we write $V^r_{n_1,n_2}$ as

$$V^r_{n_1,n_2} = \left\{V^r_{n_1,n_2}\right\}_r + \sum_{m=r+1}^{\infty}\left\{V^r_{n_1,n_2}\right\}_m$$

where $\left\{V^r_{n_1,n_2}\right\}_m$ is the m-th term of the series. Namely,

$$\left\{V^r_{n_1,n_2}\right\}_m = (m)_{r\downarrow}\prod_{j=1}^{2}\left((\gamma_j m)_{n_j}\right)^{-1}q_M(m)\,,$$

for each integer $m \geq r$. Since $q_M(r) > 0$, we can collect the first term and we get

$$V^r_{n_1,n_2} = \left\{V^r_{n_1,n_2}\right\}_r\left[1+\frac{1}{\left\{V^r_{n_1,n_2}\right\}_r}\sum_{m=r+1}^{\infty}\left\{V^r_{n_1,n_2}\right\}_m\right].$$

Following the same steps of the proof in Miller and Harrison (2018), we show that the second term in the squared brackets goes to zero. Once again, we first isolate the $\left\{V^r_{n_1,n_2}\right\}_{r+1}$ term for the infinite sum.

$$\begin{aligned}
\frac{1}{\left\{V^r_{n_1,n_2}\right\}_r}\sum_{m=r+1}^{\infty}\left\{V^r_{n_1,n_2}\right\}_m &= \frac{\left\{V^r_{n_1,n_2}\right\}_{r+1}}{\left\{V^r_{n_1,n_2}\right\}_r}+\frac{1}{\left\{V^r_{n_1,n_2}\right\}_r}\sum_{m=r+2}^{\infty}\left\{V^r_{n_1,n_2}\right\}_m\\
&= n_1^{-\gamma_1}n_2^{-\gamma_2}(r+1)(\gamma_1 r)_{\gamma_1}(\gamma_2 r)_{\gamma_2}\frac{q_M(r+1)}{q_M(r)}+\frac{1}{\left\{V^r_{n_1,n_2}\right\}_r}\sum_{m=r+2}^{\infty}\left\{V^r_{n_1,n_2}\right\}_m\,.
\end{aligned}\tag{6.57}$$

The second term in the final line of Equation (6.57) converges to 0 as $n_1, n_2 \to 0$. This follows using

(Miller and Harrison, 2018, Proposition S1.1). Moreover, it can be written as an infinite polynomial with respect to $(n_1)^{-\gamma_1}(n_2)^{-\gamma_2}$ for some polynomial coefficients depending on m. Namely

$$\begin{aligned}\frac{1}{\left\{V^r_{n_1,n_2}\right\}_r}\sum_{m=r+2}^{\infty}\left\{V^r_{n_1,n_2}\right\}_m \\ = \sum_{m=r+2}^{\infty} C_m\left((n_1)^{-\gamma_1}\right)^{m-r}\left((n_2)^{-\gamma_2}\right)^{m-r} = o((n_1)^{-\gamma_1}(n_2)^{-\gamma_2})\,.\end{aligned} \tag{6.58}$$

The statement follows combining Equations (6.57) and (6.58).

E.3.2 Proof of Proposition 2

To prove Equation (6.12) we exploit the identity

$$(m)_{(r+1)\downarrow} = \frac{\gamma_2 m + n_2}{\gamma_2}(m)_{r\downarrow} - (r + \frac{n_2}{\gamma_2})(m)_{(r)\downarrow}\,.$$

Then, we have that

$$\begin{aligned}V^{r+1}_{n_1,n_2+1} = \sum_{m=1}^{\infty}(m)_{r\downarrow}\left(\frac{1}{\gamma_2}\frac{(\gamma_2 m + n_2)}{(\gamma_2 m)_{(n_2+1)}}\right)\frac{1}{(\gamma_1 m)_{(n_1)}}q_M(m) \\ - (r + \frac{n_2}{\gamma_2})\sum_{m=1}^{\infty}(m)_{r\downarrow}\left(\frac{1}{(\gamma_2 m)_{(n_2+1)}}\right)\frac{1}{(\gamma_1 m)_{(n_1)}}q_M(m)\,.\end{aligned}$$

The statement follows after recognizing the Pochhammer symbol in the first term of Equation (6.12). Equation (6.13) can be proven analogously.

Firstly, we rewrite $(m)_{(r+2)\downarrow}$ in a convenient way.

$$(m)_{(r+2)\downarrow} = (m-r-1)(m-r)(m)_{r\downarrow} = \left\{m^2 - m(2r+1) + r(r+1)\right\}(m)_{r\downarrow}\,.$$

Then, we exploit the following identity

$$m^2 = \frac{(\gamma_1 m + n_1) - n_1}{\gamma_1}\frac{(\gamma_2 m + n_2) - n_2}{\gamma_2}\,,$$

to write

$$\begin{aligned}(m)_{(r+2)\downarrow} = \left\{\frac{(\gamma_1 m + n_1) - n_1}{\gamma_1}\frac{(\gamma_2 m + n_2) - n_2}{\gamma_2} - \frac{n_1(\gamma_2 m + n_2)}{\gamma_1\gamma_2} - \frac{n_2(\gamma_1 m + n_1)}{\gamma_1\gamma_2} + \right. \\ \left. + \frac{n_1 n_2}{\gamma_1\gamma_2} + r(r+1) - m(2r+1)\right\}(m)_{r\downarrow}\,.\end{aligned} \tag{6.59}$$

We exploit Equation (6.59) to have

$$
\begin{aligned}
V^{r+2}_{n_1+1,n_2+1} &= \sum_{m=0}^{\infty} (m)_{(r+2)\downarrow} \frac{1}{(\gamma_1 m)_{n_1+1}(\gamma_2 m)_{n_2+1}} q_M(m) \\
&= \sum_{m=0}^{\infty} \left\{ \frac{(\gamma_1 m + n_1) - n_1}{\gamma_1} \frac{(\gamma_2 m + n_2) - n_2}{\gamma_2} - \frac{n_1(\gamma_2 m + n_2)}{\gamma_1\gamma_2} - \frac{n_2(\gamma_1 m + n_1)}{\gamma_1\gamma_2} + \frac{n_1 n_2}{\gamma_1\gamma_2} \right. \\
&\qquad \left. + r(r+1) - m(2r+1) \right\} \frac{(m)_{r\downarrow}}{(\gamma_1 m)_{n_1}(\gamma_2 m)_{n_2}(\gamma_1 m + n_1)(\gamma_1 m + n_2)} q_M(m)
\end{aligned}
$$

Using the definition of V coefficients in Equation (6.6), we split the sum and recognize some terms.

$$
\begin{aligned}
V^{r+2}_{n_1+1,n_2+1} &= \frac{1}{\gamma_1\gamma_2} V^r_{n_1,n_2} - \frac{n_1}{\gamma_1\gamma_2} V^r_{n_1+1,n_2} - \frac{n_2}{\gamma_1\gamma_2} V^r_{n_1,n_2+1} + \left\{ \frac{n_1 n_2}{\gamma_1\gamma_2} + r(r+1) \right\} V^r_{n_1+1,n_2+1} \\
&\quad - (2r+1) \sum_{m=0}^{\infty} m(m)_{r\downarrow} \frac{1}{(\gamma_1 m)_{n_1+1}(\gamma_2 m)_{n_2+1}} q_M(m).
\end{aligned} \tag{6.60}
$$

Using $m(m)_{r\downarrow} = r(m)_{r\downarrow} + (m)_{(r+1)\downarrow}$ and following the same steps as before, it is easy to show that the sum in the final line of Equation (6.60) equals $rV^r_{n_1+1,n_2+1} + V^{r+1}_{n_1+1,n_2+1}$. Hence, we have

$$
\begin{aligned}
V^{r+2}_{n_1+1,n_2+1} &= \frac{1}{\gamma_1\gamma_2} \left\{ V^r_{n_1,n_2} - n_1 V^r_{n_1+1,n_2} - n_2 V^r_{n_1,n_2+1} + (n_1 n_2 + \gamma_1\gamma_2 r(r+1)) \right\} \\
&\quad - (2r+1) \left\{ rV^r_{n_1+1,n_2+1} + V^{r+1}_{n_1+1,n_2+1} \right\} .
\end{aligned}
$$

The statement follows after some trivial linear algebra.

E.4 Proof of Theorem 6.3.1

Firstly, we notice from Section 6.3 that K_{n_1,n_2}, K_{1,n_1} and K_{2,n_2} are linearly independent quantities from which we can also derive the remaining ones, i.e., $S_{n_1,n_2} = t$ and $K^*_{j,n}$, for $j = 1, 2$. From Equations (6.17), it follows that

$$
t = r_1 + r_2 - r, \quad r^*_1 = r - r_2, \quad r^*_2 = r - r_1. \tag{6.61}
$$

Moreover, the following conditions must hold: $1 \le r \le n_1 + n_2$ and $1 \le r_j \le n_j$, for $j = 1, 2$. The probability of interest may be evaluated as follows

$$
\mathbb{P}\left(K_{n_1,n_2} = r, K_{1,n_1} = r_1, K_{2,n_2} = r_2\right) = \sum_{(\star)} \frac{1}{r!} \prod_{j=1}^{2} \binom{n_j}{n_{j,1}, \ldots, n_{j,r}} \cdot \Pi^{(n)}_r(\boldsymbol{n}_1, \boldsymbol{n}_2)
$$

where $n = n_1 + n_2$ and the sum $(\star)$ is extended over all the vectors $(n_{1,1}, \ldots, n_{1,r})$ and $(n_{2,1}, \ldots, n_{2,r})$ of non-negative integers satisfying the following constraints

$$
\sum_{l=1}^{r} n_{j,l} = n_j \text{ with } n_{j,l} \ge 0, \quad l = 1, \ldots, r \; j = 1, 2,
$$

$$
n_{1,l} + n_{2,l} \ge 1 \quad l = 1, \ldots, r, \; \sum_{l=1}^{r} \delta_{\{n_{j,l}>0\}} = r_j, \; j = 1, 2.
$$

By exploiting the expression of the pEPPF in Equation (6.5), we get

$$\begin{aligned}&\mathbb{P}\left(K_{n_1,n_2}=r, K_{1,n_1}=r_1, K_{2,n_2}=r_2\right)\\&\quad= V^r_{n_1,n_2}\sum_{(\star)}\frac{1}{r!}\prod_{j=1}^{2}\left\{\binom{n_j}{n_{j,1},\ldots,n_{j,r}}\prod_{l=1}^{r}(\gamma_j)_{n_{j,l}}\right\}.\end{aligned}$$

In the following, we aim to solve the sum over the set $(\star)$. The main difficulty here is the joint condition $n_{1,j}+n_{2,l}\geq 1$, therefore we elaborate the sum trying to decouple it in two sums, only involving the local cardinalities $n_{j,l}$. To do so, t out of r species must be shared. Without loss of generality, assume that the first t species are these shared species. Moreover, we fix an ordering for them, noticing that this operation can be done in $\binom{r}{t}$ equivalent ways. Hence, we have

$$\begin{aligned}&\sum_{(\star)}\frac{1}{r!}\prod_{j=1}^{2}\left\{\binom{n_j}{n_{j,1},\ldots,n_{j,r}}\prod_{l=1}^{r}(\gamma_j)_{n_{j,l}}\right\}\\&\quad=\sum_{(\star\star)}\frac{1}{r!}\binom{r}{t}\prod_{j=1}^{2}\left\{\binom{n_j}{n_{j,1},\ldots,n_{j,r}}\prod_{l=1}^{r}(\gamma_j)_{n_{j,l}}\right\},\end{aligned}\tag{6.62}$$

where the set $(\star\star)$ must satisfy the following constraints,

$$\begin{aligned}&\sum_{l=1}^{r}n_{j,l}=n_j\ j=1,2,\\&n_{j,l}\geq 1,\quad l=1,\ldots,t,\quad n_{j,l}\geq 0,\quad j=1,2;\ l=t+1,\ldots,r,\\&n_{1,l}+n_{2,l}\geq 1\quad n_{1,l}\cdot n_{2,l}=0\quad l=t+1,\ldots,r,\\&\sum_{l=1}^{r}\delta_{\{n_{j,l}>0\}}=r_j,\quad j=1,2.\end{aligned}$$

Equation (6.62) is properly defined as long as $0\leq t\leq r$, which implies $r\leq r_1+r_2\leq 2r$.

When moving from set $(\star)$ to set $(\star\star)$, the joint condition only refers to the final $r-t$ species. We can further reorder such species, which are not shared among the two groups. Indeed, we know that r_1^* species have been observed in group 1 only while r_2^* species are specific to group 2 only. Hence, we assume the r_1^* species are in positions from $t+1$ to $t+r_1^*$ and we fix and ordering in any of the $\binom{r-t}{r^*-1}$ possible ways. The remaining species are $r-t-r_1^*$ which, from Equations (6.61), can be shown to equal r_2^*. Note that, from Equations (E.4), $n_{1,l}\geq 1$ implies $n_{2,l}=0$ for $l=t+1,\ldots,r+r_1^*$ while $n_{2,l}\geq 1$ implies $n_{1,l}=0$ for $l=t+r_1^*+1,\ldots,r$. This fully resolves the joint condition in set $(\star\star)$. Moreover, for each $j=1,2$, the number of non-zero elements in vectors $\boldsymbol{n}_j$ is $t+r_j^*$ and, from Equations (6.61), this equals r_j. This guarantees one of the conditions in Equation (E.4). As a consequence, we can discard

zero elements in vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ and we obtain

$$\begin{aligned}\sum_{(\star)} \frac{1}{r!} \prod_{j=1}^{2} \left\{ \binom{n_j}{n_{j,1}, \dots, n_{j,r}} \prod_{l=1}^{r} (\gamma_j)_{n_{j,l}} \right\} \\ = \frac{1}{r!} \binom{r}{t} \binom{r-t}{r_1^*} \prod_{j=1}^{2} \left\{ \sum_{(\star\star j)} \binom{n_j}{n_{j,1}, \dots, n_{j,r_j}} \prod_{l=1}^{r_j} (\gamma_j)_{n_{j,l}} \right\},\end{aligned} \tag{6.63}$$

where the sum over the sets $(\star\star j)$, for $j = 1, 2$, is extended over all vectors $(n_{j,1}, \dots, n_{j,r_j})$ such that $n_{j,l} \geq 1$ and $\sum_{l=1}^{r_j} n_{j,l} = n_j$. Equation (6.63) is properly defined as long as $0 \leq r_1^* \leq r - t$, which implies $r_1 \leq r$ and $r_2 \leq r$. In particular, this also ensures that $r_1 + r_2 \leq 2r$.

Finally, we use Equation (6.51) to solve final sums over the sets $(\star\star j)$ and we conclude that

$$\begin{aligned}&\mathbb{P}\left(K_{n_1,n_2} = r, K_{1,n_1} = r_1, K_{2,n_2} = r_2\right) \\ &\quad = V^r_{n_1,n_2} \frac{1}{r!} \binom{r}{t} \binom{r-t}{r_1^*} \prod_{j=1}^{2} r_j! |C(n_j, r_j; -\gamma_j)|.\end{aligned}$$

The statement follows plugging the values of t and r_1^* in terms of r, r_1 and r_2 reported in Equations (6.61) and rearranging the factorials and the binomial coefficients.

E.5 Proofs of the results in Section 6.4

E.5.1 Proof of Equation (6.24)

The proof of Equation (6.24) easily follows from Section E.2.1. Indeed, we can look at Equation (6.53) as a posterior marginal distribution with respect to M. It follows that,

$$\begin{aligned}q_M(m^\star \mid \boldsymbol{X}) &\propto \frac{(m^\star + r)!}{m^\star!} q_M(m^\star + r) \prod_{j=1}^{2} \frac{1}{\left(\gamma_j(m^\star + r)\right)_{n_j}} \prod_{j=1}^{2} \prod_{l=1}^{r} (\gamma_j)_{n_{j,l}} \\ &\propto \frac{(m^\star + r)!}{m^\star!} q_M(m^\star + r) \prod_{j=1}^{2} \frac{1}{\left(\gamma_j(m^\star + r)\right)_{n_j}}.\end{aligned} \tag{6.64}$$

Then, the normalising constant for the previous expression is

$$\sum_{m^\star=0}^{\infty} \frac{(m^\star + r)!}{m^\star!} q_M(m^\star + r) \prod_{j=1}^{2} \frac{1}{\left(\gamma_j(m^\star + r)\right)_{n_j}}, \tag{6.65}$$

which is, up to a change of variables, the $V^r_{n_1,n_2}$ coefficient defined in Equation (6.6).

E.5.2 Proof of Equation (6.25)

In the following, we compute the expected value the $M^\star$ whose probability mass function is $q^\star_{M|\boldsymbol{X}}$, defined in Equation (6.24), i.e.,

$$\begin{aligned}
E_{q^\star_{M|\boldsymbol{X}}}\left(M^\star\right) &= \frac{1}{V^r_{n_1,n_2}} \sum_{m^\star=1}^{\infty} m^\star (m^\star + r)_{r\downarrow} q_M(m^\star + r) \prod_{j=1}^{d} \frac{1}{\left(\gamma_j\,(m^\star + r)\right)_{n_j}} \\
&= \frac{1}{V^r_{n_1,n_2}} \sum_{m^{\star\star}=0}^{\infty} (m^{\star\star} + 1)(m^{\star\star} + r + 1)_{r\downarrow} q_M(m^{\star\star} + r + 1) \prod_{j=1}^{d} \frac{1}{\left(\gamma_j\,(m^{\star\star} + r + 1)\right)_{n_j}} \\
&= \frac{1}{V^r_{n_1,n_2}} \sum_{m^{\star\star}=0}^{\infty} (m^{\star\star} + r + 1)_{(r+1)\downarrow} q_M(m^{\star\star} + r + 1) \prod_{j=1}^{d} \frac{1}{\left(\gamma_j\,(m^{\star\star} + r + 1)\right)_{n_j}},
\end{aligned}$$

where we first applied the change of index $m^{\star\star} = m^\star + 1$ and then we used the following identity: $(m^{\star\star} + 1)(m^{\star\star} + r + 1)_{r\downarrow} = (m^{\star\star} + r + 1)_{(r+1)\downarrow}$. Then, we note that $(m^{\star\star} + r + 1)_{(r+1)\downarrow} = 0$ for $m^{\star\star} \leq r$. Hence, we change variables once again, setting $\bar{m} = m^{\star\star} + r + 1$. By doing so, we obtain

$$E_{q^\star_{M|\boldsymbol{X}}}\left(M^\star\right) = \frac{1}{V^r_{n_1,n_2}} \sum_{\bar{m}=r+1}^{\infty} (\bar{m})_{(r+1)\downarrow} q_M(\bar{m}) \prod_{j=1}^{d} \frac{1}{\left(\gamma_j\,(\bar{m})\right)_{n_j}} = \frac{V^{r+1}_{n_1,n_2}}{V^r_{n_1,n_2}},$$

where the final equivalence follows by definition, see Equation (6.6).

E.5.3 Proof of Equation (6.26)

The expected value in Equation (6.25) is the ratio of two V coefficients. Using the asymptotic expansion given in Equation (6.11), we have that

$$\begin{aligned}
E\left(M^\star \mid \boldsymbol{X}\right) &= \frac{V^{r+1}_{n_1,n_2}}{V^r_{n_1,n_2}} \\
&\sim (r+1) \frac{(\gamma_1(r+1))_{n_1} (\gamma_2(r+1))_{n_2}}{(\gamma_1 r)_{n_1} (\gamma_2 r)_{n_2}} \frac{q_M(r+1)}{q_M(r)} \\
&\quad \times \frac{\left\{1 + n_1^{-\gamma_1} n_2^{-\gamma_2} (r+2)(\gamma_1(r+1))_{\gamma_1} (\gamma_2(r+1))_{\gamma_2} \frac{q_M(r+2)}{q_M(r+1)} + o\left(n_1^{-\gamma_1} n_2^{-\gamma_2}\right)\right\}}{\left\{1 + n_1^{-\gamma_1} n_2^{-\gamma_2} (r+1)(\gamma_1 r)_{\gamma_1} (\gamma_2 r)_{\gamma_2} \frac{q_M(r+1)}{q_M(r)} + o\left(n_1^{-\gamma_1} n_2^{-\gamma_2}\right)\right\}}.
\end{aligned} \tag{6.66}$$

To further expand the second term, let us define $C_r = (r+1)(\gamma_1 r)_{\gamma_1} (\gamma_2 r)_{\gamma_2} q_M(r+1)/q_M(r)$ and $n_j^\star = (n_j)^{\gamma_j}$. Hence, we have that

$$\begin{aligned}
&\frac{\left(1 + C_{r+1}(n_1^\star)^{-1}(n_2^\star)^{-1} + o\left((n_1^\star)^{-1}(n_2^\star)^{-1}\right)\right)}{\left(1 + C_r(n_1^\star)^{-1}(n_2^\star)^{-1} + o\left((n_1^\star)^{-1}(n_2^\star)^{-1}\right)\right)} \\
&\quad = \left(1 + C_{r+1}(n_1^\star)^{-1}(n_2^\star)^{-1} + o\left((n_1^\star)^{-1}(n_2^\star)^{-1}\right)\right)\left(1 - C_r(n_1^\star)^{-1}(n_)^\star 2^{-1} + o\left((n_1^\star)^{-1}(n_2^\star)^{-1}\right)\right) \\
&\quad = 1 + (C_{r+1} - C_r)(n_1^\star)^{-1}(n_2^\star)^{-1} + o\left((n_1^\star)^{-1}(n_2^\star)^{-1}\right).
\end{aligned}$$

Plugging the latter expansion into (6.66), we have that,

$$\begin{aligned}
E\left(M^{\star} \mid \boldsymbol{X}\right) &= \frac{V^{r+1}_{n_1,n_2}}{V^{r}_{n_1,n_2}} \\
&\sim (r+1)\frac{q_M(r+1)}{q_M(r)}\prod_{j=1}^{2}(\gamma_j r)_{\gamma_j}\frac{\Gamma(\gamma_j r+n_j)}{\Gamma(\gamma_j r+\gamma_j+n_j)} \\
&\quad\times\left\{1+(C_{r+1}-C_r)(n_1^{\star})^{-1}(n_2^{\star})^{-1}+o\left((n_1^{\star})^{-1}(n_2^{\star})^{-1}\right)\right\} \\
&= (r+1)\frac{q_M(r+1)}{q_M(r)}(\gamma_1 r)_{\gamma_1}(\gamma_2 r)_{\gamma_2}(n_1^{\star})^{-1}(n_2^{\star})^{-1}o\left((n_1^{\star})^{-1}(n_2^{\star})^{-1}\right) \\
&\quad\times\left\{1+(C_{r+1}-C_r)(n_1^{\star})^{-1}(n_2^{\star})^{-1}+o\left((n_1^{\star})^{-1}(n_2^{\star})^{-1}\right)\right\} \\
&= (r+1)\frac{q_M(r+1)}{q_M(r)}(\gamma_1 r)_{\gamma_1}(\gamma_2 r)_{\gamma_2}(n_1^{\star})^{-1}(n_2^{\star})^{-1}\left(1+o\left((n_1^{\star})^{-1}(n_2^{\star})^{-1}\right)\right).
\end{aligned} \tag{6.67}$$

E.5.4 Proof of Theorem 6.4.1

Firstly, we notice from Section 6.4 that the number of linearly independent quantities we need to fully characterize all the posterior sample is five. It follows that $K^{(n_1,n_2)}_{m_1,m_2}, K^{(n_1)}_{1,m_1}, K^{(n_2)}_{2,m_2}$ are not enough and this would require to introduce two more random variables, that would be marginalized out. Of course, we must choose them so that the five selected quantities form a system of linearly independent variables. To achieve this, we choose $S^{*}_{m} = s^{*}$ and $K^{*(n)}_{1,m} = k^{*}_1$. Now, from Equations (6.22), it follows that

$$k^{*}_2 = k - k^{*}_1 - s^{*},\ s = k_1 + k_2 - k,\ s_{2,1} = k_1 - k^{*}_1 - s^{*},\ s_{1,2} = k_2 + k^{*}_1 - k \tag{6.68}$$

For what concerns the support of the variables, it is natural to ask for $0 \le k_j \le m_j$, for $j = 1, 2$. Furthermore, from Equation (6.68), we see that $k \le k_1 + k_2$, which is a more stringent condition with respect to $k \le m_1 + m_2$. Hence, we also ask for $0 \le k \le k_1 + k_2$.

Let (π_1, π_2) denotes the partition of the additional observations $\{(X_{n_j+1}, \ldots, X_{n_j+m_j}) : \ j = 1, 2\}$ into $r+k$ sets of distinct values, of which r coincide with already observed values in the initial sample and the remaining k are new. We also indicate by $\boldsymbol{m}_j(\pi_j) = (m_{j,1}(\pi_j), \ldots, m_{j,K+r}(\pi_j))$ the corresponding frequency counts, as $j = 1, 2$. Finally, let $\mathscr{P}_{m_1,m_2,r+k}$ be the space of all such possible partitions, so that $(\pi_1, \pi_2) \in \mathscr{P}_{m_1,m_2,r+k}$. Moreover, let $n = n_1 + n_2$ and $m = m_1 + m_2$. The posterior probability of interest can be evaluated as follows,

$$\begin{aligned}
&\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2} = k,\ K^{(n_1)}_{1,m_1} = k_1,\ K^{(n_2)}_{2,m_2} = k_2 \mid \boldsymbol{X}\right) \\
&\qquad = \sum_{\mathscr{P}_{m_1,m_2,r+k}} \frac{\Pi^{(n+m)}_{r+k}(\boldsymbol{n}_1 + \boldsymbol{m}_1(\pi_1), \boldsymbol{n}_2 + \boldsymbol{m}_2(\pi_2))}{\Pi^{(n)}_{r}(\boldsymbol{n}_1, \boldsymbol{n}_2)},
\end{aligned} \tag{6.69}$$

where the sum is extended for all $(\pi_1, \pi_2) \in \mathscr{P}_{m_1,m_2,r+k}$.

We now elaborate the numerator in Equation (6.69). Writing explicitly all terms in Equations (6.5)

and (6.6) we have that

$$
\begin{aligned}
&\Pi_{r+K}^{(n+m)}(\boldsymbol{n}_1+\boldsymbol{m}_1(\pi_1),\boldsymbol{n}_2+\boldsymbol{m}_2(\pi_2)) \\
&=\sum_{m^\star=r+k}^{\infty}(m^\star)_{(r+k)\downarrow}q_M(m^\star)\prod_{j=1}^{2}\left\{\frac{1}{(\gamma_j(m^\star))_{n_j+m_j}}\prod_{l=1}^{k+r}(\gamma_j)_{n_{j,l}+m_{j,l}(\pi_j)}\right\} \\
&=\sum_{\overline{m}=k}^{\infty}(\overline{m})_{k\downarrow}\frac{(\overline{m}+r)!}{\overline{m}!}q_M(\overline{m}+r)\prod_{j=1}^{2}\left\{\frac{1}{(\gamma_j(\overline{m}+r))_{n_j+m_j}}\prod_{l=1}^{k+r}(\gamma_j)_{n_{j,l}+m_{j,l}(\pi_j)}\right\} \\
&=\sum_{\overline{m}=k}^{\infty}\left\{(\overline{m})_{k\downarrow}\frac{(\overline{m}+r)!}{\overline{m}!}q_M(\overline{m}+r)\prod_{j=1}^{2}\frac{1}{\left(\gamma_j(r+\overline{m})\right)_{n_j}}\prod_{j=1}^{2}\prod_{l=1}^{r}(\gamma_j)_{n_{j,l}}\right. \\
&\quad\left.\times\prod_{j=1}^{2}\prod_{l=1}^{r}\frac{(\gamma_j)_{n_{j,l}+m_{j,l}(\pi_j)}}{(\gamma_j)_{n_{j,l}}}\prod_{j=1}^{2}\prod_{l=r+1}^{r+k}(\gamma_j)_{m_{j,l}(\pi_j)}\prod_{j=1}^{2}\frac{\left(\gamma_j(r+\overline{m})\right)_{n_j}}{\left(\gamma_j(r+\overline{m})\right)_{n_j+m_j}}\right\}
\end{aligned}
\tag{6.70}
$$

The second equality in Equation (6.70) follows after the change of variables $\overline{m}=m^\star-r$ and using the identity $1/(\overline{m}-k)!=(\overline{m})_{k\downarrow}/\overline{m}!$. The third equality is obtained rearranging terms after multiplying and dividing for $\left(\gamma_j(r+\overline{m})\right)_{n_j+m_j}$ as well as for $(\gamma_j)_{n_{j,l}}$, for all $j=1,2$ and $l=1,\dots,r$. Additionally, we also assumed, without loss of generality, that the first r species are the ones which have already been observed.

Then, plugging Equations (6.5) and (6.70) into Equation (6.69) we get

$$
\begin{aligned}
&\mathbb{P}\left(K_{m_1,m_2}^{(n_1,n_2)}=k,K_{1,m_1}^{(n_1)}=k_1,K_{2,m_2}^{(n_2)}=k_2\mid\boldsymbol{X}\right) \\
&=\sum_{\mathscr{P}_{m_1,m_2,r+k}}\sum_{\overline{m}=k}^{\infty}\left\{(\overline{m})_{k\downarrow}\frac{1}{V_{n_1,n_2}^{r}}\frac{(\overline{m}+r)!}{\overline{m}!}q_M(\overline{m}+r)\prod_{j=1}^{2}\frac{1}{\left(\gamma_j(r+\overline{m})\right)_{n_j}}\right. \\
&\quad\left.\times\prod_{j=1}^{2}\prod_{l=1}^{r}\frac{(\gamma_j)_{n_{j,l}+m_{j,l}(\pi_j)}}{(\gamma_j)_{n_{j,l}}}\prod_{j=1}^{2}\prod_{l=r+1}^{r+k}(\gamma_j)_{m_{j,l}(\pi_j)}\prod_{j=1}^{2}\frac{\left(\gamma_j(r+\overline{m})\right)_{n_j}}{\left(\gamma_j(r+\overline{m})\right)_{n_j+m_j}}\right\} \\
&=\sum_{\mathscr{P}_{m_1,m_2,r+k}}\sum_{\overline{m}=k}^{\infty}\left\{(\overline{m})_{k\downarrow}q_M(\overline{m}\mid\boldsymbol{X})\prod_{j=1}^{2}\frac{\left(\gamma_j(r+\overline{m})\right)_{n_j}}{\left(\gamma_j(r+\overline{m})\right)_{n_j+m_j}}\right\} \\
&\quad\times\prod_{j=1}^{2}\prod_{l=1}^{r}\frac{(\gamma_j)_{n_{j,l}+m_{j,l}(\pi_j)}}{(\gamma_j)_{n_{j,l}}}\prod_{j=1}^{2}\prod_{l=r+1}^{r+k}(\gamma_j)_{m_{j,l}(\pi_j)}.
\end{aligned}
\tag{6.71}
$$

In Equation (6.71), we recognized the posterior distribution $q_M(\cdot\mid\boldsymbol{X})$ defined in Equation (6.24).

Finally, we notice that the inner infinite sum does not depend on (π_1, π_2). In particular, we have

$$\begin{aligned}
&\sum_{m^\star=k}^{\infty} \left\{ (m^\star)_{k\downarrow} q_M(m^\star \mid \boldsymbol{X}) \prod_{j=1}^{2} \frac{\left(\gamma_j(r+m^\star)\right)_{n_j}}{\left(\gamma_j(r+m^\star)\right)_{n_j+m_j}} \right\} \\
&= \frac{1}{V^r_{n_1,n_2}} \sum_{m^\star=k}^{\infty} \left\{ \frac{m^\star!(m^\star+r)!}{(m^\star-k)!m^\star!} q_M(m^\star+r) \prod_{j=1}^{2} \frac{\Gamma\left(\gamma_j(r+m^\star)\right)}{\Gamma\left(\gamma_j(r+m^\star)+n_j\right)} \frac{\Gamma\left(\gamma_j(r+m^\star)+n_j\right)}{\Gamma\left(\gamma_j(r+m^\star)+n_j+m_j\right)} \right\} \\
&= \frac{1}{V^r_{n_1,n_2}} \sum_{m^\star=k}^{\infty} \left\{ \frac{(m^\star+r)!}{(m^\star-k)!} q_M(m^\star+r) \prod_{j=1}^{2} \frac{\Gamma\left(\gamma_j(r+m^\star)\right)}{\Gamma\left(\gamma_j(r+m^\star)+n_j+m_j\right)} \right\} \\
&= \frac{1}{V^r_{n_1,n_2}} \sum_{m^{\star\star}=k+r}^{\infty} \left\{ \frac{m^{\star\star}!}{(m^{\star\star}-r-k)!} q_M(m^{\star\star}) \prod_{j=1}^{2} \frac{1}{\left(\gamma_j m^{\star\star}\right)_{n_j+m_j}} \right\} = \frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^r_{n_1,n_2}}
\end{aligned}$$

where the final equality follows noticing that $m^{\star\star}!/(m^{\star\star}-k-r)! = (m^{\star\star})_{(k+r)\downarrow}$.

We now focus on solving the sum over the set of partitions $\mathscr{P}_{m_1,m_2,r+k}$. The partitions (π_1, π_2) only appears thought the cardinalities of the sets, hence, the quantities of interest can equivalently be computed as

$$\begin{aligned}
\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2} = k, K^{(n_1)}_{1,m_1} = k_1, K^{(n_2)}_{2,m_2} = k_2 \mid \boldsymbol{X}\right) &= \frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^r_{n_1,n_2}} \\
\times \frac{1}{k!} \sum_{(\Delta)} \binom{m_j}{m_{j,1},\ldots,m_{j,r+k}} &\prod_{j=1}^{2} \prod_{l=1}^{r} \frac{(\gamma_j)_{n_{j,l}+m_{j,l}}}{(\gamma_j)_{n_{j,l}}} \prod_{j=1}^{2} \prod_{l=r+1}^{r+k} (\gamma_j)_{m_{j,l}},
\end{aligned}$$

where the sum is extended over the set (Δ) of non-negative integers $\boldsymbol{m}_1 = \left(m_{1,1},\ldots,m_{1,r+k}\right)$ and $\boldsymbol{m}_2 = \left(m_{2,1},\ldots,m_{2,r+k}\right)$ which satisfy the following constraints,

$$\begin{aligned}
&\sum_{l=1}^{r+k} m_{j,l} = m_j\,, \quad m_{j,l} \geq 0 \quad j = 1,2;\ l = 1,\ldots,r+k\,, \\
&m_{1,l} + m_{2,l} \geq 1 \quad l = r+1,\ldots,r+k\,, \\
&\sum_{l=1}^{r} \delta_{\{m_{j,l}\geq 1,\, n_{j,l}=0\}} + \sum_{l=r+1}^{r+k} \delta_{\{m_{j,l}\geq 1\}} = k_j\,, \quad j = 1,2\,.
\end{aligned} \tag{6.72}$$

As mentioned at the beginning of the proof, the knowledge of k, k_1 and k_2 only is not enough to fully characterise $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ and therefore decouple the joint condition in Equation (6.72) as done in Section E.4. To do so, we introduce the auxiliary quantities $S^*_m = s^*$ and $K^{*(n)}_{1,m} = k^*_1$. See Section 6.4 for their interpretation. Since we are not interested in inferring such quantities, we then marginalised them out. As a consequence of this augmentation, all other posterior quantities (namely, s, k^*_2, $s_{1,2}$ and $s_{2,1}$) can be recovered and their expressions are reported in Equation (6.68). Additionally, s^* and k^*_1 to fix some ordering of the new species. We say that the new shared species among the new k new species are located in the first s^* positions. Their order is fixed in any of the $\binom{k}{s^*}$ equivalent ways. We point out that as we set s^*, we also set to $s - s^*$ the number of new shared species among the already observed r species. Then, consecutively to the s^* new shared species, we set the following k^*_1 species to be those that are found in group 1 only. The order is fixed in any of the $\binom{k-s^*}{k^*_1}$ equivalent ways. In particular, among the

r new species, we are left with $k_2^* = k - s^* - k_1^*$ species that are specific to area 2 only and which are placed, by construction, in the final positions. It follows that the target probability equals

$$\begin{aligned}
&\mathbb{P}\left(K_{m_1,m_2}^{(n_1,n_2)} = k, K_{1,m_1}^{(n_1)} = k_1, K_{2,m_2}^{(n_2)} = k_2 \mid \boldsymbol{X}\right) = \\
&\quad = \frac{V_{n_1+m_1,n_2+m_2}^{r+k}}{V_{n_1,n_2}^{r}} \frac{1}{k!} \sum_{s^*=0}^{k} \sum_{k_1^*=0}^{k-s^*} \binom{k}{s^*} \binom{k-s^*}{k_1^*} \\
&\qquad \times \sum_{(\Delta\Delta)} \binom{m_j}{m_{j,1}, \ldots, m_{j,r+k}} \prod_{j=1}^{2} \prod_{l=1}^{r} \frac{(\gamma_j)_{n_{j,l}+m_{j,l}}}{(\gamma_j)_{n_{j,l}}} \prod_{j=1}^{2} \prod_{l=r+1}^{r+k} (\gamma_j)_{m_{j,l}},
\end{aligned}$$

where the summation over $(\Delta\Delta)$ satisfies the following set of constraints:

$$\begin{aligned}
&\sum_{l=1}^{r+k} m_{j,l} = m_j\,, \quad j = 1, 2\,, \\
&m_{j,l} \geq 0 \quad j = 1, 2;\ l = 1, \ldots, r\,, \\
&\sum_{l=1}^{r} \delta_{\{n_{1,l} \cdot n_{2,l} = 0,\ n_{1,l}+m_{1,l} \geq 1,\ n_{2,l}+m_{2,l} \geq 1,\}} = s - s^*\,, \\
&m_{j,l} \geq 1 \quad j = 1, 2;\ l = r+1, \ldots, r+s^*\,, \\
&m_{1,l} \geq 1,\ m_{2,l} = 0 \quad j = 1, 2;\ l = r+s^*+1, \ldots, r+s^*+k_1^*\,, \\
&m_{1,l} = 0,\ m_{2,l} \geq 1 \quad j = 1, 2;\ l = r+s^*+k_1^*+1, \ldots, r+k\,, \\
&\sum_{l=1}^{r} \delta_{\{m_{j,l} \geq 1,\ n_{j,l} = 0\}} = s_{j',j} \quad j, j' = 1, 2\,,\ j \neq j'\,.
\end{aligned} \tag{6.73}$$

The final conditions in Equation (6.73) rise noticing that the second sum in the corresponding condition in Equation (6.72) has been set equal to $s^* + k_j^*$, for $j = 1, 2$, and therefore the first sum must be equal to $k_j - s^* - k_j^*$, which coincides $s_{j',j}$, see Equation (6.68).

It is still not possible to decouple $(\Delta\Delta)$ into two disjoint sets because of the joint condition regarding the number of new shared species among the already observed r species. See line 3 in Equation (6.73). However, $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ can be further reordered to ensure it. Indeed, from Section 6.3, we know that the observed sample $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ can be arranged so that the first t out of r species are shared, than the following r_1^* species are only present in the first group while the remaining r_2^* are only present in the second group. Let us fix j. In order to satisfy the joint condition in Equation (6.73) it is enough to reorder $m_{j,l}$, $l = 1, \ldots, r$, so that the first $s_{j',j}$ species, which where first only observed in group j' and that are then observed in group j, are the first $s_{j',j}$ species among the $r_{j'}^*$ such that $n_{j,l} = 0$. By construction, it follows that the remaining $r_{j'}^* - s_{j',j}$ must be such that $m_{j,l} = 0$. This can be done in $\binom{r_2^*}{s_{j',j}}$ equivalent

ways. We have,

$$\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2}=k, K^{(n_1)}_{1,m_1}=k_1, K^{(n_2)}_{2,m_2}=k_2 \mid \boldsymbol{X}\right) = \\ = \frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^{r}_{n_1,n_2}} \frac{1}{k!} \sum_{s^*=0}^{k} \sum_{k_1^*=0}^{k-s^*} \binom{k}{s^*} \binom{k-s^*}{k_1^*} \binom{r_1^*}{s_{1,2}} \binom{r_2^*}{s_{2,1}} \\ \times \sum_{(\Delta\Delta\Delta)} \binom{m_j}{m_{j,1},\ldots,m_{j,r+k}} \prod_{j=1}^{2} \prod_{l=1}^{r} \frac{(\gamma_j)_{n_{j,l}+m_{j,l}}}{(\gamma_j)_{n_{j,l}}} \prod_{j=1}^{2} \prod_{l=r+1}^{r+k} (\gamma_j)_{m_{j,l}},$$

where the set $(\Delta\Delta\Delta)$ describes the following set of constraints,

$$\begin{aligned} &\sum_{l=1}^{r+k} m_{j,l} = m_j\,, \\ &m_{1,l} \geq 0\,, \quad l = 1,\ldots,t+r_1^*\,, \\ &m_{1,l} \geq 1\,, \quad l = t+r_1^*+1,\ldots,t+r_1^*+s_{2,1}\,, \\ &m_{1,l} = 0\,, \quad l = t+r_1^*+s_{2,1}+1,\ldots,r\,, \\ &m_{2,l} \geq 0\,, \quad l = 1,\ldots,t \quad l = r-r_2^*,\ldots,r\,, \\ &m_{2,l} \geq 1\,, \quad l = t+1,\ldots,t+s_{1,2}\,, \\ &m_{2,l} = 0\,, \quad l = t+s_{1,2}+1,\ldots,t+r_1^*\,, \\ &m_{j,l} \geq 1\,, \quad j = 1,2;\, , l = r+1,\ldots,r+s^*\,, \\ &m_{1,l} \geq 1,\, m_{2,l} = 0\,, \quad j = 1,2;\, l = r+s^*+1,\ldots,r+s^*+k_1^*\,, \\ &m_{1,l} = 0,\, m_{2,l} \geq 1\,, \quad j = 1,2;\, l = r+s^*+k_1^*+1,\ldots,r+k\,. \end{aligned} \tag{6.74}$$

We are finally ready to decouple the set $(\Delta\Delta\Delta)$. We discard the elements such that $m_{j,l} = 0$ and note that the number of species such that $m_{j,l} \geq 1$ is $s_{j',1} + s^* + k^* j$ which equals k_j while among the first r species, the number of those such that $m_{j,l} \geq 0$ is $t + r_j^*$ which equals r_j. We have that,

$$\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2}=k, K^{(n_1)}_{1,m_1}=k_1, K^{(n_2)}_{2,m_2}=k_2 \mid \boldsymbol{X}\right) = \\ = \frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^{r}_{n_1,n_2}} \frac{1}{k!} \sum_{s^*=0}^{k} \sum_{k_1^*=0}^{k-s^*} \binom{k}{s^*} \binom{k-s^*}{k_1^*} \binom{r_1^*}{s_{1,2}} \binom{r_2^*}{s_{2,1}} \\ \times \prod_{j=1}^{2} \sum_{(\Delta j)} \binom{m_j}{m_{j,1},\ldots,m_{j,r_j+k_j}} \prod_{l=1}^{r_j} (\gamma_j + n_{j,l})_{m_j} \prod_{l=r_j+1}^{r_j+k_j} (\gamma_j)_{m_{j,l}}\,, \tag{6.75}$$

where the sum is extended over the sets (Δj), for $j = 1, 2$, of non-negative integers $\left(m_{1,1},\ldots,m_{1,r_j+k_j}\right)$ and $\left(m_{2,1},\ldots,m_{2,r_j+k_j}\right)$ which satisfy the following constraints,

$$\begin{aligned} &\sum_{l=1}^{r_j+k_j} m_{j,l} = m_j\,, \\ &m_{j,l} \geq 0\,, \quad l = 1,\ldots,r_j\,, \\ &m_{j,l} \geq 1\,, \quad l = r_j+1,\ldots,r_j+k_j\,, \end{aligned}$$

for each $j = 1, 2$. Moreover, in Equation (6.75), we leveraged on the fact that $n_{j,l} = 0$ for $l = r_j + 1, \dots, r$ and the identity $(x)_{n+m}/(x)_n = (x+n)_m$, that holds for any non-negative integers n, m and any positive real number $x > 0$. We are finally left to compute the sums over the sets (Δj). To do so, let us fix j and introduce the additional variable h_j, with $h_j \in \{k_j, k_j + 1, \dots, m_j\}$, such that

$$\sum_{j=1}^{r_j} m_{j,l} = m_j - h_j\,, \qquad \sum_{j=r_j+1}^{r_j+k_j} m_{j,l} = h_j\,.$$

If follows that

$$\begin{aligned}\sum_{(\Delta j)} \binom{m_j}{m_{j,1}, \dots, m_{j,r_j+k_j}} \prod_{l=1}^{r_j} (\gamma_j + n_{j,l})_{m_j} \prod_{l=r_j+1}^{r_j+k_j} (\gamma_j)_{m_{j,l}} \\ = \sum_{h_j=k_j}^{m_j} \sum_{(\Delta j,1)} \sum_{(\Delta j,2)} \binom{m_j}{h_j} \binom{m_j - h_j}{m_{j,1}, \dots, m_{j,r_j}} \binom{h_j}{m_{j,r_j+1}, \dots, m_{j,r_j+k_j}} \\ \times \prod_{l=1}^{r_j} (\gamma_j + n_{j,l})_{m_j} \prod_{l=r_j+1}^{r_j+k_j} (\gamma_j)_{m_{j,l}}\,,\end{aligned} \tag{6.76}$$

where the sets $(\Delta j, 1)$ and $(\Delta j, 2)$ are defined as

$$\begin{aligned}(\Delta j, 1) &= \left\{ (m_{j,1}, \dots, m_{j,r_j}) \;:\; m_{j,l} \geq 0, \text{ and } \sum_{l=1}^{r_j} m_{j,l} = m_j - h_j \right\} \\ (\Delta j, 2) &= \left\{ (m_{j,r_j+1}, \dots, m_{j,r_j+k_j}) \;:\; m_{j,l} \geq 1, \text{ and } \sum_{l=r_j+1}^{r_j+k_j} m_{j,l} = h_j \right\}\end{aligned} \tag{6.77}$$

Then, from Vandermonde's generalized identity, see Equation (6.52), it follows that

$$\sum_{(\Delta j,1)} \binom{m_j - h_j}{m_{j,1}, \dots, m_{j,r_j}} \prod_{l=1}^{r_j} (\gamma_j + n_{j,l})_{m_{j,l}} = \left(\sum_{l=1}^{r_j} (\gamma_j + n_{j,l}) \right)_{m_j - h_j} = (\gamma_j r_j + n_j)_{m_j - h_j} \tag{6.78}$$

while, thanks to Equation (6.51), we have

$$\sum_{(\Delta j,2)} \binom{h_j}{m_{j,r_j+1}, \dots, m_{j,r_j+k_j}} \prod_{l=r_j+1}^{r_j+k_j} (\gamma_j)_{m_{j,l}} = k_j!\, |C(h_j, k_j; -\gamma_j)|. \tag{6.79}$$

Finally, plugging Equations (6.78) and (6.79) into (6.76), we have

$$\begin{aligned}\sum_{(\Delta j)} \binom{m_j}{m_{j,1}, \dots, m_{j,r_j+k_j}} \prod_{l=1}^{r_j} (\gamma_j + n_{j,l})_{m_j} \prod_{l=r_j+1}^{r_j+k_j} (\gamma_j)_{m_{j,l}} \\ = k_j! \sum_{h_j=k_j}^{m_j} \binom{m_j}{h_j} |C(h_j, k_j; -\gamma_j)| \, (\gamma_j r_j + n_j)_{m_j - h_j} \\ = k_j!\, |C\left(m_j, k_j; -\gamma_j, -(\gamma_j r_j + n_j)\right)|,\end{aligned} \tag{6.80}$$

where the final equality follows from Equation (6.50). The statement follows after plugging Equation

(6.80) into (6.75) and rewriting all quantities in terms of k, k_1 and k_2.

E.5.5 Proof of Proposition 3

The best way to evaluate $\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2} = k \mid \boldsymbol{X}\right)$ is not marginalizing $K^{(n_1)}_{1,m_1}$ and $K^{(n_2)}_{2,m_2}$ out of Equation (6.27). Section 6.4 explains that the global number of new distinct species, $K^{(n_1,n_2)}_{m_1,m_2}$, can be computed using those posterior quantities that do not require information about the frequencies of species in the future sample that regard the previously observed r distinct species. Namely, $m_{j,l}$ for $l = 1,\dots,r$. In summary, the proof follows the steps of the one in Section E.5.4 but does not require reordering the first r species. Indeed, following the same steps of Section E.5.4, we can show that the quantity of interest can be computed as

$$\begin{aligned}\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2} = k \mid \boldsymbol{X}\right) &= \sum_{\mathscr{P}_{m_1,m_2,r+k}} \frac{\Pi^{(n+m)}_{r+k}(\boldsymbol{n}_1 + \boldsymbol{m}_1(\pi_1), \boldsymbol{n}_2 + \boldsymbol{m}_2(\pi_2))}{\Pi^{(n)}_r(\boldsymbol{n}_1,\boldsymbol{n}_2)} \\ &= \frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^r_{n_1,n_2}} \frac{1}{k!} \sum_{(\star)} \binom{m_j}{m_{j,1},\dots,m_{j,r+k}} \prod_{j=1}^{2}\prod_{l=1}^{r} \frac{(\gamma_j)_{n_{j,l}+m_{j,l}}}{(\gamma_j)_{n_{j,l}}} \prod_{j=1}^{2}\prod_{l=r+1}^{r+k} (\gamma_j)_{m_{j,l}},\end{aligned}$$

where the sum is extended over the set $(\star)$ of non-negative integers $\boldsymbol{m}_1 = (m_{1,1},\dots,m_{1,r+k})$ and $\boldsymbol{m}_2 = (m_{2,1},\dots,m_{2,r+k})$ which satisfy the following constraints,

$$\begin{aligned}&\sum_{l=1}^{r+k} m_{j,l} = m_j\,, \quad m_{j,l} \geq 0 \quad j = 1,2;\ l = 1,\dots,r+k\,, \\ &m_{1,l} + m_{2,l} \geq 1\,, \quad l = r+1,\dots,r+k\,.\end{aligned} \tag{6.81}$$

We now augment the space by introducing the auxiliary random variables $K^{*(n)}_{1,m} = k^*_1$ and $K^{*(n)}_{2,m} = k^*_2$. Due to Equation (6.68), this also implies that the number of new shared species among the new k distinct species is $s^* = k - k^*_1 - k^*_2$. Hence, we order the new k species so that the first s^* are shared, the following k^*_1 are present in group one only and the remaining k^*_2 are found in group 2 only. This can be done in $\binom{k}{k^*_1}\binom{k-k^*_1}{k^*_2}$ equivalent ways. As for Section E.5.4, we are not interested in $K^{*(n)}_{1,m}$ and $K^{*(n)}_{2,m}$, which can then be marginalized out. Hence, we have that

$$\begin{aligned}&\mathbb{P}\left(K^{(n_1,n_2)}_{m_1,m_2} = k \mid \boldsymbol{X}\right) \\ &\quad= \frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^r_{n_1,n_2}} \frac{1}{k!} \sum_{(\star\star)} \sum_{k^*_1=0}^{k} \sum_{k^*_2=0}^{k-k^*_1} \binom{k}{k^*_1}\binom{k-k^*_1}{k^*_2}\binom{m_j}{m_{j,1},\dots,m_{j,r+k}} \\ &\qquad\times \prod_{j=1}^{2}\prod_{l=1}^{r} \frac{(\gamma_j)_{n_{j,l}+m_{j,l}}}{(\gamma_j)_{n_{j,l}}} \prod_{j=1}^{2}\prod_{l=r+1}^{r+k} (\gamma_j)_{m_{j,l}},\end{aligned}$$

where the set $(\star\star)$ defines the following constraints,

$$\begin{aligned}
&\sum_{l=1}^{r+k} m_{j,l} = m_j\,, \quad m_{j,l} \geq 0 \quad j = 1,2;\, l = 1,\dots,r\,,\\
&m_{j,l} \geq 1\,, \quad j = 1,2;\, l = r+1,\dots,r+s^*\,,\\
&m_{1,l} \geq 1\,,\, m_{2,l} = 0\,, \quad l = r+s^*+1,\dots,r+s^*+k_1^*\,,\\
&m_{1,l} = 0,\, m_{2,l} \geq 1\,, \quad l = r+s^*+k_1^*+1,\dots,r+k\,.
\end{aligned} \tag{6.82}$$

The set $(\star\star)$ does not involve any coupled condition as we are not interested in shared quantities. Hence, it can be decoupled similarly to the set $(\Delta\Delta\Delta)$ in Section E.5.4. It follows that,

$$\begin{aligned}
\mathbb{P}\left(K_{m_1,m_2}^{(n_1,n_2)} = k \mid \boldsymbol{X}\right) \;=\; & \frac{V_{n_1+m_1,n_2+m_2}^{r+k}}{V_{n_1,n_2}^{r}} \frac{1}{k!} \sum_{k_1^*=0}^{k} \sum_{k_2^*=0}^{k-k_1^*} \binom{k}{k_1^*} \binom{k-k_1^*}{k_2^*} \\
& \times \prod_{j=1}^{2} \sum_{(\star j)} \binom{m_j}{m_{j,1},\dots,m_{j,r+s^*+k_j^*}} \prod_{l=1}^{r} (\gamma_j + n_{j,l})_{m_{j,l}} \prod_{l=r+1}^{r+s^*+k_j^*} (\gamma_j)_{m_{j,l}}
\end{aligned}$$

where the sum is extended over the sets $(\star j)$, for $j = 1,2$, of non-negative integers $\left(m_{1,1},\dots,m_{1,r+s^*+k_1^*}\right)$ and $\left(m_{2,1},\dots,m_{2,r+s^*+k_2^*}\right)$ which satisfy the following constraints,

$$\begin{aligned}
&\sum_{l=1}^{r+s^*+k_j^*} m_{j,l} = m_j\,, \quad m_{j,l} \geq 0 \quad l = 1,\dots,r\,,\\
&m_{j,l} \geq 1\,, \quad l = r+1,\dots,r+s^*+k_j^*\,,
\end{aligned}$$

for each $j = 1,2$. Let us now fix j. The sum over $(\star j)$ is solved as in Section E.5.4, that is introducing the additional variable h_j, with $h_j \in \{s^*+k_j^*, s^*+k_j^*+1,\dots,m_j\}$, such that

$$\sum_{j=1}^{r} m_{j,l} = m_j - h_j \;\text{ and }\; \sum_{j=r+1}^{r+s^*+k_j^*} m_{j,l} = h_j\,.$$

Following the same steps of Section E.5.4, it is possible to show that

$$\begin{aligned}
\mathbb{P}\left(K_{m_1,m_2}^{(n_1,n_2)} = k \mid \boldsymbol{X}\right) \;=\; & \frac{V_{n_1+m_1,n_2+m_2}^{r+k}}{V_{n_1,n_2}^{r}} \\
& \times \frac{1}{k!} \sum_{k_1^*=0}^{k} \sum_{k_2^*=0}^{k-k_1^*} \binom{k}{k_1^*} \binom{k-k_1^*}{k_2^*} \prod_{j=1}^{2} (s^*+k_j^*)!\, |C\left(m_j, s^*+k_j^*; -\gamma_j, -(\gamma_j r + n_j)\right)|.
\end{aligned}$$

Finally, we perform a change of variables to lighten the notation and so that s^* is not involved in the final expression, even though it can be written in terms of k, k_1^* and k_2^*. Let us fix j and let $z_j = k - s^* - k_j^*$. This can be interpreted as the number of new species that are not present in group j but that are present

in group j'. Hence, $z_j = k^*_{j'}$. If follows that $s^* + k^*_j = k - z_j$. Furthermore, the following identity holds

$$\binom{k}{z_2}\binom{k-z_2}{z_1} = \binom{k}{z_1}\binom{k-z_1}{z_2},$$

for $z_1 \in \{0, \ldots, k\}$ and $z_2 \in \{0, \ldots, k - z_1\}$. It is enough to rearrange the binomial and factorial coefficients to conclude the proof.

E.5.6 Proof of Equation (6.29)

Let us fix $j, j' = 1, 2$ with $j \neq j'$. The statement follows from Equation (6.28), after setting both $n_{j'}$ and $m_{j'}$ to zero. Indeed, $V(m_j, m_{j'}; k) = V(m_j; k_j)$ because, from Equation (6.22), $k = k_j$ as $k_{j'}$ and s must be zero since $m_{j'} = 0$ and $(x)_0 = 1$ for all $x > 0$. Similarly, $n_{j'} = 0$ also implies that $r = r_j$. Then, we recall that $|C(0, 0; x, y)| = 1$ and $|C(0, k; x, y)| = 0$ for all $x \neq 0$, $y \in \mathbb{R}$ and $k = 1, 2, 3, \ldots$, see Charalambides (2002). Hence, the only non-zero term in Equation (6.28) is when $z_1 = 0$ and $z_2 = k$. Finally, recalling that $r = r_j$ and $k = k_j$, it follows that

$$\mathbb{P}\left(K^{(n_j)}_{j,m_j} = k_j \mid \boldsymbol{X}\right) = \frac{V^{r_j+k_j}_{n_j+m_j}}{V^{r_j}_{n_j}} |C\left(m_j, k_j; -\gamma_j, -(\gamma_j r_j + n_j)\right)|.$$

E.6 Proofs of the results in Section 6.5

E.6.1 Proofs of Equations (6.37) - (6.38)

Firstly, we note that:

$$E\left(\sum_{m=1}^{M} w^2_{j,m}\right) = E_{q_M}\left(E\left(\sum_{m=1}^{M} w^2_{j,m} \mid M\right)\right) = E_{q_M}\left(\sum_{m=1}^{M} E\left(w^2_{j,m} \mid M\right)\right).$$

Given M, $(w_{j,1}, \ldots, w_{j,M}) \sim \mathrm{Dir}_M(\gamma_j, \ldots, \gamma_j)$, hence $E(w_{j,m} \mid M) = 1/M$ and $\mathrm{var}(w_{j,m} \mid M) = (M-1)/(M^2(\gamma_j M + 1))$. It follows that

$$E\left(\sum_{m=1}^{M} w^2_{j,m}\right) = E_{q_M}\left(\sum_{m=1}^{M} \frac{1+\gamma_j}{M(1+\gamma_j M)}\right) = (1+\gamma_j) E_{q_M}\left(\frac{1}{(1+\gamma_j M)}\right).$$

$$E\left(\sum_{m=1}^{M} w_{1,m} w_{2,m}\right) = E_{q_M}\left(\sum_{m=1}^{M} E\left(w_{1,m} w_{2,m} \mid M\right)\right) =$$
$$E_{q_M}\left(\sum_{m=1}^{M} E\left(w_{1,m} \mid M\right) E\left(w_{2,m} \mid M\right)\right) = E_{q_M}\left(1/M\right).$$

The equality holds since $(w_{1,1}, \ldots, w_{1,M})$ and $(w_{2,1}, \ldots, w_{2,M})$ are independent given M.

E.6.2 Preliminaries

In this section, we show a useful inequality that allows us to apply the dominated convergence theorem to exchange the order of the limit and the sum in the following proofs.

Lemma E.6.1. *Let $j, j' \in \{1,2\}$, $j \neq j'$ and let $n_1 \geq 1$, $n_2 \geq 1$, $0 \geq n_{j,l} <\leq n_j$ for $l = 1, \dots, r$ and $r \geq 1$. Let $q^\star_{M|\boldsymbol{X}}$ be the probability mass function defined in Equation (6.24). The following inequality holds for every $\gamma_1 > 0$, $\gamma_2 > 0$, $m^\star \geq 0$:*

$$\begin{aligned}\frac{\sum_{l=1}^r n_{j,l}(n_{j,l}+1) + \gamma_j\left(\gamma_j(r+m^\star) + r + m^\star + 2n_j\right)}{n_j(n_j+1) + \gamma_j^2(r+m^\star)^2 + \gamma_j(r+m^\star)(2n_j+1)} q^\star_{M|\boldsymbol{X}}(m^\star) \\ \lesssim \frac{(m^\star+r)!}{m^\star} q_M(m^\star+r)\,.\end{aligned} \tag{6.83}$$

$$\frac{\sum_{l=1}^t n_{1,l}n_{2,l} + \gamma_1\gamma_2(r+M^\star) + n_1\gamma_2 + n_2\gamma_1}{(\gamma_1(r+M^\star)+n_1)(\gamma_2(r+M^\star)+n_2)} q^\star_{M|\boldsymbol{X}}(m^\star) \lesssim \frac{(m^\star+r)!}{m^\star} q_M(m^\star+r)\,, \tag{6.84}$$

where we use notation $\lesssim$ to indicate that the upper bound holds up to some constant.

Proof. We prove Equation (6.83) only. The proof of Equation (6.84) follows the same steps.

$$\begin{aligned}&\frac{\sum_{l=1}^r n_{j,l}(n_{j,l}+1) + \gamma_j\left(\gamma_j(r+m^\star) + r + m^\star + 2n_j\right)}{n_j(n_j+1) + \gamma_j^2(r+m^\star)^2 + \gamma_j(r+m^\star)(2n_j+1)} q^\star_{M|\boldsymbol{X}}(m^\star) \\ &\leq \left\{\frac{r(n_j+1)^2}{n_j(n_j+1) + \gamma_j\left(\gamma_j(r+m^\star)^2 + r + m^\star + 2n_j(r+m^\star)\right)}\right. \\ &\quad \left. + \frac{\gamma_j\left(\gamma_j(r+m^\star) + r + m^\star + 2n_j\right)}{n_j(n_j+1) + \gamma_j\left(\gamma_j(r+m^\star)^2 + r + m^\star + 2n_j(r+m^\star)\right)}\right\} \frac{(m^\star+r)_{r\downarrow} q_M(m^\star+r)}{V_{n_1,n_2}(\gamma_j(m^\star+r))_{n_j}(\gamma_{j'}(m^\star+r))_{n_{j'}}} \\ &\leq \left\{\frac{r(n_j+1)^2}{n_j(n_j+1)} + \frac{\gamma_j\left(\gamma_j(r+m^\star) + r + m^\star + 2n_j\right)}{\gamma_j\left(\gamma_j(r+m^\star)^2 + r + m^\star + 2n_j(r+m^\star)\right)}\right\} \\ &\quad \times \frac{(m^\star+r)_{r\downarrow} q_M(m^\star+r)}{V_{n_1,n_2}(\gamma_j(m^\star+r))_{n_j}(\gamma_{j'}(m^\star+r))_{n_{j'}}}\,.\end{aligned}$$

In the final line, we only used the positivity of each term in the denominator. Then, we exploit that V_{n_1,n_2} is a sum of positive terms, hence it is larger than its first positive term. Moreover, note that the second term in the parenthesis is smaller that one since $r + m^\star \geq 1$.

$$\begin{aligned}&\frac{\sum_{l=1}^r n_{j,l}(n_{j,l}+1) + \gamma_j\left(\gamma_j(r+m^\star) + r + m^\star + 2n_j\right)}{n_j(n_j+1) + \gamma_j^2(r+m^\star)^2 + \gamma_j(r+m^\star)(2n_j+1)} q^\star_{M|\boldsymbol{X}}(m^\star) \\ &\leq \left\{\frac{r(n_j+1)^2}{n_j(n_j+1)} + 1\right\} \frac{(m^\star+r)_{r\downarrow} q_M(m^\star+r)}{r! q_M(r)} \frac{(\gamma_j r)_{n_j}(\gamma_{j'} r)_{n_{j'}}}{(\gamma_j(m^\star+r))_{n_j}(\gamma_{j'}(m^\star+r))_{n_{j'}}}\,.\end{aligned}$$

Then, due to monotonicity of the Pochhammer symbol, the final term is smaller or equal than one, which concludes the proof. As a corollary, note that we also proved that

$$q^\star_{M|\boldsymbol{X}}(m^\star) \lesssim \frac{(m^\star+r)!}{m^\star} q_M(m^\star+r)\,. \tag{6.85}$$

□

Moreover, under the hypothesis that q_M is a probability mass function on the positive integers for which there exists some constant $a \in (0,1)$ such that $q_M(m) \leq a^m$ for large values of m, then we have that

$$\sum_{m^\star=0}^{\infty} \frac{(m^\star + r)!}{m^\star} q_M(m^\star + r) < \infty \,.$$

The result holds since

$$\sum_{m^\star=0}^{\infty} \frac{(m^\star + r)!}{m^\star!} a^{m^\star} < \infty \,,$$

for every $a \in (0,1)$, see Argiento and De Iorio (2022).

We conclude this section about some preliminaries results giving additional details about the limiting distribution of $q^\star_{M|\boldsymbol{X}}$ when $\lim_{\gamma_j \to \gamma_0}$, where $\gamma_0 \in [0, \infty]$. In particular, $q^\star_{M|\boldsymbol{X}}$ is a well defined probability mass function on $\mathbb{N}$ for each $\gamma_1 > 0, \gamma_2 > 0$. We define the limiting distribution $q^\star_{M|\boldsymbol{X},\gamma_0}$ as the pointwise limit of each atom. Namely,

$$q^\star_{M|\boldsymbol{X},\gamma_0}(m^\star) = \lim_{\gamma_j \to \gamma_0} q^\star_{M|\boldsymbol{X}}(m^\star) \,,$$

for each $m^\star \geq 0$. Clearly, $q^\star_{M|\boldsymbol{X},\gamma_0}(m^\star) \geq 0$ for each $m^\star \geq 0$. Moreover, thanks to Equation (6.85), we apply the dominated convergence theorem to show that

$$\sum_{m^\star=0}^{\infty} q^\star_{M|\boldsymbol{X},\gamma_0}(m^\star) = \sum_{m^\star=0}^{\infty} \lim_{\gamma_j \to \gamma_0} q^\star_{M|\boldsymbol{X}}(m^\star) = \lim_{\gamma_j \to \gamma_0} \sum_{m^\star=0}^{\infty} q^\star_{M|\boldsymbol{X}}(m^\star) = 1 \,.$$

In particular, $q^\star_{M|\boldsymbol{X},\gamma_0}(m^\star) \leq 1$ for each $m^\star \geq 0$ since the sum is one and all terms are non-negative. As a consequence, we conclude that $q^\star_{M|\boldsymbol{X},\gamma_0}$ is a well defined probability mass function on $\mathbb{N}$.

Lemma E.6.2. *When γ_0 is $+\infty$, it holds that*

$$q^\star_{M|\boldsymbol{X},\infty}(m^\star) \propto \frac{(m^\star + r)_{r\downarrow}}{(m^\star + r)^{n_1}(m^\star + r)^{n_2}} q_M(m^\star + r) \,. \tag{6.86}$$

Proof. We must evaluate the following limit

$$\begin{aligned}
&\lim_{\gamma_1,\gamma_2 \to +\infty} q^\star_{M|\boldsymbol{X}}(m^\star) \\
&= \lim_{\gamma_1,\gamma_2 \to +\infty} \frac{1}{V^r_{n_1,n_2}} (m^\star + r)_{r\downarrow} q_M(m^\star + r) \prod_{j=1}^{d} \frac{1}{(\gamma_j (m^\star + r))_{n_j}} \\
&= (m^\star + r)_{r\downarrow} q_M(m^\star + r) \lim_{\gamma_1,\gamma_2 \to +\infty} \left\{ \sum_{m=0}^{\infty} (m)_{r\downarrow} q_M(m) \prod_{j=1}^{2} \frac{(\gamma_j (m^\star + r))_{n_j}}{(\gamma_j m)_{n_j}} \right\}^{-1} .
\end{aligned}$$

Exploiting the results in Lemma E.6.1, we exchange the order of the limit and the sum. Hence, we have

that

$$\begin{aligned}\lim_{\gamma_1,\gamma_2\to+\infty} q^{\star}_{M|\boldsymbol{X}}(m^{\star}) &= (m^{\star}+r)_{r\downarrow} q_M(m^{\star}+r) \\ &\times \left\{\sum_{m=0}^{\infty}(m)_{r\downarrow} q_M(m)\prod_{j=1}^{2}\lim_{\gamma_j\to+\infty}\frac{\Gamma(\gamma_j(m^{\star}+r)+n_j)}{\Gamma(\gamma_j m+n_j)}\frac{\Gamma(\gamma_j m)}{\Gamma(\gamma_j(m^{\star}+r))}\right\}^{-1} \\ &= (m^{\star}+r)_{r\downarrow} q_M(m^{\star}+r)\left\{\sum_{m=0}^{\infty}(m)_{r\downarrow} q_M(m)\prod_{j=1}^{2}\left(\frac{m^{\star}+r}{m}\right)^{n_j}\right\}^{-1},\end{aligned}$$

where the final equality holds because of the asymptotic of the ratio of gamma functions, $\Gamma(x+a)/\Gamma(x+b)\sim(x)^{a-b}$, for $x\to+\infty$. Moreover, we note that the normalising constant of $q^{\star}_{M|\boldsymbol{X},\infty}$ is the pEPPF computed in the corresponding limiting case. □

E.6.3 Proof of Equation (6.39) and Equation (6.42)

The proof follows the same steps given in Section E.6.1 but exploiting the posterior representation $(P_1,P_2)\mid\boldsymbol{X}$. The latter is provided in Colombi et al. (2024a) and reported here for sake of completeness.

$(P_1,P_2)\mid\boldsymbol{X}\overset{d}{=}(P_1^{\star},P_2^{\star})$, where each component is defined as

$$P_j^{\star}=\sum_{l=1}^{r}w_{j,l}\delta_{\tau_l^{\star}}+\sum_{l=r+1}^{r+M^{\star}}w_{j,l}\delta_{\tau_l},$$

where $n_{j,l}\geq 0$ are the observed counts and $\tau_l^{\star}$ are the corresponding species labels. Then, the random number of the unseen species $M^{\star}\sim q^{\star}_{M|\boldsymbol{X}}$ where $q^{\star}_{M|\boldsymbol{X}}$ has been defined in Section 6.4.1. The labels of the unseen species are $\tau_l\mid M^{\star}\overset{\text{iid}}{\sim}P_0(\mathrm{d}\tau)$ for $l=r+1,\ldots,r+M^{\star}$. Finally, the vector of posterior probabilities $\boldsymbol{w}_j^{\star}=(w_{j,1}^{\star},\ldots,w_{j,r+M^{\star}}^{\star})$ follows a Dirichlet distribution,

$$(w_{j,1}^{\star},\ldots,w_{j,r+M^{\star}}^{\star})\mid M^{\star}\sim\text{Dir}_{r+M^{\star}}\left(\gamma_j+n_{j,1},\ldots,\gamma_j+n_{j,r},\gamma_j,\ldots\gamma_j\right).$$

Furthermore, $P_1\perp\!\!\!\perp P_2\mid M^{\star},\boldsymbol{X}$.

The marginal distributions of $w_{j,l}^{\star}\mid M^{\star}$ are crucial in the derivation of result. Using the aggregation property of the Dirichlet distribution, we have that

$$w_{j,l}^{\star}\mid M^{\star}\sim\text{Beta}\left(a_{j,l},n_j+\gamma_j(r+M^{\star})-a_{j,l}\right),$$

where $a_{j,l}=n_{j,l}+\gamma_j$ for $l=1,\ldots,r$ and $a_{j,l}=\gamma_j$ for $l=r+1,\ldots,r+M^{\star}$.

The posterior expected value of the Simpson index is

$$\begin{aligned}E\left(\rho_j\mid\boldsymbol{X}\right) &= E\left(\sum_{l=1}^{r+M^{\star}}\left(w_{j,l}^{\star}\right)^2\mid\boldsymbol{X}\right)=E\left(E\left(\sum_{l=1}^{r+M^{\star}}\left(w_{j,l}^{\star}\right)^2\mid M^{\star},\boldsymbol{X}\right)\mid\boldsymbol{X}\right) \\ &= E_{q^{\star}_{M|\boldsymbol{X}}}\left(\sum_{l=1}^{r}E\left((w_{j,l}^{\star})^2\mid M^{\star},\boldsymbol{X}\right)+M^{\star}E\left((w_{j,r+1}^{\star})^2\mid M^{\star},\boldsymbol{X}\right)\right).\end{aligned}$$

To conclude the proof, it is enough to recall that the second moment of the marginal distributions

$w^\star_{j,l} \mid M^\star$ equals

$$E\left((w^\star_{j,r+1})^2 \mid M^\star, \boldsymbol{X}\right) = \frac{a_{j,l}}{(n_j + \gamma_j(r + M^\star))} \frac{(a_{j,l}+1)}{(n_j + \gamma_j(r + M^\star) + 1)} .$$

Moving to the posterior expected value of ρ_{12}, we follow similar steps and exploit the conditional independence of $(P_1, P_2) \mid \boldsymbol{X}$.

$$\begin{aligned} E(\rho_{12} \mid \boldsymbol{X}) &= E\left(\sum_{l=1}^{r+M^\star} w^\star_{1,l} w^\star_{2,l} \mid \boldsymbol{X}\right) = E\left(E\left(\sum_{l=1}^{r+M^\star} w^\star_{1,l} w^\star_{2,l} \mid M^\star, \boldsymbol{X}\right) \mid \boldsymbol{X}\right) \\ &= E_{q^\star_{M|\boldsymbol{X}}}\left(\sum_{l=1}^{r+M^\star} \left\{E\left(w^\star_{1,l} \mid M^\star, \boldsymbol{X}\right) E\left(w^\star_{2,l} \mid M^\star, \boldsymbol{X}\right)\right\}\right) \\ &= E_{q^\star_{M|\boldsymbol{X}}}\left(\sum_{l=1}^{r} \frac{(n_{1,l}+\gamma_1)}{(n_1 + \gamma_1(r + M^\star))} \frac{(n_{2,l}+\gamma_2)}{(n_2 + \gamma_2(r + M^\star))} + M^\star \frac{\gamma_1\gamma_2}{(n_1 + \gamma_1(r + M^\star))(n_2 + \gamma_2(r + M^\star))}\right) . \end{aligned}$$

The result follows after developing the sum in the numerator of the first term. Indeed,

$$\sum_{l=1}^{r} (\gamma_1 + n_{1,l})(\gamma_2 + n_{2,l}) = r\gamma_1\gamma_2 + \gamma_1 n_2 + \gamma_1 n_1 + \sum_{l=1}^{t} n_{1,l} n_{2,l} .$$

E.6.4 Proofs of Equation (6.40) and Equation (6.41)

Let us write the expected value with respect to $q^\star_{M|\boldsymbol{X}}$ as an infinite sum. We recall that, although it is not explicit in the notation, $q^\star_{M|\boldsymbol{X}}$ depends both of γ_1 and γ_2, see Equation (6.24). We use the dominated convergence theorem to exchange the limit and the sum, which is valid thanks to the upper bound provided in Section E.6.2.

Consider the limit for $\gamma_j \to 0$:

$$\begin{aligned} &\lim_{\gamma_j \to 0} E(\rho_j \mid \boldsymbol{X}) \\ &= \lim_{\gamma_j \to 0} \sum_{m^\star=0}^{\infty} \frac{\sum_{l=1}^{r} n_{j,l}(n_{j,l}+1) + \gamma_j\left(\gamma_j(r+m^\star) + r + m^\star + 2n_j\right)}{n_j(n_j+1) + \gamma_j^2(r+m^\star)^2 + \gamma_j(r+m^\star)(2n_j+1)} q^\star_{M|\boldsymbol{X}}(m^\star) \\ &= \sum_{m^\star=0}^{\infty} \lim_{\gamma_j \to 0} \frac{\sum_{l=1}^{r} n_{j,l}(n_{j,l}+1) + \gamma_j\left(\gamma_j(r+m^\star) + r + m^\star + 2n_j\right)}{n_j(n_j+1) + \gamma_j^2(r+m^\star)^2 + \gamma_j(r+m^\star)(2n_j+1)} q^\star_{M|\boldsymbol{X}}(m^\star) \\ &= \frac{\sum_{l=1}^{r} n_{j,l}(n_{j,l}+1)}{n_j(n_j+1)} \sum_{m^\star=0}^{\infty} \lim_{\gamma_j \to 0} q^\star_{M|\boldsymbol{X}}(m^\star) = \frac{\sum_{l=1}^{r} n_{j,l}(n_{j,l}+1)}{n_j(n_j+1)} . \end{aligned}$$

The final equality holds since $\sum_{m^\star=0}^{\infty} \lim_{\gamma_j \to 0} q^\star_{M|\boldsymbol{X}}(m^\star) = \sum_{m^\star=0}^{\infty} q^\star_{M|\boldsymbol{X},0}(m^\star) = 1$, see Section E.6.2.

Consider the limit for $\gamma_j \to \infty$:

$$\begin{aligned}
&\lim_{\gamma_j\to\infty} E\left(\rho_j \mid \boldsymbol{X}\right) \\
&= \lim_{\gamma_j\to\infty} \sum_{m^\star=0}^{\infty} \frac{\sum_{l=1}^{r} n_{j,l}(n_{j,l}+1) + \gamma_j\left(\gamma_j(r+m^\star)+r+m^\star+2n_j\right)}{n_j(n_j+1)+\gamma_j^2(r+m^\star)^2+\gamma_j(r+m^\star)(2n_j+1)} q^\star_{M|\boldsymbol{X}}(m^\star) \\
&= \sum_{m^\star=0}^{\infty} \lim_{\gamma_j\to\infty} \frac{\sum_{l=1}^{r} n_{j,l}(n_{j,l}+1) + \gamma_j\left(\gamma_j(r+m^\star)+r+m^\star+2n_j\right)}{n_j(n_j+1)+\gamma_j^2(r+m^\star)^2+\gamma_j(r+m^\star)(2n_j+1)} q^\star_{M|\boldsymbol{X}}(m^\star) \\
&= \sum_{m^\star=0}^{\infty} \frac{1}{r+m^\star} \lim_{\gamma_j\to\infty} q^\star_{M|\boldsymbol{X}}(m^\star) = E_{q^\star_{M|\boldsymbol{X}},\infty}\left(1/(r+M^\star)\right) ,
\end{aligned}$$

where $q^\star_{M|\boldsymbol{X},\infty}$, inf has been defined in Lemma E.6.2.

E.6.5 Proofs of Equation (6.43) and Equation (6.44)

We follow the same approach of Section E.6.4. Firstly, consider the limit for $\gamma_1, \gamma_2 \to 0$ following the same rate. This is equivalent to assume there exists some constants c_1, c_2 such that $\gamma_j = c_j\gamma$ and let γ goes to zero o infinity.

$$\begin{aligned}
&\lim_{\gamma_1,\gamma_2\to 0} E\left(\rho_{12} \mid \boldsymbol{X}\right) \\
&= \lim_{\gamma_1,\gamma_2\to 0} \sum_{m^\star=0}^{\infty} \left(\frac{\sum_{l=1}^{t} n_{1,l}n_{2,l} + \gamma_1\gamma_2(r+M^\star) + n_1\gamma_2 + n_2\gamma_1}{(\gamma_1(r+M^\star)+n_1)(\gamma_2(r+M^\star)+n_2)} \right) q^\star_{M|\boldsymbol{X}}(m^\star) \\
&= \sum_{m^\star=0}^{\infty} \lim_{\gamma_1,\gamma_2\to 0} \left(\frac{\sum_{l=1}^{t} n_{1,l}n_{2,l} + \gamma_1\gamma_2(r+M^\star) + n_1\gamma_2 + n_2\gamma_1}{(\gamma_1(r+M^\star)+n_1)(\gamma_2(r+M^\star)+n_2)} \right) q^\star_{M|\boldsymbol{X}}(m^\star) \\
&= \frac{\sum_{l=1}^{t} n_{1,l}n_{2,l}}{n_1n_2} \sum_{m^\star=0}^{\infty} \lim_{\gamma_j\to 0} q^\star_{M|\boldsymbol{X}}(m^\star) = \frac{\sum_{l=1}^{t} n_{1,l}n_{2,l}}{n_1n_2} .
\end{aligned}$$

Consider the limit for $\gamma_1, \gamma_2 \to \infty$ following the same rate:

$$\begin{aligned}
&\lim_{\gamma_1,\gamma_2\to\infty} E\left(\rho_{12} \mid \boldsymbol{X}\right) \\
&= \lim_{\gamma_j\to\infty} \sum_{m^\star=0}^{\infty} \left(\frac{\sum_{l=1}^{t} n_{1,l}n_{2,l} + \gamma_1\gamma_2(r+M^\star) + n_1\gamma_2 + n_2\gamma_1}{(\gamma_1(r+M^\star)+n_1)(\gamma_2(r+M^\star)+n_2)} \right) q^\star_{M|\boldsymbol{X}}(m^\star) \\
&= \sum_{m^\star=0}^{\infty} \lim_{\gamma_1,\gamma_2\to\infty} \left(\frac{\sum_{l=1}^{t} n_{1,l}n_{2,l} + \gamma_1\gamma_2(r+M^\star) + n_1\gamma_2 + n_2\gamma_1}{(\gamma_1(r+M^\star)+n_1)(\gamma_2(r+M^\star)+n_2)} \right) q^\star_{M|\boldsymbol{X}}(m^\star) \\
&= \sum_{m^\star=0}^{\infty} \frac{1}{r+m^\star} \lim_{\gamma_1,\gamma_2\to\infty} q^\star_{M|\boldsymbol{X}}(m^\star) = E_{q^\star_{M|\boldsymbol{X}},\infty}\left(1/(r+M^\star)\right) .
\end{aligned}$$

The exchange of limit and series is valid, as shown in Section E.6.2.

Chapter 7

Dynamic Features Allocation

This chapter is a joint work with Raffaele Argiento and Jim Griffin.

7.1 Exponential completely random measures

We begin introducing the notion of *exponential families* for CRMs, as defined in Broderick et al. (2018). In what follows, we use terms *factors*, *features* and *traits* interchangeably, to remain consistent with the terminology commonly used in the related literature.

Let $\mathbb{X}$ be a Polish space of traits and let G be a CRM on $\mathbb{X}$ as defined in Section 7.1. In the literature, G is typically employed as a Bayesian nonparametric prior. We now consider a sampling model that generates n i.i.d. observations Θ_i, drawn conditionally on G. The independence assumption will be relaxed in Section 7.2 to allow for time dependency. Each observation Θ_i is viewed as a discrete measure whose atoms are fixed and correspond to the support points of G, and whose jumps $\xi_{i,l}$ have a distribution with parameters given by the jumps of G. More in detail,

$$\Theta_i \mid G \overset{\text{d}}{=} \sum_{l\geq 1} \xi_{i,l}\, \delta_{\lambda_l}\,, \quad i = 1, \ldots, n \tag{7.1}$$

with $\xi_{i,l}$ are generated from some distribution $H(\cdot \mid s_l)$, independently across i and l. In this chapter, we assume that the jumps $\xi_{i,l}$ are discrete and we let $h_l(\cdot)$ be the probability mass function of $H(\cdot \mid s_l)$. The pairs $(\xi_{i,l}, \lambda_l)$ have a straightforward interpretation: $\xi_{i,l}$ represents the degree to which observation i possesses the trait λ_l, where $\xi_{i,l} = 0$ indicates that it does not possess it at all. We emphasize that every atom in Θ_i is supported on an atom of G, but not every atom of G necessarily appears in Θ_i, since $\xi_{i,l}$ may be equal to zero.

In Section 4.2.3, we stated that a crucial assumption for well-defined inference is that each Θ_i carries finite information. Broderick et al. (2018) provide a generalization of the condition in Equation (4.5), which specifies the relationship that must hold between the density $h_l(\cdot)$ and the Lévy intensity ρ_{ord} in order to guarantee that this assumption is satisfied.

$$\sum_{x=1}^{\infty} \int_0^{\infty} h(x \mid s)\rho(s)\mathrm{d}s \;=\; \int_0^{\infty} (1 - h(0 \mid s))\rho(s)\mathrm{d}s < +\infty\,. \tag{7.2}$$

The Exponential family of Completely Random Measures (ExpCRMs) provides a convenient and

powerful framework for this task, primarily due to its property of automatic conjugacy (Broderick et al., 2018). To clarify the notation, we also refer to Diaconis and Ylvisaker (1979) that defines the class of conjugate prior in the parametric framework. Firstly, we focus on the class of sampling models by assuming that

$$h_l(x) = \kappa(x) \exp\{\langle \eta(s_l), \phi(x) \rangle - A(s_l)\} \tag{7.3}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product, κ is the base density, η is the natural parameter function, ϕ is the sufficient statistics and A is the log-partition function, or, equivalently, $\exp\{A(s_l)\}$ is the normalizing constant. We write $\Theta_i \mid G \sim \text{ExpCRM}(\kappa, \eta, \phi, A)$ to refer to a CRM of the form given in Equation (7.1) with jump distribution h_l having the parametric for reported in Equation (7.3). Then, automatic conjugacy is achieved setting ρ_{ord} equal to

$$\rho_{\text{ord}}(s) = \gamma \exp\{\langle a, \eta(s) \rangle + b\,(-A(s))\} \tag{7.4}$$

where a and b are atom specific hyperparameters and η and A are the same functions of Equation (7.3). We write $G \sim \text{Conj-CRM}(\gamma, a, b, P_0)$ to refer to a CRM of the form given in Equation (4.3) without fixed atoms, with an ordinary component ρ_{ord} having the parametric for reported in Equation (7.4) and i.i.d. atoms distributes as P_0.

This specific construction guarantees that the posterior distribution of G is represented as a CRM made of a part with fixed atoms and an ordinary component that belongs to the same parametric family of the one in Equation (7.4), with updated parameters. Furthermore, the possibility of having the posterior distribution in analytical form is not the only advantage of the conjugacy. As a by-product, Broderick et al. (2018) also derive the analytical expressions of marginal and size-biased representations of the Θ_i's, which are crucial in the remainder of this work.

This class of models is quite broad and encompasses well-known examples such as the Beta–Bernoulli process (Ghahramani and Griffiths, 2005) and the Gamma–Poisson process (Titsias, 2007), both previously introduced in Section 7.1. We now show that these models belong to the ExpCRM family, thereby establishing the notation that will be used throughout this chapter.

Example (Beta-Bernoulli process)**.** The Beta-Bernoulli process describes which of the infinitely many possible features are exhibited by each individual i, for $i = 1, \ldots, n$. In particular, $\xi_{i,l} \in \{0, 1\}$ and a Bernoulli likelihood is adopted. So that $\xi_{i,l} \mid s_l \overset{\text{ind}}{\sim} \text{Be}(s_l)$. Specifically,

$$h(x \mid s) = \exp\left\{\log\left(\frac{s}{1-s}\right) x + \log(1-s)\right\}.$$

Hence, $\kappa(x) = 1$, $\eta(s) = \log(s/(1-s))$, $\phi(x) = x$, and $A(s) = -\log(1-s)$. The automatic prior specification yields

$$\rho_{\text{ord}}(s) = \gamma \exp\{a \log\left(\frac{s}{1-s}\right) + b \log(1-s)\} = \gamma\, s^a\, (1-s)^{b-a}.$$

The corresponding prior is known as the three-parameters Beta process, introduced in Teh and Gorur (2009). It is well-defined for $a \in (-2, -1]$, $b > a - 1$ and $\gamma > 0$. Following Ghahramani and Griffiths (2005), in many applications, as well as in Section 7.5, it is common to employ the two-parameters version of the Beta process, as originally introduced by Hjort (1990). The latter, which is simply known as the Beta process, is recovered by setting $a = -1$.

Example (Poisson-Gamma process)**.** When the aim is not only to account for the presence or absence of traits but also to quantify their intensity, a Poisson likelihood is adopted, $\xi_{i,l} \overset{\text{ind}}{\sim} \text{Pois}(s_l)$. Namely,

$$h(x \mid s) = \frac{1}{x!} \exp\{\log(s)x - s\} .$$

Accordingly, $\kappa(x) = 1/x!$, $\phi(x) = x$, $\eta(s) = \log(s)$ and $A(s) = s$. The automatic prior choice leads to

$$\rho_{\text{ord}}(s) = \gamma \exp\{\alpha \log(s) - \beta s\} = \gamma\, s^{\alpha}\, e^{-\beta s}, \tag{7.5}$$

and it is well-defined for $\alpha \in (-2, -1]$, $\beta > 0$ and $\gamma > 0$. The CRM defined through the Lévy intensity in Equation (7.5) is known as generalized gamma process, which admits several different parametrizations. See Hougaard (1986); Brix (1999); Lijoi et al. (2007b); Lee et al. (2016)

7.2 Time-dependent trait allocation model

Section 7.1 illustrates how to exploit conjugacy in the classical setting of a Bayesian nonparametric prior G and discrete measures Θ_i that are i.i.d. conditional on G. In the present section, we depart from the i.i.d. assumption and aim to construct a model for discrete measures of the form in Equation (7.1) that evolve over time. As anticipated in Section 4.3.1, we build the temporal evolution of the random measures by leveraging the works by Pitt et al. (2002) and Pitt and Walker (2005). We first recap their general construction in the parametric framework and then move on to generalize it to our problem.

For the sake of notation, in the following, all the random objects X are defined over a suitable space. Moreover, $p_X(\mathrm{d}x)$ denotes the probability law (i.e., the distribution) of the random object X. In the same way we write $p_X(x)\mathrm{d}x$ when the distribution is dominated by a measure on the support of X.

The goal in Pitt et al. (2002) is to define a stationary sequence (W_t) for $t = 1, \dots, T$ with a given invariant distribution $p_W(w)$. This is achieved by considering an additional latent variable Z such that $p_{W,Z}(w, z) = p_Z(z)p_{W|Z}(w \mid z) = p_W(w)p_{Z|W}(z \mid w)$. A sampling scheme to generate (W_t) works as follows: first generate $z_0 \sim p_Z(\cdot)$; then iterate, for $t = 1, \dots, T$, by sampling $w_t \sim p_{W|Z}(\cdot \mid z_{t-1})$ and $z_t \sim p_{Z|W}(\cdot \mid w_t)$. The latter is trivially a Gibbs sampling strategy to draw a Markov chain (W_t, Z_t) with invariant distribution $p_{W,Z}$. Consequently, the marginal process (W_t) is itself a Markov chain with invariant distribution p_W. The corresponding transition equation for the marginal chain is

$$Q(w_{t+1} \mid w_t) = \int p_{W|Z}(w_{t+1} \mid z_t)\, p_{Z|W}(\mathrm{d}z_t \mid w_t), \tag{7.6}$$

which ensures stationarity of (W_t) by construction, regardless of the choice of the conditional distribution $p_{Z|W}$. However, certain choices of the conditional distributions are more convenient, as they lead to transition kernels that are analytically tractable. The integral in Equation (7.6) involves two conditional distributions that can be interpreted in the Bayesian framework. Indeed, the first, $p_{W|Z}$, is a sampling model for W given the latent parameter Z. Given a prior p_Z, the second conditional distribution, $p_{Z|W}$, is the corresponding posterior. Consequently, $Q(w_{t+1} \mid w_t)$ in Equation (7.6) is the predictive distribution. Building on this interpretation Pitt et al. (2002) developed a general theory to build stationary time series with marginal corresponding to the a priori predictive $p_W(w) = \int p_{W|Z}(w \mid x)p_Z(\mathrm{d}z)$. For instance, using the Diaconis and Ylvisaker (1979) approach, it is straightforward to build time series with marginal

distribution p_W corresponding to the a priori predictive of a parametric conjugate Bayesian model.

We extend the Pitt et al. (2002) idea to our setting, to build a time series for a time varying trait profile measure Θ_t with a given marginal distribution. As latent parameter, we consider G_t to be an ExpCRM, which yields a model that is both highly flexible, thanks to the nonparametric structure induced by completely random measures, and mathematically tractable, due to the conjugacy property guaranteed by this family. To fix the notation, for any $t \geq 1$, let Θ_t denote the *trait profile measure* at time t, consisting of K_t active traits (or features), denoted by $\boldsymbol{\lambda}_t^* = \left\{\lambda_{t,1}^*, \ldots, \lambda_{t,K_t}^*\right\}$. Each trait is associated with a strictly positive weight (or jump) $\xi_{t,l}$, $l = 1, \ldots, K_t$, representing the degree to which the l-th trait is present at time t. As in Section 7.1, we assume these jumps to take integer values. Our goal is to construct a transition kernel $Q\left(\mathrm{d}\Theta_{t+1} \mid \Theta_t\right)$ that describes the temporal evolution of the trait profile measure, specifying which traits remain active, how their intensities change, and whether new traits emerge. We define $Q(\mathrm{d}\Theta_{t+1} \mid \Theta_t)$ as the predictive distribution of an underlying Bayesian nonparametric conjugate model. Namely,

$$Q(\mathrm{d}\Theta_{t+1} \mid \Theta_t) = \int p_{\Theta|G}\left(\mathrm{d}\Theta_{t+1} \mid G_t\right) p_{G|\Theta}\left(\mathrm{d}G_t \mid \Theta_t\right), \tag{7.7}$$

for an additional latent random measure G_t. As time progresses, Equation (7.7) is equivalent to introduce a latent process (G_t) with the following Markovian transitions

$$\begin{aligned} G_0 &\sim p_G(\mathrm{d}G_0) \\ \Theta_t \mid G_{t-1} &\sim p_{\Theta|G}(\mathrm{d}\Theta_t \mid G_{t-1}), \quad t = 1, \ldots, T \\ G_t \mid \Theta_t &\sim p_{G|\Theta}(\mathrm{d}G_t \mid \Theta_t), \quad t = 1, \ldots, T. \end{aligned} \tag{7.8}$$

Following the same interpretation as in the parametric case, we can regard p_G as the prior, $p_{\Theta|G}$ as the sampling model, and $p_{G|\Theta}$ as the posterior. This formulation implicitly defines p_Θ as the prior predictive of the Bayesian model, and, following Pitt et al. (2002), it is trivial that the time series process in Equation (7.8) constitutes a stationary and reversible Markov chain for any choice of the conditional laws, as it satisfies the following detailed balance condition:

$$Q(\mathrm{d}\Theta_{t+1} \mid \Theta_t) p_\Theta(\mathrm{d}\Theta_t) = Q(\mathrm{d}\Theta_t \mid \Theta_{t+1}) p_\Theta(\mathrm{d}\Theta_{t+1})$$

The analytical form of the transition kernel in Equation (7.7) is obtained by computing the posterior distribution $p_{G|\Theta}$ and the corresponding posterior predictive. A convenient specification is to assume a sampling model in the exponential family of completely random measures, namely, by $p_{\Theta|G} = \mathrm{ExpCRM}(\kappa, \eta, \phi, A)$ we denote that $p_{\Theta|G}$ is law of a ExpCRM, while $p_G = \text{Conj-CRM}(\gamma, a, b, P_0)$ is its conjugate prior. It follows that, the posterior distribution $p_{G|\Theta}$ admits the following representation: $G' \sim p_{G|\Theta}$ if and only if it is a CRM whose fixed component $G'_{\text{fix}} = \sum_{l=1}^{K_t} s_l^* \, \delta_{\lambda_{t,l}^*}$ has jumps independently distributed according to the density

$$\rho_{\text{fix},l}^*(s) = \exp\left\{\langle a + \phi(x_{t,l}), \eta(s)\rangle + (b+1)\left(-A(s)\right) - C(a + \phi(x_{t,l}), b+1)\right\}, x > 0$$

where $C(a_l, b_l)$ is a function defined as

$$\exp\{C(a_l, b_l)\} = \int_{\mathbb{R}^+} \exp\{\langle a_l, \eta(s)\rangle + b_l(-A(s))\}\, \mathrm{d}s\,. \tag{7.9}$$

Moreover, the ordinary component $G'_{\text{ord}} = \sum_{l\geq 1} s'_l\, \delta_{\lambda_{t,l}}$ has Lévy intensity

$$\rho^*_{\text{ord}}(s) = \gamma\kappa(0)\, \exp\{\langle a + \phi(0), \eta(s)\rangle + (b+1)\,(-A(s))\}\,. \tag{7.10}$$

Notably, ρ^*_{ord} belongs to the same family as ρ_{ord} with updated parameters $\gamma\kappa(0)$, $a + \phi(0)$ and $b + 1$. In the following section, we detail the analytical form of the transition kernel in Equation (7.7). Namely, we characterize the evolution of the process (Θ_t) under our choice for $p_{\Theta|G}$ and p_G.

7.2.1 Transition kernel

We begin with the first time point, $t = 1$, where no conditioning on previous observations is required. In this case, Equation (7.7) reduces to

$$p(\mathrm{d}\Theta_1) = \int p_{\Theta|G}\,(\mathrm{d}\Theta_1 \mid G)\; p_{G|\Theta}\,(\mathrm{d}G \mid \Theta_1)\,, \tag{7.11}$$

As is common in the point process framework, evaluation of the integral in Equation (7.11) is given by describing how to sample of $\Theta_1 \sim p(\mathrm{d}\Theta_1)$: first, a Poisson process Φ with Lévy intensity $\rho_{\text{ord}}(s)$ is considered. Next, for any $x \geq 1$, Φ is thinned using the retention function $p_{th,x}(s) = h(x \mid s)$; the resulting thinned process is denoted by $\Phi_{th,x}$. By the thinning property of Poisson (Kingman, 1993) processes, $\Phi_{th,x}$ is itself a Poisson process with Lévy intensity $h(x \mid s)\rho_{\text{ord}}(s)$. Let $N_{1,x}$ be the number of points that survive the thinning then, $N_{1,x} \sim \text{Pois}(M_x)$, where $M_x = \int_{\mathbb{R}^+} h(x \mid s)\rho_{\text{ord}}(s)\mathrm{d}s$, which is finite thanks to the assumption in Equation (7.2). Specifically, when $p_{\Theta|G} = \text{ExpCRM}(\kappa, \eta, \phi, A)$ and $p_G = \text{Conj-CRM}(\gamma, a, b, P_0)$, then

$$M_x = \gamma\,\kappa(x)\, \exp\{C(a + \phi(x), b + 1)\}\,.$$

Finally, the marks $\lambda_{1,l} \overset{\text{iid}}{\sim} P_0$ are attached to $\Phi_{th,x}$ resulting in the point process $\Psi_{th,x} = \sum_{l=1}^{N_1} \delta_{x,\lambda_{1,l}}$. The latter collects the traits whose degree is equal to x for all $x \geq 1$. Superposing for all possible values of x, we have

$$\Theta_1 \overset{\text{d}}{=} \sum_{x=1}^{\infty} \sum_{l=1}^{N_{1,x}} x\, \delta_{\lambda_{1,l}}\,.$$

where N_1 is the number of new traits, and $N_1 = \sum_{x\geq 1} N_{1,x} \sim \text{Pois}(\sum_{x\geq 1} M_x)$, which is well defined thanks to assumption in Equation (7.2).

We now turn to the general case of $Q(\mathrm{d}\Theta_{t+1} \mid \Theta_t)$ for any $t \geq 1$. Let the conditioning trait measure Θ_t to consist of K_t active traits, denoted by $\boldsymbol{\lambda}^*_t = \left\{\lambda^*_{t,1}, \ldots, \lambda^*_{t,K_t}\right\}$, and let $\boldsymbol{\xi}_t = \left\{\xi_{t,1}, \ldots, \xi_{t,K_t}\right\}$ be the corresponding trait degrees. Since the posterior distribution $p_{G|\Theta}$ is the superposition of two independent components, the fixed and the ordinary part, the representation of $\Theta_{t+1} \mid \Theta_t$ also decomposes into two independent contributions: (i) thinning of the active traits at time t, and (ii) innovation from traits that have never been observed before. The innovation part is analogous to the case $t = 1$, except that the integral is computed with respect to the posterior $p_{G|\Theta}$, which uses the updated hyperparameters given

in Equation (7.10). The thinning part is obtained computing finite-dimensional integrals, which can be interpreted as the parametric predictive distribution of a Bayesian model with likelihood $h(\cdot \mid s)$ and posterior density proportional to $h(\xi_{t,l} \mid s)\rho_{\text{ord}}(s)\mathrm{d}s$. Adding the two parts together, we obtain

$$\Theta_{t+1} \mid \Theta_t \overset{\text{d}}{=} \sum_{l=1}^{K_t} \xi'_{t+1,l}\, \delta_{\lambda^*_{t,l}} + \sum_{x=1}^{\infty} \sum_{l=1}^{N_{t+1,x}} x\, \delta_{\lambda_{t+1,l}}\,, \tag{7.12}$$

where $N_{t+1,x}$ is Poisson distributed with mean $M^*_x = \int_{\mathbb{R}^+} h(x \mid s)h(0 \mid s)\rho_{\text{ord}}(s)\mathrm{d}s$ and $\xi'_{t+1,l} \overset{\text{ind}}{\sim} h'(\cdot \mid \xi_{t,l}) \propto \int_{\mathbb{R}^+} h(x \mid s)h(\xi_{t,l} \mid s)\rho_{\text{ord}}(s)\mathrm{d}s$. Moreover, $N_{t+1} = \sum_{x\geq 1} N_{t+1,x}$ that is the total number of traits at time $t+1$, is Poisson distributed with mean $\sum_{x\geq 1} M^*_x$. We extended the definition of $h'(\cdot \mid \xi_{t,l})$ by setting $h'(x \mid 0) = \mathbf{1}(x = 0)$. This implies that $\xi_{t,l} = 0$ entails $\xi_{s,l} = 0$ for all $s > t$, namely, if a trait is thinned at time t, it can not appear any more with probability one. This is consistent with the fact that atoms λ_j's are i.i.d. samples from the diffuse probability measure P_0, and therefore the same trait cannot reappear more than once. The previous characterization as thinning and innovation and does not depend on the specific choice of $p_{\Theta|G}$ and p_G. Specifically, when $p_{\Theta|G} = \text{ExpCRM}(\kappa, \eta, \phi, A)$ and $p_G = \text{Conj-CRM}(\gamma, a, b, P_0)$, we have

$$h'(x \mid \xi_{t,l}) = \kappa(x)\, \exp\left\{-C(a + \Phi(\xi_{t,l}), b+1) + C(a + \Phi(x) + \Phi(\xi_{t,l}), b+2)\right\}\,, \tag{7.13}$$

$$M^*_x = \gamma\, \kappa(0)\, \kappa(x)\, \exp\left\{C(a + \phi(0) + \phi(x), b+2)\right\}\,. \tag{7.14}$$

Example (Beta-Bernoulli process, continued)**.** By exploiting the conjugacy of the Beta–Bernoulli model, it is straightforward to verify that the function $C(a_l, b_l)$, as defined in Equation (7.9), is given by $\exp C(a_l, b_l) = B(a_l + 1, b_l - a_l + 1)$, where $B(a, b)$ is the Beta function. Since the Bernoulli process likelihood has support only at $x = 0$ and $x = 1$, it follows that the marginal process for $t = 1$ is

$$\Theta_1 \overset{\text{d}}{=} \sum_{l=1}^{N_1} \delta_{\lambda_{t,l}}\,,$$

where $N_1 \sim \text{Pois}(M_1)$ with $M_1 = \gamma\, B(a+2, b-a+1)$. Moreover, for $t \geq 1$, we have

$$\Theta_{t+1} \mid \Theta_t \overset{\text{d}}{=} \sum_{l=1}^{K_t} \xi'_{t+1,l}\, \delta_{\lambda^*_{t,l}} + \sum_{l=1}^{N_{t+1}} \delta_{\lambda_{t+1,l}}\,,$$

where $N_{t+1} \sim \text{Pois}(M^*_1)$ with $M^*_1 = \gamma\, B(a+2, b-a+2)$, As far as the distribution of the jumps is concerned, note that $\xi_{t,l} = 1$ for any t, l since they are associated to active features at each time tHence, $\xi'_{t+1,l} \sim \text{Be}\left((a+2)/(b+3)\right)$. In the two-parameters Beta process, the previous quantities simplify to

$$M_1 = \frac{\gamma}{b+2}\,, \quad M^*_1 = \frac{\gamma}{b+3}\,, \quad \xi'_{t+1,l} \sim \text{Be}\left(\frac{1}{b+3}\right)\,.$$

Example (Poisson-Gamma process, continued)**.** Under the Poisson-Gamma process the function $C(a_l, b_l)$, defined in Equation (7.9), is given by $\exp\{C(a_l, b_l)\} = \Gamma(a_l + 1)/(b_l)^{a_l+1}$. Then, for $t = 1$, the marginal process is

$$\Theta_1 \overset{\text{d}}{=} \sum_{x=1}^{\infty} \sum_{l=1}^{N_{1,x}} x\, \delta_{\lambda_{t,l}}\,,$$

where $N_{1,x} \sim \text{Pois}(M_x)$, with $M_x = (\gamma\,\Gamma(\alpha+x+1))\,/\left(x!\;(\beta+1)^{\alpha+x+1}\right)$. For $t \geq 1$, it holds that

$$\Theta_{t+1} \mid \Theta_t \overset{\text{d}}{=} \sum_{l=1}^{K_t} \xi'_{t+1,l}\,\delta_{\lambda^*_{t,l}} + \sum_{x=1}^{\infty}\sum_{l=1}^{N_{t+1,x}} x\,\delta_{\lambda_{t+1,l}}\,,$$

where $N_{t+1,x} \sim \text{Pois}(M^*_x)$ with $M^*_x = (\gamma\,\Gamma(\alpha+x+1))\,/\left(x!\;(\beta+2)^{\alpha+x+1}\right)$. Regarding the jumps distribution, we have

$$\begin{aligned} h'(x \mid \xi) &= \frac{1}{x!}\frac{\Gamma(\alpha+\xi+x+1)}{\Gamma(\alpha+\xi+1)}\frac{(\beta+1)^{\alpha+\xi+1}}{(\beta+2)^{\alpha+\xi+1+x}} \\ &= \frac{\Gamma(\alpha+\xi+x+1)}{\Gamma(x+1)\Gamma(\alpha+\xi+1)}\left(\frac{\beta+1}{\beta+2}\right)^{\alpha+\xi+1}\left(\frac{1}{\beta+2}\right)^{x}. \end{aligned}$$

It follows that the jumps $\xi'_{t+1,l}$ are negative binomial distributed, with success probability parameter $1-1/(\beta+2)$ and size parameter $\alpha+\xi_{t,l}+1$. We write,

$$\xi'_{t+1,l} \mid \xi_{t,l} \sim \text{NegBin}\left(\alpha+\xi_{t,l}+1, \frac{\beta+1}{\beta+2}\right).$$

7.2.2 Traits dynamic

Let $t = 1$ and for any $x \geq 1$ let $N_{1,x} \sim \text{Pois}(M_x)$ be the number of new traits, each with degree x, independently from the probability measure P_0. Moreover, let $N_1 = \sum_{x\geq 1} N_{1,x}$ denote the total number of new traits at time $t = 1$. The resulting pairs of traits and their corresponding degrees are collected in the set $\left\{(\lambda^*_{1,1},\xi_{1,1}),\ldots,(\lambda^*_{1,K_1},\xi_{1,K_1})\right\}$, where K_1 is the total number of traits at time $t = 1$. Clearly, $K_1 = N_1$ since all traits are new at this initial stage. For notational convenience, we separate this set into $\boldsymbol{\lambda}^*_1 = \left\{\lambda^*_{1,1},\ldots,\lambda^*_{1,K_1}\right\}$ and $\boldsymbol{\xi}_1 = \left\{\xi_{1,1},\ldots,\xi_{1,K_1}\right\}$, with the implicit assumption that the order in the two sets matches the original pairing; that is, the l-th elements of the two sets correspond to the pair $(\lambda^*_{1,l},\xi_{1,l})$ in the original collection. We refer to $\boldsymbol{\lambda}^*_1$ as the set of active traits at time $t = 1$ since they all relate to positive jumps $\xi_{1,l}$.

Let us now examine the changes to the trait profile induced by the transition kernel in Equation (7.12) when moving from time t to $t+1$. Since $h'(0 \mid \xi_{t,l}) > 0$, some traits in $\boldsymbol{\lambda}^*_t$ will have $\xi'_{t+1,l} = 0$ and are thus removed from the set of active traits, whereas the remaining traits in $\boldsymbol{\lambda}^*_t$ are retained, although their degrees may change from $\xi_{t,l}$ at time t to $\xi'_{t+1,l}$ at time $t+1$. We refer to this mechanism as *thinning*. In addition, the *innovation* step introduces $N_{t+1,x}$ new traits with degree x, where $N_{t+1,x}$ follows a Poisson distribution with mean M^*_x. Let $N_{t+1} = \sum_{x\geq 1} N_{t+1,x}$ denote the total number of new traits at time $t+1$. The set of active traits at time $t+1$ is then defined as the union of: (i) the traits that remain active after the thinning step, and (ii) the newly generated traits from the innovation step. Namely,

$$\boldsymbol{\lambda}^*_{t+1} = \bigcup_{l=1}^{K_t}\left\{\lambda^*_l \,:\, \xi'_{t+1,l} > 0\right\} \cup \left\{\lambda_{t+1,1},\ldots,\lambda_{t+1,N_{t+1}}\right\} = \left\{\lambda^*_{t+1,1},\ldots,\lambda^*_{2,K_{t+1}}\right\},$$

where K_{t+1} is the total number of traits at time $t+1$. We then let $\boldsymbol{\xi}_{t+1} = \left\{\xi_{t+1,1},\ldots,\xi_{t+1,K_{t+1}}\right\}$ be the set of degrees corresponding to the traits in $\boldsymbol{\lambda}^*_{t+1}$. We recall that we are assuming that the order of the two sets matches the true pairs; hence, $\xi^*_{t+1,l} = \xi'_{t+1,l}$ for all l whose traits were not eliminated from the set of

active features at the previous time step.

Finally, let $\boldsymbol{\Theta}_{1:T}$ be the observed set of features collected over the whole time period. The latter is fully represented by the sets $\left\{(\lambda^*_{t,1}, \xi_{t,1}), \ldots, (\lambda^*_{t,K_t}, \xi_{t,K_t})\right\}$, which we split into the two sets $\boldsymbol{\lambda}^*_t$ and $\boldsymbol{\xi}_t$, for $t = 1, \ldots, T$. The trait evolution we have just described implies that the same trait may appear at multiple time points. We therefore let $\boldsymbol{\lambda}^{**}$ denote the union over all times, i.e., $\boldsymbol{\lambda}^{**} = \cup_{t=1}^{T} \boldsymbol{\lambda}^*_t = \left\{\lambda^{**}_1, \ldots, \lambda^{**}_K\right\}$, where $K = \sum_{t=1}^{T} N_t$ is the total number of observed features. Note that $\boldsymbol{\lambda}^{**}$ does not contain any information about the specific times at which each trait has been observed; it only records that it has been observed at least once.

7.2.3 Properties

In this section we provide some fundamental prior properties of the proposed model. All proofs are deferred to Section F.1.

Proposition 4. *Let $\Theta_t \mid G_t = \sum_{l \geq 1} \xi_{t,l}\, \delta_{\lambda_{t,l}}$ be a CRM as defined as in Section 7.2. Then, for any $t \geq 1$, the prior expected value of the number of active traits is*

$$E\left(\sum_{l=1}^{\infty} \mathbf{1}\left(\xi_{t,l} > 0\right)\right) = \sum_{x=1}^{\infty} M_x\,, \tag{7.15}$$

Moreover, the expected value of the total mass is

$$E\left(\Theta_t(\mathbb{X})\right) = \sum_{x=1}^{\infty} x\, M_x\,, \tag{7.16}$$

where M_x is defined in Equation (7.2.1).

The transition kernel in Equation (7.12) shows a Markovian structure as Θ_{t+1} solely depends on Θ_t. As a further analysis, we now give the one-step ahead characterization of the process.

Proposition 5. *Let $\tilde{s} = \eta(s)$ be the natural parameter of the exponential family in Equation (7.3). Suppose there exists $\tilde{a}$ and $\tilde{b}$ such that*

$$\frac{1}{\eta'(\eta^{-1}(\tilde{s}))} = \exp\left\{\langle \tilde{a}, \tilde{s}\rangle - \tilde{b} A(\eta^{-1}(\tilde{s}))\right\}\,,$$

for $\eta'(s) = \mathrm{d}\eta(s)/\mathrm{d}s$. Then, for any $t \geq 1$ and for any measurable set A, we have

$$\begin{aligned} E\left(\Theta_{t+1}(A) \mid \Theta_t\right) &= \sum_{l=1}^{K_t} \frac{a + \tilde{a} + \xi_{t,l}}{b + \tilde{b} + 1}\, \delta_{\lambda^*_{t,l}}(A) + P_0(A) \sum_{x=1}^{\infty} x\, M^*_x \\ &= \frac{1}{b + \tilde{b} + 1}\Theta_t(A) + \frac{a + \tilde{a}}{b + \tilde{b} + 1} \sum_{l=1}^{K_t} \delta_{\lambda^*_{t,l}}(A) + P_0(A) \sum_{x=1}^{\infty} x\, M^*_x\,, \end{aligned}$$

where M^*_x is defined in Equation (7.14).

A closer inspection of Proposition 5 shows that the process exhibits a quasi-AR structure: it depends linearly on the one-step lagged value of the process, Θ_t, plus a constant term. We refer to it as quasi-AR

because it also involves an additional term, $\sum_{l=1}^{K_t} \delta_{\lambda^*_{t,l}}(A)$, which is not constant, as it allocates mass at the same location as Θ_t.

We now turn our attention to the evolution of the degree of each trait, which follow a Markov process.

Proposition 6. *Let $\xi_{l,t}$ be the l-th trait associated to λ^{**}_l, which we suppose to first appear in at time τ_l. Then, it follows a Markov's process with initial distribution*

$$h_0(x) = \frac{\kappa(x)\exp\{C(a+\phi(0)+\phi(x), b+2)\}}{\exp\{C(a+\phi(0), b+1)\} - \kappa(0)\exp\{C(a+2\phi(0), b+2)\}}\mathbf{1}(x \geq 1)\,, \tag{7.17}$$

and transition equation $h'(x \mid \xi_{t,l})$ given in Equation (7.13), for ant $t > \tau_l$ and $\xi_{t,l} > 0$. In particular, zero is an absorbing state and all other states are transient.

From the previous proposition, since zero is an absorbing state, once the intensity of a trait vanishes, that trait disappears permanently. Consequently, each trait is associated with a random activity time, which we denote by L_l. Specifically, if the l-th trait first appears at time τ_l and remains active for L_l time steps, then $\xi_{\tau_l+s,l} > 0$ for all $s \in [0, L_l]$, whereas $\xi_{\tau_l+L_l+1,l} = 0$. The next proposition characterizes the distribution of L_l.

Proposition 7. *Let $\{\xi_{t,l}\}_{t\geq 0}$ denote the Markov process for the l-th trait described in Proposition 5. Define the initial distribution vector $\boldsymbol{a}$ with entries $a_0 = 0$ and $a_x = h_0(x)$ for $x \geq 1$. Let matrix S be the transition probabilities, whose entries are $S_{i,j} = h'(j \mid i)$. Denote by $S^{(m)}$ the l-th power of S, whose (i,j)-th entry $S^{(m)}_{i,j}$ represents the probability of reaching state j from state i in m steps. Then, the distribution of the trait lifetime L_l defined as above is*

$$\mathbb{P}(L_l = m) = \begin{cases} \boldsymbol{e}_0^T S^T \boldsymbol{a}\,, & \text{if } m = 1, \\ \boldsymbol{e}_0^T \left(S^{(m)} - S^{(m-1)}\right)^T \boldsymbol{a}\,, & \text{if } m \geq 1\,, \end{cases} \tag{7.18}$$

where $\boldsymbol{e}_0$ is the column vector with first entry equal to one and all other entries equal to zero.

We conclude with a remark on the behavior of the total number of traits, denoted by K. Since the number of newly introduced features at each time point is independent across t, it follows that K has a Poisson distribution:

$$K \sim \text{Pois}\left(\sum_{x\geq 1}\left[M_x + (T-1)M^*_x\right]\right). \tag{7.19}$$

In particular, the expected growth of K is linear in T. This behavior is a direct consequence of stationarity: from Equation (7.12), the number of new traits introduced at each time step is Poisson distributed with mean M^*_x, which is constant over time.

Example (Beta-Bernoulli process, continued)**.** The $\tilde{a}$ and $\tilde{b}$ parameters in Equation (5) equal $\tilde{a} = 1$ and $\tilde{b} = 2$. Then, the expected number of features at time t is

$$E\left(\sum_{l=1}^{\infty}\mathbf{1}\left(\xi_{t,l} > 0\right)\right) = \gamma\,\frac{\Gamma(a+2)\Gamma(b-a+1)}{\Gamma(b+3)}\,.$$

The expected value of the total mass coincides with the expression in Equation (7.2.3) since the Bernoulli only admits $x = 1$. Furthermore, this implies that the expectation of the random measure linearly depends

on the observed measure at the previous time step. Indeed,

$$E\left(\Theta_{t+1}(A) \mid \Theta_t\right) = \frac{a+2}{b+3}\Theta_t(A) + \gamma\frac{\Gamma(a+2)\Gamma(b-a+2)}{\Gamma(b+4)}P_0(A)\,.$$

Finally, let us consider the lifetime of the l-th feature, which we denoted as L_l. In this case, L_l represents the number of Bernoulli trials needed to reach the zero state. Following Example 7.2.1, the probability of such event is $1 - \frac{a+2}{b+3}$ at each time step. Then, L_l is geometrically distributed with mean $\frac{b+3}{b-a+1}$. The result is also retrieved from Equation (7.18) noticing that since the Markov's chain only admits two states, $x = 0$ and $x = 1$, then $\boldsymbol{a} = (0, 1)$ and the transition matrix is fully characterized by $S_{0,0} = 1$ and $S_{1,1} = \frac{a+2}{b+3}$.

Example (Poisson-Gamma process, continued)**.** The $\tilde{a}$ and $\tilde{b}$ parameters in Equation (5) equal $\tilde{a} = 1$ and $\tilde{b} = 0$. The expected number of features at time t is

$$E\left(\sum_{l=1}^{\infty}\mathbf{1}\left(\xi_{t,l} > 0\right)\right) = \frac{\gamma\Gamma(\alpha+1)}{\beta^{\alpha+1}}\left[1 - \left(\frac{\beta}{\beta+1}\right)^{\alpha+1}\right].$$

The expected value of the total mass is

$$E\left(\Theta_t(\mathbb{X})\right) = \frac{\gamma\Gamma(\alpha+2)}{\beta^{\alpha+2}}\,.$$

In this case, the linear expectation in Equation (5) reduces to

$$E\left(\Theta_{t+1}(A) \mid \Theta_t\right) = \frac{1}{\beta+1}\Theta_t(A) + \frac{\alpha+1}{\beta+1}\sum_{l=1}^{K_t}\delta_{\lambda^*_{t,l}}(A) + \gamma\frac{\Gamma(\alpha+2)}{\beta+2}\left(\frac{\beta+2}{\beta+1}\right)^{\alpha+2}.$$

7.3 Time-dependent trait allocation model with random centers

The time-dependent trait allocation model presented in Section 7.2 is characterized by the fact that once a trait becomes inactive, it cannot reappear in the future. As already emphasized in the previous section, this property follows from the combination of two assumptions. First, traits are i.i.d. realizations from the diffuse probability measure P_0, implying that the probability of sampling the same feature more than once is zero. Second, the proposed model has a quasi-AR(1) type structure, so that at each transition it depends only on the present state, discarding any information about traits that appeared in the past but are currently inactive. Both assumptions are mathematically convenient and, apparently, they are essential to preserve tractability of the model. Depending on the application of interest, this restriction may not pose a limitation; however, in general, the ability to recognize that the same trait reappears over time is often a property one would like to capture in practice. The current section addresses this issue without altering the two assumptions stated above. The key idea is to introduce a collection of random centers, not evolving in time. At each innovation step of the process, new features are then generated in the neighborhood of these centers, rather than being drawn from a common base measure. In this way, a feature appearing in two distinct time windows is still represented by two different atoms, but both are linked to the same underlying center. Hence, although mathematically distinct, such features can be quantitatively recognized as similar. Moreover, the problem of learning the number of features translates

into the problem of learning the number of centers.

Let Λ be a random probability measure on $\mathbb{X}$ with H (random) atoms and equal weights. Namely,

$$\Lambda \overset{\mathrm{d}}{=} \frac{1}{H}\sum_{h=1}^{H}\delta_{\zeta_h}\,.$$

The atoms of Λ are denoted as centers. Moreover, since the number of the latter in generally unknown, we assume H to be random and place a prior on it. Hence, $H \sim \mathrm{Pois}_1(\omega)$, where Pois_1 denotes a one-shifted Poisson distribution. Then, conditionally to H, the centers are i.i.d. distributed from the same diffuse probability measure P_0 we introduced in the previous section. Namely, $\zeta_1,\ldots,\zeta_H \mid H \overset{\text{iid}}{\sim} P_0$. Finally, we note that this is equivalent to say that Λ is distributed according to a Normalized Independent Finite Point Process(NIFPP) as defined in Argiento and De Iorio (2022) with unnormalized jumps distribution $h(s) = \delta_1$, the number of atoms distributed as $q_M = \mathrm{Pois}_1(\omega)$ and base probability measure P_0. We write $\Lambda \sim \mathrm{NIFPP}(\delta_1, \omega, P_0)$.

We now introduce a modified version of the model in Equation (7.8), where the marginal law of the latent process (G_t) is specified conditionally on Λ. As discussed in Section 7.1, the law of the CRM G_t without fixed atoms is determined by two components: the Poisson process governing the jump intensities and the distribution of the atoms, which are i.i.d. draws from a diffuse probability measure. To preserve the conjugacy property that ensures tractable computation of the temporal evolution, the parametric form of ρ_{ord} must remain unchanged. However, we are free to modify the distribution of the atoms, provided that the two previously stated assumptions are satisfied. Specifically, we define

$$P_{0|\Lambda}(\mathrm{d}\lambda) = \int_{\mathbb{X}} k_\alpha(\lambda - v)\,\mathrm{d}\lambda\Lambda(\mathrm{d}v) = \frac{1}{H}\sum_{h=1}^{H} k_\alpha(\lambda-\zeta_h)\,\mathrm{d}\lambda\,,$$

where $k_\alpha(\cdot-\zeta):\mathbb{X}\to\mathbb{R}^+$ is a symmetric probability kernel, with $\zeta\in\mathbb{X}$ denoting the center parameter controlling the location of the kernel, and $\alpha > 0$ a scale parameter regulating its dispersion. In what follows, we focus on the case where k_α is a multivariate normal distribution with mean ζ and variance–covariance matrix $\alpha\mathbf{I}$, with $\mathbf{I}$ denoting the identity matrix. Thanks to the superposition theorem of Poisson processes (Kingman, 1993), the model can equivalently be represented as $G_t \mid \Lambda \overset{\mathrm{d}}{=} \sum_{h=1}^{H} G_{t,h}$, where each $G_{t,h}$ is a CRM with Lévy intensity $\frac{1}{H}\rho_{\mathrm{ord}}(s)\mathrm{d}s$ and atoms that are i.i.d. draws from $P_{0,h}(\mathrm{d}\lambda) = k_\alpha(\lambda-\zeta_h)\mathrm{d}\lambda$. Hence, when we choose ρ_{ord} as in Equation (7.4), the model can be written as

$$\begin{aligned} G_t \mid \Lambda &= \sum_{h=1}^{H} G_{t,h} \\ G_{t,h} \mid \Lambda &\sim \text{Conj-CRM}\left(\frac{\gamma}{H}, a, b, P_{0,h}\right) \\ \Lambda &\sim \mathrm{NIFPP}(\delta_1,\omega,P_0) \end{aligned} \tag{7.20}$$

Moreover, G_t admits a double summation formulation

$$G_t = \sum_{h=1}^{H}\sum_{l\geq 1} s_{t,h,l}\,\delta_{\lambda_{t,h,l}}. \tag{7.21}$$

By exchangeability, this representation can be consistently mapped back to the formulation introduced earlier. To make this connection explicit, we introduce indicator variables $t_{t,l}$, one for each atom $\lambda_{t,l}$, specifying the center to which it belongs. In particular, we set $t_{t,l} = h$ if and only if there exists an index l' such that $\lambda_{t,l} = \lambda_{t,h,l'}$. In other words, $t_{t,l}$ identifies the center h that generated the atom $\lambda_{t,l}$ through the double-sum representation of G_t. The construction in Equation (7.20) is equivalent to say that G_t is a particular case of a Shot-Noise Cox processes (Møller, 2003) which have recently been investigated as a richer alternative to Poisson processes in the construction of random measures to be used as Bayesian nonparametric prior. See Beraha et al. (2025a) for an application to mixture model with repulsive and attractive atoms, Carminati et al. (2025) for an extension to the hierarchical case and Beraha et al. (2025c) for further theoretical investigation.

7.3.1 Traits dynamic with centers

Consider the trait profile measure Θ_t defined as in Section 7.2, that is $\Theta_t \mid G_t \overset{\mathrm{d}}{=} \sum_{l\geq 1} \xi_{t,l}\delta_{\lambda_{t,l}}$. The double-sum representation of G_t, see Equation (7.21), allows us to highlight additional structural properties of the emission model Θ_t. Indeed, we have that

$$\Theta_t \mid G_t, \Lambda \overset{\mathrm{d}}{=} \sum_{l\geq 1} \xi_{t,l}\,\delta_{\lambda_{t,l}} \overset{\mathrm{d}}{=} \sum_{h=1}^{H}\sum_{l\geq 1} \xi_{t,h,l}\,\delta_{\lambda_{t,h,l}} \overset{\mathrm{d}}{=} \sum_{h=1}^{H} \Theta_{t,h}$$

where we defined $\Theta_{t,h} = \sum_{l\geq 1} \xi_{t,h,l}\,\delta_{\lambda_{t,h,l}}$. Regarding temporal evolution, we extend the Markovian transitions of Equation (7.8), now conditional on Λ, as follows:

$$\begin{aligned} G_{0,h} &\sim p_{G|\Lambda}(\mathrm{d}G_{0,h}) \\ \Theta_{t,h} \mid G_{t-1,h} &\sim p_{\Theta|G}(\mathrm{d}\Theta_t \mid G_{t-1,h}, \Lambda)\,, \quad t = 1,\dots,T \\ G_{t,h} \mid \Theta_{t,h}, \Lambda &\sim p_{G|\Theta,\Lambda}(\mathrm{d}G_{t,h} \mid \Theta_{t,h}, \Lambda)\,, \quad t = 1,\dots,T-1\,, \end{aligned} \tag{7.22}$$

for each $h = 1,\dots,H$. We take $p_{\Theta|G} = \mathrm{ExpCRM}(\kappa, \eta, \phi, A)$ so that conjugacy is preserved at the level of each h, while $p_{G|\Theta,\Lambda}$ remains of the same form described in Section 7.2, with the only modifications being the replacement of P_0 by $P_{0,h}$ and the rescaling of γ by $\gamma_h = \gamma/H$.

In view of this construction, it is natural to introduce additional notation for the trait dynamics that explicitly records the center to which each feature is assigned. Let the set of active features at time t in group h be denoted by $\boldsymbol{\lambda}^*_{t,h} = \{\lambda^*_{t,h,1},\dots,\lambda^*_{t,h,K_{t,h}}\}$ with corresponding intensities $\boldsymbol{\xi}_{t,h} = \{\xi_{t,h,1},\dots,\xi_{t,h,K_{t,h}}\}$ where $K_{t,h}$ is the number of active features in group h at time t. Analogously, we let $N_{t,h}$ denote the total number of new traits at time t related to the h-th center.

Without loss of generality, we order the groups so that the global set of active features and intensities is obtained by concatenating the group-specific ones. That is, $\boldsymbol{\lambda}^*_t = (\boldsymbol{\lambda}^*_{t,1},\dots,\boldsymbol{\lambda}^*_{t,H})$ and, as described in Section 7.2.2, the vector $\boldsymbol{\xi}_t$ is reordered accordingly to preserve the original pairings $(\lambda^*_{t,h,l}, \xi_{t,h,l})$. The total number of active features at time t, previously denoted by K_t in Section 7.2.2, satisfies $K_t = \sum_{h=1}^{H} K_{t,h}$ since replicates cannot occur across different groups, a consequence of the diffuseness of the base measures $P_{0,h}$. Similarly, we define the set of all traits that have appeared at least once in group h over the observation horizon as $\boldsymbol{\lambda}^{**}_h = \cup_{t=1}^{T}\boldsymbol{\lambda}^*_{t,h}$. The set $\boldsymbol{\lambda}^{**}$ that collect all those traits that appeared at least one is defined as the union over all centers and all times, i.e., $\boldsymbol{\lambda}^{**} = \cup_{h=1}^{H}\cup_{t=1}^{T}\boldsymbol{\lambda}^*_{t,H} = \{\lambda^{**}_1,\dots,\lambda^{**}_K\}$, where here $K = \sum_{h=1}^{H}\sum_{t=1}^{T} N_{t,h}$ is the total number of observed features.

The center–specific transition kernel $Q_h(\mathrm{d}\Theta_{t+1,h} \mid \Theta_{t,h}, \Lambda)$ is such that, at $t = 1$, we have

$$\Theta_{1,h} \overset{\mathrm{d}}{=} \sum_{x\geq 1} \sum_{l=1}^{N_{1,h,x}} x, \delta_{\lambda_{1,h,l}} ,$$

where the atoms are i.i.d. draws from $P_{0,h}$ and $N_{1,h,x} \sim \mathrm{Pois}(M_x/H)$, with M_x defined in Equation (7.2.1). Consequently, the total number of features at time $t = 1$ in group h is $N_{1,h} \sim \mathrm{Pois}\left(\frac{1}{H}\sum_{x\geq 1} M_x\right)$. For $t \geq 1$, the transition is given by

$$\Theta_{t+1,h} \mid \Theta_{t,h} \overset{\mathrm{d}}{=} \sum_{l=1}^{K_{t,h}} \xi'_{t+1,h,l}\, \delta_{\lambda^*_{t,h,l}} + \sum_{x=1}^{\infty} \sum_{l=1}^{N_{t+1,h,x}} x\, \delta_{\lambda_{t+1,h,l}} ,$$

where $\xi'_{t+1,h,l} \mid \xi_{t,h,l} \sim h'(\cdot \mid \xi_{t,h,l})$, as in Equation (7.13), since this distribution does not depend on either $P_{0,h}$ or γ_h. The innovation term consists of $N_{t+1,h,x}$ new atoms, drawn i.i.d. from $P_{0,h}$. $N_{t+1,h,x}$ is Poisson distributed with mean equal to M^*_x/H, where M^*_x is given in Equation (7.14). Hence, the total number of features at time $t + 1$ in the h-th group is $N_{t+1,h} \sim \mathrm{Pois}\left(\frac{1}{H}\sum_{x\geq 1} M^*_x\right)$.

Conditional independence with respect to Λ implies that the global transition kernel factorizes into the product of the center–specific ones:

$$Q(\mathrm{d}\Theta_{t+1} \mid \Theta_t, \Lambda) = \prod_{h=1}^{H} Q_h(\mathrm{d}\Theta_{t+1,h} \mid \Theta_{t,h}, \Lambda) . \tag{7.23}$$

A direct implication is that the total number of new traits at time t, namely $N_t = \sum_{h=1}^{H} N_{t,h}$, follows a Poisson distribution with mean $\sum_{x\geq 1} M_x$ when $t = 1$, and with mean $\sum_{x\geq 1} M^*_x$ for $t > 1$. This coincides with the result obtained in Section 7.2.1. In other words, Λ does not alter the total number of traits K which evolves exactly as in the previous case. Its role is instead confined to governing the allocation of traits across centers, determining where features are activated, while leaving unchanged how many features appear overall.

Finally, we turn to the full conditional distribution of $\Lambda \mid \boldsymbol{\Theta}_{1:T}$ or, equivalently, how to update $\{\zeta_1, \ldots, \zeta_H, H\}$ given the entire emission process over time. This distribution follows directly from Theorem 5.1 in Argiento and De Iorio (2022), as it coincides with the posterior distribution of the mixing measure in a mixture model with observations $\lambda^{**}_1, \ldots, \lambda^{**}_K$. Indeed, from Equation (7.23), each transition kernel $Q_h(\mathrm{d}\Theta_{t+1,h} \mid \Theta_{t,h}, \Lambda)$ decomposes into two independent components: the thinning of existing traits and the innovation of new ones. As discussed above, Λ does not play any role in the propagation mechanism, so the thinning part can be disregarded. Likewise, total number of new traits is unaffected by Λ. Therefore, conditionally on $\boldsymbol{\Theta}_{1:T}$, and, in particular on $\{\lambda^{**}_1, \ldots, \lambda^{**}_K, K\}$, the full conditional of interest coincides with the posterior of the following mixture model:

$$\begin{aligned} \lambda^{**}_1, \ldots, \lambda^{**}_K \mid K, \Lambda &\overset{\text{iid}}{\sim} \frac{1}{H}\sum_{h=1}^{H} k_\alpha(\cdot - \zeta_h) \\ \zeta_1, \ldots, \zeta_H \mid H &\sim P_0(\cdot) \\ H &\sim \mathrm{Pois}_1(\omega) . \end{aligned}$$

An explicit expression of the full conditional distribution is reported in Section F.2.

7.4 Dynamic generalized latent trait model

The time-dependent trait profile (Θ_t) defines a dynamic process of almost surely discrete random measures, which are not dominated by any σ-finite measure. Consequently, such measures are typically employed as mixing distributions in Bayesian models. In particular, discrete random probability measures are central to problems such as density estimation and clustering, whereas more general random measures often serve as the fundamental building blocks of nonparametric factor models. Since our focus lies in this latter setting, the present section introduces a nonparametric generalized latent trait model that evolves over time.

Generalized latent trait models (Moustaki and Knott, 2000) assume that X_t follows a parametric exponential family distribution with natural parameter θ_t and dispersion parameter τ^2 (Mcculloch and Neuhaus, 2013). Specifically, $X_t \sim f\left(\cdot \mid \theta_t, \tau^2\right)$, where

$$f\left(x_t \mid \theta_t, \tau^2\right) = \kappa(x_t) \exp\left\{\tau^2 \left(\langle \theta_t, x_t \rangle - A(\theta_t)\right)\right\} . \tag{7.24}$$

Standard properties of the exponential family imply $\mu_t = E\left(X_t \mid \theta_t, \tau^2\right) = \dfrac{\mathrm{d}A(\theta_t)}{\mathrm{d}\theta_t}$ and $\nu_t = \mathrm{Var}\left(X_t \mid \theta_t, \tau^2\right) = \dfrac{1}{\tau^2}\dfrac{\mathrm{d}^2 A(\theta_t)}{\mathrm{d}\theta_t^2}$. The model is completed by introducing a monotonic differentiable link function $g(\mu_t) = \eta_t$ to model the mean through the linear predictor η_t, defined as

$$\eta_t = \sum_{l=1}^{p} \xi_{t,l}\, \lambda_j ,$$

where $\xi_{t,1}, \ldots, \xi_{t,p}$ denotes a vector of loadings and $\lambda_1, \ldots, \lambda_p$ a collection of factors. The linear predictor in Equation (7.4) can easily be extended to incorporate covariates.

In generalized latent trait models, selecting the number of factors p is a challenging problem, and no universally accepted strategy exists for fixing this number; see, for example, Lopes and West (2004) for a comprehensive discussion. An alternative approach postulates a countably infinite collection of potential factors, with each observation involving only finitely many of them. This framework can be naturally formulated by employing completely random measures as priors within a Bayesian nonparametric setting. Indeed, the linear predictor η_t can be written as

$$\eta_t = \sum_{l=1}^{\infty} \xi_{t,l}\, \lambda_{t,l} = \int_{\mathbb{X}} \lambda\, \Theta_t(\mathrm{d}\lambda) ,$$

where $\Theta_t = \sum_{l \geq 1} \xi_{t,l}\, \delta_{\lambda_{t,l}}$ is the process introduced in Section 7.2 which is now used as mixing measure. Notably, the number of non-zero elements in Θ_t is finite by assumption in Equation (7.2), which guarantees the hypothesis about the finite number of factors for each observation X_t.

So far, we have described the statistical model at a given time t. Before turning to the temporal dynamics, we pause to emphasize that the static version of this model already encompasses several well-known frameworks. Specifically, let $X_1, \ldots, X_n$ be an i.i.d. sample of size n from the static version of the model in Equation (7.24). For any fixed number of factors, classical Gaussian factor analysis is

recovered when $f(\cdot \mid \theta, \tau)$ corresponds to the multivariate normal distribution. Similarly, when allowing for a potentially infinite number of factors, the model includes as special cases the widely studied infinite latent feature model of Ghahramani and Griffiths (2005) and the latent Poisson factor analysis of Zhou et al. (2009). More generally, the formulation in Equation (7.24) accommodates samples with mixed data types and, to the best of our knowledge, provides a unified representation that has not previously appeared in the literature.

7.4.1 Latent traits changepoint model

Change point detection has long been a central topic in statistics, with early contributions dating back to Page (1954, 1957) from a frequentist standpoint and to Chernoff and Zacks (1964) in a Bayesian framework. A major development for Bayesian nonparametric methods came with the works of Barry and Hartigan (1992, 1993), who proposed modeling multiple change points through a restricted version of the product partition model (Hartigan, 1990), limiting attention to partitions that respect the temporal ordering of the data. In this class of models, temporal observations are divided into disjoint intervals within which the data are assumed to be homogeneous and possess simpler statistical properties. This structure simplifies inference and avoids the need to specify complex temporal dynamics. However, a key limitation is that information is not shared across intervals, which can be restrictive in certain applications. For instance, consider modeling how the frequency of words evolves over time. A common approach, known as topic modeling, assumes that words can be organized into latent, hidden topics. It is appealing to construct a model that captures the introduction of new topics over time. However, a changepoint model would be inappropriate here: because it treats intervals as independent, the emergence of a single new topic would imply that all topics are new in the subsequent interval. This limitation motivates the development of models that enable information sharing across changepoints. The nonparametric generalized latent trait model introduced in the previous section fits naturally within this framework. It represents observations through latent factors whose number can evolve over time as factors are added or removed at each changepoint. This construction allows some factors to persist across intervals, thus enabling information sharing through time.

Consider a time series $X_1, \ldots, X_T$ is governed by the latent parameters $\theta_1, \ldots, \theta_T$, which are deterministically linked to the linear predictors $\eta_1, \ldots, \eta_T$. At each time point, the linear predictor is expressed as a finite combination of traits,

$$\eta_t = \sum_{l=1}^{K_t} \xi_{t,l}\, \lambda^*_{t,l}\,.$$

As described in Section 7.2.2, each trait l either persists to the next time step with probability $1-h'(0 \mid \xi_{t,l})$ or disappears with probability $h'(0 \mid \xi_{t,l})$, for $l = 1, \ldots, K_t$. As a result, $\mathbb{P}(\eta_t = \eta_{t+1}) > 0$ which allows us to introduce the random times $\tau_1, \ldots, \tau_{C_T}$ such that $\boldsymbol{\lambda}^*_t = \boldsymbol{\lambda}^*_s$ for each couple of times $t, s \in [\tau_j, \tau_{j+1})$. If this condition holds, we refer to τ_j and τ_{j+1} as two consecutive changepoints, and we denote by C_T their total count over the time horizon T. It is important to note that these temporal boundaries refer solely to the configuration of active traits: changepoints capture the appearance or disappearance of traits, not fluctuations in their magnitudes, provided the latter remain strictly positive.

Wrapping up, the full model is

$$\begin{aligned} X_t \mid \theta_t, \boldsymbol{\tau}^2 &\overset{\text{ind}}{\sim} f\left(X_t \mid \theta_t, \tau^2\right), \quad t = 1, \dots, T \\ g(\mu_t(\theta_t)) &= \eta_t = \int_{\mathbb{X}} \lambda \Theta_t(\mathrm{d}\lambda) = \sum_{l \geq 1} \xi_{t,l} \lambda_{t,l} \\ \Theta_{t+1} \mid \Theta_t &\sim Q(\mathrm{d}\Theta_{t+1} \mid \Theta_t), \quad t = 1, \dots, T \\ \tau^2 &\sim p(\cdot) \end{aligned} \tag{7.25}$$

where the transition kernel $Q(\mathrm{d}\Theta_{t+1} \mid \Theta_t)$ is defined either in Equation (7.7) for the model without centers or in Equation (7.23) if centers are included.

For sake of notation, we do not provide a general version of the algorithm we use to perform inference on the model in Equation (7.25). Instead, we only provide the specialized versions for the application in subsequent section. In particular, in this thesis we limit ourselves to fit a dynamic version of the feature allocation problem as presented in Ghahramani and Griffiths (2005). In Section 7.6 we briefly discuss a second possible application regarding a Poisson Factor Analysis for time dependent topic modeling.

7.5 Dynamic features allocation

In this experiment, we adapt the well-known toy example of Griffiths and Ghahramani (2011) to the setting of time-evolving images. As in the original setup, gray-scale images are constructed by linearly superimposing 4 distinct features and adding Gaussian noise. Each image has dimension 8×8, which we vectorize into a vector of length $D = 64$ containing real-valued pixel intensities. The generating features are shown in Figure 7.1. Unlike the original example, however, features are not assigned independently to each image. Instead, we let them evolve over time across $T = 24$ time steps. The objective of the modeling task is to recover both the number and identities of the features, as well as their temporal dynamics, i.e., when each feature appears and how long it remains active.

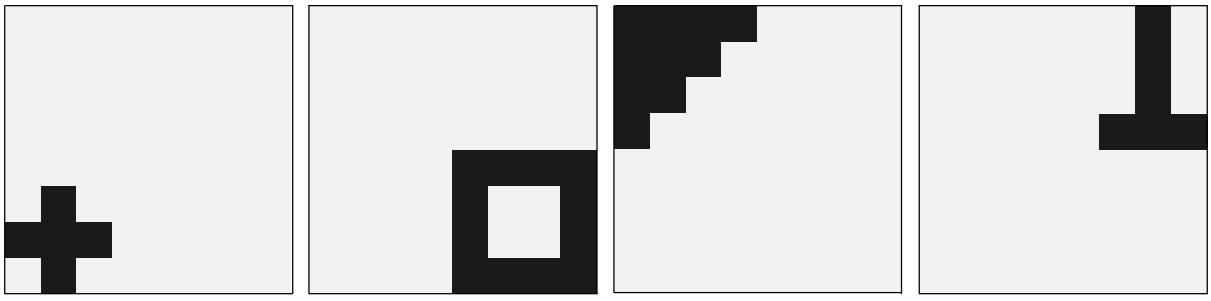

Figure 7.1: Image elements corresponding to the four latent features used to generate the data.

Specifically, let $\boldsymbol{X}_t$ be a D-dimensional vector representing the image at time t and collect all images in a $T \times D$ matrix named $\boldsymbol{X}$, such that $\boldsymbol{X} = \left(\boldsymbol{X}_1^T, \dots, \boldsymbol{X}_T^T\right)$. The model likelihood is

$$\boldsymbol{X}_t \mid \Theta_t, \sigma_X^2 \overset{\text{ind}}{\sim} \mathrm{N}_D\left(\boldsymbol{\mu}_t, \sigma_X^2 \mathbf{I}_D\right)$$

where $\boldsymbol{\mu}_t = \sum_{l \geq 1} \xi_{t,l} \lambda_{t,l}$ and σ_X^2 represents the noise. The model is completed setting a Beta-Bernoulli process for the random measure Θ_t on the space of features, that is $\mathbb{X} = \mathbb{R}^D$. In particular, we can collect

all the possible features λ_l in a matrix denoted as A having D columns and potentially infinitely many rows. For what concerns the binary indicator variables $\xi_{t,l}$, these are collected in a binary latent variable Z_t such that the l-th entry equal $\xi_{t,l}$. Consequently, Z_t has infinitely many entries as we do not upper bound the number of possible features in each image. These vectors are collected in a binary matrix Z with T rows and an unbounded number of columns. In the data generating process, we fix $\sigma^2_{X,\text{true}} = 0.001$ and set the true number of feature $K_{\text{true}} = 4$. It follows that A_{true} is a $K_{\text{true}} \times D$ matrix and Z_{true} is a $T \times K_{\text{true}}$ matrix. A sample from the model is given in Figure 7.2. This setup falls naturally within the general framework of dynamic latent trait models presented in Section 7.4. For the base measure P_0, we take a multivariate Gaussian distribution on $\mathbb{R}^D$ with mean vector 0 and diagonal variance-covariance matrix $\sigma^2_A \mathbf{I}_D$.

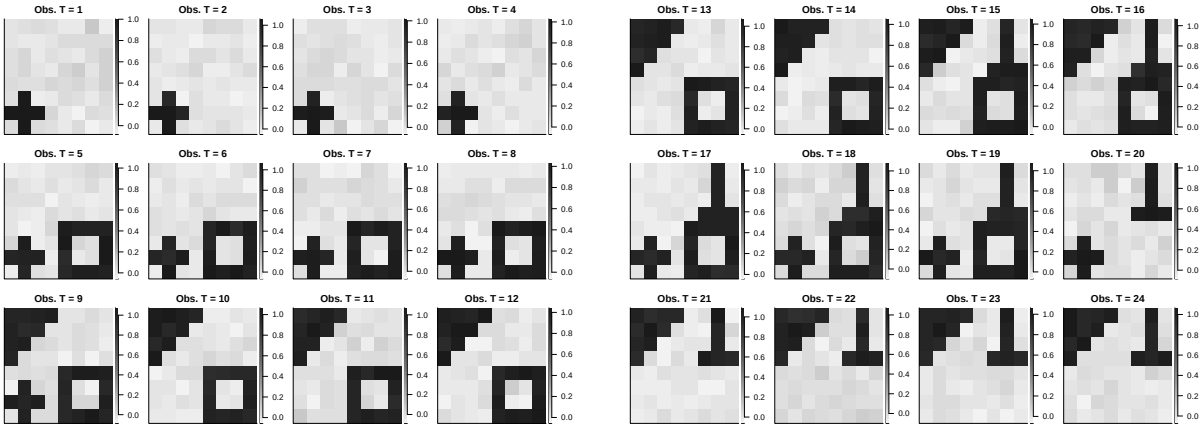

Figure 7.2: Simulated data. Each image is 8×8 and it displays some of the four latent features.

7.5.1 Sampling strategy

Sequential Monte Carlo (SMC) methods represent a widely used class of algorithms for sampling from general high-dimensional probability distributions. They have proven to be highly efficient when dealing with state-space (or hidden Markov) models (Doucet and Johansen, 2009), which is the case we deal with here.

Our objective is to sample from a distribution of the form

$$p\left(Z, A, \sigma^2_X, c, \gamma \mid \boldsymbol{X}\right)$$

where the hidden Markov state process is defined over an (ideally) infinite-dimensional vector of binary variables, each indicating the presence or absence of specific features at a given time. We denote by $\boldsymbol{\theta} = (\sigma^2_X, a, b, \gamma)$ the vector of static hyperparameters. To achieve this, we adopt the particle Gibbs sampler introduced by Andrieu et al. (2010), which provides a valid particle-based approximation to the Gibbs sampler by incorporating a conditional SMC step.

In a nutshell, the algorithm alternates between two main steps: updating the static parameters $\boldsymbol{\theta}$ through their full conditionals, $p\left(\boldsymbol{\theta} \mid Z, A, \boldsymbol{X}\right)$, and running a conditional SMC algorithm targeting $p_{\boldsymbol{\theta}}\left(Z, A \mid \boldsymbol{X}\right)$, given the previously sampled $\boldsymbol{\theta}$ and its ancestral lineage. After this step, we obtain N particle–weight pairs $\left(Z^{(i)}, A^{(i)}, w^i\right)_{i=1,\ldots,N}$ that approximate the target distribution, each associated with $K^{(i)}$ features. Sampling from this approximation simply requires drawing an index n from a discrete

distribution with probabilities proportional to the weights $\left(w^1, \ldots, w^N\right)$ and selecting the corresponding particle $\left(Z^{(n)}, A^{(n)}\right)$. In what follows, we detail the steps of the algorithm as adapted to our specific model. For a comprehensive overview of the general method, see Andrieu et al. (2010). We here provide an overview of the conditional SMC, while the full conditionals of the elements in $\boldsymbol{\theta}$ are reported in Section F.3.1.

We recall that A is the feature matrix, namely, the l-th row of A is a vector of length D that represents the l-th feature $\boldsymbol{\lambda}_l^{**}$. Hence, A is a $K \times D$ matrix, where the number of rows K is the total number of features discovered throughout the whole time period and it is itself a random quantity. We let $A^{1:t}$ be the $K_t \times D$ submatrix of A matrix collecting all the features observed up to time t. Similarly, we let $Z^{1:t}$ be the $t \times K_t$ submatrix of Z that collects all the binary indicator variables addressing the activeness of all features appeared up to time t. Clearly, $A^{1:T} = A$ and $Z^{1:T} = Z$. Finally, we let $P_k^{1:T} = \left((Z^1, A^1)_{B_1^k}, \ldots, (Z^{1:t}, A^{1:t})_{B_t^k}, \ldots, (Z^{1:T}, A^{1:T})_{B_T^k}\right)$ be path of the k-th particle, $k = 1, \ldots, N$, with ancestral lineage $B_k^{1:T} = \left(B_k^1, \ldots, B_k^T\right)$. Recall that, by construction, $B_k^T = k$.

This representation helps us in devising a particle Gibbs sampler and we now provide details for the conditional SMC step: this update is similar to a standard SMC algorithm but it ensures that a prespecified path $P^{1:T}$ with ancestral lineage $B^{1:T}$ survives all the resampling steps, whereas the remaining $N - 1$ particles are generated according to a proposal distribution which is hopefully similar to the target. In the following, we present the skeleton of the algorithm. Details about the proposal distributions are deferred to Section F.3.

Time $t = 1$

1. For $i = 1, \ldots, N, i \neq B_1$, sample

$$\left(K_1^{(i)}, A_{(i)}^1, Z_{(i)}^1\right) \sim q\left(K_1^{(i)} \mid \gamma, c\right) p\left(Z_{(i)}^1 \mid K_1^{(i)}\right) p\left(A_{(i)}^1 \mid K_1^{(i)}, \boldsymbol{X}_1\right),$$

2. Compute the unnormalized weight of each particle $i = 1 \ldots N$

$$w_1^{(i)} = p\left(X_1 \mid K_1^{(i)}\right). \tag{7.26}$$

Time $t \in \{2, \ldots, T\}$

1. Let $\mathcal{F}(\boldsymbol{w}_{t-1})$ be the discrete probability distribution on $\{1, \ldots, N\}$ obtained by normalizing the weights $w_{t-1}^{(i)}, i = 1, \ldots, N$.

2. For $i \neq B_t$, draw $\aleph_{t-1}^i \sim \mathcal{F}(\boldsymbol{w}_{t-1})$. $\aleph_{t-1}^i$ is the parent at time $t-1$ of particle i. Consequently, update the path of the i-th particle, $\aleph^i = (\aleph_1^i, \ldots, \aleph_{t-1}^i)$.

3. For $i = 1, \ldots, N, i \neq B_1$, sample

$$\left(N_t^{(i)}, A_{(i)}^{1:t}, Z_{(i)}^{1:t}\right) \sim q\left(N_t^{(i)} \mid \gamma, c\right) p\left(Z_{(i)}^{1:t} \mid N_t^{(i)}, Z_{\aleph_{t-1}^i}^{1:t-1}\right) p\left(A_{(i)}^{1:t} \mid A_{\aleph_{t-1}^i}^{1:t-1}, N_t^{(i)}, \boldsymbol{X}_t\right),$$

   where $N_t^{(i)}$ represents the number of new features introduced at time t in the i-th particle.

4. Compute the unnormalized weight of each particle $i = 1 \ldots N$

$$w_t^{(i)} = p\left(X_t - (A_{\aleph_{t-1}^i}^{1:t-1})^T \mathbf{1}_{N_t} \mid N_t^{(i)}\right).$$

7.5.2 Empirical analysis

We start fitting the model without centers described in Section 7.2. We consider a $\sigma_X^2 \sim$ Inv-gamma$(3, 0.02)$ hyperprior for the noise. For what concern the Beta-Bernoulli process, we set $a = -2$ so that it boils down to the two parameters case. Then, we place a $\gamma \sim$ Gamma$(0.08, 0.2)$ hyperprior while we consider the following alternative parametrization: $c = b + 2$, where $c \sim$ Inv-gamma$(2, 1)$. Finally, we set $\sigma_A^2 = 0.5$. The model is fitted using the conditional SMC algorithm presented in the previous section with and use $N = 500$ particles and $G = 1000$ Gibbs sampling iterations. The filtered density is basically identical to the observed values, showing a good recovering of the images. See Figure 7.3.

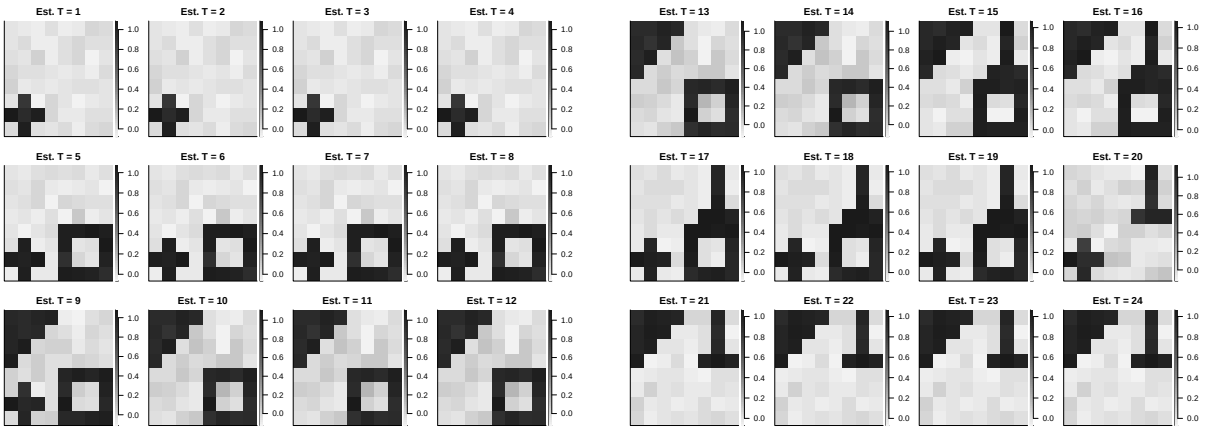

Figure 7.3: Estimated means using the first model, fitted using $N = 500$ particles.

The total number of estimated features is equal to $K_{\text{est}} = 6$. These are reported in Figure 7.4 alongside the different times they are active. The reason why two additional features are selected compared to the true number is due to the fact that the first and third images reported in Figure 7.1 disappear at times $t = 9$ and $t = 16$ and then reappear at times $t = 17$ and $t = 21$, respectively. However, because of the quasi-AR(1) structure of our model, once a feature disappears it can no longer be selected in future times. For this reason, although the estimated features λ_1^{**} and λ_5^{**} as well as λ_3^{**} and λ_6^{**} are almost indistinguishable, see Figure 7.4, the model treats them as two entirely separate images.

We then fit the data using the model with the random centers described in Section 7.3. Because of non-identifiability issues, it is crucial to get a good first guess on the possible values of the centers ζ_h's. For this reason, we first run a single iteration of the first model and then, given the estimated features (see Figure 7.4), we with a k-means clustering with four different centers. The initial values of the ζ's closely resemble the four images in Figure 7.1. Then, we choose a Pois$_1(\omega)$ prior for the number of centers, setting $\omega = 10$. Finally, the centering measure is a multivariate normal distribution with variance $\alpha = 1$. We run the conditional SMC for $G = 1000$ iterations using the same number of particles as before. The most important difference regards the estimated features and their group allocation. We still get a total of $K_{\text{est}} \geq 4$ estimated features. However, if we do not look at the features level but at the estimated centers, see Figure 7.5, we see that the same center can be associated to features which disappear and come back. Indeed, because of the existence of the underlying centers, λ_1^{**} and λ_5^{**} (λ_3^{**} and λ_6^{**} as well), are now linked together since we are able to recognized that they are both generated from the same underlying image.

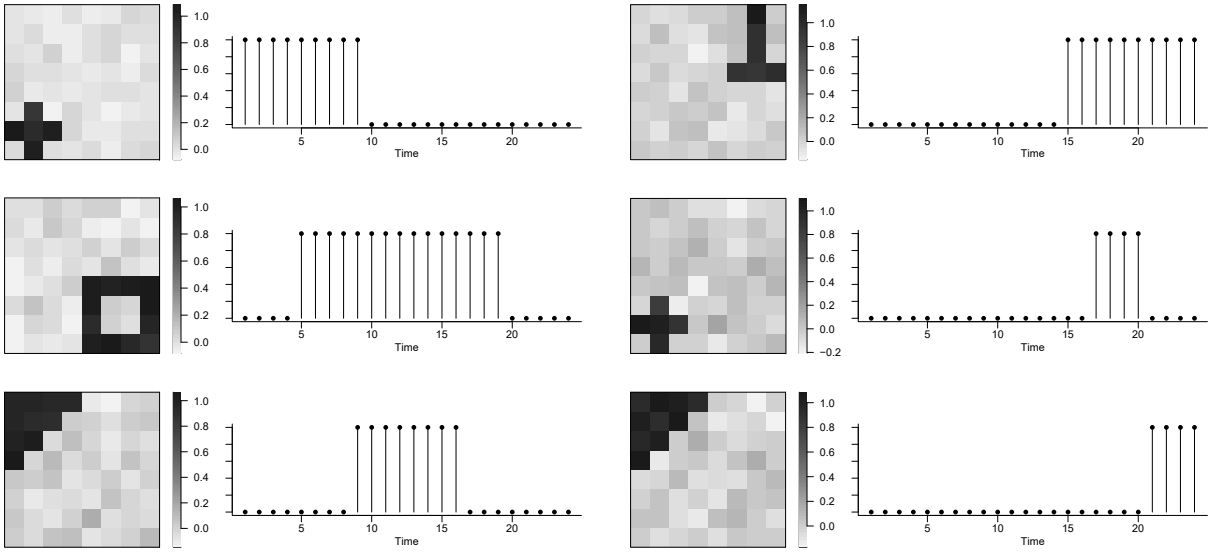

Figure 7.4: The estimated feature (left panels) and a barplot of the different times they were selected (right panels).

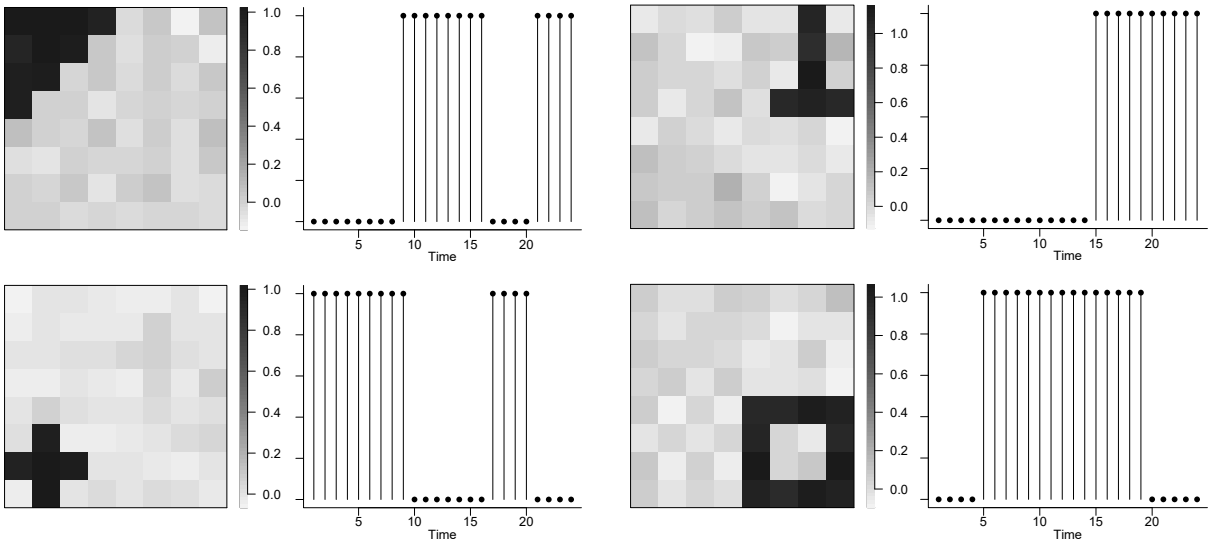

Figure 7.5: The estimated center (left panels) and a barplot of the different times they were selected by at least one feature (right panels).

7.6 Poisson Factor Analysis for time dependent topic modeling

We conclude this chapter discussing a possible application about learning latent topics in a time series of documents, where each document can be viewed as arising from an unknown number of latent topics whose popularity evolves over time. The goal is to uncover these latent topics and to study how their relative importance changes dynamically. In the static case, topic modeling assumes that the data represent word counts: for a vocabulary of size V and N documents, $D_{i,n}$ denotes the frequency of the i-th word in the n-th document. By assuming the existence of K hidden topics, each count can be represented as

$$D_{i,n} = \sum_{l=1}^{K} D_{i,n}^{l},$$

where $D^l_{i,n}$ represents the number of times word i is used in the document n within the context of topic l. Each topic corresponds to a discrete probability distribution over the vocabulary. In a Bayesian framework, these topic-specific word distributions are typically assigned symmetric Dirichlet priors of dimension V and parameter δ. That is, $\boldsymbol{\lambda}_l = (\lambda_{1,l}, \dots, \lambda_{V,l})$ with $\boldsymbol{\lambda}_l \overset{\text{iid}}{\sim} \text{Dir}_V(\delta, \dots, \delta)$ for $l = 1, \dots, K$.

A key advantage of the Bayesian nonparametric approach is that it eliminates the need to fix the number of topics in advance, allowing the model complexity to grow with the data. This flexibility has motivated several important developments, including latent Dirichlet allocation (Blei et al., 2003), the hierarchical Dirichlet process (Teh et al., 2006), and extensions based on Poisson and negative binomial processes (Zhou et al., 2012; Broderick et al., 2015). For a comprehensive overview, see Blei and Lafferty (2009); Blei (2012).

In the case of dynamic topic modeling, we do not observe a static collection of N documents, but rather a time series of documents. Namely, $\boldsymbol{D}_t = (D_{1,t}, \dots, D_{V,t})$ where $D_{i,t}$ denote the frequency of the word i in the document observed at time t, for $t = 1, \dots, T$. By explicitly modeling temporal dependence, we allow for the possibility that new topics may emerge or that existing topics may become irrelevant over time. Previous attempts in this direction include Blei and Lafferty (2006), Zhang et al. (2016), and Perrone et al. (2017). To achieve this, we take the emission process Θ_t and the latent process G_t to be a Poisson-Gamma process, as described in Example 7.1. In particular, since the space of traits $\mathbb{X}$ is defined as the set of all possible discrete distributions over the vocabulary of size V, we choose the base probability measure P_0 to be a symmetric Dirichlet distribution of dimension V, with hyperparameter δ. In the following, we write each $\boldsymbol{\lambda}_{t,l}$ as $\boldsymbol{\lambda}_{t,l} = \left(\lambda^l_{1,t}, \dots, \lambda^l_{V,t}\right)$.

We link the emission process Θ_t to the observable data as follow. Firstly, we let N^l_t be the number of words at time t that refers to topic l, which we draw as

$$N^l_t \overset{\text{ind}}{\sim} \text{Pois}\left(\xi_{t,l}\right), \quad l \geq 1,$$

for each time instant t. By convention, we define $\text{Pois}(0) = \delta_0$. Alternative approaches consider continuous Poisson rates for generating N^l_t. However, in that case, it is necessary to introduce an additional binary parameter to indicate whether the l-th topic is active at time t or not. We believe that our construction is more parsimonious, and that the cost of using integer-valued rather than continuous Poisson rates is not a substantial limitation. Moreover, we define $\boldsymbol{D}^l_t = \left(D^l_{1,t}, \dots, D^l_{V,t}\right)$ as the vectors collecting the frequencies of all words at time t in the l-th topic. Given N^l_t and the topic distribution at time t, we assume that

$$\boldsymbol{D}^l_t \mid N^l_t, \boldsymbol{\lambda}_{t,l} \overset{\text{ind}}{\sim} \text{Multi}_V\left(N^l_t; \boldsymbol{\lambda}_{t,l}\right), \quad l = 1, \dots, K_t,$$

where $\text{Multi}_V(n; \boldsymbol{p})$ represents the Multinomial distribution of dimension V with n trials and probabilities $p_1, \dots, p_V$. Finally, the observed counts of the i-th word at time t are recovered as $D_{i,t} = \sum_{l=1}^{K_t} D^l_{i,t}$. This construction, for each time t and conditionally to Θ_t, is summarized as

$$\begin{aligned} D_{i,t} &= \sum_{l=1}^{K_t} D^l_{i,t}, \quad i = 1, \dots, V, \\ \boldsymbol{D}^l_t \mid \Theta_t &\overset{\text{ind}}{\sim} \text{Multi}_V\left(N^t_l; \boldsymbol{\lambda}_{t,l}\right), \quad l = 1, \dots, K_t, \\ N^l_t \mid \Theta_t &\overset{\text{ind}}{\sim} \text{Pois}\left(\xi_{t,l}\right), \quad l = 1, \dots, K_t. \end{aligned} \tag{7.27}$$

Finally, the time evolution of latent process Θ_t may either follow the dynamic described in Section 7.2 or the one in Section 7.3, if centers are included in the model.

The following equivalent representation of this model is known as Poisson Factor Analysis, a well-known technique that has been successfully applied to static and dynamic topic modeling (Zhou et al., 2012; Zhou and Carin, 2015; Acharya et al., 2015; Roychowdhury and Kulis, 2015).

$$D_{i,t} \mid \Theta_t \overset{\text{ind}}{\sim} \text{Pois}\left(\sum_{l=1}^{K_t} \lambda^l_{i,t}\xi_{l,t}\right), \quad i = 1, \dots, V \tag{7.28}$$

By leveraging standard properties of the Poisson distribution, the model in Equation (7.28) can also be formulated as

$$D_{i,t} = \sum_{l=1}^{K_t} D^l_{i,t}, \quad D^l_{i,t} \mid \Theta_t \overset{\text{ind}}{\sim} \text{Pois}\left(\lambda^l_{i,t}\xi_{l,t}\right), \quad l = 1, \dots, K_t, \quad i = 1, \dots, V. \tag{7.29}$$

The equivalence between Equation (7.27) and (7.29) can be assessed as follow: from Equation (7.27), we have

$$\begin{aligned}
&p\left(N^1_t, \dots, N^{K_t}_t, D^l_{1,t}, \dots, D^l_{V,t} \mid \Theta_t\right) \\
&\quad = p\left(N^1_t, \dots, N^{K_t}_t \mid \Theta_t\right) p\left(D^l_{1,t}, \dots, D^l_{V,t} \mid N^1_t, \dots, N^{K_t}_t, \Theta_t\right) \\
&\quad = \left\{\prod_{l=1}^{K_t} N^l_t! \prod_{i=1}^{V} \frac{\left(\lambda^l_{i,t}\right)^{D^l_{i,t}}}{D^l_{i,t}!}\right\} \left\{\prod_{l=1}^{K_t} \frac{\exp\left(-\xi_{t,l}\right)}{N_{t,l}!} \xi^{N_{t,l}}_{t,l}\right\} \\
&\quad = \exp\left(-\sum_{l=1}^{K_t} \xi_{t,l}\right) \prod_{i=1}^{V} \prod_{l=1}^{K_t} \frac{\left(\lambda^l_{i,t}\xi_{t,l}\right)^{D^l_{i,t}}}{D^l_{i,t}!},
\end{aligned} \tag{7.30}$$

where the final equality follows noticing that $N^l_t = \sum_{i=1}^{V} D^l_{i,t}$. Similarly, from Equation (7.29) we first note that N^l_t is deterministically derived as $N^l_t = \sum_{i=1}^{V} D^l_{i,t}$. Consequently, we set $p(N^l_t = n) = \delta_{\sum_{i=1}^{V} D^l_{i,t}}(n)$. It follows that, from Equation (7.29), the likelihood can be written as

$$\begin{aligned}
&p\left(N^1_t, \dots, N^{K_t}_t, D^l_{1,t}, \dots, D^l_{V,t} \mid \Theta_t\right) \\
&\quad = p\left(D^l_{1,t}, \dots, D^l_{V,t} \mid \Theta_t\right) p\left(N^1_t, \dots, N^{K_t}_t \mid D^l_{1,t}, \dots, D^l_{V,t}, \Theta_t\right) \\
&\quad = \prod_{l=1}^{K_t} \left\{\exp(-\xi_{l,t}) \prod_{i=1}^{V} \frac{\left(\lambda^l_{i,t}\xi_{t,l}\right)^{D^l_{i,t}}}{D^l_{i,t}!}\right\} = \exp\left(-\sum_{l=1}^{K_t} \xi_{l,t}\right) \prod_{l=1}^{K_t} \prod_{i=1}^{V} \frac{\left(\lambda^l_{i,t}\xi_{t,l}\right)^{D^l_{i,t}}}{D^l_{i,t}!}.
\end{aligned}$$

The latter distribution matches with the one in Equation (7.30), hence we conclude that the two model formulation carries the same information. As a concluding remark, from Equation (7.30) it follows that, by introducing suitable latent variables, the likelihood can be expressed as a product over the active topics and the words in the vocabulary. This property is crucial for further developments and for the construction of inference algorithms.

Appendix of Chapter 7

F.1 Proof of the main results

F.1.1 Proof of Proposition 4

The proof builds on the size-biased representation of an ExpCRM, as presented in Broderick et al. (2018). We write

$$G_t = \sum_{x=1}^{\infty} \sum_{l=1}^{N_{t,x}} \theta_{x,l} \delta_{\lambda_{t,l}} ,$$

where $\theta_{x,l}$ are the jumps that generate $\xi_{t,l} = x$. Their distribution does not depend on l and it is given in (Broderick et al., 2018, Theorem 5.1). $N_{t,x}$ is Poisson distributed with mean M_x. Hence, we use the tower property of conditional expectation to write

$$E\left(\sum_{l=1}^{\infty} \mathbf{1}\left(\xi_{t,l} > 0\right)\right) = E\left(E\left(\sum_{l=1}^{\infty} \mathbf{1}\left(\xi_{t,l} > 0\right) \mid G_t = \sum_{x=1}^{\infty} \sum_{l=1}^{N_{t,x}} \theta_{x,l} \delta_{\lambda_{t,l}}\right)\right) .$$

The positive jumps $\xi_{t,l}$ are all such jumps which are generated for $x \geq 1$. Hence,

$$E\left(\sum_{l=1}^{\infty} \mathbf{1}\left(\xi_{t,l} > 0\right)\right) = E_G\left(\sum_{x \geq 1} N_{t,x}\right) = \sum_{x \geq 1} M_x ,$$

which concludes the proof of Equation (7.15). Regarding Equation (7.16), we first prove a slightly more general result as we evaluate the random measure on a measurable A set. Then,

$$E\left(\Theta_t(A)\right) = E\left(\sum_{l \geq 1} \xi_{t,l} \delta_{\lambda_{t,l}}(A)\right) = E\left(\sum_{l \geq 1} \xi_{t,l}\right) P_0(A) ,$$

where the final equality holds since the jumps and the atoms are independent. We now leverage the size-biased representation and write,

$$E\left(\Theta_t(A)\right) = P_0(A) E_G\left(\left(\sum_{x \geq 1} \sum_{l=1}^{N_{t,x}} x \mid G_t\right)\right) = P_0(A) E_G\left(\sum_{x \geq 1} x N_{t,x}\right) = P_0(A) \sum_{x \geq 1} x M_x .$$

F.1.2 Proof of Proposition 5

We want to compute the following expectation,

$$E\left(\Theta_{t+1}(A) \mid \Theta_t\right) = E\left(\sum_{l=1}^{K_t} \xi'_{t+1,l}\delta_{\lambda^*_{t,l}}(A) + \sum_{x=1}^{\infty}\sum_{l=1}^{N_{t,x}} x\delta_{\lambda_{t,l}}(A) \mid \Theta_t\right).$$

By linearity, we can divide the expectations of the two sums. The latter is straightforward since the innovation does not depend on the past. Moreover, $N_{t,x}$ is independent of the atoms. Hence,

$$E\left(\sum_{x=1}^{\infty}\sum_{l=1}^{N_{t,x}} x\delta_{\lambda_{t,l}}(A)\right) = \sum_{x\geq 1} xE\left(\sum_{l=1}^{N_{t,x}} \delta_{\lambda_{t,l}}(A)\right) = \sum_{x\geq 1} xE\left(P_0(A)N_{t,x}\right) = \sum_{x\geq 1} xM^*_x.$$

We are left with the computation of

$$E\left(\sum_{l=1}^{K_t} \xi'_{t+1,l}\delta_{\lambda^*_{t,l}}(A) \mid \Theta_t\right) = \sum_{l=1}^{K_t} E\left(\xi'_{t+1,l} \mid \xi_{t,l}\right)\delta_{\lambda^*_{t,l}}(A).$$

The expectation $E\left(\xi'_{t+1,l} \mid \xi_{t,l}\right)$ is computed with respect to the posterior predictive $h'(\cdot \mid \xi_{t,l})$. Consider a realization $\xi_{t,l} = y$, we recall that

$$h'(x \mid y) = \int_0^{\infty} h(x \mid s)\pi(s \mid y)\mathrm{d}s,$$

where $\pi(s \mid y) \propto h(y \mid s)\rho_{\text{ord}}(s)$ is the posterior distribution. Hence,

$$\begin{aligned} E\left(\xi'_{t+1,l} \mid \xi_{t,l}\right) &= \sum_{x\geq 1} xh'(x \mid y) = \sum_{x\geq 1} x\int_0^{\infty} h(x \mid s)\pi(s \mid y)\mathrm{d}s \\ &= \int_0^{\infty}\sum_{x\geq 1} xh(x \mid s)\,\pi(s \mid y)\mathrm{d}s = E\left(E(X) \mid \xi_{t,l} = y\right), \end{aligned} \tag{7.31}$$

where $X \sim h(\cdot \mid s)$. The distribution $h(\cdot \mid s)$ is given in Equation (7.3) is not parametrized according to the natural parameter $\tilde{s} = \eta(s)$. Hence, we perform the following change of variables in the previous integral,

$$\begin{aligned} \tilde{s} &= \eta(s), \\ A(s) &= A(\eta^{-1}(\tilde{s})) = M(\tilde{s}), \\ \mathrm{d}s &= \frac{1}{\eta'(\eta^{-1}(s))}\mathrm{d}\tilde{s}, \end{aligned}$$

where $\eta'(s) = \mathrm{d}\eta(s)/\mathrm{d}s$. The integral in Equation (7.31) is then equal to

$$\begin{aligned} E\left(\xi'_{t+1,l} \mid \xi_{t,l}\right) &= \int_0^{\infty}\sum_{x\geq 1} xh(x \mid s)\,\pi(s \mid y)\mathrm{d}s \\ &= \int_0^{\infty}\sum_{x\geq 1} xh(x \mid \tilde{s})\,\pi(\eta^{-1}(\tilde{s}) \mid y)\frac{1}{\eta'(\eta^{-1}(s))}\mathrm{d}\tilde{s}, \end{aligned}$$

where $\tilde{\pi}(\tilde{s} \mid y) = \pi(\eta^{-1}(\tilde{s}) \mid y)\frac{1}{\eta'(\eta^{-1}(s))}$ represents the induced posterior on the natural parameter and $h(\cdot \mid \tilde{s})$ is an element of the exponential family in natural form. Hence, by standard properties of such a family, the expected value of $X \sim h(\cdot \mid \tilde{s})$ is equal to the derivative of the cumulant $M(\tilde{s}) = A(\eta^{-1}(\tilde{s}))$.

$$E_{h(\cdot|\tilde{s})}(X) \;=\; \frac{\mathrm{d}A(\eta^{-1}(\tilde{s}))}{\mathrm{d}\tilde{s}}\,.$$

We are only left to compute

$$E\left(\xi'_{t+1,l} \mid \xi_{t,l}\right) \;=\; E_{\tilde{\pi}(\cdot|y)}\left(\frac{\mathrm{d}A(\eta^{-1}(\tilde{s}))}{\mathrm{d}\tilde{s}} \mid \xi_{t,l} = y\right)\,.$$

Under the hypothesis that,

$$\frac{1}{\eta'(\eta^{-1}(\tilde{s}))} \;=\; \exp\left\{\langle\tilde{a},\tilde{s}\rangle - \tilde{b}A(\eta^{-1}(\tilde{s}))\right\}\,,$$

we can write $\tilde{\pi}(\tilde{s} \mid y)$ proportional to

$$\exp\left\{(b+\tilde{b}+1)\frac{(a+\tilde{a}+y)}{b+\tilde{b}+1}\tilde{s} \;-\; (b+\tilde{b}+1)A(\eta^{-1}(\tilde{s}))\right\}\,.$$

Then, the result follows using the celebrated result in (Diaconis and Ylvisaker, 1979, Equation (2.10)).

F.1.3 Proof of Proposition 6

Firstly, we must compute the distribution of the jumps $\xi_{t,l}$ when they are selected from the innovation part in Equation (7.12). This is,

$$\begin{aligned} h^*(x) \;&=\; \int h(x \mid s)\frac{h(0 \mid s)\rho_{\mathrm{ord}}(s)}{\int h(0 \mid \theta)\rho_{\mathrm{ord}}(\theta)\mathrm{d}\theta}\mathrm{d}s \\ &=\; \kappa(x)\,\exp\{C(a+\phi(x)+\phi(0),b+2) - C(a+\phi(0),b+1)\}, \quad x \geq 0. \end{aligned}$$

However, we are conditioning on the l-th trait to be first active at time τ_l. Consequently, the jump $\xi_{\tau_l,l}$ can not be zero. It follows that $\xi_{\tau_l,l} \sim h_0$, where

$$h_0(x) \;=\; \mathbb{P}\left(\xi_{\tau_l,l} > 0 \mid \xi_{\tau_l,l} > 0\right) \;=\; \frac{h^*(x)}{1-h^*(0)}, \quad x \geq 1\,.$$

The result in Equation (7.17) follows directly.

F.1.4 Proof of Proposition 7

Consider the Markov's chain for the intensity of the l-th trait. To simplify the notation, we discard the dependence on label l and write $\{\xi_t\}_{t\geq 0}$. Without loss of generality, we consider a trait that first appears at time $t = 1$. Let $\mathbb{P}_i(\cdot) = \mathbb{P}(\cdot \mid \xi_1 = i)$, be the conditional probability with respect to the initial state. We recall the following important properties of the process: $a_0 = 0$, $S_{0,0} = 1$ and $S_{0,j} = 0$ for any $j \geq 1$. Then,

$$\mathbb{P}(L = l) \;=\; \sum_{x\geq 1}\mathbb{P}(\xi_1 = i)\mathbb{P}_i(L = l)\,,$$

hence, in the following the object of interest is $\mathbb{P}_i\,(L = l)$. The latter can also be interpreted as the first hitting probability of the state $x = 0$ given the initial condition $x = i$. The result for $l = 1$ is trivial, while for $l \geq 2$ the probability of interest can be expressed as

$$\mathbb{P}_i\,(L = l) \;=\; \mathbb{P}_i\left(\xi_{l+1} = 0, \bigcap_{m=2}^{l} \xi_m > 0\right).$$

Hence, for $l = 2$, we have

$$\begin{aligned}\mathbb{P}_i\,(L = l) \;&=\; \sum_{j=1}^{\infty} \mathbb{P}_i\,(\xi_3 = 0, \xi_2 = j) \;=\; \sum_{j=1}^{\infty} \mathbb{P}\,(\xi_3 = 0 \mid \xi_2 = j)\,\mathbb{P}\,(\xi_2 = j \mid \xi_1 = i) \\ &=\; \sum_{j=1}^{\infty} S_{i,j} S_{j,0} \;=\; \sum_{j=0}^{\infty} S_{i,j} S_{j,0} - S_{i,0} S_{0,0}\,.\end{aligned}$$

The result follows by recognizing that the first sum is $S^{(2)}_{i,0}$ and $S_{0,0} = 1$. Then, as induction hypothesis, assume

$$\mathbb{P}_i(L = l-1) = \mathbb{P}_i\left(\xi_{l+1} = 0, \bigcap_{m=2}^{l} \xi_m > 0\right) = S^{(l-1)} - S^{(l-2)}\,.$$

In order to prove the result for $L = l$, we introduce the set of labels $\Delta_1^{1:l}$ such that

$$\Delta_{1:l} = \{j_1 = 1,\, j_m \in \{1, 2, \ldots\},\, m = 1, \ldots, l\}$$

$$\begin{aligned}\mathbb{P}_i(L = l-1) &= \mathbb{P}_i\left(\xi_{l+1} = 0, \bigcap_{m=2}^{l} \xi_m > 0\right) \\ &= \sum_{j_1,\ldots,j_l \in \Delta_{1:l}} \prod_{m=2}^{l} \mathbb{P}\,(\xi_m = j_m \mid \xi_{m-1} = j_{m-1})\,\mathbb{P}\,(\xi_{l+1} = 0 \mid \xi_l = j_l) \\ &= \sum_{j_1,\ldots,j_l \in \Delta_{1:l}} \prod_{m=2}^{l} S_{j_{m-1}, j_m}\, S_{j_l, 0}\,.\end{aligned}$$

Working at the final step of the summation, we explicitly write the transition including the state $x = 0$.

$$\begin{aligned}\mathbb{P}_i(L = l-1) &= \sum_{j_1,\ldots,j_{l-1} \in \Delta_{1:(l-1)}} \prod_{m=2}^{l-1} S_{j_{m-1}, j_m} \left(\sum_{j_l=0}^{\infty} S_{j_{l-1}, j_l}\, S_{j_l, 0} \;-\; S_{j_{l-1}, 0} S_{0,0}\right) \\ &= \sum_{j_1,\ldots,j_{l-1} \in \Delta_{1:(l-1)}} \prod_{m=2}^{l-1} S_{j_{m-1}, j_m} \sum_{j_l=0}^{\infty} S_{j_{l-1}, j_l}\, S_{j_l, 0} \;-\; \sum_{j_1,\ldots,j_{l-1} \in \Delta_{1:(l-1)}} \prod_{m=2}^{l-1} S_{j_{m-1}, j_m} S_{j_{l-1}, 0}\,.\end{aligned}$$

We note that the final sum is $\mathbb{P}_i\left(\xi_l = 0, \bigcap_{m=2}^{l-1} \xi_m > 0\right) = \mathbb{P}_i(L = l-1)$, which equals $S^{(l-1)} - S^{(l-2)}$ for inductive hypothesis. Moreover, the first term marginalizes out the value out the Markov's chain at time $t = l$. Hence, it equals $\mathbb{P}_i\left(\xi_{l+1} = 0, \bigcap_{m=2}^{l-1} \xi_m = 0\right)$. Repeating the same steps as above with respect to the summation at time $l - 1$, we get

$$\mathbb{P}_i\left(\xi_{l+1} = 0, \bigcap_{m=2}^{l-1} \xi_m = 0\right) \;=\; \mathbb{P}_i\left(\xi_{l+1} = 0, \bigcap_{m=2}^{l-2} \xi_m = 0\right) \;-\; \mathbb{P}_i\left(\xi_{l-1} = 0, \bigcap_{m=2}^{l-2} \xi_m = 0\right).$$

Iterating the reasoning, it follows that

$$\mathbb{P}_i\left(L = l\right) \;=\; \mathbb{P}_i\left(\xi_{l+1} = 0\right) \;-\; \sum_{j=3}^{l} \mathbb{P}_i\left(\xi_j = 0, \bigcap_{m=2}^{j-1} \xi_m > 0\right) \;-\; \mathbb{P}_i\left(\xi_2 = 0\right)$$
$$= \; S_{i,0}^{(l)} - \sum_{j=3}^{l} \left(S_{i,0}^{(j-1)} - S_{i,0}^{(j-2)}\right) - S_{i,0} \;=\; S_{i,0}^{(l)} - S_{i,0}^{(l-1)} \,.$$

The induction hypothesis holds and the statement follows.

F.2 Posterior distribution of the centers

We introduce the latent variables $\theta_1, \ldots, \theta_K$ such that λ_l^{**} belongs to the h-th center if and only if $\theta_l = \zeta_h$. Moreover, let $\boldsymbol{\theta} = (\theta_1^{**}, \ldots, \theta_H^{**})$ be the corresponding unique values, whose frequencies are $(n_1, \ldots, n_H)$. Then, the posterior distribution of Λ admits the following representation

$$\Lambda \mid \boldsymbol{\Theta}_{1:T} \overset{\mathrm{d}}{=} \frac{1}{H + M^\star} \left\{ \sum_{h=1}^{H} \delta_{\theta_h} \;+\; \sum_{h=1}^{M^\star} \delta_{\theta'_h} \,, \right\}$$

where

- $M^\star \geq 0$ such that $M^\star \;\propto\; \omega^{M^\star} / (M^\star! (M^\star + K)^{K-1})$
- $\theta'_h \overset{\text{iid}}{\sim} P_0(\cdot)$.

F.3 Additional details about the sampling strategy

For sake of notation, we drop the particle label. Consider time $t = 1$.

- $q\left(K_1 \mid \gamma, c\right) = \text{Pois}\left(M_1\right)$, M_1 is defined in Example 7.2.1.
- $p\left(Z^1 \mid K_1\right) = \delta_{(1,\ldots,1)}$ is simply a point mass on a vector of length K_1 made of all ones, since at the first time step there is only the innovation part.
- $p\left(A^1 \mid K_1, \boldsymbol{X}_1\right) \;=\; \text{MatN}_{K_1 \times D}\left(\tilde{A}, \Sigma_{K_1}, \mathbf{I}_D\right)$, where $\text{MatN}_{K \times D}\left(\tilde{A}, \Sigma_K, V\right)$ is a matrix normal distribution of size $K \times D$. Specifically, the mean $\tilde{A}$ is a $K_1 \times D$ matrix such that the rows $\tilde{A}_l \in \mathbb{R}^D$ are equal to

$$\tilde{A}_l \;=\; \frac{1}{\sigma_X^2 + K_1 \sigma_A^2} \left(\sigma_A^2 \boldsymbol{X}_1 + \sigma_X^2 \boldsymbol{\mu}_0\right) \,,$$

  where $\boldsymbol{\mu}_0$ is the mean vector of the base measure. Hence, $\boldsymbol{\mu}_0 = (0, \ldots, 0)$ when fitting the model without centers presented in Section 7.2, while $\boldsymbol{\mu}_0 = \zeta_h$ when fitting the model with centers presented in Section 7.3.

  Moreover, the $K_1 \times K_1$ scale matrix Σ_{K_1} equals

$$\Sigma_{K_1} \;=\; \sigma_A^2 \left(\mathbf{I}_{K_1} - \frac{\sigma_A^2}{\sigma_X^2 + K_1 \sigma_A^2} \mathbf{1}_{K_1, K_1}\right) \,, \tag{7.32}$$

  where $\mathbf{I}_K$ is the identity matrix of size K and $\mathbf{1}_{K,K}$ is a $K \times K$ matrix with all entries equal to one.

- $p\left(X_1 \mid K_1\right)$ is the following marginal distribution

$$\begin{aligned} p\left(X_1 \mid K_1\right) &= \int p\left(\boldsymbol{X}_1 \mid A^1, K_1\right) p\left(A^1 \mid K_1, \boldsymbol{X}_1\right) \mathrm{d}A^1 \\ &= \int \mathrm{N}_D\left(\boldsymbol{X}_1 \mid (A^1)^T \mathbf{1}_{K_1}, \sigma_X^2 \mathbf{I}_{K_1}\right) \mathrm{MatN}_{K\times D}\left(A^1 \mid \tilde{A}, \Sigma_{K_1}, \mathbf{I}_{K_1}\right) \mathrm{d}A^1 \\ &= \mathrm{N}_D\left(\boldsymbol{X}_1 \mid K_1\boldsymbol{\mu}_0, (\sigma_X^2 + K_1\sigma_A^2)\mathbf{I}_{K_1}\right), \end{aligned}$$

  where $\mathbf{1}_K = (1, \ldots, 1)$ is a vector of K elements all equal to one.

The unnormalized weight in Equation (7.26) is derived as follows:

$$\begin{aligned} w_1 &= \frac{p\left(\boldsymbol{X}_1 \mid A^1, K_1\right)\, p\left(K_1 \mid \gamma, c\right) p\left(A^1 \mid K_1, Z^1, \sigma_A^2, \boldsymbol{\mu}_0\right)\, p\left(Z^1 \mid K_1\right)}{q\left(K_1 \mid \gamma, c\right)\, p\left(Z^1 \mid K_1\right)\, p\left(A^1 \mid K_1, \boldsymbol{X}_1\right)}, \\ &= \frac{p\left(\boldsymbol{X}_1 \mid A^1, K_1\right)\, p\left(A^1 \mid K_1, Z^1, \sigma_A^2, \boldsymbol{\mu}_0\right)}{p\left(\boldsymbol{X}_1 \mid A^1, K_1\right)\, p\left(A^1 \mid K_1, Z^1, \sigma_A^2, \boldsymbol{\mu}_0\right) / p\left(\boldsymbol{X}_1 \mid A^1, K_1\right)} \\ &= p\left(\boldsymbol{X}_1 \mid A^1, K_1\right), \end{aligned}$$

where we simply used Bayes' theorem to explicit $p\left(A^1 \mid K_1, \boldsymbol{X}_1\right)$.

For subsequent times $t \in \{2, \ldots, T\}$ we must consider the thinning and innovation mechanism described in Section 7.2.2.

- $q\left(N_t \mid \gamma, c\right) = \mathrm{Pois}\left(M_1^*\right)$, where $M_1^\star$ is defined in Example 7.2.1.
- $p\left(Z^{1:t} \mid N_t, Z^{1:t-1}\right)$ is composed of the thinning of the first $t-1$ rows and K_{t-1} columns and the innovation in the t-th row with N_t new columns.

$$\begin{aligned} p(Z_{t,l}) &= Z_{t-1,l}\left\{\left(\frac{1}{c+1}\right)^{Z_{t,l}} + \left(\frac{c}{c+1}\right)^{1-Z_{t,l}}\right\} + (1 - Z_{t-1,l})\delta_0, \quad l = 1, \ldots, K_{t-1}, \\ p(Z_{t,l}) &= \delta_1, \quad l = K_{t-1}+1, \ldots, K_t. \end{aligned}$$

- $p\left(A^{1:t} \mid A^{1:t-1}, N_t, \boldsymbol{X}_t\right)$ is such that the first K_{t-1} rows of $A^{1:t}$ are equal to $A^{1:t-1}$ while the new N_t rows are sampled from $\mathrm{MatN}_{N_t\times D}\left(A^\star, \Sigma_{N_t}, \mathbf{I}_D\right)$, where $A^\star$ is a $N_t \times D$ matrix such that the rows $A_l^\star$ are

$$A_l^\star = \frac{1}{\sigma_X^2 + N_t\sigma_A^2}\left(\sigma_A^2\left[(A^{1:t-1})^T\mathbf{1}_{K_{t-1}} - \boldsymbol{X}_1\right] + \sigma_X^2\boldsymbol{\mu}_0\right).$$

F.3.1 Full conditionals of static parameters

Consider the static parameters $\boldsymbol{\theta} = \left(\sigma_X^2, \gamma, c\right)$, hence assuming the two-parameters Beta process, i.e., $a = -1$. The components of $\boldsymbol{\theta}$ are assumed to be a prior independent. We consider the following set of hyperpriors and we give the corresponding full conditionals.

- If $\sigma_X^2 \sim \text{Inv-gamma}(a_x, b_x)$, then

$$\sigma_X^2 \mid \text{rest} \sim \text{Inv-gamma}\left(a_X + \frac{TD}{2},\, b_X + \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{X}_t - \boldsymbol{\mu}_t)^T(\boldsymbol{X}_t - \boldsymbol{\mu}_t)\right),$$

where we recall that $\boldsymbol{\mu}_t = \sum_{l \geq 1} \xi_{t,l} \lambda_{t,l}$.

- If $\gamma \sim \text{Gamma}(a_\gamma, b_\gamma)$, then

$$\gamma \mid \text{rest} \sim \text{Gamma}\left(a_\gamma + K,\ b_\gamma + \frac{cT+1}{c(c+1)}\right).$$

- If $c \sim \text{Gamma}(a_c, b_c)$, then $c \mid \text{rest} \sim p(c \mid \text{rest})$, which is given by

$$p(c \mid \text{rest}) \propto c^{N_T - S_T + a_c - 1}(c+1)^{-N_T^{\text{new}} - N_T} \exp\left(-\frac{\gamma T}{c+1} - b_c c\right),$$

where $N_T = \sum_{t=2}^{T} \rho_{t-1}$, $S_T = \sum_{t=2}^{T} \sum_{l=1}^{\rho_{t-1}} \xi_{t,l}^{th}$.

Conclusions

The first part of this thesis has explored a broad range of techniques for modeling spectrometric data through the framework of Gaussian graphical models, with the goal of uncovering the dependence structure among the chemical components present in strawberry purees. To this end, we have reinterpreted the original problem as a functional data smoothing task, by means of suitable B-spline basis functions. This formulation provides a flexible representation of the absorbance spectra while enabling a probabilistic treatment of the underlying dependencies. Within this framework, our Bayesian smoothing model achieves two key objectives: on the one hand, it facilitates the sharing of information across different spectral curves; on the other hand, it induces sparsity in the shared precision matrix of the smoothing coefficients, thus revealing a latent graphical structure that encodes conditional dependencies.

Our findings in Chapter 1 revealed a graph structure characterized by strong associations between spectral bands that are not necessarily closely located in the spectrum. In particular, the estimated dependence structure exhibits a clear block organization: connections occur primarily between groups of nodes rather than being uniformly distributed across the spectrum. This intriguing block pattern was independently validated by Waghmare and Panaretos (2024), who analyzed the same dataset with a completely different methodology and obtained comparable results. These observations motivated us to extend our analysis and to explicitly consider possible dependencies among groups of nodes. To this end, in Chapter 2 we introduced a new class of priors, called block graph priors, which make it possible to incorporate available prior knowledge about a given node partition into the Gaussian graphical model. Specifically, we worked conditionally on an empirical partition provided by domain experts. The idea of investigating block structures has been studied in parallel, though independently, in the Southern Italian Community Structure model of van den Boom et al. (2022b), which, however, was not designed for spectrometric data and therefore does not account for the natural ordering of the spectral bands. Related approaches that encourage block structures in graphical models can also be found in Liu et al. (2025). To further demonstrate the potential of our method in real life problems, we plan to analyze samples from different types of fruit purees in future research, with the aim of identifying key nutritional differences. This approach follows the spirit of Casa et al. (2022), who used infrared spectrometric data from milk samples to distinguish between different feeding strategies of dairy cows.

Finally, motivated by the need to learn the partition of the nodes from the data rather than assuming it known in advance, in Chapter 3 we introduced a Product Partition Model to infer such a block structure. The main methodological contribution lies in interpreting the natural ordering of the nodes, imposed by the functional nature of the absorbance spectra, as a changepoint model, and in generalizing the extended Stochastic Block Model framework of Legramanti et al. (2022) from exchangeable Gibbs-type priors to their ordered counterpart, as formalized in Martínez and Mena (2014). This final chapter is still a work in progress. The primary line of research is to inform the model with prior knowledge from an empirical

partition while still allowing the partition to be estimated from the data. Recent contributions by Page et al. (2022) and Paganin et al. (2025) move in this direction, but additional methodological advances are required to adapt these approaches to the ordered partition setting.

The building blocks of Bayesian Gaussian graphical models are the prior distributions for the graph and, conditionally on the graph, the prior for the precision matrix. In Chapters 2 and 3, we focused primarily on the former, that is, the prior on the graph, while consistently adopting the G-Wishart distribution for the precision matrix. The G-Wishart, however, is not the only option, nor is it the most scalable one. In fact, Wang (2015) demonstrated this by introducing the spike-and-slab prior for Gaussian graphical models. This prior shrinks each off-diagonal element of the precision matrix through a product of univariate mixtures: a spike (a normal distribution with small variance) assigned to coefficients corresponding to absent edges, and a slab (a normal distribution with larger variance) for those corresponding to present edges. While the resulting precision matrix is not sparse in the strict sense, the spike induces coefficients so small that they become negligible compared to those generated by the slab component. For the purposes of this thesis, we continued to develop our methodology under the G-Wishart prior. This choice was dictated by the modeling constraint introduced in Chapter 2, where blocks are required to be either fully connected or completely disconnected. Under such a constraint, the spike-and-slab prior of Wang (2015) would be impractical, as it remains unclear how to set up reversible jump moves in this framework. Nevertheless, exploring whether such a prior could be employed in the context of Chapter 3 would be both interesting and potentially very beneficial.

The main computational challenge lies in efficiently exploring the vast discrete spaces: block-structured graphs in Chapter 2, and admissible partitions in Chapter 3. In the former case, we introduced the BDRJ algorithm, a trans-dimensional MCMC sampler on the joint space of graphs and precision matrices, which jointly modifies an arbitrary number of edges at each iteration. By employing a double reversible move, the algorithm circumvents the computation of the normalizing constant $I_G(b, D)$, thereby overcoming the primary inferential obstacle associated with G-Wishart priors. In the latter case, we relied on the adaptive split-merge algorithm originally proposed by Benson and Friel (2018) for changepoint models, which we adapted to the specific requirements of our framework.

The second part of the thesis mainly concerns the introduction and the study of the Vec-FDP prior in the context of partially exchangeable data. In particular, in Chapter 5 we presented an innovative Bayesian nonparametric model for the analysis of grouped data. Furthermore, we provided a comprehensive Bayesian analysis of this novel model class, delving into the examination of the pEPPF, posterior distributions and predictive distributions. A special emphasis has been placed on vectors of finite Dirichlet processes, which stand out as a noteworthy example in this context. Besides, we have also defined the HMFM as a natural extension of the work by Miller and Harrison (2018). Based on our theory, marginal and conditional algorithms have been developed, which we theoretically showed that achieve better scaling with respect to current implementation of the HDP as the number of data increases. This improvement comes from a simpler restaurant franchise-like representation of the Vec-FDP prior with respect to the HDP's complex distinction between tables and menus. In this regard, we point out that the recent work by Catalano and Sole (2025) studies a new representation of the HDP which removes the need for tables, enhancing computational methods to sample from it. The chapter includes a simulation study in which we empirically compared the proposed HMFM, which assumes an almost surely finite number of mixture components, with the HDP, which instead operates under the assumption of infinitely

many mixture components. As a direction for future work, we plan to compare HMFM with the earlier version of the HMFM (Miller, 2014), in which each group is endowed with a random discrete measure distributed according to a nonparametric prior with a finite number of components. In this setting, we expect to observe the same computational advantages already highlighted in the comparison with the HDP. However, it would be particularly interesting to provide a more detailed discussion and empirical comparisons to clarify when each formulation is more suitable, whether they yield comparable inferential performance, and to what extent the benefits of the proposed approach are primarily computational and modeling-driven. In particular, our formulation may be more accessible to practitioners, as it avoids the need for a hierarchical structure of nonparametric priors while retaining modeling flexibility.

In Chapter 6, the Vec-FDP prior has been employed to introduce a model-based approach to the problem of discovering the number of distinct and shared species in two different areas. The novel approach relied entirely on closed-form expressions and exact calculations, and it is focused on two primary objectives: estimating the discovery probability of new shared species and predicting the number of new shared species in an additional, unobserved future samples. In Chapter 7, we extended the framework from static to dynamic settings by introducing a general and flexible class of time-dependent random measures with arbitrary marginal distributions. In contrast to Chapter 6, where the discrete random structure was directly employed to model the data, here it serves as a latent process governing temporal evolution. Our focus shifted to time-evolving latent feature models, in which each observation can express multiple features that vary across time. In particular, we illustrated how the proposed framework naturally accommodates dynamic feature allocation, allowing features to emerge, disappear, and reappear over time. Furthermore, by introducing random centers, we enhanced the model's ability to capture feature reappearance and to maintain temporal coherence in feature identification.

We now pinpoint several open problems related to our works, which are left for future research. A possible direction of research aims to enhance between-group dependence based on our approach. An intriguing extension we plan to investigate is to adopt a construction akin to compound random measures (Griffin and Leisen, 2017). Following this idea, we could replace the unnormalized group specific weights by replacing them with a product between a shared component across groups and an idiosyncratic component, which both depends on the specific group as well as the specific atom. This modification would break the conditional independence of the unnormalized weights. However, it could be advantageous in situations where additional information sharing is desired, as for the overlapping communities in modular graphs (Todeschini et al., 2020).

Additionally, we assumed that the atoms are interdependent and identically distributed according to a diffuse base measure P_0. However, recent research lines have explored the use of repulsive point processes to setup Bayesian nonparametric priors (Petralia et al., 2012; Beraha et al., 2021, 2025a) and applied them both to model-based clustering for high-dimensional data (Ghilotti et al., 2024) as well as in ecology problems Beraha et al. (2025b).

Regarding the problem of discovering new species, the most stringent assumption we made is that the two observed samples are generated from a single underlying population, manifesting with different proportions across the two groups. An intriguing direction for future research would be to relax this assumption, thus allowing the number of species to be group-specific. Alternatively, a similar effect could be achieved by placing an asymmetric Dirichlet distribution on the species proportions, while allowing some of the Dirichlet parameters to shrink towards zero. This second approach would serve as a Bayesian

nonparametric counterpart to the model employed in Chao et al. (2000) and its subsequent extensions. Moreover, the exact formulas presented in Chapter 6 require the computation of the generalized factorial coefficients, whose evaluation scales quadratically with the sample size. We hope that a more in-depth investigation into the asymptotic behavior of our model may pave the way for suitable approximations that make our solutions more scalable, similar to how Favaro et al. (2009) and Contardi et al. (2025) improved upon Lijoi et al. (2007a) in the exchangeable case.

Finally, several open questions remain regarding the time-evolving model introduced in Chapter 7. In the parametric framework, Pitt and Walker (2005) demonstrated that the construction proposed by Pitt et al. (2002) can be extended to define more elaborate temporal models beyond the simple first-order autoregressive case. An interesting direction for future research is to investigate whether such generalizations can be carried over to the nonparametric setting as well. Moreover, the implementation and in-depth study of the dynamic Poisson factor analysis model introduced in Section 7.6 remain important steps toward fully assessing the empirical performance and practical applicability of our proposed framework.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume No. 55 of *National Bureau of Standards Applied Mathematics Series*. U. S. Government Printing Office, Washington, DC. 177

Acharya, A., Ghosh, J., and Zhou, M. (2015). Nonparametric Bayesian factor analysis for dynamic count matrices. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 220

Aiello, L., Argiento, R., Legramanti, S., and Paci, L. (2025). Bayesian nonparametric clustering for spatio-temporal data, with an application to air pollution. *arXiv 2505.24694*. 83

Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238. 46

Amongero, M. and Blasi, P. D. (2025). Bayesian community detection in assortative stochastic block model with unknown number of communities. *arXiv 2506.19576*. 50

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, Chichester, 3rd edition. 81

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342. 215, 216

Antoniano-Villalobos, I. and Walker, S. G. (2016). A nonparametric model for stationary time series. *Journal of Time Series Analysis*, 37(1):126–142. 85

Archer, E., Park, I. M., and Pillow, J. W. (2014). Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research*, 15(81):2833–2868. 166

Argiento, R., Cremaschi, A., and Vannucci, M. (2020). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association*, 115(529):318–333. 85, 95, 124

Argiento, R. and De Iorio, M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics*, 50(5):2641–2663. 4, 75, 76, 79, 80, 93, 95, 97, 100, 101, 107, 125, 136, 137, 143, 154, 156, 158, 194, 209, 211

Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92:317–335. 24, 25, 28, 30, 41

Atkins, P. and De Paula, J. (2013). *Elements of physical chemistry*. Oxford University Press, USA. 37

Bacallado, S., Favaro, S., and Trippa, L. (2015). Bayesian nonparametric inference for shared species richness in multiple populations. *Journal of Statistical Planning and Inference*, 166:14–23. 86

Baccelli, F., Błaszczyszyn, B., and Karray, M. (2020). *Random measures, point processes, and stochastic geometry*. Inria. 73, 90, 91, 92, 127, 128, 131, 134

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221. 81

Balocchi, C., Camerlenghi, F., and Favaro, S. (2024a). A Bayesian nonparametric approach to species sampling problems with ordering. *Bayesian Analysis*, 1:1–26. 165

Balocchi, C., Favaro, S., and Naulet, Z. (2024b). Bayesian nonparametric inference for "species-sampling" problems. *Statistical Science (forthcoming)*. 79

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897. 17

Barry, D. and Hartigan, J. A. (1992). Product Partition Models for Change Point Problems. *The Annals of Statistics*, 20(1):260 – 279. 213

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319. 4, 48, 52, 213

Bassetti, F., Casarin, R., and Rossini, L. (2020). Hierarchical species sampling models. *Bayesian Analysis*, 15(3):809–838. 85, 93, 107, 113

Beal, M., Ghahramani, Z., and Rasmussen, C. (2001). The infinite hidden Markov model. *Advances in neural information processing systems*, 14. 81

Benson, A. and Friel, N. (2018). Adaptive MCMC for multiple changepoint analysis with applications to large datasets. *Electronic Journal of Statistics*, 12(2):3365–3396. 4, 55, 57, 230

Beraha, M., Argiento, R., Camerlenghi, F., and Guglielmi, A. (2025a). Bayesian mixture models with repulsive and attractive atoms. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 138, 210, 231

Beraha, M., Argiento, R., Møller, J., and Guglielmi, A. (2021). MCMC computations for Bayesian mixture models using repulsive point processes. *Journal of Computational and Graphical Statistics*, 31:1–37. 231

Beraha, M., Camerlenghi, F., and Ghilotti, L. (2025b). Bayesian calculus and predictive characterizations of extended feature allocation models. *arXiv: 2502.10257*. 231

Beraha, M., Camerlenghi, F., and Ghilotti, L. (2025c). Palm distributions of superposed point processes for statistical inference. *arXiv :2508.20924v1*. 210

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69(2):447–457. 15

Blackwell, D. (1973). Discreteness of Ferguson selections. *The Annals of Statistics*, 1(2):356–358. 72

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84. 219

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 113–120. 219

Blei, D. M. and Lafferty, J. D. (2009). Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC. 219

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3:993–1022. 81, 219

Borgatti, S. P. and Everett, M. G. (2000). Models of core/periphery structures. *Social Networks*, 21(4):375–395. 50

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31(4):929–953. 74, 201

Broderick, T., Mackey, L., Paisley, J., and Jordan, M. I. (2015). Combinatorial clustering and the beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:290–306. 219

Broderick, T., Wilson, A. C., and Jordan, M. I. (2018). Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 24(4B):3181 – 3221. 5, 84, 199, 200, 221

Cai, Q., Kang, J., and Yu, T. (2020). Bayesian network marker selection via the thresholded graph Laplacian Gaussian prior. *Bayesian Analysis*, 15(1):79–102. 11

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019a). Latent nested nonparametric priors (with discussion). *Bayesian Analysis*, 14(4):1303–1356. 86

Camerlenghi, F., Favaro, S., Masoero, L., and Broderick, T. (2024). Scaled process priors for Bayesian nonparametric estimation of the unseen genetic variation. *Journal of the American Statistical Association*, 119(545):320–331. 165

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019b). Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92. 85, 95, 97, 100, 125, 130, 152

Camerlenghi, F., Lijoi, A., and Prünster, I. (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis*, 156:18–28. 86

Cappé, O., Robert, C. P., and Rydén, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):679–700. 17

Carminati, A., Beraha, M., Camerlenghi, F., and Guglielmi, A. (2025). Hierarchical shot-noise cox process mixtures for clustering across groups. *arXiv 2510.14681*. 210

Caron, F. and Teh, Y. (2012). Bayesian nonparametric models for ranked data. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. 84

Casa, A., O'Callaghan, T. F., and Murphy, T. B. (2022). Parsimonious Bayesian factor analysis for modelling latent structures in spectroscopy data. *The Annals of Applied Statistics*, 16(4):2417–2436. 9, 229

Catalano, M. and Sole, C. D. (2025). Hierarchical random measures without tables. *arXiv 2505.02653*. 230

Chakraborty, S. and Lozano, A. C. (2019). A graph Laplacian prior for Bayesian variable selection and grouping. *Computational Statistics & Data Analysis*, 136:72–91. 11

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):265–270. 149

Chao, A., Chiu, C.-H., Colwell, R. K., Magnago, L. F. S., Chazdon, R. L., and Gotelli, N. J. (2017). Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on good-turing theory. *Ecology*, 98(11):2914–2929. 150, 164, 168, 170, 171

Chao, A., Hwang, W.-H., Chen, Y.-C., and Kuo, C.-Y. (2000). Estimating the number of shared species in two communities. *Statistica Sinica*, 10(1):227–246. 150, 169, 170, 172, 232

Chao, A., Ma, K., Hsieh, T., Chiu, C.-H., and Chao, M. A. (2016). Package SpadeR. *Species-richness prediction and diversity estimation with R*. 169

Chao, A., Shen, T.-J., and Hwang, W.-H. (2006). Application of Laplace's boundary-mode approximations to estimate species and shared species richness. *Australian & New Zealand Journal of Statistics*, 48(2):117–128. 150

Chao, A. and Yang, M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, 80(1):193–201. 76

Charalambides, C. A. (2002). *Enumerative combinatorics*. CRC Press. 53, 66, 98, 141, 157, 161, 175, 192

Chernoff, H. and Zacks, S. (1964). Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time. *The Annals of Mathematical Statistics*, 35(3):999 – 1018. 213

Christen, J. A. and Fox, C. (2005). Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810. 25

Chu, W. and Ghahramani, Z. (2009). Probabilistic models for incomplete multi-dimensional arrays. In *Artificial Intelligence and Statistics*, pages 89–96. PMLR. 81

Chuang, C., Shen, T., and Hwang, W. (2015). Estimating the number of shared species by a jackknife procedure. *Environmental and Ecological Statistics*, 22:759–778. 150, 170

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660. 83

Codazzi, L., Colombi, A., Gianella, M., Argiento, R., Paci, L., and Pini, A. (2022). Gaussian graphical modeling for spectrometric data analysis. *Computational Statististics & Data Analysis*, 174:e107416. xiii, 9, 11, 46

Cohen, M., Allamandola, L., Tielens, A., Bregman, J., Simpson, J., Witteborn, F. C., Wooden, D., and Rank, D. (1986). The infrared emission bands. i-correlation studies and the dependence on c/o ratio. *The Astrophysical Journal*, 302:737–749. 9

Colombi, A., Argiento, R., Camerlenghi, F., and Paci, L. (2024a). Hierarchical mixture of finite mixtures. *Bayesian Analysis*, 20(4):1 – 29. 93, 176, 195

Colombi, A., Argiento, R., Camerlenghi, F., and Paci, L. (2025a). How many unseen species are in multiple areas? *arXiv 2502.04122*. 149

Colombi, A., Argiento, R., Paci, L., and Pini, A. (2024b). Learning block structured graphs in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 33(1):152–165. 23

Colombi, A., Paci, L., and Pini, A. (2025b). Learning block structures in Gaussian graphical models for spectrometric data analysis. In Pollice, A. and Mariani, P., editors, *Methodological and Applied Statistics and Demography III. SIS 2024*, Italian Statistical Society Series on Advances in Statistics, pages 444–449. Springer. 45

Colwell, R. K. et al. (2009). Biodiversity: concepts, patterns, and measurement. *The Princeton guide to ecology*, 663:257–263. 78, 163, 164

Contardi, C., Dolera, E., and Favaro, S. (2025). Gaussian credible intervals in Bayesian nonparametric estimation of the unseen. *arXiv 2501.16008*. 78, 232

Corradin, R., Danese, L., and Ongaro, A. (2022). Bayesian nonparametric change point detection for multivariate time series with missing observations. *International Journal of Approximate Reasoning*, 143:26–43. 55, 57, 58

Crainiceanu, C. and Goldsmith, A. (2010). Bayesian functional data analysis using WinBUGS. *Journal of Statistical Software*, 32(11):1–33. 10

Cremaschi, A., Argiento, R., Iorio, M. D., Shirong, C., Chong, Y. S., Meaney, M., and Kee, M. (2022). Seemingly unrelated multi-state processes: A Bayesian semiparametric approach. *Bayesian Analysis*, pages 1–23. 27

Cremaschi, A., Argiento, R., Shoemaker, K., Peterson, C., and Vannucci, M. (2019). Hierarchical normalized completely random measures for robust graphical modeling. *Bayesian Analysis*, 14(4):1271–1301. 10, 33

Dahl, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture model. *In Do, K.-A., Müller, P., and Vannucci, M. (eds.), Bayesian Inference for Gene Expression and Proteomics, Cambridge University Press*, page 201–218. 107

Dahl, D. B., Johnson, D. J., and Müller, P. (2022). Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, 31:1189–1201. 107, 117

Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II: General Theory and Structure*. Probability and Its Applications (New York). Springer, New York, second edition. 73

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272 – 1317. 25

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229. 48, 52, 53, 75, 95, 154, 162

de Finetti, B. (1937). La prévision : ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, 7(1):1–68. 71, 72, 85

De Finetti, B. (1938). Sur la condition d'" equivalence partielle.". *Actualités scentiques et industrielles*, 739:5–18. 85

De Iorio, M., Favaro, S., Guglielmi, A., and Ye, L. (2023). Bayesian nonparametric mixture modeling for temporal dynamics of gender stereotypes. *The Annals of Applied Statistics*, 17(3):2256–2278. 84

Defernez, M., Kemsley, E. K., and Wilson, R. H. (1995). Use of infrared spectroscopy and chemometrics for the authentication of fruit purees. *Journal of Agricultural and Food Chemistry*, 43(1):109–113. 37

Dellaportas, P., Giudici, P., and Roberts, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā: The Indian Journal of Statistics*, 65(1):43–55. 25

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281. 84, 102, 200, 201, 223

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212. 10, 14

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433. 25, 28, 29

Dolmeta, P., Argiento, R., and Montagna, S. (2023). Bayesian GARCH modeling of functional sports data. *Statistical Methods & Applications*, 32:401–423. 114, 115

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. In *Handbook of Nonlinear Filtering*, volume 12, pages 656–704. 215

Downey, G., Briandet, R., Wilson, R. H., and Kemsley, E. K. (1997). Near- and mid-infrared spectroscopies in food authentication: Coffee varietal identification. *Journal of Agricultural and Food Chemistry*, 45(11):4357–4361. 9, 18

Dunson, D. B. and Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11–25. 81

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18. 58

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: how many words did shakespeare know? *Biometrika*, 63(3):435–447. 76

Erdös, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290–298. 47

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588. 79, 80

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–112. 53

Fang, H. and Lai, T.-Y. (1997). Co-kurtosis and capital asset pricing. *Financial Review*, 32(2):293–307. 124

Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):993–1008. 78, 165, 232

Favaro, S. and Teh, Y. W. (2013). MCMC for normalized random measure mixture models. *Statistical Science*, 28(3):335–359. 80, 101, 143

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230. 72, 80

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Elsevier. 79

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12(1):42–58. 149

Fortini, S. and Petrone, S. (2025). Exchangeability, prediction and predictive modeling in Bayesian statistics. *Statistical Science*, 40(1):40–67. 72

Franzolini, B., Iorio, M. D., and Eriksson, J. (2025a). Conditional partial exchangeability: a probabilistic framework for multi-view clustering. *arXiv 2307.01152*. 84

Franzolini, B., Lijoi, A., Prünster, I., and Rebaudo, G. (2025b). Multivariate species sampling models. *arXiv:2503.24004*. 86

Friend, I. and Westerfield, R. (1980). Co-skewness and capital asset pricing. *The Journal of Finance*, 35(4):897–913. 124

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in data analysis and classification*, 13:33–64. 102

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16(4):1279–1307. 86, 96

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035. 83

Geng, J., Bhattacharya, A., and Pati, D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526):893–905. 48

Ghahramani, Z. and Griffiths, T. (2005). Infinite latent feature models and the indian buffet process. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press. 82, 200, 213, 214

Ghilotti, L., Beraha, M., and Guglielmi, A. (2024). Bayesian clustering of high-dimensional data via latent repulsive mixtures. *Biometrika*, 112(2). 231

Ghilotti, L., Camerlenghi, F., and Rigon, T. (2025). Bayesian analysis of product feature allocation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 81

Giudici, P. and Castelo, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Machine learning*, 50(1-2):127–158. 15, 24

Giudici, P. and Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801. 25, 28, 29

Gnedin, A. (2010). A species sampling model with finitely many types. *Electronic Communications in Probability*, 15:79 – 88. 48, 79

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138:5674–5685. 51, 75, 95, 153, 154, 158

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264. 5, 149

Good, I. J. and Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63. 149

Gopalan, P. K., Gerrish, S., Freedman, M., Blei, D., and Mimno, D. (2012). Scalable inference of overlapping communities. *Advances in Neural Information Processing Systems*, 25. 51

Gotelli, N. J. and Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4(4):379–391. 78

Grazian, C. (2025). Advances in Bayesian random partition models: A comprehensive review. *arXiv 2303.17182*. 81

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732. 4, 48

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. 86, 231

Griffin, J. E. and Walker, S. G. (2011). Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20:241–259. 80

Griffiths, T. L. and Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(32):1185–1224. 214

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242. 55

Hartigan, J. (1990). Partition models. *Communications in Statistics - Theory and Methods*, 19(8):2745–2756. 48, 52, 83, 213

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80:470–501. 71

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(1):1259–1294. 75, 200

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer. 116

Holland, J. K., Kemsley, E. K., and Wilson, R. H. (1998). Use of Fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purées. *Journal of the Science of Food and Agriculture*, 76(2):263–269. 9, 18

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137. 4, 47, 81

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–396. 74, 201

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218. 107

Iorio, M. D., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An anova model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215. 83

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173. 80

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87. 93

James, L. F., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97. 80, 100, 101, 125, 137, 138, 143

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20:388–400. 14, 15, 23, 25, 31, 37

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*. Probability and its Applications. Springer, New York. 96

Kallenberg, O. (2017). *Random measures, theory and applications*, volume 77 of *Probability Theory and Stochastic Modelling*. Springer, Cham. 74

Kallenberg, O. (2021). *Foundations of modern probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer Cham. 91

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5. 48

Kemsley, E. K., Holland, J. K., Defernez, M., and Wilson, R. H. (1996). Detection of adulteration of raspberry purees using infrared spectroscopy and chemometrics. *Journal of Agricultural and Food Chemistry*, 44(12):3864–3870. 9, 18

Kingman, J. F. C. (1975). Random discrete distributions (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 37:1–22. 77

Kingman, J. F. C. (1993). *Poisson Processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications. 73, 203, 209

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer, New York. 47

Kumar, S., Ying, J., de Miranda Cardoso, J. V., and Palomar, D. P. (2020). A unified framework for structured graph learning via spectral constraints. *Journal of Machine Learning Research*, 21(22):1–60. 46

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212. 10

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford. 10, 23

Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50. 48

Lee, J., James, L. F., and Choi, S. (2016). Finite-dimensional BFRY priors and variational Bayesian inference for power law models. In *Advances in Neural Information Processing Systems*, pages 3162–3170. 201

Lee, S. Y. and Song, X.-Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika*, 29:23–39. 81

Lee, S.-Y., Zhang, W., and Song, X.-Y. (2002). Estimating the covariance function with functional data. *British Journal of Mathematical and Statistical Psychology*, 55(2):247–261. 9

Legramanti, S., Rigon, T., Durante, D., and Dunson, D. B. (2022). Extended stochastic block models with application to criminal networks. *The annals of applied statistics*, 16(4):2369–2395. 4, 48, 52, 81, 229

Lenkoski, A. (2013). A direct sampler for G-Wishart variates. *Stat*, 2(1):119–128. 4, 16, 25, 28, 29, 30, 31, 32, 33

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157. 25, 28

Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Annals of Statistics*, 35(3):1278–1323. 13

Li, B. and Solea, E. (2018). A nonparametric graphical model for functional data with application to brain networks based on fMRI. *Journal of the American Statistical Association*, 113(524):1637–1655. 11

Liang, X., Caron, A., Livingstone, S., and Griffin, J. (2023a). Structure learning with adaptive random neighborhood informed MCMC. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 40760–40772. 55

Liang, X., Livingstone, S., and Griffin, J. (2022). Adaptive random neighbourhood informed Markov chain Monte Carlo for high-dimensional bayesian variable selection. *Statistics and Computing*, 32(5):84. 55

Liang, X., Livingstone, S., and Griffin, J. (2023b). Adaptive MCMC for Bayesian variable selection in generalised linear models and survival models. *Entropy*, 25(9). 55

Lijoi, A., Mena, R. H., and Prünster, I. (2007a). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786. 78, 156, 165, 175, 232

Lijoi, A., Mena, R. H., and Prünster, I. (2007b). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):715–740. 74, 83, 116, 201

Lijoi, A., Nipoti, B., and Prünster, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260–1291. 86, 95

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *Bayesian Nonparametrics*, pages 80–136. Cambridge University Press, Cambridge. 73

Lijoi, A., Prünster, I., and Rigon, T. (2020). The Pitman–Yor multinomial process for mixture modelling. *Biometrika*, 107(4):891–906. 76

Liu, C., Kowal, D. R., Doss-Gollin, J., and Vannucci, M. (2025). Bayesian functional graphical models with change-point detection. *Computational Statistics & Data Analysis*, 206:108–122. 229

Liu, F., Chakraborty, S., Li, F., Liu, Y., and Lozano, A. C. (2014). Bayesian regularization via graph Laplacian. *Bayesian Analysis*, 9(2):449–474. 11

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. i. density estimates. *The Annals of Statistics*, 12(1):351–357. 79

Lopes, H. F. and West, M. (2004). Bayesian model assesment in factor analysis. *Statistica Sinica*, 14(1):41–67. 212

Lu, X. and Szymanski, B. K. (2019). A regularized stochastic block model for the robust community detection in complex networks. *Scientific reports*, 9(1):13247. 51

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, volume 1, pages 50–55. Alexandria, VA. 82, 83

MacEachern, S. N. (2000). Dependent Dirichlet processes. *Technical Report, Department of Statistics, Ohio State University, Columbus, OH.*, 5. 82, 83

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232. 29

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and computing*, 26(1-2):303–324. 102

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2):285–295. 102

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press. 81

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *Journal of the American Statistical Association*, 99(468):1108–1118. 76

Martínez, A. F. and Mena, R. H. (2014). On a nonparametric change point detection model in Markovian regimes. *Bayesian Analysis*, 9(4):823–857. 4, 48, 52, 53, 54, 66, 83, 229

Masoero, L., Camerlenghi, F., Favaro, S., and Broderick, T. (2021). More for less: predicting and maximizing genomic variant discovery via Bayesian nonparametrics. *Biometrika*, 109(1):17–32. 81

Matechou, E. and Argiento, R. (2023). Capture-recapture models with heterogeneous temporary emigration. *Journal of the American Statistical Association*, 118(541):56–69. 55

Mcculloch, C. E. and Neuhaus, J. M. (2013). *Generalized Linear Mixed Models*, chapter 4, pages 28–33. John Wiley & Sons, Ltd. 212

Mena, R. H. and Walker, S. G. (2005). Stationary autoregressive models via a Bayesian nonparametric approach. *Journal of Time Series Analysis*, 26(6):789–805. 84, 85

Meza-Márquez, O. G., Gallardo-Velázquez, T., and Osorio-Revilla, G. (2010). Application of mid-infrared spectroscopy with multivariate analysis and soft independent modeling of class analogies (SIMCA) for the detection of adulterants in minced beef. *Meat Science*, 86(2):511–519. 9, 18

Miller, J. W. (2014). *Nonparametric and variable-dimension Bayesian mixture models: Analysis, comparison, and new methods*. PhD thesis, Brown University. 93, 231

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356. 4, 75, 80, 86, 96, 154, 158, 178, 179, 230

Moghaddam, B., Khan, E., Murphy, K. P., and Marlin, B. M. (2009). Accelerating Bayesian structural inference for non-decomposable Gaussian graphical models. *Advances in Neural Information Processing Systems*, 22:1285–1293. 25, 28

Mohammadi, A. and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138. 4, 15, 16, 28, 33, 54

Mohammadi, R., Massam, H., and Letac, G. (2021). Accelerating Bayesian structure learning in sparse Gaussian graphical models. *Journal of the American Statistical Association*, pages 1–14. 28, 33, 39

Mohammadi, R. and Wit, E. C. (2019). BDgraph: An R package for Bayesian structure learning in graphical models. *Journal of Statistical Software*, 89(3):1–30. 16, 25, 33

Morisita, M. (1959). Measuring of dispersion of individuals and analysis of the distributional patterns. *Memories of the Faculty of Science, Kyushu University. Series E: Biology*, pages 215–235. 164

Morrissey, E. R., Juárez, M. A., Denby, K. J., and Burroughs, N. J. (2011). Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression. *Biostatistics*, 12(4):682–694. 11

Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65:391–411. 212

Muliere, P. and Secchi, P. (1995). A note on a proper Bayesian bootstrap. Technical Report 18, Università degli Studi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi. 76

Müller, P., Parmigiani, G., and Rice, K. (2007). FDR and Bayesian multiple comparisons rules. In Bernardo, J. M., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 8*, pages 1–19. Oxford University Press, Oxford. 17, 33

Müller, P. and Quintana, F. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801–2808. 83

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(3):735–749. 86

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278. 83

Murray, I., Ghahramani, Z., and MacKay, D. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 359–366. 4, 25, 28, 29, 31

Møller, J. (2003). Shot noise Cox processes. *Advances in Applied Probability*, 35(3):614–640. 210

Naik, C., Caron, F., Rousseau, J., Teh, Y. W., and Palla, K. (2022). Bayesian nonparametrics for sparse dynamic networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 191–206. Springer. 84

Nayak, T. K. (1988). A note on estimating the number of errors in a system by recapture sampling. *Statistics & Probability Letters*, 7(3):191–194. 76

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265. 80

Newman, M. (2018). *Networks*. Oxford university press. 47

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2015). Bayesian nonlinear model selection for gene regulatory networks. *Biometrics*, 71(3):585–595. 11

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2017). Sparse multi-dimensional graphical models: A unified Bayesian framework. *Journal of the American Statistical Association*, 112(518):779–793. 10

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics*, 32(5):2044–2073. 97

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087. 47

Orbanz, P. and Teh, Y. W. (2011). Bayesian nonparametric models. In *Encyclopedia of machine learning*, pages 81–89. Springer. 72

Orlitsky, A., Suresh, A. T., and Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47):13283–13288. 149

Osborne, N., Peterson, C. B., and Vannucci, M. (2021). Latent network estimation and variable selection for compositional data via variational EM. *Journal of Computational and Graphical Statistics*, 31:1–22. 24, 34

Owen, A. B. and Wang, J. (2016). Bi-cross-validation for factor analysis. *Statistical Science*, 31(1):119–139. 81

Paci, L. and Consonni, G. (2020). Structural learning of contemporaneous dependencies in graphical VAR models. *Computational Statistics & Data Analysis*, 144:106880. 10, 14

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. (2021). Centered partition processes: Informative priors for clustering (with discussion). *Bayesian analysis*, 16(1):301–370. 60

Paganin, S., Page, G. L., and Quintana, F. A. (2025). Informed random partition models with temporal dependence. *arXiv 2311.14502*. 60, 230

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1-2):100–115. 213

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, 44(1-2):248–252. 213

Page, G., Barney, B., and McGuire, A. (2013). Effect of position, usage rate, and per game minutes played on NBA player production curves. *Journal of Quantitative Analysis in Sports*, 9:337–345. 94

Page, G. L. and Quintana, F. A. (2016). Spatial product partition models. *Bayesian Analysis*, 11(1):265–298. 83

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. 60, 61, 84, 120, 230

Palla, K., Ghahramani, Z., and Knowles, D. (2012). A nonparametric variable clustering model. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*. 46

Pan, H.-Y., Chao, A., and Foissner, W. (2009). A nonparametric lower bound for the number of species shared by multiple communities. *Journal of Agricultural, Biological and Environmental Statistics*, 14(4):452–468. 150, 170

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95:169–186. 80

Park, J.-H. and Dunson, D. B. (2010). Bayesian generalized product partition model. *Statistica Sinica*, pages 1203–1226. 83

Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39. 53

Perrone, V., Jenkins, P. A., Spanò, D., and Teh, Y. W. (2017). Poisson random fields for dynamic feature models. *Journal of Machine Learning Research*, 18(127):1–45. 84, 219

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174. 10, 17, 33

Peterson, C. B., Stingo, F. C., and Vannucci, M. (2016). Joint Bayesian variable and graph selection for regression models with network-structured predictors. *Statistics in Medicine*, 35(7):1017–1031. 13

Petralia, F., Rao, V., and Dunson, D. (2012). Repulsive mixtures. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. 231

Pini, A. and Vantini, S. (2016). The interval testing procedure: a general framework for inference in functional data analysis. *Biometrics*, 72(3):835–845. 10

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158. 77

Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, 30:245–267. 76, 77, 152

Pitman, J. (2006). *Combinatorial stochastic processes: Ecole d'eté de probabilités de saint-flour xxxii-2002*. Springer. 48, 52, 75, 77, 78

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900. 78

Pitt, M. K., Chatfield, C., and Walker, S. G. (2002). Constructing first order stationary autoregressive models via latent processes. *Scandinavian Journal of Statistics*, 29(4):657–663. 5, 84, 201, 202, 232

Pitt, M. K. and Walker, S. G. (2005). Constructing stationary time series models using auxiliary variables with applications. *Journal of the American Statistical Association*, 100(470):554–564. 84, 201, 232

Preston, C. (1977). Spatial birth and death processes. *Bulletin of the International Statistical Institute*, 46:371–391. 16

Qiao, X., Guo, S., and James, G. M. (2019). Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222. 11

Qiao, X., Qian, C., James, G. M., and Guo, S. (2020). Doubly functional graphical models in high dimensions. *Biometrika*, 107(2):415–431. 11

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574. 83

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. 83

Quintana, F. A., Müller, P., and Papoila, A. L. (2015). Cluster-specific variable selection for product partition models. *Scandinavian Journal of Statistics*, 42:1065–1077. 83

Rabiner, L. R. and Juang, B.-H. (1986). An introduction to hidden Markov models. *IEEE Acoustics, Speech, and Signal Processing Magazine*, 3:4–16. 81

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer. 9

Rasmussen, S. L. and Starr, N. (1979). Optimal and adaptive stopping in the search for new species. *Journal of the American Statistical Association*, 74(367):661–667. 149

Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585. 73, 74, 95

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4):731–792. 116

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(1):255–268. 106

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1144. 86

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(5):689–710. 102

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411. 13, 24, 25, 28, 30, 41

Roychowdhury, A. and Kulis, B. (2015). Gamma processes, stick-breaking, and variational inference. In *Artificial Intelligence and Statistics*, pages 800–808. PMLR. 220

Scott, J. and Carvalho, C. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):790–808. 14, 23

Scutari, M. (2013). On the prior and posterior distributions used in graphical modelling. *Bayesian Analysis*, 8(3):505–532. 23

Simpson, E. (1949). Measurement of diversity. *Nature*, 688:163. 163, 164

Smith, A. F. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416. 4, 48

Smith, A. N. and Allenby, G. M. (2020). Demand models with random partitions. *Journal of the American Statistical Association*, 115(529):47–65. 60

Srebro, N. and Roweis, S. (2005). Time-varying topic models using dependent Dirichlet processes. Technical report, Department of Computer Science, University of Toronto. 84

Sun, S., Wang, H., and Xu, J. (2015). Inferring block structure of graphical models in exponential families. In Lebanon, G. and Vishwanathan, S. V. N., editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 939–947. 46

Tan, K. M., Witten, D., and Shojaie, A. (2015). The cluster graphical lasso for improved estimation of Gaussian graphical models. *Computational statistics & data analysis*, 85:23–36. 46

Tan, L. S. and Iorio, M. D. (2019). Dynamic degree-corrected blockmodels for social networks: A nonparametric approach. *Statistical Modelling*, 19(4):386–411. 48

Tan, L. S., Jasra, A., De Iorio, M., and Ebbels, T. M. (2017). Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *The Annals of Applied Statistics*, 11(4):2222–2251. 10

Teh, Y. and Gorur, D. (2009). Indian buffet processes with power-law behavior. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc. 74, 200

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581. 85, 101, 106, 107, 152, 219

Telesca, D. and Inoue, L. Y. T. (2008). Bayesian hierarchical curve registration. *Journal of the American Statistical Association*, 103(481):328–339. 10

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2, pages 564–571. 82

Thurstone, L. L. (1947). *Multiple Factor Analysis*. University of Chicago Press, Chicago. 81

Titsias, M. (2007). The infinite gamma-Poisson feature model. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc. 200

Todeschini, A., Miscouridou, X., and Caron, F. (2020). Exchangeable random measures for sparse and modular graphs with overlapping communities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2):487–520. 231

Tsamardinos, I., Brown, L. E., Aliferis, C. F., and Moore, A. W. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78. 34

Uhler, C., Lenkoski, A., and Richards, D. (2018). Exact formulas for the normalizing constants of Wishart distributions for graphical models. *The Annals of Statistics*, 46(1):90–118. 14, 25

van den Boom, W., Beskos, A., and Iorio, M. D. (2022a). The G-Wishart weighted proposal algorithm: efficient posterior computation for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 31:1215–1224. 25

van den Boom, W., De Iorio, M., and Beskos, A. (2022b). Bayesian learning of graph substructures. *Bayesian Analysis*, 1(1):1–29. 4, 46, 47, 50, 229

Van Havre, Z., White, N., Rousseau, J., and Mengersen, K. (2015). Overfitting Bayesian mixture models with an unknown number of components. *PloS one*, 10(7):e0131739. 102

Wade, S. (2023). Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A*, 381. 81

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626. 107, 117

Waghmare, K. G. and Panaretos, V. M. (2024). Continuously indexed graphical models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1):211–231. 18, 45, 58, 229

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36:45–54. 80

Walker, S. G. (2010). Bayesian nonparametric methods: Motivation and ideas. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probabilistic Mathematics, pages 22–34. Cambridge University Press. 72

Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377. 230

Wang, H. and Li, S. Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198. 25, 30, 33

Williamson, S., Orbanz, P., and Ghahramani, Z. (2010). Dependent indian buffet processes. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 924–931, Chia Laguna Resort, Sardinia, Italy. PMLR. 84

Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data. *Statistics and Computing*, 26(1-2):409–421. 9

Xiong, L., Chen, X., Huang, T.-K., Schneider, J., and Carbonell, J. G. (2010). Temporal collaborative filtering with Bayesian probabilistic tensor factorization. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 211–222. SIAM. 81

Yang, J., Cox, D. D., Lee, J. S., Ren, P., and Choi, T. (2017). Efficient Bayesian hierarchical functional data analysis with basis function approximations using Gaussian-Wishart processes. *Biometrics*, 73(4):1082–1091. 10, 13

Yang, J., Zhu, H., Choi, T., and Cox, D. D. (2016). Smoothing and mean-covariance estimation of functional data with a Bayesian hierarchical model. *Bayesian Analysis*, 11(3):649–670. 9

Yook, S.-H., Oltvai, Z. N., and Barabási, A.-L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942. 24

Yue, J. C. and Clayton, M. K. (2012). Sequential sampling in the search for new shared species. *Journal of Statistical Planning and Inference*, 142(5):1031–1039. 150, 167, 168, 171, 172

Yue, J. C., Clayton, M. K., and Hung, C.-R. (2022). Comparing nonparametric estimators for the number of shared species in two populations. *Diversity*, 14(4). 168

Zanella, G. (2020). Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865. 25

Zara, L., Tordoni, E., Castro-Delgado, S., Colla, A., Maccherini, S., Marignani, M., Panepinto, F., Trittoni, M., and Bacaro, G. (2021). Cross-taxon relationships in mediterranean urban ecosystem: A case study from the city of trieste. *Ecological Indicators*, 125:107538. 170, 172

Zhang, Y., Zhao, Y., David, L., Henao, R., and Carin, L. (2016). Dynamic Poisson factor analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1359–1364. 84, 219

Zhou, M. and Carin, L. (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:307–320. 220

Zhou, M., Chen, H., Ren, L., Sapiro, G., Carin, L., and Paisley, J. (2009). Non-parametric Bayesian dictionary learning for sparse image representations. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc. 213

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). Beta-negative binomial process and Poisson factor analysis. In Lawrence, N. D. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1462–1471, La Palma, Canary Islands. 81, 219, 220

Zhou, M., Padilla, O. H. M., and Scott, J. G. (2016). Priors for random count matrices derived from a family of negative binomial processes. *Journal of the American Statistical Association*, 111:1144–1156. 81

Zhu, H., Strawn, N., and Dunson, D. B. (2016). Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17:1–27. 11

Zito, A., Rigon, T., Ovaskainen, O., and Dunson, D. B. (2023). Bayesian modeling of sequential discoveries. *Journal of the American Statistical Association*, 118(544):2521–2532. 79

Zuur, A. F., Fryer, R. J., Jolliffe, I. T., Dekker, R., and Beukema, J. J. (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, 14(7):665–685. 84