UNIVERSITÀ COMMERCIALE “LUIGI BOCCONI”

PhD SCHOOL

PhD program in Statistics

Cycle: XXXII

Disciplinary Field: SECS-S/01

Bayesian Inference for Complex Data Structures: Theoretical and Computational Advances

Advisor: Igor Prünster

Co-Advisor: Antonio Lijoi

PhD Thesis by

Giovanni Rebaudo

ID number: 3025272

Year 2021

Abstract

In Bayesian Statistics, the modeling of data with complex dependence structures is often obtained by a composition of simple dependence assumptions. Such representations facilitate the probabilistic assessment and ease the derivation of analytical and computational results in complex models. In the present thesis, we derive novel theoretical and computational results on Bayesian inference for probabilistic clustering and flexible dependence models for complex data structures. We focus on models arising from hierarchical specifications in both parametric and nonparametric frameworks.

More precisely, we derive novel conjugacy results for one of the most applied dynamic regression models for binary time series: the dynamic probit model. Exploiting such theoretical results we derive new efficient sampling schemes improving state-of-the-art approximate or sequential Monte Carlo inference. Motivated by an issue of the well-known nested Dirichlet process, we also introduce a novel model, arising from the composition of Dirichlet processes, to cluster populations and observations across populations simultaneously. We derive a closed form expression for the induced distribution of the random partition which allows to gain a deeper understanding of the theoretical properties and inferential implications of the model and we propose a conditional Markov Chain Monte Carlo (MCMC) algorithm to effectively perform inference. Moreover, we generalize the previous composition of discrete random probabilities defining a novel wide class of species sampling priors which allows to predict future observations in different groups and test for homogeneity among sub-populations. Posterior inference is feasible thanks to a marginal MCMC routine and urn schemes that allow to evaluate posterior and predictive functionals of interest. Finally, we prove a surprising consistency result for the number of clusters in the most famous nonparametric model for clustering, that is the Dirichlet process mixture model. In this way we partially answer an open question in the methodological literature.

Acknowledgments

At the end of this enriching journey, I feel the need to thank the people that made this experience possible and fun.

First of all, I want to thank my two great mentors, Antonio and Igor, without whom this thesis would not have been possible. I am extremely grateful to them for patiently guiding me through my research with their immense knowledge and inspiring me with their passion. I always count on them, and I owe them a lot.

During my Ph.D. work I was lucky to find a stimulating environment with amazing people and top-level researchers. In particular, I want to thank Daniele, Giacomo, and Sonia, with whom I had the pleasure to work. Their enthusiasm for research and brilliant ideas taught me a lot and showed me different perspectives on how to solve research challenges.

I am also extremely grateful to all my friends and colleagues. I wish you the best. In particular, I want to thank Augusto and Filippo, with whom I shared this journey. We had a lot of fun doing research together and enjoying Ph.D. life in general. I hope we will continue to run in Milan, hike in Corio, and just have fun together around the world.

Finally, I want to thank my family for always supporting me, encouraging my studies, and giving me the privilege to make my passion a career.

Contents

Chapter 1

Introduction

Probability estimation is naturally approached and justified in the Bayesian nonparametric framework. Indeed, when we judge an extendable sequence of observable variables exchangeable, a random probability measure arises from de Finetti's representation theorem and the observations can be seen as independent identically distributed given such a random probability. If a subject wants to make inference and prediction using the Bayes-Laplace paradigm they can interpret the law of the random probability measure as a prior. When the support of the prior does not degenerate on a finite-dimensional parameter space, we are in the Bayesian nonparametric framework.

In real world applications, the homogeneity assumption of exchangeability is often too restrictive when we want to model complex data structures. To quote de Finetti, Bruno (1938): "*But the case of exchangeability can only be considered as a limiting case: the case in which this 'analogy' is, in a certain sense, absolute for all events under consideration. [..] To get from the case of exchangeability to other cases which are more general but still tractable, we must take up the case where we still encounter 'analogies' among the events under consideration, but without attaining the limiting case of exchangeability.*" Indeed, exchangeability entails that the order of the observations does not count in the inferential procedures. According to the specific application, the type of data and the availability of covariates different dependence assumptions can be assessed more reasonable by a subject. For instance, when modeling time series (Chapter 2) it is reasonable to exploit the time information to perform inference and prediction. Likewise, when data are collected in different studies or populations (Chapters 3 and 4), it is sound to perform inference effectively borrowing information across them without degenerating to the exchangeable case. Though in such aforementioned cases we clearly need to go beyond the assumption of exchangeability, exchangeability still remains the fundamental building block of a major part of more flexible Bayesian models. More generally, the idea of combining simple conditional independent structures to characterize complex dependence relationships in the data is ubiquitous in Bayesian Statistics. For instance, in the time series setting introducing an hidden state process with a simple Markovian dependence allows to set far more general dependence assumptions on the observable process. In the same spirit, thanks to de Finetti's representation Theorem for partially exchangeable case, we can flexibly model partial exchangeable arrays by assigning a distribution on the vectors of the underlying random probabilities, that is the de Finetti's measure, which takes the role of the prior. Note that such hierarchical compositions can, at least in principle, be extended to an arbitrary level of depth. Such conditional independence assumptions

have also several practical advantages. Indeed, they facilitate the elicitation of the subject's prior opinion and also ease the derivation of analytical and computational results in complex models. It is important to stress that the simpler prior elicitation on latent quantities can be also made "coherent" to the de Finetti's idea of assessing just observable quantities if we derive analytical results that allow to understand the model linking the assumptions on latent quantities to observable ones. In the present thesis we derive novel theoretical and computational results on Bayesian inference for probabilistic clustering and complex dependence models both in the parametric and nonparametric settings. As said we focus on Bayesian models arising from the different hierarchical specifications of simple dependence structures that combined together allow to characterize and flexibly model complex data structures preserving mathematical and computational tractability.

More precisely, in Chapter 2 we analyze the dynamic probit model which allows to asses complex dependence in binary time series by exploiting the conditional independence structure of hidden Markov models. We prove that the filtering, predictive and smoothing distributions in dynamic probit models with Gaussian state variables are, in fact, available and belong to a class of unified skew-normals (SUN) whose parameters can be updated recursively in time via analytical expressions. Also the functionals of these distributions depend on known functions, but their calculation requires intractable numerical integration. Leveraging the SUN properties, we address this point via new Monte Carlo methods based on independent and identically distributed samples from the smoothing distribution, which can naturally be adapted to the filtering and predictive case, thereby improving state-of-the-art approximate or sequential Monte Carlo inference in small-to-moderate dimensional studies. A scalable and optimal particle filter which exploits the SUN properties is also developed to deal with online inference in high dimensions.

In Chapters 3 and 4, the core of the present thesis, we focus on the case where the data arises from different, though related, populations or studies and can be naturally modeled in the partially exchangeable framework to borrow information across them. Roughly speaking, partially exchangeable extendable arrays can be thought of, thanks to de Finetti representation theorem, as decomposable into different conditionally independent exchangeable subpopulations. More precisely, in Chapter 3 we propose a novel Bayesian nonparametric prior arising from the composition of Dirichlet processes that allows to perform inference in the partially exchangeable framework when we are interested in clustering populations and observations simultaneously and/or perform density estimation borrowing information across populations. A well-known Bayesian nonparametric prior to perform such tasks is the nested Dirichlet process which is known to group distributions in a single cluster when there are ties across populations. We propose a novel hybrid nonparametric prior which solves the problem by hierarchically combining two different Dirichlet processes structures. We derive a closed form expression for the induced distribution of the random partition which allows to gain a deeper understanding of the theoretical properties and inferential implications of the model and, further, yields a MCMC algorithm for evaluating Bayesian inference of interest. However, such an algorithm becomes infeasible when the number of populations is larger than two. Therefore, we also propose a different MCMC algorithm to perform inference for a larger number of populations and to test homogeneity between different populations as a by-product.

In Chapter 4 we generalize the previous composition of Dirichlet processes to a wider class of composition of Gibbs type priors in order to face species sampling problems in heterogeneous populations while simultaneously identifying sub-groups of populations and borrowing information across them. Indeed, our goal is two-fold: predict future discrete observations in different groups and test for homogeneity among sub-populations. The

former is usually the main focus in species-sampling problems, while the latter task is not feasible with the state-of-the art methods available in the literature, since they generally consider all the populations distributions different almost surely. In order to do so, we extend what is arguably the most popular species sampling model in Bayesian nonparametrics in this partially exchangeable framework, that is the hierarchical Pitman-Yor process. Adding a latent structure on the distributions, we allow to have ties across the sub-populations distributions, performing both the above-mentioned tasks at the same time. We show that the distribution of the induced random partition admits a closed form expression and we derive the asymptotic behavior of the number of species and homogeneous subpopulations, allowing to gain a deeper understanding of the theoretical properties and inferential implications of the model. Moreover, we derive computational results to perform inference via a marginal Gibbs sampler and predictive urn schemes.

In Chapter 5 we study the asymptotic behavior of the number of clusters under Dirichlet process mixture models, arguably the most famous Bayesian nonparametric model to perform clustering and density estimation. It has been recently shown that, when data are generated from a finite mixture, this posterior is inconsistent as it does not concentrate around the "true" value of the number of components. We show that, in contrast to existing conjectures in the literature, placing a prior on the concentration parameter of the Dirichlet process drastically changes the asymptotics of the number of clusters, possibly allowing to overcome the inconsistency issue.

Chapter 2

A closed–form filter for binary time series[1]

2.1 Introduction

Despite the availability of several alternative approaches for dynamic inference and prediction of binary time series (MacDonald and Zucchini, 1997), state-space models are a source of continuous interest due to their flexibility in accommodating a variety of representations and dependence structures via an interpretable formulation (West and Harrison, 2006; Petris et al., 2009; Durbin and Koopman, 2012). Let $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})^\intercal \in \{0;1\}^m$ be a vector of binary event data at time t, and define with $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{pt})^\intercal \in \Re^p$ the corresponding vector of state variables. Adapting the notation in Petris et al. (2009) to our setting, we aim to provide closed-form expressions for the filtering, predictive and smoothing distributions in the multivariate dynamic probit model

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t), \tag{2.1}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t), \; t = 1 \ldots, n, \tag{2.2}$$

with $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$, and dependence structure as defined by the directed acyclic graph displayed in Figure 2.1. In (2.1), $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ is the cumulative distribution function of the $\mathrm{N}_m(\mathbf{0}, \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ evaluated at $\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t$, with $\mathbf{B}_t = \mathrm{diag}(2y_{1t} - 1, \ldots, 2y_{mt} - 1)$ denoting the $m \times m$ sign matrix associated with $\mathbf{y}_t$, which defines the multivariate probit likelihood in equation (2.1). Model (2.1)–(2.2) is a natural generalization of univariate probit models to multivariate settings, as we will clarify in equations (2.3)–(2.5). The quantities $\mathbf{F}_t, \mathbf{V}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{a}_0$ and $\mathbf{P}_0$ define, instead, known matrices controlling the location, scale and dependence structure in the state-space model (2.1)–(2.2). Estimation and inference for these matrices is, itself, a relevant problem which can be addressed both from a frequentist and a Bayesian perspective. Yet our focus is on providing exact results for inference on state variables and prediction of future binary events under (2.1)–(2.2). Hence, consistent with the classical Kalman filter (Kalman, 1960), we rely on known system matrices $\mathbf{F}_t, \mathbf{V}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{a}_0$ and $\mathbf{P}_0$. Nonetheless, results on marginal likelihoods, which can be used in parameter estimation, are provided

[1] Joint work with Augusto Fasano, Daniele Durante and Sonia Petrone. Department of Decision Sciences, Bocconi University.

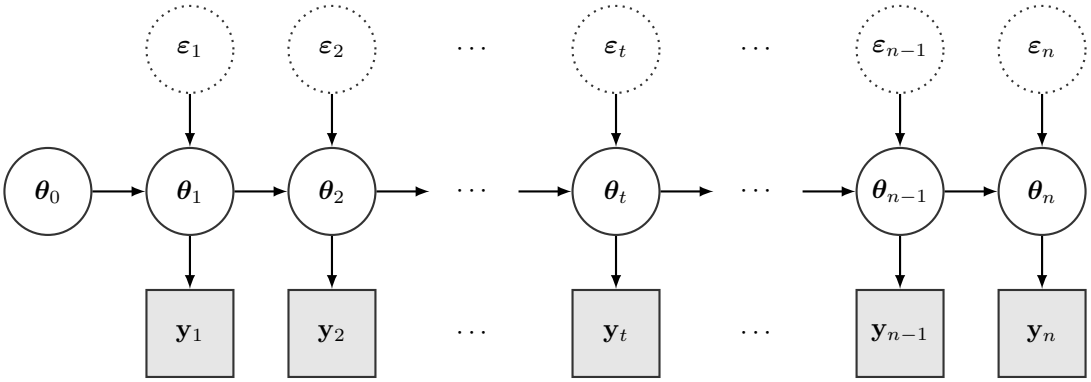

Figure 2.1: Representation of model (2.1)–(2.2). Dashed circles, solid circles and grey squares denote Gaussian errors, Gaussian states and observed binary data, respectively.

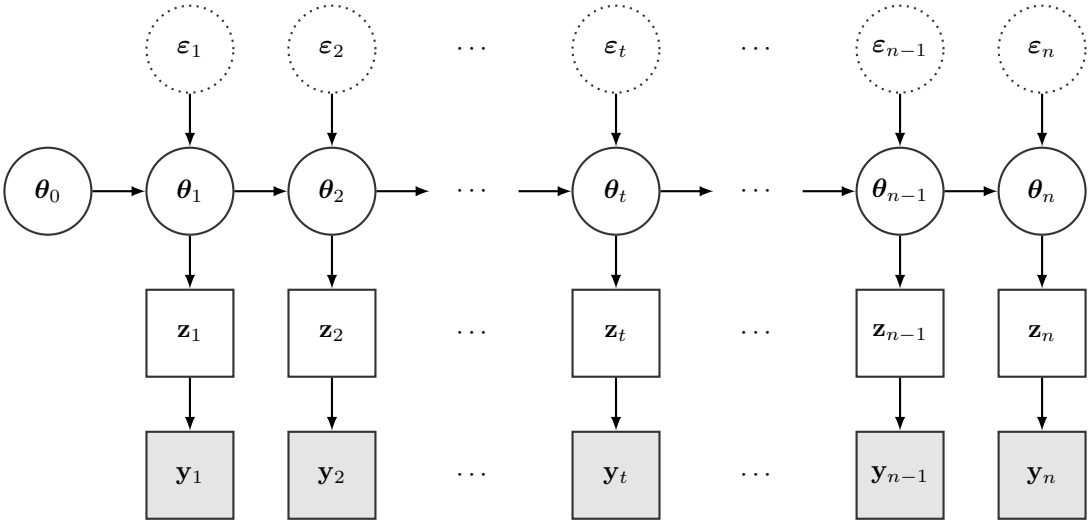

Figure 2.2: Representation of model (2.3)–(2.5). Dashed circles, solid circles, white squares and grey squares denote Gaussian errors, Gaussian states, latent Gaussian data and observed binary data, respectively.

in Section 2.3.2.

Model (2.1)–(2.2) provides a general representation encompassing a variety of formulations. For example, setting $\mathbf{V}_t = \mathbf{I}_m$ in (2.1) yields a standard probit regression, for $t = 1, \ldots, n$, which includes the classical univariate dynamic probit model when $m = 1$. These representations have appeared in several applications, especially within the econometrics literature, due to a direct connection between (2.1)–(2.2) and dynamic discrete choice models (Keane and Wolpin, 2009). This is due to the fact that representation (2.1)–(2.2) can be alternatively obtained via the dichotomization of an underlying state-space model for the m-variate Gaussian time series $\mathbf{z}_t = (z_{1t}, \ldots, z_{mt})^\intercal \in \Re^m$, $t = 1, \ldots, n$, which is regarded, in econometric applications, as a set of time-varying utilities. Indeed, adapting the classical results from probit regression (Albert and Chib, 1993; Chib and Greenberg, 1998), model (2.1)–(2.2) is equivalent to

$$\begin{aligned} \mathbf{y}_t &= (y_{1t}, \ldots, y_{mt})^\intercal = \mathbb{1}(\mathbf{z}_t > \mathbf{0}) \\ &= [\mathbb{1}(z_{1t} > 0), \ldots, \mathbb{1}(z_{mt} > 0)]^\intercal, \quad t = 1, \ldots, n, \end{aligned} \tag{2.3}$$

with $\mathbf{z}_1, \ldots, \mathbf{z}_n$ evolving in time according to the Gaussian state-space model

$$p(\mathbf{z}_t \mid \boldsymbol{\theta}_t) = \phi_m(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{V}_t), \tag{2.4}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t), \ t = 1 \ldots, n, \tag{2.5}$$

having $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$ and dependence structure as defined by the directed acyclic graph displayed in Figure 2.2. In (2.4), $\phi_m(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{V}_t)$ denotes the density function of the Gaussian $\mathrm{N}_m(\mathbf{F}_t\boldsymbol{\theta}_t, \mathbf{V}_t)$ at point $\mathbf{z}_t \in \Re^m$. To clarify the connection between (2.1)–(2.2) and (2.3)–(2.5), note that the generic element $y_{lt} = \mathbb{1}(z_{lt} > 0)$ of $\mathbf{y}_t$ is 1 or 0 depending on whether $z_{lt} > 0$ or $z_{lt} \leq 0$. Therefore, $p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \mathrm{pr}(\mathbf{B}_t\mathbf{z}_t > 0) = \mathrm{pr}[-\mathbf{B}_t(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\theta}_t) \leq \mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t] = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$, provided that $-\mathbf{B}_t(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\theta}_t) \sim \mathrm{N}_m(\mathbf{0}, \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ under (2.4).

As is clear from model (2.4)–(2.5), if $\mathbf{z}_{1:t} = (\mathbf{z}_1, \ldots, \mathbf{z}_t)$ were observed, dynamic inference on the states $\boldsymbol{\theta}_t$, for $t = 1, \ldots, n$, would be possible via direct application of the Kalman filter (Kalman, 1960). Indeed, exploiting Gaussian-Gaussian conjugacy and the conditional independence properties displayed in Figure 2.2, filtering $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t})$ and predictive $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t-1})$ distributions are also Gaussian and have parameters which can be computed recursively via simple expressions relying on the previous updates. Moreover, also the smoothing distribution $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{z}_{1:n})$ and its marginals $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:n})$, $t \leq n$, can be obtained in closed-form leveraging the Gaussian-Gaussian conjugacy. However, in (2.3)–(2.5) only a dichotomized version $\mathbf{y}_t$ of $\mathbf{z}_t$ is available. Therefore the filtering, predictive and smoothing distributions of interest are $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, respectively. Recalling model (2.1)–(2.2) and Bayes rule, the calculation of these quantities proceeds by updating the Gaussian distribution for the states in (2.2) with the probit likelihood in (2.1), thereby providing conditional distributions which seem not available in closed-form (Albert and Chib, 1993; Chib and Greenberg, 1998).

When the focus is online inference for filtering and prediction, a common solution to the above issue is to rely on approximations of model (2.1)–(2.2) which allow the implementation of standard Kalman filter updates, thus leading to approximate dynamic inference on the state variables via extended (Uhlmann, 1992) or unscented (Julier and Uhlmann, 1997) Kalman filters, among others. However, in different studies these approximations may lead to unreliable inference (Andrieu and Doucet, 2002). Markov chain Monte Carlo (MCMC) strategies (Carlin et al., 1992; Shephard, 1994; Soyer and Sung, 2013) address this problem, but, unlike the classical Kalman filter updates, these methods are suitable for batch learning of the smoothing distribution. Moreover, as discussed by Johndrow et al. (2019), common MCMC strategies face mixing or time-inefficiency issues, especially for imbalanced binary datasets. Sequential Monte Carlo solutions (Doucet et al., 2001) partially address MCMC issues and are specifically developed for online inference via particle-based representations of the conditional states distributions, which are propagated in time for dynamic filtering and prediction (Gordon et al., 1993; Kitagawa, 1996; Liu and Chen, 1998; Pitt and Shephard, 1999; Doucet et al., 2000; Andrieu and Doucet, 2002). These methods provide state-of-the-art solutions in non-Gaussian state-space models, and can be also adapted to provide batch learning of the smoothing distribution; see Doucet and Johansen (2009) for a discussion on particle degeneracy issues that may arise in this setting. Nonetheless, sequential Monte Carlo is clearly still sub-optimal compared to the case in which $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are available in closed-form and belong to a tractable class of distributions whose parameters can be sequentially updated via analytical expressions.

In Section 2.3, we prove that for the dynamic probit model defined in (2.1)–(2.2) the quantities $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are unified skew-normal (SUN) distributions (Arellano-Valle and Azzalini, 2006) having tractable expressions for the recursive computation of the corresponding parameters. To the best of our knowledge, this result provides the first closed-form filter and smoother for binary time series, and allows improvements both in online and in batch inference within this framework. As highlighted in Section 2.2, the multivariate SUN distribution has several closure properties (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014) in addition to explicit formulas—involving the cumulative distribution function of multivariate normals—for the moments (Azzalini and Bacchieri, 2010; Gupta et al., 2013) and the normalizing constant (Arellano-Valle and Azzalini, 2006). In Sections 2.3, we exploit these properties to derive closed-form expressions for key functionals of $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, including, in particular, the observations' predictive distribution $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ and the marginal likelihood $p(\mathbf{y}_{1:n})$. Besides these analytical results, we further propose in Section 2.4.1 an exact Monte Carlo scheme to compute complex functionals of the smoothing distribution. This routine relies on a stochastic representation of the SUN via a linear combination of Gaussians and truncated Gaussians (Arellano-Valle and Azzalini, 2006), and can be also applied effectively to calculate complex functionals of filtering and predictive distributions when the dimension of the time series is small-to-moderate, a common situation in several studies. As discussed in Section 2.4.2, the aforementioned strategies face computational bottlenecks in higher dimensional settings (Botev, 2017), due to challenges in computing cumulative distribution functions of multivariate Gaussians and in sampling from multivariate truncated normals. In these contexts, we propose a novel particle filter which exploits the SUN properties to obtain an optimal (Doucet et al., 2000) sequential Monte Carlo which effectively scales with t; see Section 2.4.2. As outlined in an illustrative study in Section 2.5, the methods developed in this work improve current strategies for batch and online inference in dynamic probit models. Future directions of research are discussed in Section 2.6.

2.2 The unified skew-normal distribution

Before deriving filtering, predictive and smoothing distributions in model (2.1)–(2.2), let us first briefly review the SUN family. Arellano-Valle and Azzalini (2006) proposed this class to unify different generalizations (Arnold and Beaver, 2000; Arnold et al., 2002; Gupta et al., 2004; González-Farías et al., 2004) of the original multivariate skew-normal (Azzalini and Dalla Valle, 1996), whose density is obtained as the product of a multivariate Gaussian density and the cumulative distribution function of a standard normal evaluated at a value which depends on a skewness inducing vector of parameters. Motivated by the success of this formulation and of its generalizations (Azzalini and Capitanio, 1999), Arellano-Valle and Azzalini (2006) developed a unifying representation, namely the unified skew-normal distribution. A random vector $\boldsymbol{\theta} \in \Re^p$ has a unified skew-normal distribution, $\boldsymbol{\theta} \sim \text{SUN}_{p,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, if its density function $p(\boldsymbol{\theta})$ can be expressed as

$$\phi_p(\boldsymbol{\theta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_h[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\theta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta}]}{\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}, \tag{2.6}$$

where the covariance matrix of the Gaussian density $\phi_p(\boldsymbol{\theta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ is obtained as $\boldsymbol{\Omega} = \boldsymbol{\omega} \bar{\boldsymbol{\Omega}} \boldsymbol{\omega}$, that is by re-scaling a correlation matrix $\bar{\boldsymbol{\Omega}}$ via a positive diagonal scale matrix $\boldsymbol{\omega} = (\boldsymbol{\Omega} \circ \mathbf{I}_p)^{1/2}$, with $\circ$ denoting the element-wise

Hadamard product. Observe that the quantities p and h are not parameters, but define the dimensions of the multivariate Gaussian density and cumulative distribution function appearing in (2.6), respectively. Moreover, the dimensionality of the former coincides with that of the vector $\boldsymbol{\theta}$. In (2.6), the skewness inducing mechanism is driven by the cumulative distribution function of the $\mathrm{N}_h(\mathbf{0}, \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})$ computed at $\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1} (\boldsymbol{\theta} - \boldsymbol{\xi})$, whereas $\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})$ denotes the normalizing constant obtained by evaluating the cumulative distribution function of a $\mathrm{N}_h(\mathbf{0}, \boldsymbol{\Gamma})$ at $\boldsymbol{\gamma}$. Arellano-Valle and Azzalini (2006) added a further identifiability condition which restricts the matrix $\boldsymbol{\Omega}^*$, with blocks $\boldsymbol{\Omega}^*_{[11]} = \boldsymbol{\Gamma}$, $\boldsymbol{\Omega}^*_{[22]} = \bar{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}^*_{[21]} = \boldsymbol{\Omega}^{*\intercal}_{[12]} = \boldsymbol{\Delta}$, to be a full–rank correlation matrix.

To clarify the role of the parameters in (2.6), let us discuss a generative representation of the SUN. In particular, if $\mathbf{z}_0 \in \Re^h$ and $\boldsymbol{\theta}_0 \in \Re^p$ characterize two random vectors jointly distributed as a $\mathrm{N}_{h+p}(\mathbf{0}, \boldsymbol{\Omega}^*)$, then $\boldsymbol{\xi} + \boldsymbol{\omega}(\boldsymbol{\theta}_0 \mid \mathbf{z}_0 + \boldsymbol{\gamma} > \mathbf{0}) \sim \mathrm{SUN}_{p,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ (Arellano-Valle and Azzalini, 2006). Hence, $\boldsymbol{\xi}$ and $\boldsymbol{\omega}$ control location and scale, respectively, whereas $\boldsymbol{\Gamma}$, $\bar{\boldsymbol{\Omega}}$ and $\boldsymbol{\Delta}$ define the dependence within $\mathbf{z}_0 \in \Re^h$, within $\boldsymbol{\theta}_0 \in \Re^p$ and between these two random vectors, respectively. Finally, $\boldsymbol{\gamma}$ controls the truncation in the partially observed Gaussian vector $\mathbf{z}_0 \in \Re^h$. The above representation provides also key insights on our closed-form filter for the dynamic probit model (2.1)–(2.2). Indeed, according to (2.3)–(2.5), the filtering, predictive and smoothing distributions induced by model (2.1)–(2.2) can be also defined as $p[\boldsymbol{\theta}_t \mid \mathbb{1}(\mathbf{z}_1 > \mathbf{0}), \ldots, \mathbb{1}(\mathbf{z}_t > \mathbf{0})]$, $p[\boldsymbol{\theta}_t \mid \mathbb{1}(\mathbf{z}_1 > \mathbf{0}), \ldots, \mathbb{1}(\mathbf{z}_{t-1} > \mathbf{0})]$ and $p[\boldsymbol{\theta}_{1:n} \mid \mathbb{1}(\mathbf{z}_1 > \mathbf{0}), \ldots, \mathbb{1}(\mathbf{z}_n > \mathbf{0})]$, respectively, with $(\mathbf{z}_t, \boldsymbol{\theta}_t)$ from the Gaussian state-space model (2.4)–(2.5) for $t = 1, \ldots, n$, thus highlighting the direct connection between these distributions and the generative representation of the SUN.

Another fundamental stochastic representation of the SUN distribution relies on linear combinations of Gaussian and truncated Gaussian random variables, thereby facilitating sampling from the SUN. In particular, recalling Azzalini and Capitanio (2014, Section 7.1.2) and Arellano-Valle and Azzalini (2006), if $\boldsymbol{\theta} \sim \mathrm{SUN}_{p,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, then

$$\boldsymbol{\theta} \overset{\mathrm{d}}{=} \boldsymbol{\xi} + \boldsymbol{\omega}(\mathbf{U}_0 + \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\mathbf{U}_1), \quad \mathbf{U}_0 \perp \mathbf{U}_1, \tag{2.7}$$

with $\mathbf{U}_0 \sim \mathrm{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}} - \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}^\intercal)$ and $\mathbf{U}_1$ from a $\mathrm{N}_h(\mathbf{0}, \boldsymbol{\Gamma})$ truncated below $-\boldsymbol{\gamma}$. As we will clarify in Section 2.4, this result can facilitate efficient Monte Carlo inference on complex functionals of filtering, predictive and smoothing distributions in model (2.1)–(2.2), based on sampling from the corresponding SUN variable. Indeed, although key moments can be explicitly derived via direct differentiation of the SUN moment generating function (Gupta et al., 2013; Arellano-Valle and Azzalini, 2006), such a strategy requires tedious calculations in the unified skew-normal framework, when the focus is on complex functionals. Moreover, recalling Azzalini and Bacchieri (2010) and Gupta et al. (2013), the first and second order moments further require the evaluation of h-variate Gaussian cumulative distribution functions $\Phi_h(\cdot)$, thus affecting computational tractability in large h settings (Botev, 2017). In these situations, Monte Carlo integration provides an effective solution, especially when independent samples can be generated efficiently. Therefore, we mostly focus on improved Monte Carlo inference in model (2.1)–(2.2) exploiting the SUN properties, and refer to Azzalini and Bacchieri (2010) and Gupta et al. (2013) for a closed-form expression of the expectation, variance and cumulative distribution function of SUN variables. As clarified in Section 2.3, h increases linearly with time t in the SUN filtering and predictive distributions.

Before concluding our overview, we shall emphasize that unified skew-normal random variables are also closed

under marginalization, linear combinations and conditioning (Azzalini and Capitanio, 2014). These properties facilitate the derivation of the SUN filtering, predictive and smoothing distributions in model (2.1)–(2.2).

2.3 Filtering, prediction and smoothing

In this section, we prove that all the distributions of interest admit a closed-form SUN representation. In particular, in Section 2.3.1 we prove that closed-form filters—meant here as exact updating schemes for predictive and filtering distributions based on simple recursive expressions for the associated parameters—can be derived for model (2.1)–(2.2), whereas in Section 2.3.2 we present the form of the SUN smoothing distribution and some important consequences. The associated computational methods are then discussed in Section 2.4.

2.3.1 Filtering and predictive distributions

To obtain the exact form of the filtering and predictive distributions under (2.1)–(2.2), let us start from $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$. This first quantity characterizes the initial step of the filter recursion, and its derivation in Lemma 1 provides key intuitions to obtain the state predictive $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and filtering $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, for every $t \geq 2$. Lemma 1 states that $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is a SUN distribution. In the following, consistent with the notation of Section 2.2, whenever $\boldsymbol{\Omega}$ is a $p \times p$ covariance matrix, the associated matrices $\boldsymbol{\omega}$ and $\bar{\boldsymbol{\Omega}}$ are defined as $\boldsymbol{\omega} = (\boldsymbol{\Omega} \circ \mathbf{I}_p)^{1/2}$ and $\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$, respectively. All proofs can be found in the Appendix and consider conjugacy properties of the SUN in probit models. Early findings on this result have been explored by Durante (2019) in the context of static univariate Bayesian probit regression. Here, we take a substantially different perspective by focusing on online inference in both multivariate and time–varying probit models that require novel and non–straightforward extensions. As seen in Soyer and Sung (2013) and Chib and Greenberg (1998), the increased complexity of this endeavor typically motivates a separate treatment relative to the static univariate case.

Lemma 1. *Under the dynamic probit model in (2.1)–(2.2), the first-step filtering distribution is*

$$(\boldsymbol{\theta}_1 \mid \mathbf{y}_1) \sim \text{SUN}_{p,m}(\boldsymbol{\xi}_{1|1}, \boldsymbol{\Omega}_{1|1}, \boldsymbol{\Delta}_{1|1}, \boldsymbol{\gamma}_{1|1}, \boldsymbol{\Gamma}_{1|1}), \tag{2.8}$$

with parameters defined by the recursive equations

$$\begin{aligned}
\boldsymbol{\xi}_{1|1} &= \mathbf{G}_1\mathbf{a}_0, \quad \boldsymbol{\Omega}_{1|1} = \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1, \\
\boldsymbol{\Delta}_{1|1} &= \bar{\boldsymbol{\Omega}}_{1|1}\boldsymbol{\omega}_{1|1}\mathbf{F}_1^\intercal\mathbf{B}_1\mathbf{s}_1^{-1}, \quad \boldsymbol{\gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1}, \\
\boldsymbol{\Gamma}_{1|1} &= \mathbf{s}_1^{-1}\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1\mathbf{s}_1^{-1},
\end{aligned}$$

where $\mathbf{s}_1 = [(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1) \circ \mathbf{I}_m]^{1/2}$.

Hence $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is a SUN distribution and its parameters can be obtained via tractable arithmetic expressions applied to the quantities characterizing model (2.1)–(2.2). Exploiting the results in Lemma 1, the general filter updates for the multivariate probit model can be obtained by induction for $t \geq 2$ and are presented in Theorem 1.

Theorem 1. *Let* $(\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1}) \sim \text{SUN}_{p,m\cdot(t-1)}(\boldsymbol{\xi}_{t-1|t-1},$

$\boldsymbol{\Omega}_{t-1|t-1}, \boldsymbol{\Delta}_{t-1|t-1}, \boldsymbol{\gamma}_{t-1|t-1}, \boldsymbol{\Gamma}_{t-1|t-1})$ *be the filtering distribution at* $t-1$ *under* (2.1)–(2.2)*. Then the one-step-ahead state predictive distribution at* t *is*

$$(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1}) \sim \mathrm{SUN}_{p,m\cdot(t-1)}(\boldsymbol{\xi}_{t|t-1}, \boldsymbol{\Omega}_{t|t-1}, \boldsymbol{\Delta}_{t|t-1}, \boldsymbol{\gamma}_{t|t-1}, \boldsymbol{\Gamma}_{t|t-1}), \tag{2.9}$$

with parameters defined by the recursive equations

$$\begin{aligned}
\boldsymbol{\xi}_{t|t-1} &= \mathbf{G}_t \boldsymbol{\xi}_{t-1|t-1}, \quad \boldsymbol{\Omega}_{t|t-1} = \mathbf{G}_t \boldsymbol{\Omega}_{t-1|t-1} \mathbf{G}_t^\intercal + \mathbf{W}_t, \\
\boldsymbol{\Delta}_{t|t-1} &= \boldsymbol{\omega}_{t|t-1}^{-1} \mathbf{G}_t \boldsymbol{\omega}_{t-1|t-1} \boldsymbol{\Delta}_{t-1|t-1}, \\
\boldsymbol{\gamma}_{t|t-1} &= \boldsymbol{\gamma}_{t-1|t-1}, \quad \boldsymbol{\Gamma}_{t|t-1} = \boldsymbol{\Gamma}_{t-1|t-1}.
\end{aligned}$$

Moreover, the filtering distribution at time t *is*

$$(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}) \sim \mathrm{SUN}_{p,m\cdot t}(\boldsymbol{\xi}_{t|t}, \boldsymbol{\Omega}_{t|t}, \boldsymbol{\Delta}_{t|t}, \boldsymbol{\gamma}_{t|t}, \boldsymbol{\Gamma}_{t|t}), \tag{2.10}$$

with parameters defined by the recursive equations

$$\begin{aligned}
\boldsymbol{\xi}_{t|t} &= \boldsymbol{\xi}_{t|t-1}, \quad \boldsymbol{\Omega}_{t|t} = \boldsymbol{\Omega}_{t|t-1}, \\
\boldsymbol{\Delta}_{t|t} &= [\boldsymbol{\Delta}_{t|t-1}, \bar{\boldsymbol{\Omega}}_{t|t} \boldsymbol{\omega}_{t|t} \mathbf{F}_t^\intercal \mathbf{B}_t \mathbf{s}_t^{-1}], \\
\boldsymbol{\gamma}_{t|t} &= [\boldsymbol{\gamma}_{t|t-1}^\intercal, \boldsymbol{\xi}_{t|t}^\intercal \mathbf{F}_t^\intercal \mathbf{B}_t \mathbf{s}_t^{-1}]^\intercal,
\end{aligned}$$

and $\boldsymbol{\Gamma}_{t|t}$ *is a full-rank correlation matrix having blocks* $\boldsymbol{\Gamma}_{t|t[11]} = \boldsymbol{\Gamma}_{t|t-1}$, $\boldsymbol{\Gamma}_{t|t[22]} = \mathbf{s}_t^{-1}\mathbf{B}_t(\mathbf{F}_t\boldsymbol{\Omega}_{t|t}\mathbf{F}_t^\intercal + \mathbf{V}_t)\mathbf{B}_t\mathbf{s}_t^{-1}$ *and* $\boldsymbol{\Gamma}_{t|t[21]} = \boldsymbol{\Gamma}_{t|t[12]}^\intercal = \mathbf{s}_t^{-1}\mathbf{B}_t\mathbf{F}_t\boldsymbol{\omega}_{t|t-1}\boldsymbol{\Delta}_{t|t-1}$*, where* $\mathbf{s}_t$ *is defined as* $\mathbf{s}_t = [(\mathbf{F}_t\boldsymbol{\Omega}_{t|t}\mathbf{F}_t^\intercal + \mathbf{V}_t) \circ \mathbf{I}_m]^{1/2}$.

Consistent with Theorem 1, online prediction and filtering in the multivariate dynamic probit model (2.1)–(2.2) proceeds by iterating between equations (2.9) and (2.10) as new observations stream in with time t. Both steps are based on closed-form distributions and rely on analytical expressions for recursive updating of the corresponding parameters as a function of the previous ones, thus providing an analog of the classical Kalman filter.

We also provide closed-form results for the predictive distribution of the multivariate binary data. Indeed, the prediction of future events $\mathbf{y}_t \in \{0;1\}^m$ given the current data $\mathbf{y}_{1:t-1}$, is a primary goal in applications of dynamic probit models. In our setting, this task requires the derivation of the predictive distribution $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ which coincides, under (2.1)–(2.2), with $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\theta}_t$, where $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ is the state predictive distribution in (2.9). Corollary 1 shows that this quantity has an explicit form.

Corollary 1. *Under model* (2.1)–(2.2)*, the observation predictive distribution* $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ *is*

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) = \frac{\Phi_{m\cdot t}(\boldsymbol{\gamma}_{t|t}; \boldsymbol{\Gamma}_{t|t})}{\Phi_{m\cdot(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})}, \tag{2.11}$$

for every time t*, with parameters* $\boldsymbol{\gamma}_{t|t}$, $\boldsymbol{\Gamma}_{t|t}$, $\boldsymbol{\gamma}_{t|t-1}$ *and* $\boldsymbol{\Gamma}_{t|t-1}$*, defined as in Theorem 1.*

Hence, the evaluation of probabilities of future events is possible via explicit calculations after marginalizing out analytically the predictive distribution of the states. As is clear from (2.11), this approach requires the

calculation of Gaussian cumulative distribution functions whose dimension increases with t and m. Efficient evaluation of these integrals is possible for small-to-moderate t and m via recent minimax tilting (Botev, 2017), but such methods are impractical for large t and m. In Section 2.4, we develop new Monte Carlo methods based on independent samples and sequential Monte Carlo strategies to overcome this issue and allow scalable inference exploiting Theorem 1 to improve current solutions.

2.3.2 Smoothing distribution

We now consider smoothing distributions. In this case, the whole data $\mathbf{y}_{1:n}$ are available and the interest is on the distribution of either the whole states' sequence $\boldsymbol{\theta}_{1:n}$ or a subset of it, given $\mathbf{y}_{1:n}$. Theorem 2 shows that also the smoothing distribution $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ belongs to the SUN family, and direct consequences of such a result, involving marginal smoothing and marginal likelihoods are reported in Corollaries 2 and 3, respectively.

Before stating the result, let us first introduce two block-diagonal matrices, $\mathbf{D}$ and $\mathbf{V}$, having dimensions $(m \cdot n) \times (p \cdot n)$ and $(m \cdot n) \times (m \cdot n)$ respectively, with diagonal blocks $\mathbf{D}_{[t,t]} = \mathbf{B}_t\mathbf{F}_t \in \Re^{m \times p}$ and $\mathbf{V}_{[t,t]} = \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t \in \Re^{m \times m}$, for every $t = 1, \ldots, n$. Moreover, let $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ denote the mean and covariance matrix of the multivariate Gaussian for $\boldsymbol{\theta}_{1:n}$ induced by the state equations. Under (2.2), $\boldsymbol{\xi}$ is a $p \cdot n$ column vector obtained by stacking the p-dimensional blocks $\boldsymbol{\xi}_{[t]} = \mathrm{E}(\boldsymbol{\theta}_t) = \mathbf{G}_1^t\mathbf{a}_0 \in \Re^p$ for every $t = 1, \ldots, n$, with $\mathbf{G}_1^t = \mathbf{G}_t \cdots \mathbf{G}_1$. Similarly, letting $\mathbf{G}_q^t = \mathbf{G}_t \cdots \mathbf{G}_q$, also the $(p \cdot n) \times (p \cdot n)$ covariance matrix $\boldsymbol{\Omega}$ has a block structure with $(p \times p)$-dimensional blocks $\boldsymbol{\Omega}_{[t,t]} = \mathrm{var}(\boldsymbol{\theta}_t) = \mathbf{G}_1^t\mathbf{P}_0\mathbf{G}_1^{t\intercal} + \sum_{q=2}^{t} \mathbf{G}_q^t\mathbf{W}_{q-1}\mathbf{G}_q^{t\intercal} + \mathbf{W}_t$, for $t = 1, \ldots, n$, and $\boldsymbol{\Omega}_{[t,q]} = \boldsymbol{\Omega}_{[q,t]}^{\intercal} = \mathrm{cov}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_q) = \mathbf{G}_{q+1}^t\boldsymbol{\Omega}_{[q,q]}$, for $t > q$.

Theorem 2. *Under model (2.1)–(2.2), the joint smoothing distribution is*

$$(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \sim \mathrm{SUN}_{p\cdot n, m\cdot n}(\boldsymbol{\xi}_{1:n|n}, \boldsymbol{\Omega}_{1:n|n}, \boldsymbol{\Delta}_{1:n|n}, \boldsymbol{\gamma}_{1:n|n}, \boldsymbol{\Gamma}_{1:n|n}), \tag{2.12}$$

with parameters defined as

$$\begin{aligned} \boldsymbol{\xi}_{1:n|n} &= \boldsymbol{\xi}, \quad \boldsymbol{\Omega}_{1:n|n} = \boldsymbol{\Omega}, \quad \boldsymbol{\Delta}_{1:n|n} = \bar{\boldsymbol{\Omega}}\boldsymbol{\omega}\mathbf{D}^{\intercal}\mathbf{s}^{-1} \\ \boldsymbol{\gamma}_{1:n|n} &= \mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}, \quad \boldsymbol{\Gamma}_{1:n|n} = \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^{\intercal} + \mathbf{V})\mathbf{s}^{-1}, \end{aligned}$$

where $\mathbf{s} = [(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^{\intercal} + \mathbf{V}) \circ \mathbf{I}_{m\cdot n}]^{1/2}$.

Since the SUN is closed under marginalization and linear combinations, it follows from Theorem 2 that the smoothing distribution for any combination of states is still a SUN. In particular, direct application of the results in Azzalini and Capitanio (2014, Section 7.1.2) provides the marginal smoothing distribution for any state $\boldsymbol{\theta}_t$ reported in Corollary 2.

Corollary 2. *Under model (2.1)–(2.2), the marginal smoothing distribution at time t is*

$$(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:n}) \sim \mathrm{SUN}_{p, m\cdot n}(\boldsymbol{\xi}_{t|n}, \boldsymbol{\Omega}_{t|n}, \boldsymbol{\Delta}_{t|n}, \boldsymbol{\gamma}_{t|n}, \boldsymbol{\Gamma}_{t|n}) \tag{2.13}$$

Algorithm 1: Independent and identically distributed sampling from $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$

Sample $\mathbf{U}_{0\ 1:n|n}^{(1)}, \ldots, \mathbf{U}_{0\ 1:n|n}^{(R)}$ independently from a $\mathrm{N}_{p \cdot n}(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{1:n|n} - \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}_{1:n|n}^{-1}\boldsymbol{\Delta}_{1:n|n}^{\intercal})$.

Sample $\mathbf{U}_{1\ 1:n|n}^{(1)}, \ldots, \mathbf{U}_{1\ 1:n|n}^{(R)}$ independently from a $\mathrm{N}_{m \cdot n}(\mathbf{0}, \boldsymbol{\Gamma}_{1:n|n})$ truncated below $-\boldsymbol{\gamma}_{1:n|n}$.

Compute $\boldsymbol{\theta}_{1:n|n}^{(1)}, \ldots, \boldsymbol{\theta}_{1:n|n}^{(R)}$ via $\boldsymbol{\theta}_{1:n|n}^{(r)} = \boldsymbol{\xi}_{1:n|n} + \boldsymbol{\omega}_{1:n|n}(\mathbf{U}_{0\ 1:n|n}^{(r)} + \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}_{1:n|n}^{-1}\mathbf{U}_{1\ 1:n|n}^{(r)})$ for each r.

with parameters defined as

$$\boldsymbol{\xi}_{t|n} = \boldsymbol{\xi}_{[t]}, \quad \boldsymbol{\Omega}_{t|n} = \boldsymbol{\Omega}_{[t,t]}, \quad \boldsymbol{\Delta}_{t|n} = \boldsymbol{\Delta}_{1:n|n[t]}$$
$$\boldsymbol{\gamma}_{t|n} = \boldsymbol{\gamma}_{1:n|n}, \quad \boldsymbol{\Gamma}_{t|n} = \boldsymbol{\Gamma}_{1:n|n},$$

where $\boldsymbol{\Delta}_{1:n|n[t]}$ denotes the t-th block of p rows in $\boldsymbol{\Delta}_{1:n|n}$. When $t = n$, (2.13) gives the filtering distribution at n.

Another important consequence of Theorem 2 is the availability of a closed-form expression for the marginal likelihood $p(\mathbf{y}_{1:n})$, which is provided in Corollary 3.

Corollary 3. *Under model (2.1)–(2.2), the marginal likelihood is $p(\mathbf{y}_{1:n}) = \Phi_{m \cdot n}(\boldsymbol{\gamma}_{1:n|n}; \boldsymbol{\Gamma}_{1:n|n})$, with $\boldsymbol{\gamma}_{1:n|n}$ and $\boldsymbol{\Gamma}_{1:n|n}$ as in Theorem 2.*

This result is useful in several contexts, including empirical Bayes and estimation of unknown system parameters via maximization of the marginal likelihood.

2.4 Inference via Monte Carlo methods

As discussed in Sections 2.2 and 2.3, inference without sampling from (2.9)–(2.10) or (2.12) is, theoretically, possible. Indeed, since the SUN densities of the filtering, predictive and smoothing distributions are available from Theorems 1– 2, the main functionals of interest can be computed via closed-form expressions (Arellano-Valle and Azzalini, 2006; Azzalini and Bacchieri, 2010; Gupta et al., 2013; Azzalini and Capitanio, 2014) or by relying on numerical integration. However, these strategies require evaluations of multivariate Gaussian cumulative distribution functions, which tend to be impractical as t grows or when the focus is on complex functionals.

In such situations, Monte Carlo integration provides a tractable solution which allows accurate evaluation of generic functionals $\mathrm{E}[g(\boldsymbol{\theta}_t) \mid \mathbf{y}_{1:t}]$, $\mathrm{E}[g(\boldsymbol{\theta}_{t+1}) \mid \mathbf{y}_{1:t}]$ and $\mathrm{E}[g(\boldsymbol{\theta}_{1:n}) \mid \mathbf{y}_{1:n}]$ for the filtering, predictive and smoothing distribution via

$$\frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{t|t}^{(r)}), \quad \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{t+1|t}^{(r)}), \text{ and } \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{1:n|n}^{(r)}),$$

with $\boldsymbol{\theta}_{t|t}^{(r)}$, $\boldsymbol{\theta}_{t+1|t}^{(r)}$ and $\boldsymbol{\theta}_{1:n|n}^{(r)}$ sampled from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_{t+1} \mid \mathbf{y}_{1:t})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, respectively. For example, if the evaluation of (2.11) is demanding, the observations predictive distribution can be computed as $\sum_{r=1}^{R} \Phi_m(\mathbf{B}_{t+1}\mathbf{F}_{t+1}\boldsymbol{\theta}_{t+1|t}^{(r)}; \mathbf{B}_{t+1}\mathbf{V}_{t+1}\mathbf{B}_{t+1})/R$.

To be implemented, the above approach requires an efficient strategy to sample from (2.9)–(2.10) and (2.12). Exploiting the SUN properties and recent results in Botev (2017), an algorithm to draw independent and identically distributed samples from the exact SUN distributions in (2.9)–(2.10) and (2.12) is developed in Section 2.4.1. As illustrated in Section 2.5, this technique is more accurate than state-of-the-art methods and can be efficiently implemented in a variety of small-to-moderate dimensional time series. In Section 2.4.2 we develop, instead, a scalable sequential Monte Carlo scheme for high dimensional settings, that has optimality properties.

2.4.1 Independent and identically distributed sampling

As discussed in Section 2.1, MCMC and sequential Monte Carlo methodologies to sample from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_{t+1} \mid \mathbf{y}_{1:t})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are available. However, the optimal solution, when possible, is to rely on independent and identically distributed (i.i.d.) samples. Here, we develop a Monte Carlo algorithm to address this goal with a main focus on the smoothing distribution, and discuss immediate modifications to allow sampling also in the filtering and predictive case. Indeed, Monte Carlo inference is particularly suitable in batch settings, although, as discussed later, the proposed routine is useful in practice also when the focus is on filtering and predictive distributions, since i.i.d. samples are simulated rapidly, for each t, in small-to-moderate dimensional time series.

Exploiting the closed-form expression of the smoothing distribution in Theorem 2 and the additive representation (2.7) of the SUN, i.i.d. samples $\boldsymbol{\theta}^{(1)}_{1:n|n}, \ldots, \boldsymbol{\theta}^{(R)}_{1:n|n}$ from the smoothing distribution (2.12) can be obtained via a linear combination between independent samples from $(p \cdot n)$-variate Gaussians and $(m \cdot n)$-variate truncated normals. Algorithm 1 provides the pseudo-code for this novel routine, whose outputs are i.i.d. samples from the joint smoothing distribution. Here, the most computationally intensive step is the sampling from the multivariate truncated normal. In fact, although efficient Hamiltonian Monte Carlo solutions are available (Pakman and Paninski, 2014), these strategies do not provide independent samples. More recently, an accept-reject method based on minimax tilting has been proposed by Botev (2017) to improve the acceptance rate of classical rejection sampling, while avoiding convergence and mixing issues of MCMC. Such a routine is available in the `R` library `TruncatedNormal` and allows efficient sampling from multivariate truncated normals having a dimension of few hundreds, thereby providing effective Monte Carlo inference via Algorithm 1 in small-to-moderate dimensional time series.

Clearly, the availability of i.i.d. sampling schemes from the smoothing distribution overcomes the need of MCMC methods and particle smoothers. The first set of strategies face mixing or time-inefficiency issues, especially for imbalanced binary datasets (Johndrow et al., 2019), whereas the second class of routines tend to be computationally intensive and subject to particles degeneracy (Doucet and Johansen, 2009).

When the focus is on Monte Carlo inference for the marginal smoothing distribution $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:n})$ at a specific time t, Algorithm 1 requires minor adaptations relying again on the additive representation of the SUN in equation (2.13), under similar arguments considered for the joint smoothing setting. This latter routine can be also used to sample from the filtering distribution by applying such a scheme with $n = t$ to obtain i.i.d. samples $\boldsymbol{\theta}^{(1)}_{t|t}, \ldots, \boldsymbol{\theta}^{(R)}_{t|t}$ from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. Based on these realizations, i.i.d. samples from the predictive distribution can be simply generated via direct application of equation (2.2) to obtain $\boldsymbol{\theta}^{(1)}_{t+1|t} = \mathbf{G}_{t+1}\boldsymbol{\theta}^{(1)}_{t|t} + \boldsymbol{\varepsilon}^{(1)}_{t+1}, \ldots, \boldsymbol{\theta}^{(R)}_{t+1|t} = \mathbf{G}_{t+1}\boldsymbol{\theta}^{(R)}_{t|t} + \boldsymbol{\varepsilon}^{(R)}_{t+1}$, with $\boldsymbol{\varepsilon}^{(1)}_{t+1}, \ldots, \boldsymbol{\varepsilon}^{(R)}_{t+1}$ denoting independent samples from a $\mathrm{N}_p(\mathbf{0}, \mathbf{W}_{t+1})$. Therefore, efficient

Algorithm 2: Optimal particle filter to sample from $p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t})$, for $t = 1, \dots, n$

for t *from* 1 *to* n **do**
  for r *from* 1 *to* R **do**
    1. Propose a value $\bar{\boldsymbol{\theta}}_{t|t}^{(r)}$ by sampling from (2.14) conditioned on $\boldsymbol{\theta}_{t-1} = \boldsymbol{\theta}_{t-1|t-1}^{(r)}$, as described below.
    - Sample $\mathbf{U}_{0\ t|t}^{(r)}$ from a $\mathrm{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{t|t,t-1} - \boldsymbol{\Delta}_{t|t,t-1}\boldsymbol{\Gamma}_{t|t,t-1}^{-1}\boldsymbol{\Delta}_{t|t,t-1}^{\intercal})$.
    - Sample $\mathbf{U}_{1\ t|t}^{(r)}$ from a $\mathrm{N}_m(\mathbf{0}, \boldsymbol{\Gamma}_{t|t,t-1})$ truncated below $-\boldsymbol{\gamma}_{t|t,t-1}^{(r)} = -\mathbf{c}_t^{-1}\mathbf{B}_t\mathbf{F}_t\mathbf{G}_t\boldsymbol{\theta}_{t-1|t-1}^{(r)}$.
    - Compute $\bar{\boldsymbol{\theta}}_{t|t}^{(r)} = \mathbf{G}_t\boldsymbol{\theta}_{t-1|t-1}^{(r)} + \boldsymbol{\omega}_{t|t,t-1}(\mathbf{U}_{0\ t|t}^{(r)} + \boldsymbol{\Delta}_{t|t,t-1}\boldsymbol{\Gamma}_{t|t,t-1}^{-1}\mathbf{U}_{1\ t|t}^{(r)})$.
    2. Calculate the associated importance weight $w_t^{(r)}$ via (2.15) and normalize them.
  3. Obtain $\boldsymbol{\theta}_{t|t}^{(1)}, \dots, \boldsymbol{\theta}_{t|t}^{(R)}$ by resampling from $\bar{\boldsymbol{\theta}}_{t|t}^{(1)}, \dots, \bar{\boldsymbol{\theta}}_{t|t}^{(R)}$ with weights $w_t^{(1)}, \dots, w_t^{(R)}$.

Monte Carlo inference in small-to-moderate dimensional dynamic probit models is possible also for the filtering and predictive distributions.

2.4.2 Optimal particle filtering

When the dimension of the dynamic probit model (2.1)–(2.2) increases, sampling from multivariate truncated Gaussians in Algorithm 1 can face computational bottlenecks (Botev, 2017). This is particularly likely to occur in series monitored on a fine time grid. Indeed, in several applications, the number of time series m is small-to-moderate, whereas the length of the time window can be large. To address this issue and allow scalable online inference for filtering and prediction also in large t settings, we propose a particle filter which exploits the SUN results to obtain optimality properties.

The proposed algorithm belongs to the class of sequential importance sampling-resampling (SISR) algorithms which provide default strategies in particle filtering (e.g., Doucet et al., 2000, 2001; Durbin and Koopman, 2012). For each time t, these routines sample R trajectories $\boldsymbol{\theta}_{1:t|t}^{(1)}, \dots, \boldsymbol{\theta}_{1:t|t}^{(R)}$, known as *particles*, conditioned on those produced at $t-1$, by iterating, in time, between the two steps summarized below.

1. Importance sampling. Let $\boldsymbol{\theta}_{1:t-1|t-1}^{(1)}, \dots, \boldsymbol{\theta}_{1:t-1|t-1}^{(R)}$ be the particles' trajectories at $t-1$, and denote with $\pi(\boldsymbol{\theta}_{t|t}|\boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$ the proposal. Then, for $r = 1, \dots, R$

[1.a] Sample $\bar{\boldsymbol{\theta}}_{t|t}^{(r)}$ from $\pi(\boldsymbol{\theta}_{t|t} \mid \boldsymbol{\theta}_{1:t-1|t-1}^{(r)}, \mathbf{y}_{1:t})$ and set

$$\bar{\boldsymbol{\theta}}_{1:t|t}^{(r)} = (\boldsymbol{\theta}_{1:t-1|t-1}^{(r)\intercal}, \bar{\boldsymbol{\theta}}_{t|t}^{(r)\intercal})^{\intercal}.$$

[1.b] Set $w_t^{(r)} = w_t(\bar{\boldsymbol{\theta}}_{1:t|t}^{(r)})$, with

$$w_t(\bar{\boldsymbol{\theta}}_{1:t|t}^{(r)}) \propto \frac{p(\mathbf{y}_t \mid \bar{\boldsymbol{\theta}}_{t|t}^{(r)})p(\bar{\boldsymbol{\theta}}_{t|t}^{(r)} \mid \boldsymbol{\theta}_{t-1|t-1}^{(r)})}{\pi(\bar{\boldsymbol{\theta}}_{t|t}^{(r)} \mid \boldsymbol{\theta}_{1:t-1|t-1}^{(r)}, \mathbf{y}_{1:t})},$$

and normalize the weights, so that their sum is 1.

2. Resampling. For $r = 1, \dots, R$, sample new particles $\boldsymbol{\theta}_{1:t|t}^{(1)}, \dots, \boldsymbol{\theta}_{1:t|t}^{(R)}$ from $\sum_{l=1}^{R} w_t^{(l)}\delta_{\bar{\boldsymbol{\theta}}_{1:t|t}^{(l)}}$.

Based on these particles, functionals of the filtering distribution $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ can be computed exploiting the

terminal values $\boldsymbol{\theta}_{t|t}^{(1)}, \ldots, \boldsymbol{\theta}_{t|t}^{(R)}$ of each particles' trajectory. It is worth noting that in point [1.a], we presented the general formulation of SISR algorithms, where the proposal distribution $\pi(\boldsymbol{\theta}_{t|t} \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$ can, in principle, depend on the whole trajectory $\boldsymbol{\theta}_{1:t-1}$ (Durbin and Koopman, 2012, Section 12.3).

As is clear from the above steps, the performance of SISR relies on the proposal $\pi(\boldsymbol{\theta}_{t|t} \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$. This importance function should allow tractable sampling along with efficient evaluation of the importance weights, and should be also carefully specified to propose effective candidate samples. Recalling Doucet et al. (2000), the optimal importance density is $\pi(\boldsymbol{\theta}_{t|t} \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t}) = p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ with weights $w_t(\boldsymbol{\theta}_{1:t}) \propto p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$. Indeed, conditionally on $\boldsymbol{\theta}_{1:t-1|t-1}^{(r)}$ and $\mathbf{y}_{1:t}$, this choice minimizes the variance of the importance weights, thus limiting degeneracy issues and improving mixing. Unfortunately, in several dynamic models, tractable sampling from $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ and direct calculation of $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ is not possible (Doucet et al., 2000). As outlined in Corollary 4, this is not the case for multivariate dynamic probit models. In particular, as a direct consequence of Theorem 1 and of the closure properties of the SUN, sampling from $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ is straightforward and $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ has a simple expression.

Corollary 4. *Under (2.1)–(2.2), the following results give the form of the optimal importance density and the optimal weights respectively (in the sense of Doucet et al. (2000)), for each $t = 1, \ldots, n$.*

$$\begin{aligned}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t) \quad &\sim \quad \mathrm{SUN}_{p,m}(\boldsymbol{\xi}_{t|t,t-1}, \boldsymbol{\Omega}_{t|t,t-1}, \boldsymbol{\Delta}_{t|t,t-1}, \\ &\qquad\qquad \boldsymbol{\gamma}_{t|t,t-1}, \boldsymbol{\Gamma}_{t|t,t-1}), \qquad (2.14)\\ p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1}) \quad &= \quad \Phi_m(\boldsymbol{\gamma}_{t|t,t-1}; \boldsymbol{\Gamma}_{t|t,t-1}), \qquad (2.15)\end{aligned}$$

with parameters defined by the recursive equations

$$\begin{aligned}\boldsymbol{\xi}_{t|t,t-1} &= \mathbf{G}_t \boldsymbol{\theta}_{t-1}, \quad \boldsymbol{\Omega}_{t|t,t-1} = \mathbf{W}_t,\\ \boldsymbol{\Delta}_{t|t,t-1} &= \bar{\boldsymbol{\Omega}}_{t|t,t-1} \boldsymbol{\omega}_{t|t,t-1} \mathbf{F}_t^\intercal \mathbf{B}_t \mathbf{c}_t^{-1},\\ \boldsymbol{\gamma}_{t|t,t-1} &= \mathbf{c}_t^{-1} \mathbf{B}_t \mathbf{F}_t \boldsymbol{\xi}_{t|t,t-1},\\ \boldsymbol{\Gamma}_{t|t,t-1} &= \mathbf{c}_t^{-1} \mathbf{B}_t \left(\mathbf{F}_t \boldsymbol{\Omega}_{t|t,t-1} \mathbf{F}_t^\intercal + \mathbf{V}_t\right) \mathbf{B}_t \mathbf{c}_t^{-1},\end{aligned}$$

where $\mathbf{c}_t = \left[\left(\mathbf{F}_t \boldsymbol{\Omega}_{t|t,t-1} \mathbf{F}_t^\intercal + \mathbf{V}_t\right) \circ \mathbf{I}_m\right]^{1/2}$.

Algorithm 2 illustrates the pseudo-code of the proposed optimal filter, which exploits the additive representation of the SUN and Corollary 4. Comparing Algorithms 1 and 2 it can be noticed that now the computational complexity of the different steps does not depend on t, thus facilitating scalable sequential inference in large t studies. Samples from the predictive distribution can be obtained from those of the filtering as in Section 2.4.1.

2.5 Illustration on financial time series

We study the performance of the methods in Sections 2.3 and 2.4 on a dynamic probit regression for the daily opening directions of the French CAC40 stock market index from January 4th, 2018 to March 29th, 2019. Consistent with this focus, the variable y_t is on a binary scale, with $y_t = 1$ if the opening value of the CAC40 on day t is greater than the corresponding closing value in the previous day, and $y_t = 0$ otherwise. Financial

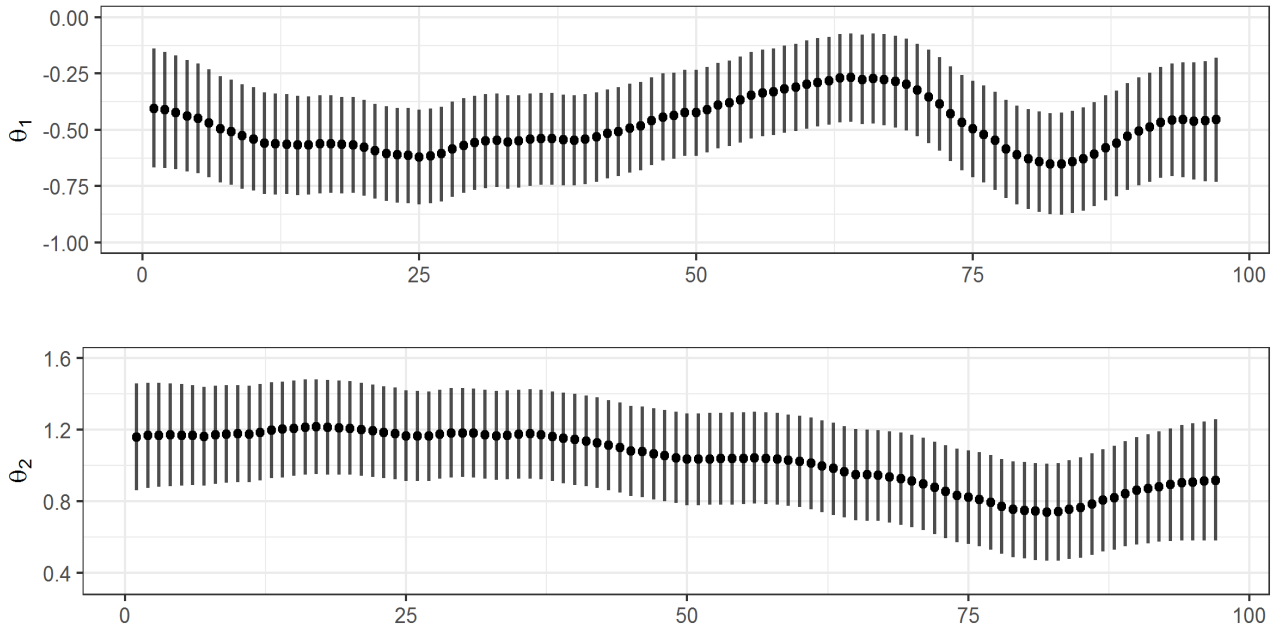

Figure 2.3: Pointwise median and interquartile range for the smoothing distributions of θ_{1t} and θ_{2t} under the dynamic probit regression in (2.16), for the time window from January 4th, 2018 to May 31st, 2018.

applications of this type have been a source of particular interest in past and recent years (e.g., Kim and Han, 2000; Kara et al., 2011; Atkins et al., 2018), with common approaches combining a wide variety of technical indicators and news information to predict stock markets directions via complex machine learning methods. Here, we show how a similar predictive performance can be obtained via a simple and interpretable dynamic probit regression for y_t, that combines past information on the opening directions of CAC40 with those of the NIKKEI225, regarded as binary covariates x_t with dynamic coefficients. Since the Japanese market opens before the French one, x_t is available before y_t and, hence, provides a valid predictor for each day t.

Recalling the above discussion and leveraging default specifications in these settings (e.g., Soyer and Sung, 2013), we rely on a dynamic probit regression for y_t with two independent random walk processes for the coefficients $\boldsymbol{\theta}_t = (\theta_{1t}, \theta_{2t})^\intercal$. Letting $\mathbf{F}_t = (1, x_t)$ and $\text{pr}(y_t = 1 \mid \boldsymbol{\theta}_t) = \Phi(\theta_{1t} + \theta_{2t} x_t; 1)$, such a model can be expressed as in equation (2.1) via

$$\begin{aligned} &p(y_t \mid \boldsymbol{\theta}_t) = \Phi[(2y_t - 1)\mathbf{F}_t\boldsymbol{\theta}_t; 1], \\ &\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} \text{N}_2(\mathbf{0}, \mathbf{W}), \quad t = 1, \dots n, \end{aligned} \tag{2.16}$$

where $\boldsymbol{\theta}_0 \sim \text{N}_2(\mathbf{a}_0, \mathbf{P}_0)$, whereas $\mathbf{W}$ is a time-invariant diagonal matrix. In (2.16), the element θ_{1t} of $\boldsymbol{\theta}_t$ measures the trend in the directions of the CAC40 when the NIKKEI225 has a negative opening on day t, whereas θ_{2t} characterizes the shift in such a trend if the opening of the NIKKEI225 index is positive, thereby providing an interpretable probit model with dynamic coefficients.

To evaluate performance in smoothing, filtering and prediction, we consider a situation in which the study starts on May 31st, 2018, with daily batch data available for the window from January 4th, 2018 to May 31st, 2018, and online observations streaming in from June 1st, 2018 to March 29th, 2019. This motivates smoothing

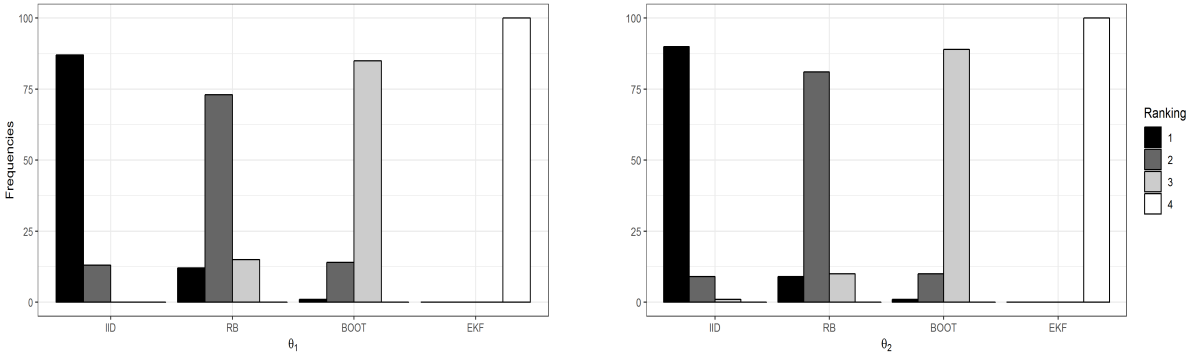

Figure 2.4: Ranking of the four sampling schemes in 100 replicated experiments according to the Wasserstein distance between the empirical smoothing distribution computed, at time $t = 97$, from 10^5 particles and the one obtained by direct evaluation of the exact density (2.10) on two grids of 2000 equally spaced values in $[-2.5, 1.5]$ and $[-1.5, 3]$ for θ_{1t} and θ_{2t}, respectively, with $t = 97$.

State	IID	RB	BOOT	EKF
θ_{1t} at $t = 97$	**0.00173**	0.00331	0.00670	0.01845
θ_{2t} at $t = 97$	**0.00221**	0.00428	0.01010	0.06245

Table 2.1: For each sampling scheme, Wasserstein distance—averaged across 100 different experiments—between the empirical smoothing distribution computed, at time $t = 97$, from 10^5 particles and the one obtained by direct evaluation of the exact density (2.10) on two grids of 2000 equally spaced values in $[-2.5, 1.5]$ and $[-1.5, 3]$ for θ_{1t} and θ_{2t}, respectively, with $t = 97$. The lowest distance for each state is bolded.

methods for the first $t = 1, \dots, 97$ times and online filters for the subsequent $t = 98, \dots, 299$ days.

Figure 2.3 shows the pointwise median and interquartile range of the smoothing distribution for θ_{1t} and θ_{2t}, $t = 1, \dots, 97$, based on 10^5 samples from Algorithm 1. To implement such a routine, we set $\mathbf{a}_0 = (0, 0)^\intercal$ and $\mathbf{P}_0 = \text{diag}(3, 3)$ following the guidelines in Gelman et al. (2008) and Chopin and Ridgway (2017) for probit regression. The states variances in the diagonal matrix $\mathbf{W}$ are instead set equal to 0.01 as suggested by a graphical search of the maximum for the marginal likelihood computed under different combinations of $(\text{W}_{11}, \text{W}_{22})$ via the analytical formula in Corollary 3.

As shown in Figure 2.3, the dynamic states θ_{1t} and θ_{2t} tend to concentrate around negative and positive values, respectively, for the entire smoothing window, thus highlighting a general concordance between CAC40 and NIKKEI225 opening patterns. However, the strength of this association varies in time, supporting our proposed dynamic probit over static specifications. For example, it is possible to observe a decay in θ_{1t} and θ_{2t} on April–May, 2018 which reduces the association among CAC40 and NIKKEI225, while inducing a general negative trend for the opening directions of the French market. Such a result could be due to the overall instability in the Eurozone on April–May, 2018 caused by the uncertainty after the Italian and British elections during those months.

To clarify the computational improvements provided by the methods in Section 2.4.1, we also compare, in Figure 2.4 and in Table 2.1, their performance against the competing strategies mentioned in Section 2.1. Here, the focus is on the marginal smoothing distribution of θ_{1t} and θ_{2t} at the last day among those available for batch smoothing. Such a distribution of interest coincides with the filtering at time $t = 97$, thereby allowing the implementation of the filters discussed in Section 2.1, to evaluate performance both in terms of smoothing and filtering. The competing methods include the extended Kalman filter (Uhlmann, 1992), the bootstrap particle filter (Gordon et al., 1993) and the Rao-Blackwellized sequential Monte Carlo by Andrieu and Doucet

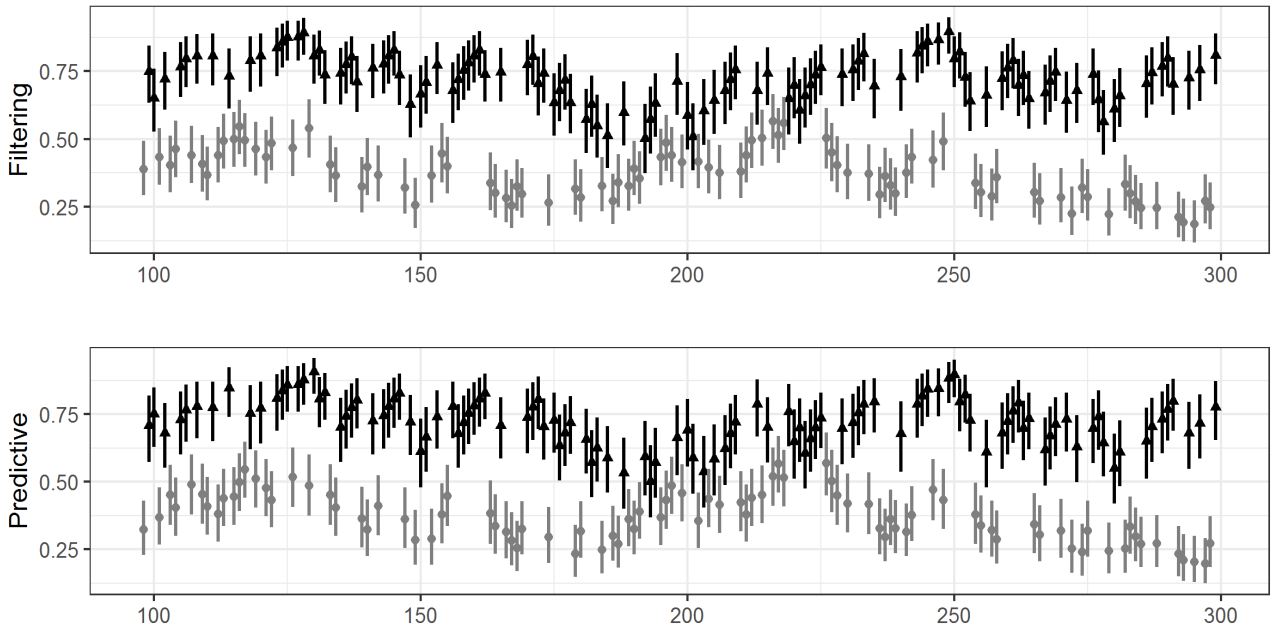

Figure 2.5: Median and interquartile range of the filtering and predictive distributions for $\Phi(\theta_{1t} + x_t\theta_{2t}; 1)$ computed from 10^5 particles produced by the optimal particle filter in Algorithm 2. Black and grey segments denote days in which $x_t = 1$ and $x_t = 0$, respectively.

(2002) which leverages the hierarchical representation (2.3)–(2.5) of model (2.1)–(2.2). Although being a popular solution in routine implementations, the extended Kalman filter relies on a quadratic approximation of the probit log-likelihood which leads to a Gaussian filtering distribution, thereby affecting the quality of online learning when imbalances in the data induce skewness. The bootstrap particle filter (Gordon et al., 1993) is, instead, motivated by the apparent absence of a tractable optimal proposal $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}^{(r)}_{t-1|t-1}, \mathbf{y}_t)$ (Doucet et al., 2000) and, thus, proposes values from $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}^{(r)}_{t-1|t-1})$. Also the Rao-Blackwellized sequential Monte Carlo (Andrieu and Doucet, 2002) aims at providing an alternative particle filter, which addresses the apparent unavailability of an analytical expression for the optimal proposal and the corresponding importance weights. The authors overcome this key issue by proposing a sequential Monte Carlo strategy for the Rao-Blackwellized filtering distribution $p(\mathbf{z}_t \mid \mathbf{y}_{1:t})$ of the partially observed Gaussian data $\mathbf{z}_t$ in model (2.3)–(2.5) and compute, for each trajectory $\mathbf{z}^{(r)}_{1:t}$, relevant moments of $p(\boldsymbol{\theta}_t \mid \mathbf{z}^{(r)}_{1:t})$ via classical Kalman filter updates—applied to model (2.4)–(2.5)—which are then averaged across particles to obtain Monte Carlo estimates for moments of $(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$.

Although the above methods provide state-of-the-art solutions, the proposed strategies are motivated by the apparent absence of a closed-form filter for (2.1)–(2.2), which is, in fact, available according to our results in Section 2.3. Figure 2.4 and Table 2.1 highlight to what extent this novel finding improves the existing methods. More specifically, Figure 2.4 compares the rankings of the different sampling schemes, in 100 replicated experiments, according to the Wasserstein distances (e.g., Villani, 2008) between the empirical smoothing distribution induced by the particles generated from each sampling method under analysis and the one obtained by direct evaluation of the exact density (2.10) on an appropriate grid. Table 2.1 shows, instead, these distances averaged across the 100 replicated experiments. For the sake of clarity, with a little abuse of terminology, the term *particle* is used to denote both the samples of the sequential Monte Carlo methods and those obtained

under i.i.d. sampling from the SUN. The Wasserstein distance is computed via the `R` function `wasserstein1d`. Note also that, although the extended Kalman filter and the Rao-Blackwellized sequential Monte Carlo focus, mostly, on the first two central moments of $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, these strategies can be adapted to draw samples from an approximation of the marginal smoothing density.

Figure 2.4 confirms that the sampling scheme in Section 2.4.1 over-performs all the competitors, since its ranking is 1 in most of the 100 experiments. The averaged Wasserstein distances in Table 2.1 yield the same conclusion. Such a result is due to the fact that the extended Kalman filter relies on an approximation of the filtering distribution, whereas, unlike the proposed exact sampler, the bootstrap and the Rao-Blackwellized particle filters consider sub-optimal dependent sampling strategies. Not surprisingly, the Rao-Blackwellized particle filter is the second best choice. Nonetheless, as expected, exact i.i.d. sampling remains the optimal solution and provides a viable strategy in any small-to-moderate study.

Motivated by the accurate performance of the Monte Carlo methods based on SUN results, we also apply the optimal particle filter in Algorithm 2 to provide scalable online filtering and prediction for model (2.16) from June 1st, 2018 to March 29th, 2019. Following the idea of sequential inference, the particles are initialized with the marginal smoothing distribution of May 31, 2018 from the batch analysis. Figure 2.5 outlines median and interquartile range for the filtering and predictive distribution of the probability that the CAC40 index has a positive opening in each day of the window considered for online inference. These two distributions can be easily obtained by applying the function $\Phi(\theta_{1t} + x_t\theta_{2t}; 1)$ to the particles of the states filtering and predictive distribution. In line with Figure 2.3, a positive opening of the NIKKEI225 provides, in general, an high estimate for the probability that $y_t = 1$, whereas a negative opening tends to favor the event $y_t = 0$. However, the strength of this result evolves over time with some periods showing less evident shifts in the probabilities process when x_t changes from 1 to 0. One-step-ahead prediction, leveraging the samples of the predictive distribution for the probability process, led to a correct classification rate of 66.34% which is comparable to those obtained under more complex procedures combining a wide variety of inputs to predict stock markets directions via state-of-the-art machine learning methods (e.g., Kim and Han, 2000; Kara et al., 2011; Atkins et al., 2018).

2.6 Discussion

This contribution shows that filtering, predictive and smoothing distributions in dynamic probit models for multivariate binary data have a SUN kernel and the associated parameters can be computed via tractable expressions. As discussed in Sections 2.3–2.5, this result provides advances in online inference and facilitates the implementation of tractable methods to draw i.i.d. samples from the exact filtering, predictive and smoothing distributions, thus allowing improved Monte Carlo inference in small-to-moderate time series. Filtering in high dimensions can be, instead, implemented via a scalable sequential Monte Carlo which exploits SUN properties to provide a particle filter with optimal proposal.

These results motivate additional future research. For instance, a relevant direction is to adapt or generalize the derivations in Section 2.3 to dynamic tobit, binomial and multinomial probit models, for which closed-form filters are unavailable. Joint filtering and prediction of continuous and binary time series is also of interest (Liu et al., 2009). A natural state-space model for these multivariate data can be obtained by generalizing (2.3)–(2.5) to allow only the subset of Gaussian variables associated with the binary data to be partially observed.

However, also in this case, closed-form filters are not available. By combining our results in Section 2.3 with the classical Kalman filter for Gaussian state-space models, such a gap could be possibly covered. As discussed in Sections 2.1 and 2.3.2, estimation and inference for possible unknown parameters characterizing the state-space model in (2.1)–(2.2) is another interesting problem which can be addressed by maximizing the marginal likelihood derived in Section 2.3.2. Finally, additional quantitative studies beyond those in Section 2.5 can be useful for obtaining a more comprehensive overview on the performance of our proposed computational methods compared to state-of-the-art strategies.

Data and Codes. The dataset considered in Section 2.5 is available at `Yahoo Finance`. Simple pseudo–codes that can be easily implemented with any software are provided in Algorithms 1 and 2.

Appendix A: Proofs of the main results

A.1. Proof of Lemma 1

To prove Lemma 1, note that, by applying the Bayes rule, we obtain

$$p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1) \propto p(\boldsymbol{\theta}_1)p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1),$$

where $p(\boldsymbol{\theta}_1) = \phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)$ and $p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1) = \Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)$. The expression for $p(\boldsymbol{\theta}_1)$ can be easily obtained by noticing that $\boldsymbol{\theta}_1 = \mathbf{G}_1\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_1$ in (2.2), with $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$ and $\boldsymbol{\varepsilon}_1 \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_1)$. The form for the probability mass function of $(\mathbf{y}_1 \mid \boldsymbol{\theta}_1)$ is instead a direct consequence of equation (2.1). Hence, combining these expressions and recalling (2.6), it is clear that $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is proportional to the density of a SUN with suitably–specified parameters, such that the kernel of (2.6) coincides with $\phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)$. In particular, letting

$$\begin{aligned}
\boldsymbol{\xi}_{1|1} &= \mathbf{G}_1\mathbf{a}_0, \quad \boldsymbol{\Omega}_{1|1} = \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1, \\
\boldsymbol{\Delta}_{1|1} &= \bar{\boldsymbol{\Omega}}_{1|1}\boldsymbol{\omega}_{1|1}\mathbf{F}_1^\intercal\mathbf{B}_1\mathbf{s}_1^{-1}, \quad \boldsymbol{\gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1}, \\
\boldsymbol{\Gamma}_{1|1} &= \mathbf{s}_1^{-1}\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1\mathbf{s}_1^{-1},
\end{aligned}$$

we have that

$$\begin{aligned}
&\boldsymbol{\gamma}_{1|1} + \boldsymbol{\Delta}_{1|1}^\intercal\bar{\boldsymbol{\Omega}}_{1|1}^{-1}\boldsymbol{\omega}_{1|1}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\xi}_{1|1}) \\
&= \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1} + \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1(\boldsymbol{\theta}_1 - \boldsymbol{\xi}_{1|1}) = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1, \\
&\boldsymbol{\Gamma}_{1|1} - \boldsymbol{\Delta}_{1|1}^\intercal\bar{\boldsymbol{\Omega}}_{1|1}^{-1}\boldsymbol{\Delta}_{1|1} \\
&= \mathbf{s}_1^{-1}[\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1 - \mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal)\mathbf{B}_1]\mathbf{s}_1^{-1} \\
&= \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1\mathbf{s}_1^{-1}.
\end{aligned}$$

with $\mathbf{s}_1^{-1}$ as in Lemma 1. Now, substituting these quantities in the kernel of the SUN density (2.6), we have

$$\begin{aligned}
&\phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1) \\
&\quad \cdot \Phi_m(\mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1\mathbf{s}_1^{-1}) \\
&= \phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1) \\
&= p(\boldsymbol{\theta}_1)p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1) \propto p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1),
\end{aligned}$$

thus proving Lemma 1. To prove that $\boldsymbol{\Omega}^*_{1|1}$ is a correlation matrix, replace the identity $\mathbf{I}_m$ with $\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1$ in the proof of Theorem 1 by Durante (2019). □

A.2. Proof of Theorem 1

Recalling (2.2), the proof for $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ in (2.9) requires studying the variable $\mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t$, given $\mathbf{y}_{1:t-1}$, where

$$\begin{aligned}
(\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1}) \sim \text{SUN}_{p,m\cdot(t-1)}(&\boldsymbol{\xi}_{t-1|t-1}, \boldsymbol{\Omega}_{t-1|t-1}, \\
&\boldsymbol{\Delta}_{t-1|t-1}, \boldsymbol{\gamma}_{t-1|t-1}, \boldsymbol{\Gamma}_{t-1|t-1}),
\end{aligned}$$

and $\boldsymbol{\varepsilon}_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t)$, with $\boldsymbol{\varepsilon}_t \perp \mathbf{y}_{1:t-1}$. To address this goal, first note that, by the closure properties of the SUN under linear transformations (Azzalini and Capitanio, 2014, Section 7.1.2), the variable $(\mathbf{G}_t\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1})$ is still a SUN with parameters $\mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal$, $[(\mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal) \circ \mathbf{I}_p]^{-1/2}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t-1|t-1}$ and $\boldsymbol{\Gamma}_{t-1|t-1}$. Hence, to conclude the proof of equation (2.9), we only need to obtain the distribution of the sum among this variable and the noise $\boldsymbol{\varepsilon}_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t)$. This can be accomplished by considering the moment generating function of such a sum—as done by Azzalini and Capitanio (2014, Section 7.1.2) to prove closure under convolution. Indeed, it is straightforward to note that the product of the moment generating functions for $\boldsymbol{\varepsilon}_t$ and $(\mathbf{G}_t\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1})$ leads to the moment generating function of the SUN with parameters $\boldsymbol{\xi}_{t|t-1} = \mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\boldsymbol{\Omega}_{t|t-1} = \mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal + \mathbf{W}_t$, $\boldsymbol{\Delta}_{t|t-1} = \boldsymbol{\omega}_{t|t-1}^{-1}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t|t-1} = \boldsymbol{\gamma}_{t-1|t-1}$ and $\boldsymbol{\Gamma}_{t|t-1} = \boldsymbol{\Gamma}_{t-1|t-1}$. To prove (2.10) note that

$$p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}) \propto \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$$

coincides with the posterior distribution in a probit model with likelihood $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ and SUN prior $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ from (2.9). Hence, expression (2.10) can be derived from Corollary 4 in Durante (2019), replacing the matrix $\mathbf{I}_m$ in the classical probit likelihood with $\mathbf{B}_t\mathbf{V}_t\mathbf{B}_t$. □

A.3. Proof of Corollary 1

To prove Corollary 1, first note that $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\text{d}\boldsymbol{\theta}_t$ can be re-written as

$$\frac{\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\text{d}\boldsymbol{\theta}_t}{\Phi_{m\cdot(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})},$$

where

$$K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1}) = p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\Phi_{m\cdot(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})$$

is the kernel of the predictive distribution in (2.9). Consistent with this result, Corollary 1 follows after noticing that $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ is the kernel of the filtering distribution in (2.10), whose normalizing constant $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\theta}_t$ is equal to $\Phi_{m\cdot t}(\boldsymbol{\gamma}_{t|t}; \boldsymbol{\Gamma}_{t|t})$. □

A.4. Proof of Theorem 2

First note that $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \propto p(\boldsymbol{\theta}_{1:n})p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$. Therefore, $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ can be interpreted as the posterior distribution in the Bayesian model having likelihood $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$ and prior $p(\boldsymbol{\theta}_{1:n})$ for the $(p \cdot n) \times 1$ vector $\boldsymbol{\theta}_{1:n} = (\boldsymbol{\theta}_1^\intercal, \ldots, \boldsymbol{\theta}_n^\intercal)^\intercal$. As already pointed out in Section 2.3.2, it immediately follows from model (2.2) that $\boldsymbol{\theta}_{1:n} \sim \mathrm{N}_{p\cdot n}(\boldsymbol{\xi}, \boldsymbol{\Omega})$, with $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ as in Section 2.3.2. The form of $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$ can be instead obtained from (2.1), by noticing that given $\boldsymbol{\theta}_{1:n}$ the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are conditionally independent, thus providing the joint likelihood $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n}) = \prod_{t=1}^n \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$. Such a quantity can be also expressed as $\Phi_{m\cdot n}(\mathbf{D}\boldsymbol{\theta}_{1:n}; \mathbf{V})$ with $\mathbf{D}$ and $\mathbf{V}$ as in Section 2.3.2. Combining these results, the joint smoothing distribution $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ is proportional to $\phi_{p\cdot n}(\boldsymbol{\theta}_{1:n} - \boldsymbol{\xi}; \boldsymbol{\Omega})\Phi_{m\cdot n}(\mathbf{D}\boldsymbol{\theta}_{1:n}; \mathbf{V})$, which is the kernel of a SUN variable with parameters as in (2.12). □

A.5. Proof of Corollary 3

The expression for the marginal likelihood follows after noticing that $p(\mathbf{y}_{1:n})$ is the normalizing constant of the joint smoothing distribution. Indeed, $p(\mathbf{y}_{1:n})$ is formally defined as $\int p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})p(\boldsymbol{\theta}_{1:n})d\boldsymbol{\theta}_{1:n}$. Hence, the integrand function coincides with the kernel of the smoothing density, so that the whole integral is equal to $\Phi_{m\cdot n}(\boldsymbol{\gamma}_{1:n|n}; \boldsymbol{\Gamma}_{1:n|n})$. □

A.6. Proof of Corollary 4

The proof of Corollary 4 is similar to the one of Lemma 1. Indeed, the proposal $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ is proportional to the product between the likelihood $p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ and the prior $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) = \phi_p(\boldsymbol{\theta}_t - \mathbf{G}_t\boldsymbol{\theta}_{t-1}; \mathbf{W}_t)$. To derive the importance weights in (2.15), it suffices to note that the marginal likelihood $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ coincides with the normalizing constant of the SUN in (2.14). □

Chapter 3

Hidden hierarchical Dirichlet process for clustering[1]

3.1 Introduction

Dirichlet process (DP) mixtures are well-established and highly successful Bayesian nonparametric models for density estimation and clustering, which also enjoy appealing frequentist asymptotic properties (Lo, 1984; Escobar, 1994; Escobar and West, 1995; Ghosal and Van Der Vaart, 2017). However, they are not suitable to model data $\{(X_{j,1}, \dots, X_{j,I_j}) : j = 1, \dots, J\}$ that are recorded under J different, though related, experimental conditions. This is due to exchangeability implying a common underlying distribution across populations, a homogeneity assumption which is clearly too restrictive. To make things concrete consider the Collaborative Perinatal Project, a large prospective epidemiologic study conducted from 1959 to 1974 (analyzed in Section 3.5.1), where pregnant women were enrolled in 12 hospitals and followed over time. Using a DP mixture would correspond to ignoring the information on the specific center j where the data are collected and, thus, the heterogeneity across samples. The opposite, also unrealistic, extreme case corresponds to modeling data from each hospital independently, thus ignoring possible similarities between centers.

A natural compromise between the aforementioned extreme cases is *partial exchangeability* (de Finetti, Bruno, 1938), which entails exchangeability within each experimental condition (but not across) and *dependent* population–specific distributions (thus allowing borrowing of information). See Kallenberg (2005) for a detailed account on the topic. In this framework the proposal of dependent versions of the DP date back to the seminal papers of Cifarelli and Regazzini (1978) and MacEachern (1999, 2000). Dependent DPs can be readily used within mixtures leading to several success stories in topic modeling, biostatistics, speaker diarization, genetics, fMRI analysis and so forth. See Dunson (2010); Teh and Jordan (2010); Foti and Williamson (2015) and references therein.

Two hugely popular dependent nonparametric priors, which will also represent the key ingredients of the present contribution, are the hierarchical Dirichlet process (HDP) (Teh et al., 2006) and the nested Dirichlet process (NDP) (Rodríguez et al., 2008). The HDP clusters observations within and across populations. The

[1] Joint work with Antonio Lijoi and Igor Prünster. Department of Decision Sciences, Bocconi University.

NDP aims to cluster both population distributions and observations, but as shown in Camerlenghi et al. (2019), does not achieve this goal. In fact, if there is a cluster of observations shared by different samples, the model degenerates to exchangeability across samples. This issue is successfully overcome in Camerlenghi et al. (2019) by introducing *latent nested nonparametric priors*. However, while this proposal has the merit of being the first to solve the degeneracy problem, it suffers from other limitations in terms of implementation and modeling: (a) with data from more than two populations the analytical and computational burden implied by the additive structure becomes overwhelming; (b) the model lacks the flexibility needed to capture different weights that common clusters may feature across different populations. More details can be found in the discussion to Camerlenghi et al. (2019).

The goal of this work is thus to introduce a nonparametric prior, which allows to cluster simultaneously distributions and observations (within and across populations). We achieve this by blending peculiar features of both the NDP and the HDP into a novel model, which we term *Hidden Hierarchical Dirichlet Process* (HHDP). Importantly, our proposal overcomes the above-mentioned theoretical, modeling and computational limitations given it, respectively, does not suffer from the degeneracy flaw, is able to effectively capture different weights of shared clusters and allows to handle several populations as showcased in the real data application.

Section 3.2 concisely reviews the HDP and the NDP with focus on the random partitions they induce. In Section 3.3 we introduce the HHDP and investigate its properties, foremost its clustering structure (induced by a partially exchangeable array of observations). These findings lead to the development of marginal and conditional Gibbs sampling schemes in Section 3.4. In Section 3.5 we draw a comparison between HHDP and NDP on synthetic data and present a real data application for our model. Finally, Section 3.6 is devoted to some concluding remarks and possible future research.

3.2 Bayesian nonparametric priors for clustering

The assumption of exchangeability that characterizes widely used Bayesian inferential procedures is equivalent to assuming data homogeneity. This is not realistic in many applied contexts, for instance, for data recorded under J different experimental conditions inducing heterogeneity. A natural assumption that relaxes exchangeability and is suited for arrays of random variables $\{(X_{j,i})_{i\geq 1} : j = 1, \dots, J\}$ is *partial exchangeability*, which amounts to assuming homogeneity within each population, though not across different populations. This is characterized by

$$\{(X_{j,i})_{i\geq 1} : j = 1, \dots, J\} \stackrel{\mathrm{d}}{=} \{(X_{j,\sigma_j(i)})_{i\geq 1} : j = 1, \dots, J\},$$

for every finitary permutations $\{\sigma_j : j = 1, \dots, J\}$ with $\stackrel{\mathrm{d}}{=}$ henceforth denoting equality in distribution. The dependence structure is effectively visualized by the hierarchical formulation

$$\begin{aligned} X_{j,i} \mid (G_1, \dots, G_J) &\stackrel{\text{ind}}{\sim} G_j, \\ (G_1, \dots, G_J) &\sim \mathcal{L}. \end{aligned} \tag{3.1}$$

Here we focus on priors $\mathscr{L}$ defined as composition of discrete random structures and including, as special cases, both the HDP and the NDP. More specifically, we consider $\mathscr{L}$ in (3.1) that arise as

$$G_j \,|\, Q \overset{\text{iid}}{\sim} \mathscr{L}(G_j|\, Q) \quad (j = 1, \ldots, J); \qquad Q \,|\, G_0 \sim \mathscr{L}(Q|\, G_0); \qquad G_0 \sim \mathscr{L}(G_0), \tag{3.2}$$

with discrete random probability measures G_j $(j = 1, \ldots, J)$, Q and G_0. The data are denoted by $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J\}$ with $\boldsymbol{X}_j = (X_{j,1}, \ldots, X_{j,I_j})$ and I_j the size of the jth sample. Discreteness of these random structures entails that with positive probability there are ties within each sample $\boldsymbol{X}_j$ and also across samples $j = 1, \ldots, J$, i.e. $\mathbb{P}(X_{j,i} = X_{j,\ell}) > 0$ for any $i \neq \ell$, and $\mathbb{P}(X_{j,i} = X_{\kappa,\ell}) > 0$ for any $j \neq \kappa$. Hence, $\boldsymbol{X}$ induces a random partition of the integers $\{1, 2, \ldots, n\}$ with $n = I_1 + \cdots + I_J$, whose distribution encapsulates the whole probabilistic clustering of the model and is therefore the key quantity to study. Importantly, the random partition can be characterized in terms of the partially exchangeable partition probability function (pEPPF) as defined in Camerlenghi et al. (2019). The pEPPF is the natural generalization of the concept of exchangeable partition probability function (EPPF) for the exchangeable case (see e.g. Pitman, 2006). More precisely, D is the number of distinct values among the $n = \sum_{j=1}^{J} I_j$ observations in the overall sample $\boldsymbol{X}$. The vector of frequency counts is denoted by $\boldsymbol{n}_j = (n_{j,1}, \ldots, n_{j,D})$ with $n_{j,d}$ indicating the number of elements in the jth sample that coincide with the dth distinct value in order of arrival. The dth distinct value is shared by any two samples j and j' if and only if $n_{j,d}\, n_{j',d} \geq 1$. The probability law of the random partition is characterized by the pEPPF defined as

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) = \mathbb{E} \int_{\mathbb{X}_*^D} \prod_{d=1}^{D} G_1^{n_{1,d}}(\mathrm{d}x_d) \ldots G_J^{n_{J,d}}(\mathrm{d}x_d), \tag{3.3}$$

with the constraint $\sum_{d=1}^{D} n_{j,d} = I_j$, for each $j = 1, \ldots, J$ and where $\mathbb{X}$ is the space in which the $X_{j,i}$'s take values and $\mathbb{X}_*^D$ is the product space $\mathbb{X}^D$ minus the set of points $\boldsymbol{x} \in \mathbb{X}^D$ that have a tie in at least two coordinates. Obviously for a single population, that is $J = 1$, the standard EPPF is recovered. Note that (3.3) is interpretable as an extension of a product partition model to a multiple samples framework and, hence, represents an alternative approach to popular covariate–dependent product partition models. See, e.g., Müller et al. (2011), Page and Quintana (2016) and Page and Quintana (2018).

If we further specify $\mathscr{L}(\,\cdot\,|Q)$ and Q such that they give rise to an NDP, then one may have ties also among the population probability distributions $G_1, \ldots, G_J$, i.e. $\mathbb{P}(G_j = G_\kappa) > 0$ for any $j \neq \kappa$. Therefore, in the framework of (3.1) and (3.2), one may investigate two types of clustering: (i) *sample clustering*, which is related to $G_1, \ldots, G_J$ and (ii) *observation clustering*, which refers to $\boldsymbol{X}$. The composition of these two clustering structures is the main tool we rely on to devise a simple, yet effective, model that considerably improves over existing alternatives.

3.2.1 Hierarchical Dirichlet process

Probably the most popular nonparametric prior for the partially exchangeable case is the HDP of Teh et al. (2006), which can be nicely framed in the composition scheme (3.2) as

$$\mathscr{L}(G_j|Q) = \textsc{dp}(G_j|\beta, Q), \quad \mathscr{L}(Q|G_0) = \delta_{G_0}(Q), \quad \mathscr{L}(G_0) = \textsc{dp}(G_0|\beta_0; H), \tag{3.4}$$

where $\text{DP}(\cdot|\alpha, P)$ denotes the law of a DP with concentration parameter $\alpha > 0$ and baseline probability measure P. Here we assume that H is a non–atomic probability measure on $\mathbb{X}$ and we refer to such prior as an J-dimensional HDP denoted by $(G_1, \ldots, G_J) \sim \text{HDP}(\beta, \beta_0; H)$. Hence, the G_j's share the atoms through G_0 and this leads to the creation of shared clusters of observations (or latent features) across the J groups. The pEPPF induced by a partially exchangeable array in (3.1) with $\mathcal{L} = \text{HDP}(\beta, \beta_0; H)$ has been determined in Camerlenghi et al. (2019, Ex. 3). It is important to stress that the model is not suited for comparing populations distributions since $\mathbb{P}(G_j = G_\kappa) = 0$ for any $j \neq \kappa$ (unless the G_j's are degenerate at G_0, in which case all distributions are equal). Similar compositions are considered in Camerlenghi et al. (2019) and, more recently, in Argiento et al. (2020) and Bassetti et al. (2020). Anyhow, the HDP and its variations cannot be used to cluster both populations and observations. To achieve this one has to rely on priors induced by nested structures, the most popular being the NDP.

3.2.2 Nested Dirichlet process

The NDP, introduced by Rodríguez et al. (2008), is the most widely used nonparametric prior allowing to cluster both observations and populations. However, as proved in Camerlenghi et al. (2019), it suffers from a *degeneracy issue*, because even a single tie shared across samples is enough to group the J population distributions into a single cluster.

Like the HDP, also the NDP can be framed in the composition structure (3.2) as

$$\mathcal{L}(G_j|Q) = Q(G_j), \quad \mathcal{L}(Q|G_0) = \text{DP}(Q|\alpha; G_0), \quad \mathcal{L}(G_0) = \delta_{\text{DP}(\beta;H)}(G_0), \tag{3.5}$$

where Q is a random probability measure on the space $\mathsf{P}_{\mathbb{X}}$ of probability measures on $\mathbb{X}$ and G_0 is degenerate on the atom $\text{DP}(\beta; H)$, which is is the law of a DP on the sample space $\mathbb{X}$. As in (3.4), H is assumed to be a non-atomic probability measure on $\mathbb{X}$. Henceforth, we write $(G_1, \ldots, G_J) \sim \text{NDP}(\alpha, \beta; H)$. By virtue of the well–known stick–breaking representation of the DP (Sethuraman, 1994) one has

$$Q = \sum_{k \geq 1} \pi_k^* \delta_{G_k^*}, \quad (\pi_k^*)_{k \geq 1} \sim \text{GEM}(\alpha), \quad G_k^* \overset{\text{iid}}{\sim} \text{DP}(\beta; H), \tag{3.6}$$

where the weights $(\pi_k^*)_{k \geq 1}$ and the random distributions $(G_k^*)_{k \geq 1}$ are independent. Recall that GEM stands for the distribution of probability weights after Griffiths, Engen and McCloskey, according to the well-established terminology of Ewens (1990). Given a sequence $(V_i)_{i \geq 1}$ such that $V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, this means that $\pi_1^* = V_1$ and $\pi_k^* = V_k \prod_{i=1}^{k-1}(1 - V_i)$, for any $k \geq 2$. Since $\mathbb{P}(G_j = G_\kappa) = 1/(\alpha + 1)$ for any $j \neq \kappa$, Q generates ties among the random distributions G_j's with positive probability and, thus, it clusters populations. Furthermore, a structure similar to the one displayed in (3.6) holds true for each G_k^*, i.e.

$$G_k^* = \sum_{l \geq 1} \omega_{k,l} \delta_{X_{k,l}^*}, \quad (\omega_{k,l})_{l \geq 1} \overset{\text{iid}}{\sim} \text{GEM}(\beta), \quad X_{k,l}^* \overset{\text{iid}}{\sim} H,$$

and, due to the non–atomicity of H, the $X_{k,l}^*$ are all distinct values.

The discrete structure of the G_k^*'s generates ties across the samples $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ with positive probability. For example, $\mathbb{P}(X_{j,i} = X_{j',i'}) = 1/\{(\alpha + 1)(\beta + 1)\}$ for any $j \neq j'$. Hence, the G_k^*'s induce a

clustering of the observations $\boldsymbol{X}$.

If the data $\boldsymbol{X}$ are modeled as in (3.1), with $(G_1, \ldots, G_J) \sim \text{NDP}(\alpha, \beta; H)$, conditional on a partition of the G_j's the observations from populations allocated to the same cluster are exchangeable and those from populations allocated to distinct clusters are independent. This potentially appealing feature of the NDP is however the one responsible for the above mentioned *degeneracy issue*. For exposition clarity, consider the case of $J = 2$ populations. If the two populations belong to different clusters, i.e. $G_1 \neq G_2$, they cannot share even a single atom $X^*_{k,l}$ due to the non–atomicity of H. Hence, $\mathbb{P}(X_{1,l} = X_{2,l'}|G_1 \neq G_2) = 0$ for any l and l'. Therefore there is neither clustering of observations nor borrowing of information across different populations. On the contrary, $\mathbb{P}(X_{1,i} = X_{2,i'}|G_1 = G_2) = 1/(\beta + 1) > 0$. These two findings are quite intuitive. Indeed, $G_1 \neq G_2$ means they are independent realizations of a DP with atoms iid from the same non-atomic probability distribution H and, thus, they are almost surely different. Instead, $G_1 = G_2$ corresponds to all observations coming from the same population distribution, more precisely from the same DP, and ties occur (with positive probability). A far less intuitive fact is that when a single atom, say $X^*_{k,l}$, is shared between G_1 and G_2 the model degenerates to the exchangeable case, namely $\mathbb{P}(G_1 = G_2|X_{1,i} = X_{2,i'}) = 1$ and the two populations have (almost surely) equal distributions. Hence, the NDP is not an appropriate specification when aiming at clustering both populations and observations across different populations. This was shown in Camerlenghi et al. (2019) where, spurred by this anomaly of the NDP, a novel class of priors named *latent nested processes* (LNP) designed to ensure that $\mathbb{P}(G_1 \neq G_2|X_{1,i} = X_{2,i'}) > 0$ is proposed. However, while this represents in principle a solution to the problem, it has computational and modeling limitations. On the one hand the implementation of LNPs with more than two samples is not feasible due to severe computational hurdles. On the other hand LNPs have limited flexibility since weights of the common clusters of observations across different populations are the same. This feature is not suited to several applications and the discussion to Camerlenghi et al. (2019) provides interesting examples. See also Soriano and Ma (2019) and in Christensen and Ma (2020) for stimulating contributions to this literature.

Hence, within the composition structure framework (3.2), our goal is to obtain a novel prior able to infer the clustering structure of both populations and observations, which is highly flexible and implementable for a large number of populations and associated samples.

3.3 Hidden hierarchical Dirichlet process

Our proposal consists in blending the HDP and the LDP in a way to leverage on their strengths, namely clustering data across multiple heterogeneous samples for the HDP and clustering different populations (or probability distributions) for the NDP. More precisely we combine these two models in a structure (3.2) as

$$\mathcal{L}(G_j|Q) = Q(G_j), \quad \mathcal{L}(Q|G_0) = \text{DP}(Q|\alpha; \text{DP}(\beta; G_0)), \quad \mathcal{L}(G_0) = \text{DP}(G_0|\beta_0; H).$$

This leads to the following definition.

Definition 1. The vector of random probability measures $(G_1, \ldots, G_J)$ is a *hidden hierarchical Dirichlet process* (HHDP) if

$$G_j \mid Q \overset{\text{iid}}{\sim} Q, \quad Q = \sum_{k\geq 1} \pi^*_k \delta_{G^*_k}, \quad (\pi^*_k)_{k\geq 1} \sim \text{GEM}(\alpha), \quad (G^*_k)_{k\geq 1} \sim \text{HDP}(\beta, \beta_0; H),$$

with $(\pi_k^*)_{k\geq 1}$ and $(G_k^*)_{k\geq 1}$ independent. In the sequel we write $(G_1, \ldots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$.

In terms of a graphical model, the HHDP can be represented as in Figure 3.1.

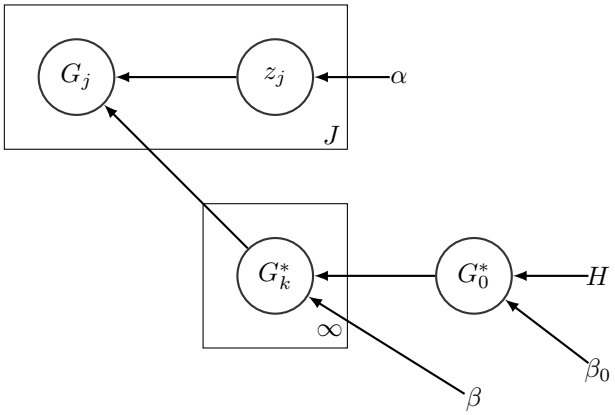

Figure 3.1: Graphical model representing the dependencies for a $\text{HHDP}(\alpha, \beta, \beta_0; H)$. Here the z_j's are auxiliary integer–valued random variables that assign each G_j to a specific atom G_k^* of Q.

The sequence $(G_k^*)_{k\geq 1}$ acts as a hidden, or latent, component that is crucial to avoid the bug of the NDP, namely clustering of populations when they share some observations. Moreover, by extending (3.4) to $J = \infty$, it can be more conveniently represented as

$$\begin{gathered} G_k^* = \sum_{l\geq 1} \omega_{k,l}\, \delta_{Z_{k,l}}, \quad Z_{k,l}|G_0^* \overset{\text{iid}}{\sim} G_0^*, \quad G_0^* = \sum_{l\geq 1} \omega_{0,l}\, \delta_{X_l^*}, \quad X_\ell^* \overset{\text{iid}}{\sim} H, \\ (\omega_{k,l})_{l\geq 1} \overset{\text{iid}}{\sim} \text{GEM}(\beta), \quad (\omega_{0,l})_{l\geq 1} \sim \text{GEM}(\beta_0), \end{gathered} \tag{3.7}$$

where independence holds true between the sequences $(\omega_{k,l})_{l\geq 1}$ and $(Z_{k,l})_{l\geq 1}$ and between $(\omega_{0,l})_{l\geq 1}$ and $(X_l^*)_{l\geq 1}$. Combining the stick-breaking representation and a closure property of the DP with respect to grouping, one further has

$$G_k^* = \sum_{l\geq 1} \omega_{k,l}^* \delta_{X_l^*}, G_0^* = \sum_{l\geq 1} \omega_{0,l} \delta_{X_l^*},$$

where $((\omega_{k,l}^*)_{l\geq 1} \mid \boldsymbol{\omega}_0) \overset{\text{iid}}{\sim} \text{DP}(\beta; \boldsymbol{\omega}_0)$, $\boldsymbol{\omega}_0 = (\omega_{0,l})_{l\geq 1} \sim \text{GEM}(\beta_0)$ and $X_l^* \overset{\text{iid}}{\sim} H$, for $l \geq 1$.

In this scheme the clustering of populations is governed, *a priori*, by the NDP layer Q through $(\pi_k^*)_{k\geq 1} \sim \text{GEM}(\alpha)$. However, the aforementioned degeneracy issue of the NDP, *a posteriori*, is successfully avoided. The intuition is quite straightforward: unlike for the NDP, the distinct distributions G_k^* in the HHDP are dependent and have a common random discrete base measure G_0^*, which leads to shared atoms across the G_k^*'s and thus borrowing of information, similarly to the HDP case.

3.3.1 Some distributional properties

Given the discreteness of $(G_1, \ldots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$, the key quantity to derive is the induced random partition, which controls the clustering mechanism of the model. However, it is useful to start with a description of pairwise dependence of the elements of the vector $(G_1, \ldots, G_J)$, which allows a better understanding of the model and intuitive parameter elicitation. To this end, as customary, we evaluate the correlation between $G_j(A)$

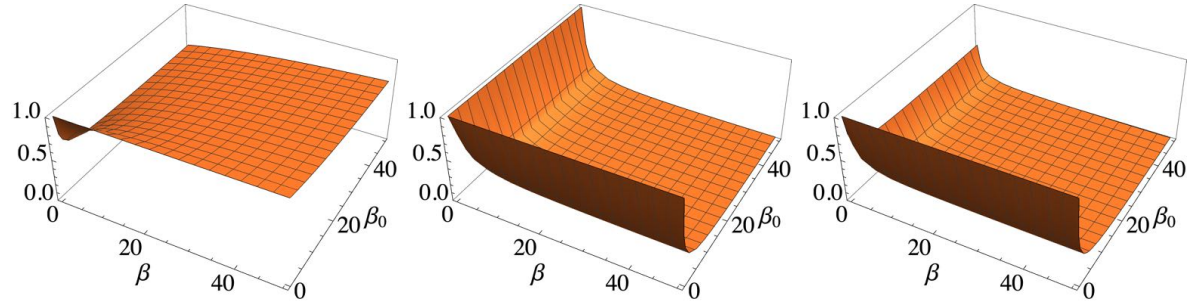

Figure 3.2: Correlations as function of the hyperparameters β and β_0 with $\alpha = 1$. The left plot represents the correlation between random probabilities $G_j(A)$, the middle one between observations collected in the same population and the right one between observations from different populations.

and $G_{j'}(A)$: whenever it does not depend on the specific set $A \subset \mathbb{X}$, it is used as a measure of overall dependence between G_j and $G_{j'}$.

Proposition 1. *If* $(G_1, \ldots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ *and* A *is a Borel subset of* $\mathbb{X}$*, then*

$$Var[G_j(A)] = \frac{H(A)[1 - H(A)](\beta_0 + \beta + 1)}{(\beta + 1)(\beta_0 + 1)} \qquad (j = 1, \ldots, J),$$

$$\text{Corr}[G_j(A), G_{j'}(A)] = 1 - \frac{\alpha\beta_0}{(\alpha + 1)(\beta + \beta_0 + 1)} \qquad (j \neq j').$$

Arguments similar to those in the proof of Proposition 1 lead to determine the correlation between observations, either from the same or from different samples.

Proposition 2. *If* $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ *are drawn from* $(G_1, \ldots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$*:*

$$\text{Corr}(X_{j,i}, X_{j',i'}) = \mathbb{P}(X_{j,i} = X_{j,i'}) = \begin{cases} \dfrac{1}{\beta_0 + 1} + \dfrac{\beta_0}{(1 + \alpha)(1 + \beta)(1 + \beta_0)} & (j \neq j') \\ \dfrac{\beta + \beta_0 + 1}{(\beta + 1)(\beta_0 + 1)} & (j = j'). \end{cases}$$

The correlation between observations of the same sample depends only on the parameters of the underlying $\text{HDP}(\beta, \beta_0; H)$ that governs the atoms G_k^*: this is not surprising since, whatever the value of the parameter α at the NDP layer, observations from the same sample are exchangeable. Moreover, an appealing feature is that such a correlation is higher than for the case of observations from different samples, i.e. $j \neq j'$. As for the dependence on the hyperparatmeters (α, β_0, β), $\alpha \to \infty$ forces the G_j's to equal different unique distributions G_k^*, similarly to the NDP case. However, unlike the NDP, this does not imply that the distributions are independent, and the correlation is controlled by the hyperparameters β and β_0 (increasing in β and decreasing in β_0). In Fig. 3.2 we report the aforementioned correlations as functions of β and β_0 with α set equal 1. Finally, if $\alpha \to 0$ the a priori probability to degenerate to the exchangeable case, i.e. all G_j's coincide a.s., tends to 1 and so does also $\text{Cor}[G_j(A), G_{j'}(A)]$.

We now investigate the random partition structure associated to a HHDP, namely the partition of $\{1, \ldots, n\}$, with $n = \sum_{j=1}^{J} I_j$, induced by a partially exchangeable sample $\boldsymbol{X}$ modeled as in (3.1). Since a $\text{HHDP}(\alpha, \beta, \beta_0; H)$ arises from the composition of two discrete random structures, it is clear that the partition induced by $\boldsymbol{X}$ will depend on the partition, say $\Psi^{(J)}$, of the random probability measures $G_1, \ldots, G_J$. As for the latter, the G_i's

are drawn from a discrete random probability measure on $\mathsf{P}_{\mathbb{X}}$ whose weights have a GEM(α) distribution and whose atoms are almost surely different since they are sampled from an HDP($\beta, \beta_0; H$). Then the probability distribution of $\Psi^{(J)}$ is the well–known Ewens sampling formula, namely

$$\mathbb{P}[\Psi^{(J)} = \{B_1, \ldots, B_R\}] = \phi_R^{(J)}(m_1, \ldots, m_R) = \frac{\alpha^R}{\alpha_{(J)}} \prod_{r=1}^{R} (m_r - 1)!,$$

where $1 \leq R \leq J$, the frequencies $m_r = \text{card}(B_r)$ are such that $\sum_{r=1}^{R} m_r = J$ and $\alpha_{(J)} = \Gamma(\alpha + J)/\Gamma(\alpha)$. This structure *a priori* implies, as in the NDP case, that $\mathbb{P}(G_j = G_\kappa) \in (0, 1)$ for any $j \neq \kappa$. However, unlike the NDP, *a posteriori* the HHDP yields $\mathbb{P}(G_j = G_\kappa \mid \boldsymbol{X}) < 1$, regardless of the shared clusters across the samples $\boldsymbol{X}$. Moreover, let $\Phi_{D,R}^{(n)}(\cdots; \beta, \beta_0)$ denote the pEPPF of a HDP($\beta, \beta_0; H$), namely

$$\Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \beta, \beta_0) = \mathbb{E} \int_{\mathbb{X}_*^D} \prod_{d=1}^{D} \hat{G}_1^{n_{1,d}^*}(\mathrm{d}x_d) \cdots \hat{G}_R^{n_{R,d}^*}(\mathrm{d}x_R),$$

where $(\hat{G}_1, \ldots, \hat{G}_R) \sim \text{HDP}(\beta, \beta_0; H)$, $D \in \{1, \ldots, n\}$ and $\sum_{r=1}^{R} \sum_{d=1}^{D} n_{r,d}^* = n$. An explicit expression of $\Phi_{D,R}^{(n)}$ has been established in Camerlenghi et al. (2019), even beyond the DP case. Recalling the interpretation and notation of the pEPPF in (3.3), $n_{r,d}^*$ can be seen as the number of the dth overall unique value arising from the rth distribution. Now we can state the pEPPF induced by $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ in (3.1), where $\mathscr{L}$ is the law of a HHDP($\alpha, \beta, \beta_0; H$).

Theorem 3. *The random partition induced by the partially exchangeable array $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ drawn from $(G_1, \ldots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ is characterized by the following pEPPF*

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) = \sum \phi_R^{(J)}(m_1, \ldots, m_R; \alpha) \Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \beta, \beta_0), \tag{3.8}$$

where the sum runs over all partitions $\{B_1, \ldots, B_R\}$ of $\{1, \ldots, J\}$ and $n_{r,d}^* = \sum_{j \in B_r} n_{j,d}$ for each $r \in \{1, \ldots, R\}$, $d \in \{1, \ldots, D\}$.

Given the composition structure underlying the HHDP($\alpha, \beta, \beta_0; H$) unsurprisingly the pEPPF (3.8) is a mixture of pEPPF's induced by different HDPs. For ease of interpretation consider the case of $J = 2$ populations and note that the pEPPF boils down to

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{1}{\alpha + 1} \Phi_{D,1}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \frac{\alpha}{\alpha + 1} \Phi_{D,2}(\boldsymbol{n}_1, \boldsymbol{n}_2), \tag{3.9}$$

where $\Phi_{D,2}^{(n)}$ and $\Phi_{D,1}^{(n)}$ are the pEPPF and EPPF of a bivariate and univariate HDP($\beta, \beta_0; H$), respectively. Clearly (3.9) arises from mixing with respect to partitions of $\{G_1, G_2\}$ in either $R = 1$ and $R = 2$ groups, where the former corresponds to exchangeability across the two populations. Still for the case $J = 2$, a straightforward application of the pEPPF leads to the posterior probability of gathering the two probability curves, G_1 and G_2, in the same cluster thus making the two samples exchangeable, or homogeneous.

Proposition 3. *If the sample $\{\boldsymbol{X}_j : j = 1, 2\}$ is drawn from $(G_1, G_2) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ the posterior*

probability of degeneracy is

$$\mathbb{P}(G_1 = G_2 \mid \boldsymbol{X}) = \frac{\Phi^{(n)}_{D,1}(\boldsymbol{n}_1 + \boldsymbol{n}_2)}{\Phi^{(n)}_{D,1}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \alpha\, \Phi^{(n)}_{D,2}(\boldsymbol{n}_1, \boldsymbol{n}_2)}, \tag{3.10}$$

where $\Phi^{(n)}_{D,2}$ *and* $\Phi^{(n)}_{D,1}$ *are the pEPPF and EPPF induced by a bivariate and univariate* $\text{HDP}(\beta, \beta_0; H)$*, respectively.*

The pEPPF is a fundamental tool in Bayesian calculus and it plays, in the partially exchangeable framework, the same role of the EPPF in the exchangeable case. Indeed, the pEPPF governs the learning mechanism, *e.g.* the strength of the borrowing information, clustering, and, in view of Proposition 3, it allows to perform hypothesis testing for distributional homogeneity between populations. Finally, one can obtain a Pólya urn scheme for performing inference (see Section 3.4.1). To develop this sampler and further understand the model behavior, we provide a characterization of the $\text{HHDP}(\alpha, \beta, \beta_0; H)$, which is reminiscent of the popular Chinese restaurant franchise metaphor for the HDP.

3.3.2 The hidden Chinese restaurant franchise

The marginalization of the underlying random probability measures, as displayed in Theorem 3, can be characterized in terms of a *hidden Chinese restaurant franchise* (HCRF) metaphor. This scheme sheds further light on the HHDP and clarifies the sense in which it generalizes the well-known Chinese restaurant (CRP) and franchise (CRF) processes induced by the DP and the HDP, respectively. For simplicity we consider the case $J = 2$.

As with simpler sampling schemes, all restaurants of the franchise share the same menu, which has an infinite number of dishes generated by the non–atomic base measure H. However, unlike the standard CRF, the restaurants of the franchise are merged into a single one if $G_1 = G_2$, while they stay distinct if $G_1 \neq G_2$. Moreover, each $X_{j,i}$ identifies the label of the dish that customer i from the j–th population chooses from the shared menu $(X^*_d)_{d\geq 1}$, with the unique dishes $X^*_d \overset{\text{iid}}{\sim} H$. All customers are either assigned to different restaurants, if $G_1 \neq G_2$, or to the same restaurant, if $G_1 = G_2$. Then given such a grouping of the restaurants, the customers are seated according to the CRF applied either to a single restaurant or to two distinct restaurants (Teh et al., 2006; Camerlenghi et al., 2018). Furthermore, each restaurant has infinitely many tables. The first customer i who arrives in a previously unoccupied table chooses a dish that is shared by all the costumers who will join the table afterwards. It is to be noted that distinct tables within each restaurant and across restaurants may share the same dish. An additional distinctive feature, compared to the CRF, is that tables can be shared across populations when they are assigned to the same restaurant, i.e. when $G_1 = G_2$. Accordingly the allocation of each customer $X_{j,i}$ to a specific restaurant clearly depends on having either $G_1 = G_2$ or $G_1 \neq G_2$.

The sampling scheme simplifies by introducing latent variables $T_{j,i}$'s denoting the tables' labels for customer i from population j. We stress that, if $G_1 \neq G_2$, the number of shared tables across the two populations is zero, given the populations $j = 1, 2$ are assigned to different restaurants, labeled $r = 1, 2$, respectively. Conversely, if $G_1 = G_2$, one may have shared tables across populations, since they are assigned to the same restaurant $r = 1$.

Now define $q_{r,t,d}$ as the frequencies of observations sitting at table t eating the dth dish, for a table specific to restaurant r. Moreover, D_t is the dish label corresponding to table t and $\ell_{r,d}$ the frequency of tables serving dish d in restaurant r. Marginal frequencies are represented with dots, e.g. $\ell_{r,\cdot}$ is the number of tables in restaurant

r. Throughout the symbol $\boldsymbol{x}^{-i}$ identifies either a set or a frequency obtained upon removing the element i from $\boldsymbol{x}$. Finally, Δ stands for an indicator function such that $\Delta = 1$ if $G_1 = G_2$, while $\Delta = 0$ if $G_1 \neq G_2$.

The stepwise structure of the sampling procedure reflects the composition of the three layers $\mathcal{L}(G_j|Q)$, $\mathcal{L}(Q|G_0)$ and $\mathcal{L}(G_0)$ in (3.7) relying on a conditional CRF. First one samples the populations' clustering Δ and, given the allocations of the populations to the restaurants, one has a CRF. Hence, the algorithm becomes

(1) Sample the population assignments to the restaurants from $\mathbb{P}(\Delta = 1) = 1/(\alpha + 1)$.

(2) Sequentially sample the table assignments $T_{j,i}$ and corresponding dishes $D_{T_{j,i}}$ from

$$p(T_{j,i}, D_{T_{j,i}} \mid \boldsymbol{T}^{-(ji+)}, \boldsymbol{X}^{-(ji+)}, \Delta) \propto \begin{cases} T_{j,i} = t & \frac{q_{r,t,\cdot}^{-(ji+)}}{q_{r,\cdot,\cdot}^{-(ji+)}+\beta} \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji+)}+\beta} \frac{\ell_{\cdot,d}^{-(ji+)}}{\ell_{\cdot,\cdot}^{-(ji+)}+\beta_0} \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d^{\text{new}} & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji+)}+\beta} \frac{\beta_0}{\ell_{\cdot,\cdot}^{-(ji+)}+\beta_0}, \end{cases}$$

where $(ji+) = \{(ji') : i' \geq i\} \cup \{(j'i') : j' \geq j\}$ is the index set associated to the future random variables not yet sampled.

3.4 Posterior Inference for HHDP mixture models

Thanks to the results of Section 3.3, we now devise MCMC algorithms for drawing posterior inferences with mixture models driven by a HHDP. Though the samplers are tailored to mixture models, they are easily adapted to other inferential problems such as e.g. survival analysis and species sampling. Henceforth, $\mathcal{K}$ is a density kernel and we consider

$$\begin{aligned} X_{j,i} \mid \theta_{j,i} &\overset{\text{ind}}{\sim} \mathcal{K}(\cdot|\theta_{j,i}), && (i = 1, \ldots, I_j \quad j = 1, \ldots, J), \\ \theta_{j,i} \mid G_j &\overset{\text{ind}}{\sim} G_j, && (i = 1, \ldots, I_j, \quad j = 1, \ldots, J), \\ (G_1, \ldots, G_J) &\sim \text{HHDP}(\alpha, \beta, \beta_0; H). \end{aligned} \tag{3.11}$$

We develop two samplers: (i) a marginal algorithm that relies on the posterior degeneracy probability (Proposition 3) in Section 3.4.1; (ii) a conditional blocked Gibbs sampler, in the same spirit of the sampler proposed for the NDP by Rodríguez et al. (2008), in Section 3.4.2. As for (i), the underlying random probability measures G_0^* and G_k^*'s are integrated out leading to urn schemes that extend the class of Blackwell-MacQueen Pólya urn processes. In such a way we generalize the *a posteriori* sampling scheme of the Chinese restaurant process for the DP mixture Neal (2000) and the one of the Chinese restaurant franchise for the HDP mixture (Teh et al., 2006). We present the marginal sampler for the case of $J = 2$ populations. Even if in principle it can be generalized in a straightforward way, it is computationally intractable for a larger number of populations. Similarly to the hidden Chinese restaurant franchise situation, one has to evaluate the posterior probability of all possible groupings of $G_1, \ldots, G_J$, which boils down to $\mathbb{P}(G_1 = G_2|\boldsymbol{X})$ when $J = 2$ but becomes involved for $J > 2$.

This shortcoming is overcome by the conditional algorithm we derive in Section 3.4.2, which relies on finite–dimensional approximations of the trajectories of the underlying random probability measure. Its effectiveness

in handling $J > 2$ populations is further illustrated in the application of Section 3.5.1.

3.4.1 A marginal Gibbs sampler

The marginal Gibbs sampler that updates Δ, the table dish assignments $T_{j,i}$ and D_t can be deduced from the hidden Chinese restaurant franchise presented in Section 3.3.2. Let $\boldsymbol{S} = \{\Delta, (T_{j,i})_{j,i}, (D_t)_t, (X_{j,i})_{j,i}\}$. Hence, the algorithm can be summarized as follows

(1) Sample the population assignments to the restaurants

$$\mathbb{P}(\Delta = 1 \mid \boldsymbol{X}) = \frac{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2)}{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \alpha\, \Phi_{D,2}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2)},$$

where $\Phi_{D,2}^{(n)}$, $\Phi_{D,1}^{(n)}$ are the pEPPF and EPPF of a bivariate and univariate $\text{HDP}(\beta, \beta_0; H)$, respectively.

(2) Sample the table assignments $T_{j,i}$ and corresponding dishes $D_{T_{j,i}}$ from $p(T_{j,i}, D_{T_{j,i}} \mid \boldsymbol{S}^{-(T_{j,i}, D_{T_{j,i}})}) \propto$

$$\propto \begin{cases} T_{j,i} = t & \frac{q_{r,t,\cdot}^{-(ji)}}{q_{r,\cdot,\cdot}^{-(ji)} + \beta}\ p_{D_t}(\{X_{j,i}\}) \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji)} + \beta} \frac{\ell_{\cdot,d}^{-(ji)}}{\ell_{\cdot,\cdot}^{-(ji)} + \beta_0}\ p_d(\{X_{j,i}\}) \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d^{\text{new}} & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji)} + \beta} \frac{\beta_0}{\ell_{\cdot,\cdot}^{-(ji)} + \beta_0}\ p_{d^{\text{new}}}(\{X_{j,i}\}), \end{cases}$$

where $p_d(\{X_{j,i}\})$ is defined by the following equation. For every index set $\mathcal{J}$

$$p_d(\{X_{j,i}\}_{(j,i)\in\mathcal{J}}) = \frac{\int \prod_{j'i'\in\mathcal{J}\cup\mathcal{J}_d} \mathcal{K}(X_{j,i} \mid \theta) \mathrm{d}H(\theta)}{\int \prod_{j'i'\in\mathcal{J}_d\setminus\mathcal{J}} \mathcal{K}(X_{j,i} \mid \theta) \mathrm{d}H(\theta)},$$

where $\mathcal{J}_d = \{(j,i) : D_{T_{j,i}} = d\}$. For instance, $p_d(\{X_{j,i}\})$ is the marginal conditional probability of $X_{j,i}$ in cluster d given the other observation assigned to cluster d.

(3) Sample the dish assignments D_t from

$$p(D_t \mid \boldsymbol{S}^{-t}) \propto \begin{cases} d & \frac{\ell_{\cdot,d}^{-t}}{\ell_{\cdot,\cdot}^{-t} + \beta_0} p_d(\{x_{j,i} : T_{j,i} = t\}) \\ d^{\text{new}} & \frac{\beta_0}{\ell_{\cdot,\cdot}^{-t} + \beta_0} p_{d^{\text{new}}}(\{x_{j,i} : T_{j,i} = t\}). \end{cases}$$

3.4.2 A conditional blocked Gibbs sampler

A more effective algorithm is based on simple blocked conditional procedure. To this end we use a finite approximation of the DP in the spirit of Muliere and Tardella (1998) and Ishwaran and James (2001). However, instead of truncating the stick–breaking representation of the DP, we use a finite Dirichlet approximation. See Ishwaran and Zarepour (2002). Therefore, we approximate $\boldsymbol{\pi}^*, \boldsymbol{\omega}_0^*$, with a K– and an L–dimensional Dirichlet distribution, respectively. More precisely, we consider the following approximation

$$\boldsymbol{\pi}^* \sim \text{DIR}(\alpha/K, \ldots, \alpha/K), \qquad \boldsymbol{\omega}_0^* \sim \text{DIR}\big(\beta_0/L, \ldots, \beta_0/L\big) \tag{3.12}$$

implying that $(\boldsymbol{\omega}_k^* \mid \boldsymbol{\omega}_0^*) \overset{\text{iid}}{\sim} \text{DIR}(\beta\,\boldsymbol{\omega}_0^*)$, for $k \geq 1$.

Introduce the auxiliary variables z_j and $\zeta_{j,i}$ which represent the distributional and observational cluster memberships, respectively, such that $z_j = k$ and $\zeta_{j,i} = l$ if and only if $G_j = G_k^*$ and $\theta_{j,i} = \theta_l^*$. Henceforth, $\boldsymbol{S} = \{(\theta_l^*)_{l=1}^L, \boldsymbol{\pi}^*, \boldsymbol{\omega}_0^*, (\boldsymbol{\omega}_k^*)_{k=1}^K, (z_j)_{j=1}^J, (\zeta_{j,i})_{j,i}, (X_{j,i})_{j,i}\}$ and, in order to identify the full conditionals of the Gibbs sampler, we note that under the finite Dirichlet approximation (3.12)

$$p(\boldsymbol{S}) = p(\boldsymbol{\pi}^*)p(\boldsymbol{\omega}_0^*)\left[\prod_{l=1}^{L} p(\theta_l^*)\right]\left[\prod_{k=1}^{K} p(\boldsymbol{\omega}_k^* \mid \boldsymbol{\omega}_0^*)\right] \times \left\{\prod_{j=1}^{J} p(z_j \mid \boldsymbol{\pi}^*)\left[\prod_{i=1}^{I_j} p(X_{j,i} \mid \theta_{\zeta_{j,i}}^*)p(\zeta_{j,i} \mid \boldsymbol{\omega}_{z_j}^*)\right]\right\}.$$

This leads to the following

(1) Sample the unique θ_l^* from

$$p(\theta_l^* \mid \boldsymbol{S}^{-\theta_l^*}) \propto H(\theta_l^*) \prod_{\{j,i:\zeta_{j,i}=l\}} \mathcal{K}(X_{j,i} \mid \theta_l^*).$$

(2) Sample distributional cluster probabilities from

$$p(\boldsymbol{\pi}^* \mid \boldsymbol{S}^{-\boldsymbol{\pi}^*}) = \text{DIR}(\boldsymbol{\pi}^* \mid \alpha/K + m_1, \ldots, \alpha/K + m_K),$$

with $m_k = \sum_{j=1}^J \mathbb{1}\{z_j = k\}$.

(3) Sample probability weights of the base DP from

$$p(\boldsymbol{\omega}_0^* \mid \boldsymbol{S}^{-\boldsymbol{\omega}_0^*}) \propto \prod_{l=1}^{L}\left[\frac{(\omega_{0,l}^*)^{\beta_0/L-1}\xi_l^{\beta\omega_{0,l}^*}}{\Gamma(\beta_0\omega_{0,l}^*)^K}\right], \tag{3.13}$$

with $\xi_l = \prod_{k=1}^K \omega_{k,l}^*$.

(4) Sample the observational cluster probabilities independently fromfrom

$$p(\boldsymbol{\omega}_k^* \mid \boldsymbol{S}^{-\boldsymbol{\omega}_k^*}) = \text{DIR}(\boldsymbol{\omega}_k^* \mid \beta\boldsymbol{\omega}_0^* + \boldsymbol{n}_k),$$

with $n_{k,l} = \sum_{\{j:z_j=k\}} \sum_{i=1}^{I_j} \mathbb{1}\{\zeta_{j,i} = l\}$.

(5) Sample distributional and observational cluster membership from

$$p(z_j = k \mid \boldsymbol{S}^{-\{z_j,\boldsymbol{\zeta}_j\}}) \propto \pi_k^* \prod_{i=1}^{I_j}\sum_{l=1}^{L} \omega_{k,l}^* \mathcal{K}(X_{j,i} \mid \theta_l^*) \qquad (k = 1, \ldots, K),$$

$$p(\zeta_{j,i} = l \mid \boldsymbol{S}^{-\zeta_{j,i}}) \propto \omega_{z_j l}^* \mathcal{K}(X_{j,i} \mid \theta_l^*) \qquad (l = 1, \ldots, L).$$

Importantly, all the full conditional distributions are available in simple closed forms, with the exception of the distributions of $\boldsymbol{\omega}_0^*$ and, possibly, of θ_l^*. To update $\boldsymbol{\omega}_0^*$ we perform a Metropolis-Hastings step, where we

work on the unconstrained space $\mathbb{R}^{L-1}$ after the transformation $[\log(\omega_{0,1}/\omega_{0,L}), \dots, \log(\omega_{0,L-1}/\omega_{0,L})]$ and we adopt a component–wise adaptive random walk proposal following Roberts and Rosenthal (2009). The update of the unique atoms θ_l^* is standard, as with the DP mixture model in the exchangeable case.

In Section 3.5 we assume a Gaussian kernel $\mathcal{K}(\cdot|\theta) = \mathrm{N}(\cdot|\mu, \sigma^2)$ and a conjugate Normal-inverse-Gamma base measure $H(\cdot) = \textsc{nig}(\cdot \mid \mu_0, \lambda_0, s_0, S_0)$ and obtain

$$p(\theta_l^* \mid \boldsymbol{S}^{-\theta_l^*}) = \textsc{nig}(\theta_l^* \mid \mu_l, \lambda_l, s_l, S_l),$$

with $\mu_l = \dfrac{n_l \bar{y}_l + \lambda_0 \mu_0}{\lambda_0 + n_l}$, $S_l = S_0 + \dfrac{1}{2}\left(e_l^2 + \dfrac{n_l \lambda_0 (\bar{y}_l - \mu_0)^2}{\lambda_0 + n_l}\right)$, $\lambda_l = \lambda_0 + n_l$, and $s_l = n_l/2 + s_0$, where $n_l = \sum_{j=1}^{J} \sum_{i=1}^{I_j} \mathbb{1}\{\zeta_{j,i} = l\}$, $\bar{y}_l = \sum_{\{j,i:\zeta_{j,i}=l\}} X_{j,i}/n_l$, and $e_l^2 = \sum_{\{j,i:\zeta_{j,i}=l\}} (X_{j,i} - \bar{y}_l)^2$ are the observational cluster sizes, means and deviances, respectively.

3.5 Illustration

We compare the performance of our proposal (3.11) with the same model where the HHDP is replaced by a NDP as in (3.5), on synthetic data. In doing so we rely on blocked Gibbs sampler of Section 3.4. The data are simulated from the same scenarios considered in Camerlenghi et al. (2019). More precisely, we consider two populations and the data in each population are iid from a mixture of two normals:

Scen 1. We simulate the data from the two populations independently from the same density

$$X_{1,i} \overset{\mathrm{d}}{=} X_{2,i'} \sim 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(0,1).$$

Scen 2. We simulate the data in the two populations independently from a mixture of two normals with one shared component

$$X_{1,i} \sim 0.9\mathrm{N}(5, 0.6) + 0.1\mathrm{N}(10, 0.6) \qquad X_{2,i'} \sim 0.1\mathrm{N}(5, 0.6) + 0.9\mathrm{N}(0, 0.6).$$

Scen 3. We simulate the data in the two populations independently from a mixture of two normals having the same components with different weights

$$X_{1,i} \sim 0.8\mathrm{N}(5,1) + 0.2\mathrm{N}(0,1) \qquad X_{2,i'} \sim 0.2\mathrm{N}(5,1) + 0.8\mathrm{N}(0,1).$$

In all these scenarios we consider balanced sample sizes $I_1 = I_2 = 100$ and an HHDP mixture model (3.11), with $\alpha = 1$, $\beta = 1$, $\beta_0 = 1$ and $H(\cdot) = \textsc{nig}(\cdot \mid \mu_0, \lambda_0, s_0, S_0)$. We set standard values of the hyperparameters in terms of the mean $\bar{y}$ and variance $\mathrm{Var}(y)$ of the data, i.e. $\mu_0 = \bar{y}$, $\lambda_0 = 1/(3\,\mathrm{Var}(y))$, $s_0 = 1$ and $S_0 = 4$. In drawing the comparison between (3.11) and the $\textsc{ndp}(\alpha, \beta; H)$, we further set $\alpha = \beta = 1$. Furthermore, we set the concentration parameters all equal to 1. In Section 3.6 we perform a sensitivity analysis with respect to hyperparameters' specifications as done, for instance, by Zuanetti et al. (2018) for the NDP. The mean measure of the marginal underlying random distributions $\mathbb{E}[G_j(A)] = H(A)$ is the same for all populations. Also variances are comparable (see Proposition 1) since $\mathrm{Var}[G_j(A)]$ equals $H(A)[1 - H(A)]/2$ for the NDP and $3H(A)[1 - H(A)]/4$ for the HHDP. The sensitivity analysis leads, for all the considered settings, to the

same conclusions in terms of comparison of the two models. Moreover, we fix the dimensions of the finite approximations $L = K = 50$ in (3.12) and we do the same for the truncation levels in the algorithm of Rodríguez et al. (2008). In the Appendix we perform an empirical analysis trying different levels of L and K which corroborates the fact that the approximation error is negligible in term of inferential results.

Inference is based on 10 000 iterations with the first half discarded as burn-in. As for the output, besides obtaining density estimates for the two populations we also determine the point estimate of the clustering of observations that minimizes the variation of information (VI) loss function. See Meilă (2007) and Wade and Ghahramani (2018) for detailed discussions on VI and point summaries of probabilistic clustering. Additionally, we estimate the probability that observations co-cluster, namely $\mathbb{P}(\zeta_{j,i} = \zeta_{j',i'} \mid \boldsymbol{X})$ through the average over MCMC draws

$$\frac{\sum_{b=1}^{B} \mathbb{1}\{\zeta_{j,i}^b = \zeta_{j',i'}^b\}}{B},$$

where B is the number of the MCMC iterations. These are visualized through heat maps in Fig. 3.5, with colors ranging from white, if the probability is 0, to dark, if the probability is 1. Our analysis is completed by reporting the estimated distributions of the numbers of mixture components in each scenario.

As expected, both models yield accurate estimates of the true densities in all scenarios. In Fig. 3.3 we report the true and estimated models under the third scenario. In terms of clustering, in the first scenario both models correctly cluster together the two populations, thus degenerating to the exchangeable case as they should. However, in the second and third scenarios the NDP makes the two samples $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ independent, therefore preventing borrowing of information across the two populations. As the distributions have a shared component, the only way for the NDP to recover correctly the true densities is by missing such a component. Had it been detected, the density estimates of the two populations would have been equal and, thus, far from the truth. The point estimate of the observations' clustering in Table 3.2, the heat maps of the posterior co-clustering probabilities in Fig. 3.4 and the posterior distributions of the overall number of components in Table 3.1 showcase the theoretical findings, namely that the NDP in the second and third scenarios cannot learn the shared components. Hence, it overestimates the total number of components and does not cluster observations across populations. In contrast, the HHDP model is able to cluster observations across populations, learns the shared components and borrows information also when the model does not degenerate to the exchangeable case.

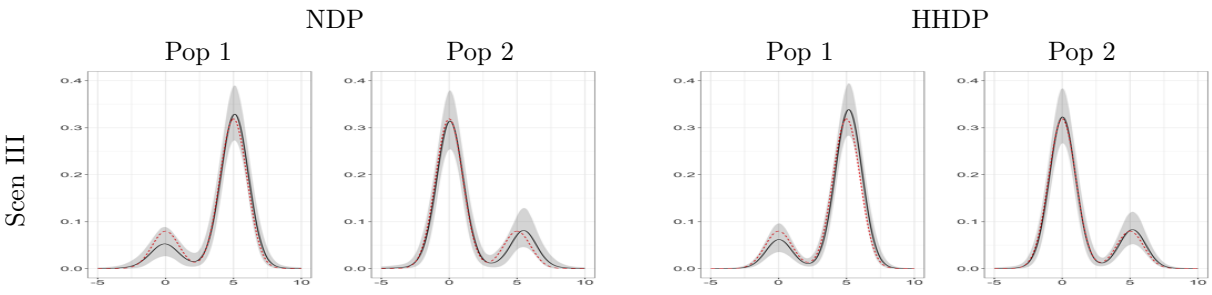

Figure 3.3: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under the third scenario.

Scen	Model	Overall number of components									
		1	2	3	4	5	6	7	8	9	≥10
I	NDP	0	0.4090	0.3615	0.1647	0.0492	0.0136	0.0020	0	0	0
	HHDP	0	0.5374	0.3743	0.0788	0.0080	0.0016	0	0	0	0
II	NDP	0	0	0	0.2959	0.3906	0.2151	0.0700	0.0256	0.0024	0.0004
	HHDP	0	0	0.5742	0.3339	0.0796	0.0116	0.0008	0	0	0
III	NDP	0	0	0	0.1331	0.3055	0.2947	0.1743	0.0608	0.0232	0.0084
	HHDP	0	0.5010	0.3966	0.0856	0.0164	0.0004	0	0	0	0

Table 3.1: Posterior distributions of the number of overall components estimated with the two models under different scenarios.

	Scenario I				Scenario II							Scenario III					
	NDP		HHDP		NDP				HHDP			NDP				HHDP	
Population	1	2	1	2	1	2	3	4	1	2	3	1	2	3	4	1	2
1	56	44	56	44	87	13	0	0	87	13	0	85	15	0	0	85	15
2	48	52	48	52	0	0	88	12	12	0	88	0	0	80	20	21	79

Table 3.2: Frequencies of observations in the two populations allocated to the point estimate of the clustering that minimizes the VI loss with the two models under different scenarios.

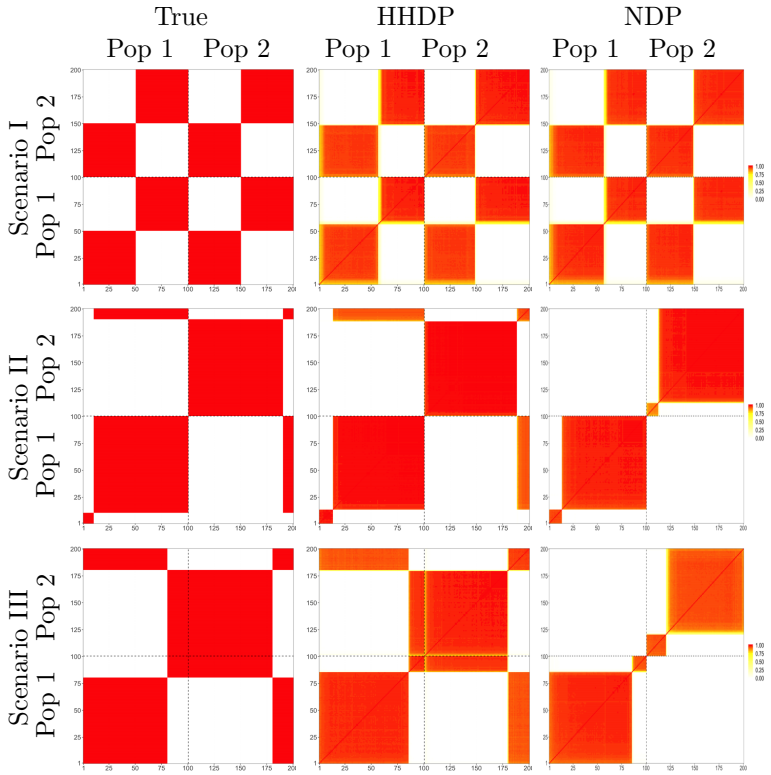

Figure 3.4: Heat maps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, with the two models under different scenarios.

3.5.1 Collaborative perinatal project data

A multi-center application is the focus of this section. We consider a data set from the Collaborative Perinatal Project (CPP), a large prospective epidemiologic study conducted from 1959 to 1974. Pregnant women were enrolled in 12 hospitals between 1959 and 1966 and were followed over time. Among several pre–pregnancy measurements, we focus on the birth weight $X_{j,i}$ for non-smoking woman i in center j. We assume the following

Gaussian mixture model:

$$X_{j,i} \mid \mu_{j,i}, \sigma_{j,i} \overset{\text{ind}}{\sim} \mathrm{N}(\mu_{j,i}, \sigma_{j,i}) \qquad (i = 1, \ldots, I_j, \quad j = 1; \ldots, 12),$$
$$\mu_{j,i}, \sigma_{j,i} \mid G_j \overset{\text{ind}}{\sim} G_j \qquad (i = 1, \ldots, I_j, \quad j = 1; \ldots, 12).$$

The same HHDP prior used for the previous synthetic data is placed the vector of random distributions. This model specification is coherent with what is suggested by Dunson (2010) for the CPP data. Indeed, it is known that the pregnancy outcome can vary substantially for women from different ethnicity and socioeconomic groups. Therefore, we specify a model allowing to capture differences between the centers since different groups of hospitals can serve different women. Canale et al. (2019) provide a further analysis of the CPP data.

The heat map of the co-clustering posterior probability for the 12 hospitals is shown in Fig. 3.5. Such probabilities imply that the clustering point estimate of the hospitals that minimizes the VI loss has two blocks and, in the same figure, the mean posterior densities associated to the two clusters are reported. Note that these mean densities are similar in the two clusters of populations. Coherently the proposed model allows to borrow information across clusters of hospitals for estimating the posterior mean densities of the birth weight.

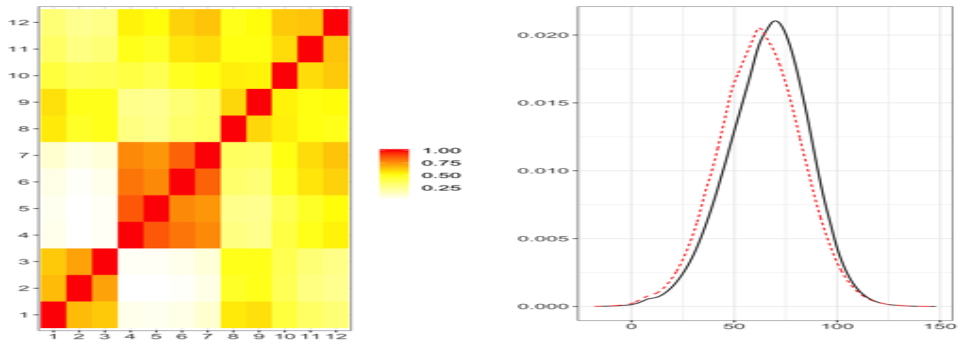

Figure 3.5: Heat map of the estimated posterior probability of co-clustering of hospitals and estimated population cluster specific posterior densities for the CPP data.

3.6 Discussion

As highlighted in the recent literature, NDP mixture models are not an appropriate tool for clustering simultaneously population distributions and observations. Our new proposal, the HHDP, overcomes the issues plaguing the NDP, while preserving tractability and clustering flexibility. We also derive sampling schemes allowing efficient inference when the number of populations increases. This work lays the foundation for future research. First, it is natural to move beyond DPs combining other discrete nonparametric priors, as the Pitman-Yor process and normalized completely random measures, and studying the induced clustering. Moreover, it is of interest to investigate and tailor the HHDP to perform inference on survival and functional data.

Appendix B

B.1. Proof of Proposition 1

Note that $G_1^*(A) \mid G_0^* \sim \text{BETA}(\beta G_0^*(A), \beta(1 - G_0^*(A)))$ and $G_0^*(A) \sim \text{BETA}(\beta_0 H(A), \beta_0(1 - H(A)))$. Hence,

$$\mathbb{E}G_0^*(A) = H(A), \quad \text{Var}[G_0^*(A)] = \frac{H(A)[1 - H(A)]}{\beta_0 + 1}$$

and since $G_j \overset{\text{d}}{=} G_1^*$,

$$\begin{aligned}
\mathbb{E}G_j(A) &= \mathbb{E}\mathbb{E}[G_1^*(A) \mid G_0^*] = \mathbb{E}G_0^*(A) = H(A) \\
\text{Var}[G_j(A)] &= \mathbb{E}\text{Var}[G_1^*(A) \mid G_0^*] + \text{Var}[G_0^*(A)] = \frac{H(A)[1 - H(A)](\beta_0 + \beta + 1)}{(\beta + 1)(\beta_0 + 1)}.
\end{aligned}$$

Mixed moments are also easy to determine, as $\mathbb{E}G_1^*(A)G_2^*(A) = \mathbb{E}\mathbb{E}[G_1^*(A) \mid G_0^*]\mathbb{E}[G_2^*(A) \mid G_0^*] = \mathbb{E}G_0^*(A)^2$ and

$$\begin{aligned}
\mathbb{E}G_j(A)G_{j'}(A) &= \mathbb{E}[G_1(A)G_2(A) \mid G_1 = G_2]\,\mathbb{P}(G_1 = G_2) + \mathbb{E}[G_1(A)G_2(A) \mid G_1 \neq G_2]\,\mathbb{P}(G_1 \neq G_2) \\
&= \frac{1}{1+\alpha}\mathbb{E}[G_1^*(A)^2] + \frac{\alpha}{\alpha+1}\mathbb{E}[G_1^*(A)G_2^*(A)] \\
&= \frac{1}{1+\alpha}\mathbb{E}[G_1^*(A)^2] + \frac{\alpha}{\alpha+1}\mathbb{E}[G_0^*(A)^2].
\end{aligned}$$

One, then, obtains

$$\text{Cov}[G_j(A), G_{j'}(A)] = \mathbb{E}[G_j(A)G_{j'}(A)] - H(A)^2 = \frac{1}{1+\alpha}\text{Var}[G_1^*(A)] + \frac{\alpha}{\alpha+1}\text{Var}[G_0^*(A)]$$

and

$$\text{Cor}[G_j(A), G_{j'}(A)] = \frac{1}{1+\alpha} + \frac{\alpha}{\alpha+1}\frac{\text{Var}[G_0^*(A)]}{\text{Var}[G_1^*(A)]} = \frac{\beta_0 + \beta + 1 + \alpha\beta + \alpha}{(1+\alpha)(\beta_0 + \beta + 1)}$$

so that the conclusion follows. □

B.2. Proof of Proposition 2

Note that $X_{j,i} \overset{\text{d}}{=} X_d^*$. Thus,

$$\text{Cov}(X_{j,i}, X_{j',i'}) = \mathbb{P}(X_{j,i} = X_{j',i'})\,\text{Var}(X_d^*)$$

Moreover, if $j = j'$, then

$$\begin{aligned}
\mathbb{P}(X_{j,i'} = X_{j,i}) &= \mathbb{P}(X_{j,i'} = X_{j,i} \mid T_{j,i} = T_{j,i'})\,\mathbb{P}(T_{j,i} = T_{j,i'}) + \mathbb{P}(X_{j,i'} = X_{j,i} \mid T_{j,i} \neq T_{j,i'})\,\mathbb{P}(T_{j,i} \neq T_{j,i'}) \\
&= \frac{1}{\beta+1} + \mathbb{P}(D_{T_{j,i'}} = D_{T_{j,i}} \mid T_{j,i} \neq T_{j,i'})\,\frac{\beta}{\beta+1} = \frac{\beta + \beta_0 + 1}{(\beta+1)(\beta_0+1)}
\end{aligned}$$

If $j \neq j'$, then

$$\begin{aligned}\mathbb{P}(X_{j,i} = X_{j',i'}) &= \mathbb{P}(X_{j,i} = X_{j',i'} \mid G_j = G_{j'})\,\mathbb{P}(G_j = G_{j'}) + \mathbb{P}(X_{j,i} = X_{j',i'} \mid G_j \neq G_{j'})\,\mathbb{P}(G_j \neq G_{j'}) \\ &= \mathbb{P}(X_{j,i'} = X_{j,i})\mathbb{P}(G_j = G_{j'}) + \mathbb{P}(D_{T_{j',i'}} = D_{T_{j,i}} \mid T_{j,i} \neq T_{j',i'})\,\mathbb{P}(G_j \neq G_{j'}) \\ &= \frac{1}{\beta_0 + 1} + \frac{\beta_0}{(1+\alpha)(1+\beta)(1+\beta_0)}\end{aligned}$$

and the conclusion follows. □

B.3. Proof of Theorem 3

In order to prove Theorem 3 we first state the following Lemma.

Lemma 2. *The random partition induced by the samples* $\{\boldsymbol{X}_j : j = 1, \dots, J\}$ *drawn from* $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ *given a particular partition of distributions* $\Psi^{(J)} = \{B_1, \dots, B_R\}$ *is characterized by the pEPPF*

$$\Pi_D^{(n)}\left(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \dots, B_R\}\right) = \Phi_{D,R}^{(n)}\left(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0\right),$$

where $n_{r,d}^* = \sum_{j \in B_r} n_{j,d}$ *for each* $r = 1, \dots, R$, $d = 1, \dots, D$ *and* $\Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0)$ *is the pEPPF associated to a R-dimensional* $\text{HDP}(\beta, \beta_0; H)$.

Now we can write

$$\begin{aligned}&\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \dots, B_R\}) = \\ &= \mathbb{E}\left[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} G_1^{n_{1,d}}(\mathrm{d}x_d) \dots G_J^{n_{J,d}}(\mathrm{d}x_d) \;\middle|\; \Psi^{(J)} = \{B_1, \dots, B_R\}\right] = \\ &= \mathbb{E}\left[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} G_1^{*n_{1,d}^*}(\mathrm{d}x_d) \dots G_R^{*n_{R,d}^*}(\mathrm{d}x_d)\right] = \Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0),\end{aligned} \tag{3.14}$$

with $\mathbb{X}_*^D = \mathbb{X}^D \setminus \{\boldsymbol{x} : x_i = x_j \text{ for some } i \neq j\}$ and $(G_1^*, \dots, G_R^*) \sim \text{HDP}(\beta, \beta_0; H)$. Moreover, note that the R unique values among $(G_1, \dots, G_J)$ are not necessarily the first $(G_1^*, \dots, G_R^*)$ but since $(G_k^*)_{k \geq 1}$ are exchangeable the third equality holds.

Therefore, by applying Lemma 3

$$\begin{aligned}\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J) &= \sum p\big(\Psi^{(J)} = \{B_1, \dots, B_R\}\big)\Pi_D^{(n)}\big(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \dots, B_R\}\big) = \\ &= \sum \phi_R^{(J)}\big(m_1, \dots, m_R; \alpha,\big)\Phi_{D,R}^{(n)}\big(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0\big)\end{aligned} \tag{3.15}$$

B.4. Proof of Proposition 3

In order to derive the posterior probability of degeneracy we write the marginal likelihood as

$$p(\boldsymbol{X}) = \Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) \prod_{d=1}^{D} H(\mathrm{d}X_d^*),$$

where $\{X_1^*, \ldots, X_D^*\}$ are the D unique values among $\boldsymbol{X}$ and $\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)$ is the pEPPF associated to the proposed model (3.8), that is

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \mathbb{P}(G_1 = G_2)\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \mathbb{P}(G_1 \neq G_2)\Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2),$$

Finally, we prove the proposition by applying Bayes theorem

$$\mathbb{P}(G_1 = G_2 \mid \boldsymbol{X}) = \frac{\mathbb{P}(G_1 = G_2)p(\boldsymbol{X} \mid G_1 = G_2)}{p(\boldsymbol{X})} = \frac{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2)}{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \alpha\Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)},$$

where $\Phi_{D,1}^{(n)}$ and $\Phi_{D,2}^{(n)}$ are the pEPPF and the EPPF of a bivariate and univariate $\text{HDP}(\beta, \beta_0; H)$, respectively.

More precisely, following Camerlenghi et al. (2019, 2018) we can derive the pEPPF $\Phi_{D,2}^{(n)}$ and the EPPF $\Phi_{D,1}^{(n)}$ of a bivariate and univariate $\text{HDP}(\beta, \beta_0; H)$, respectively. In particular

$$\Phi_{D,1}^{(n)}(\boldsymbol{n}^*) = \frac{\beta_0^D}{(\beta)_n} \sum_{\boldsymbol{\ell}^*} \frac{\beta^{|\boldsymbol{\ell}^*|}}{(\beta_0)_{|\boldsymbol{\ell}^*|}} \prod_{d=1}^{D} (\ell_d^* - 1)! |\jmath(n_d^*, \ell_d^*)|, \tag{3.16}$$

where $|\jmath(n, \ell)|$ is the signless Stirling numbers of the first kind and the sum runs over all vectors $\boldsymbol{\ell}^* = (\ell_1^*, \ldots, \ell_D^*)$ such that $\ell_d^* \in \{1, \ldots, n_d^*\}$, $|\ell^*| = \sum_{d=1}^D \ell_d^*$ and

$$\Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{\beta_0^D}{\prod_{j=1}^J (\beta)_{I_j}} \sum_{\boldsymbol{\ell}} \frac{\beta^{|\boldsymbol{\ell}|}}{(\beta_0)_{|\boldsymbol{\ell}|}} \prod_{d=1}^{D} (\ell_{\cdot l} - 1)! \prod_{j=1}^{2} |\jmath(n_{j,d}, \ell_{j,d})|, \tag{3.17}$$

where $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$, with each $\boldsymbol{\ell}_j = (\ell_{j,1}, \ldots, \ell_{j,D}) \in \times_{d=1}^D \{1, \ldots, n_{j,d}\}$ and $|\ell| = \sum_{j=1}^2 \sum_{d=1}^D \ell_{j,d}$.

B.5. Sensitivity analysis for the hyperparameters specification

Here we study the robustness with respect to the specification of hyperparameters in relation to the comparison between the NDP and the HHDP mixture models presented in Section 3.5. The results are reported in terms of density estimates in Fig. 3.6 and probabilities of co-clustering of the observations in Fig. 3.7 using the finite–dimensional approximations of the DPs with $L = K = 50$ and different hyperparameter specifications. The sensitivity analysis is performed by selecting different values for the concentration parameters. This allows to verify the robustness of the results comparing the two models. We report the results for the data simulated according to scenario III, in which the two populations share both the Gaussian components, but with different mixture weights.

We perform inference with the model as in Section 3.5 with the following specifications for the concentration parameters:

- **Parameters 1:** all the concentration parameters are set equal to 1, that is $(G_1, G_2) \sim \text{NDP}(\alpha = 1, \beta = 1; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 1, \beta = 1, \beta_0 = 1; H)$, respectively.
- **Parameters 0.1:** all the concentration parameters are set equal to 0.1, that is $(G_1, G_2) \sim \text{NDP}(\alpha = 0.1, \beta = 0.1; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 0.1, \beta = 0.1, \beta_0 = 0.1; H)$, respectively.
- **Parameters 3:** all the concentration parameters are set equal to 3, that is

$(G_1, G_2) \sim \text{NDP}(\alpha = 3, \beta = 3; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 3, \beta = 3, \beta_0 = 3; H)$, respectively.

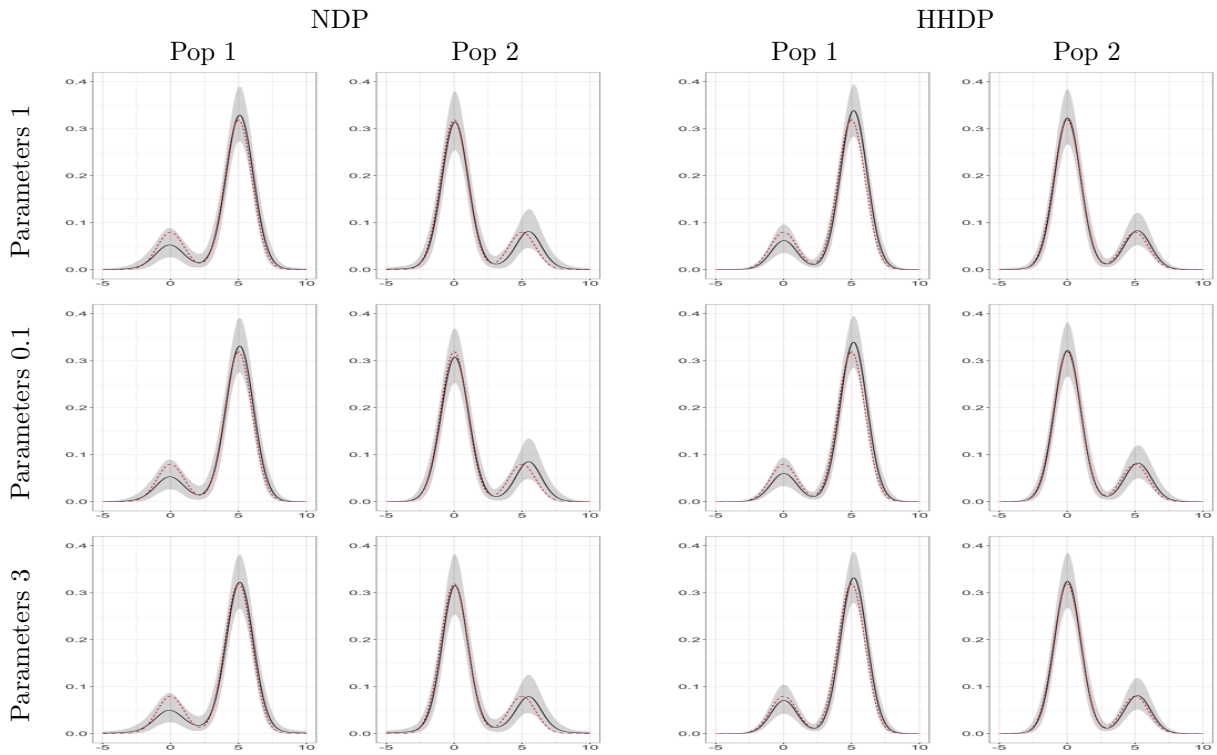

Figure 3.6: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under different hyperparameters specifications.

Importantly the density estimates are essentially the same under the different hyperparameters specifications. Probabilities of co-clustering change under the different hyperparameter settings coherently with the theory developed in Section 3.3.1. However, in all the scenarios both models do not degenerate to the exchangeable case. This implies that the NDP cannot cluster observations across populations, while the HHDP overcomes this issue. Therefore, the results of the comparison between the two models presented in Section 3.5 are essentially the same.

B.6. Choice of the finite dimensional approximations

We now present the inferential results in terms of density estimates in Fig. 3.6 and probability of co-clustering of the observations in Fig. 3.7 for the two specifications in Section 3.5. We report the results for the data simulated according to scenario III with the following finite dimensional approximations of the DPs:

- $L = K = 50$;
- $L = K = 30$;
- $L = K = 70$.

Under all the different finite dimensional approximations the inference is qualitatively the same, corroborating the idea that the finite dimensional approximations $L = K = 50$ proposed for the comparison of the NDP and HHDP in Section 3.5 induce a negligible error in our analysis.

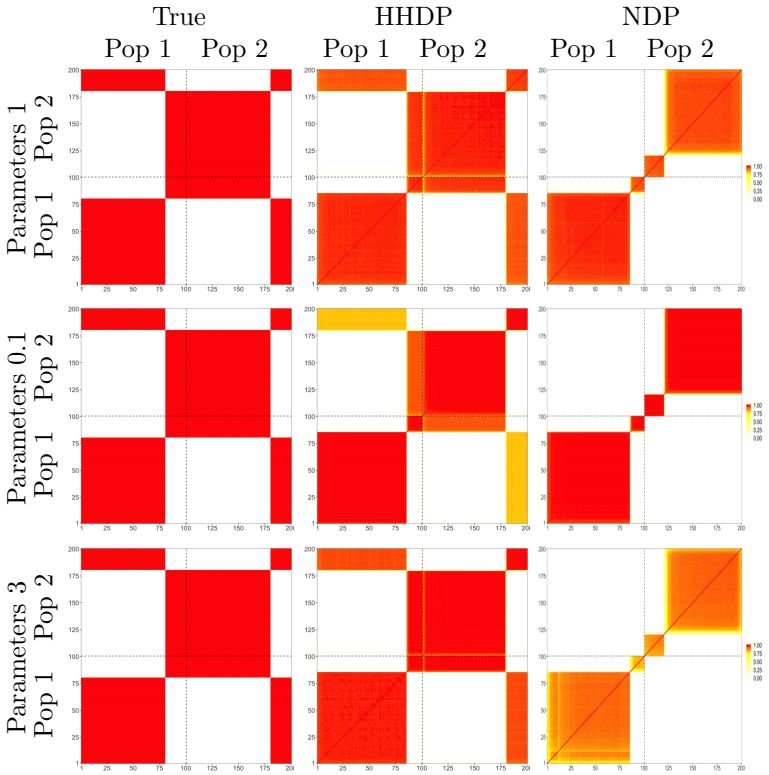

Figure 3.7: Heat maps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, with the two models under different hyperparameters specifications.

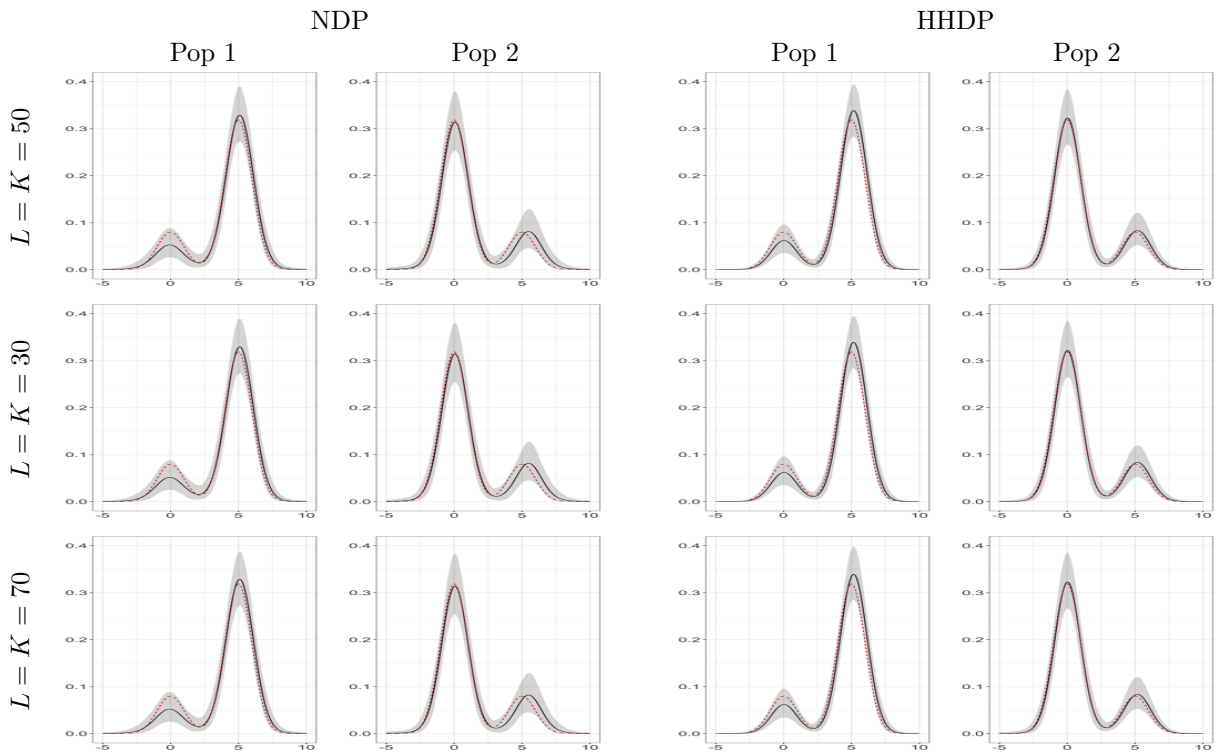

Figure 3.8: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under different truncation levels.

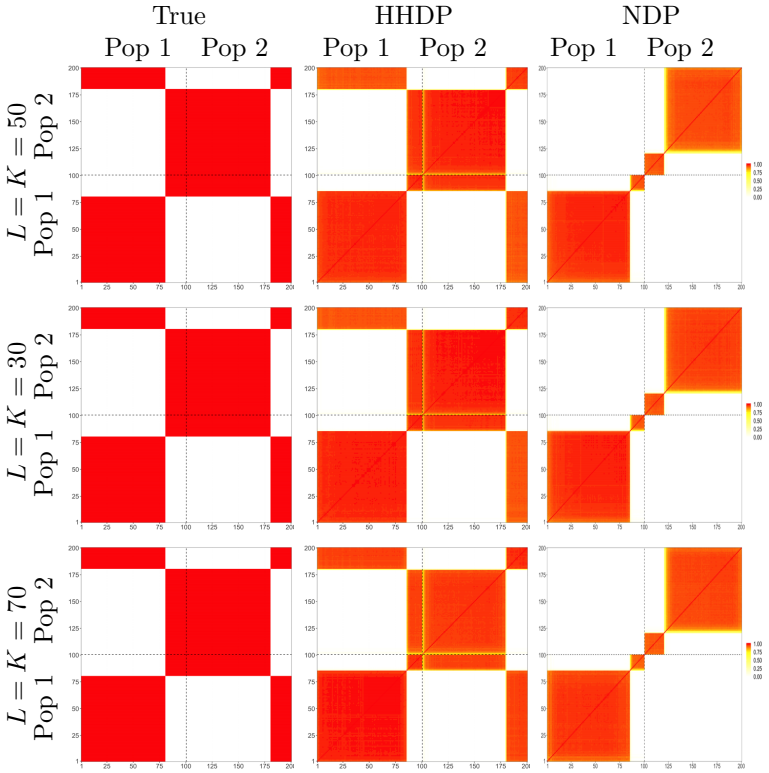

Figure 3.9: Heat maps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, with the two models under different finite dimensional approximations.

Chapter 4

Probabilistic discovery of new species and homogeneous subpopulations[1]

4.1 Introduction

Species sampling models have been widely applied to face one of the most important problems in Statistics: prediction. They owe their name to the seminal contributions by Good (1953) and Good and Toulmin (1956), who focused, among other, on studying the number of new species one would observe if additional observations are sampled. Such models find their natural fit in Ecology and Biology, where they were originally developed, but an increasing number of applications is developing. Since the original formulation, the term 'species sampling model' has been broadly used for a wide range of discrete distributions, not necessarily linked to biological applications, while maintaining the original terminology and denoting as 'species' the unique values that the observations can take (Pitman, 1996). In the single-sample or exchangeable setting, they allow to perform inference on the values of the future observations given a sample from a discrete population. The focus is typically on the prediction of the number of new species one would discover if one is allowed to sample additional observations, or, similarly, on the assessment of the number of unobserved species in the original sample Efron and Ronald (1976); Chao (1981); Chao and Lee (1992); Bunge and Fitzpatrick (1993); Mao (2004).

Lately, species sampling models faced a growing interest from both applied and theoretical perspective. In addition to the original ecological applications (Bunge and Fitzpatrick, 1993; Stockwell and Peterson, 2002), they have been applied in several fields such as genetics (Mao and Lindsay, 2002; Lijoi et al., 2007; Favaro et al., 2009), machine learning and privacy data (Samuels, 1998) just to mention a few. See also De Blasi et al. (2015) for an extensive overview and other possible applications. In a Bayesian setting, these constructions have been further generalized to effectively tackle the problem of prediction when the data arise from different related experiments or populations, i.e. when we are in the so-called partially-exchangeable framework. In such a scenario, Bayesian hierarchical models can be successfully applied to naturally borrow information across the

[1] Joint work with Augusto Fasano, Antonio Lijoi and Igor Prünster. Department of Decision Sciences, Bocconi University.

different populations to improve the predictive performance of the model. This is the underlying idea of some of the most popular Bayesian nonparametrics constructions as the hierarchical and nested formulations for the Dirichlet Process (DP) (Ferguson, 1973) and their generalizations to the Pitman-Yor process (PYP) and beyond (Teh, 2006; Teh et al., 2006; Rodríguez et al., 2008; Camerlenghi et al., 2017, 2019).

Despite the availability of a large numbers of works in literature to face the species sampling problem in a single population framework, just a few works treat the more challenging case of multiple populations. Camerlenghi et al. (2017) exploits a hierarchical Pitman-Yor process (HPYP) construction to effectively face the problem of prediction combining different populations. The choice of the HPYP arises naturally in the species-sampling framework, as the random partition structure induced by the PYP is governed by two parameters and is such that the probability of observing a new species in an additional observation depends on the number of distinct species observed so far, while in the DP case there is only one parameter governing the clustering structure and the above mentioned probability depends only on the global sample size.

This different behavior gives rise to different asymptotic distributions for the number of cluster observed as the population size diverges, with the PYP showing a power-law behavior, which is observed in many empirical studies, while the DP shows only a logarithmic growth, which appears too restrictive. However, the hierarchical construction exploited in the two above-mentioned works does not allow to naturally test homogeneity of subpopulations and cluster the populations with the same *species* distributions. We define a novel hierarchical construction based on PYPs which allows to effectively face also the aforementioned task. This model is obtained by adding a latent nonparametric discrete prior distribution on the population distributions, so that ties among the different population distributions are allowed. In such a setting, testing for homogeneity of population distributions arises naturally, as the model allows to perform probabilistic clustering of the distributions of the groups.

4.2 Preliminaries

Before presenting the proposed model in Section 4.3, we shortly review the literature involved in such construction. Following Pitman (1996), a random probability P is said to be distributed according to a proper species sampling process if it admits the series representation

$$P = \sum_{i\geq 1} \pi_i \delta_{X_i^*}, \quad (X_i^*)_{i\geq 1} \overset{iid}{\sim} H \perp (\pi_i)_{i\geq 1}, \tag{4.1}$$

with H non-atomic. The law of P is completely specified after one fixes the law of the vector of weights $(\pi_i)_{i\geq 1}$. In particular, when the π_i's are such that $\pi_i = v_i \prod_{l=1}^{i-1} v_l$, with $v_i \sim \text{Beta}(1-\sigma, \theta + i\sigma)$, $i \geq 1$, $\sigma \in [0,1)$ and $\theta > -\sigma$, then P is distributed according to a PYP with parameters (θ, σ, H), denoted $P \sim \text{PYP}(\theta, \sigma; H)$. This process is also called two-parameter Poisson-Dirichlet process, and its particular case $\sigma = 0$ boils down to the DP. Observe that, although in species sampling processes the base measure H is nonatomic, in the general PYP formulation this is not required. A vector of weights $(\pi_i)_{i\geq 1}$ constructed with the process just described is said to be $\text{GEM}(\sigma, \theta)$ distributed, after Griffiths, Engen, and McCloskey. A well-known urn scheme allows to sequentially sample observations from P since if $\boldsymbol{U}_n = (U_1, \ldots, U_n)$ is a conditionally independent sample from

P, i.e. $U_i \mid P \overset{iid}{\sim} P$, then a new observation U_{n+1} will have predictive distribution

$$U_{n+1} \mid \boldsymbol{U}_n \sim \sum_{i=1}^{K_n} \frac{n_i - \sigma}{\theta + n} \delta_{U_i^*}(\cdot) + \frac{\theta + K_n \sigma}{\theta + n} H(\cdot), \tag{4.2}$$

where K_n is the number of distinct values $(U_1^*, \dots, U_{K_n}^*)$ in the sample $\boldsymbol{U}_n$, and n_i are their multiplicities, so that $\sum_{i=1}^{K_n} n_i = n$.

This single–sample scenario is well established in the literature (see De Blasi et al. (2015) for a review), however in many applications the data are collected in J different, but related, experiments or populations. In the following we denote with $\boldsymbol{X} = \{(X_{j,i})_{i \geq 1} : j = 1, \dots, J\}$ the data matrix. In such a framework the assumption of a common underlying distribution (*exchangeability*) is too restrictive since it does not take into account the possible differences of the populations. On the other hand, the assumption of independence across populations does not allow to borrow information across experiments in the Bayesian learning.

A natural compromise between the aforementioned extreme cases is partial exchangeability (de Finetti, Bruno, 1938), that entails exchangeability within but not across the different groups. Thanks to de Finetti theorem, we can characterize the array $\boldsymbol{X}$ as arising from a vector of J dependent random probabilities. More precisely, for every vector of population sample sizes $(I_1, \dots, I_J)$, it holds

$$\begin{aligned} X_{j,i} \mid (P_1, \dots, P_J) &\overset{\text{ind}}{\sim} P_j \quad (i = 1, \dots, I_j; j = 1, \dots, J) \\ (P_1, \dots, P_J) &\sim \mathscr{L}, \end{aligned} \tag{4.3}$$

where $\mathscr{L}$ takes the role of the prior in the Bayes-Laplace paradigm and controls the dependence, thus the borrowing of information, across the different populations.

Many possible prior specifications for the vector $(P_1, \dots, P_J)$ are possible. When dealing with species sampling problems, one of the most famous priors in a single–population framework is arguably the PYP. This is due to the fact that, as apparent from equation (4.2), when sampling a new out-of-sample observation, the probability to allocate it to a new cluster depends on the number of already created cluster, and not only on the total number of observations, as happens instead in the case of a DP prior. For this reason, together with the asymptotic power law shown by the number of clusters as n diverges, the PYP is usually the first choice in species sampling problems, being the DP a valuable choice for density estimation under mixture models, but not flexible enough for species sampling processes. Consistently, a common prior specification in multiple-sample cases for $(P_1, \dots, P_J)$ is the HPYP (Teh, 2006; Teh et al., 2006; Camerlenghi et al., 2017). This construction is shortly reviewed in Section 4.2.1: although being well-suited for multiple-sample prediction, it does not allow to test for distribution homogeneity across different populations. This is one of the two tasks of interest in the present work, and, to the best of the authors' knowledge, its treatment in the species sampling framework is lacking, aside from early attempts by Lijoi et al. (2008). In order to achieve this, a nested structure is added, allowing for possible ties in the group distributions P_j. This is done exploiting a nested Pitman-Yor process (NPYP), which is introduced in Section 4.2.2 and follows from the nested Dirichlet Process (NDP) (Rodríguez et al., 2008), after replacing the DP with a PYP.

4.2.1 Hierarchical Pitman-Yor process

A well-known Bayesian nonparametric prior for a vector of dependent discrete random probabilities $(P_1, \ldots, P_J)$ is the hierarchical Pitman-Yor process (HPYP) (Teh, 2006; Teh and Jordan, 2010), which extends the definition of the hierarchical DP (Teh et al., 2006).

The idea is to introduce dependence across the random probabilities $P_1, \ldots, P_J$ via a common random discrete base measure P_0. More precisely we say that $(P_1, \ldots, P_J)$ follows a HPYP with parameter vector $(\sigma, \theta, \sigma_0, \theta_0, H)$, denoted $(P_1, \ldots, P_J) \sim \text{HPYP}(\sigma, \theta, \sigma_0, \theta_0; H)$ if

$$P_j \mid P_0 \overset{\text{iid}}{\sim} \text{PYP}(\sigma, \theta; P_0) \;\; j = 1, \ldots, J, \qquad P_0 \sim \text{PYP}(\sigma_0, \theta_0; H). \tag{4.4}$$

Thanks to the discreteness of P_j we will observe ties with positive probability between the observations recorded in each population $\boldsymbol{X}_j = \{X_{j,i} : i = 1, \ldots, I_j\}$. Furthermore, the discreteness of the common random base measure P_0 allows to share species (cluster observations) across the random probabilities. This feature is essential to perform clustering with mixture models as well as species sampling under heterogeneous populations (Teh et al., 2006; Camerlenghi et al., 2017).

This random partition structure induced by the ties is the core element of species sampling models and from a statistical perspective it can be interpreted as a random clustering. The probability distribution of such a random partition structure can be characterized via the *partially exchangeable partition probability function* (pEPPF) marginalizing out the vector of random probabilities. The pEPPF is an essential object to understand the model and perform inference. For instance, from the pEPPF we can derive closed form results for the joint moments of the observations, both in the same or different populations. Moreover, it can also be used to derive urn schemes that allow to develop marginal Monte Carlo Markov Chain routines which constitute the basis to perform predictive inference. See Camerlenghi et al. (2019) for results on the pEPPF for a large class of models.

However, when the goal is to test population homogeneity, the HPYP has a huge drawback, as it does not allow two groups to share the same distribution. Indeed, in the HPYP, $\text{pr}(P_j = P_k) = 0$ for any $j \neq k$. In order to allow for homogeneous subgroups of populations we will rely on nested structures, extending the HPYP in order to allow $P_j = P_k$, for $j \neq k$, with positive probability. Thus, before moving to the presentation of the proposed model, we introduce the nested Pitman-Yor process (NPYP).

4.2.2 Nested Pitman-Yor process

The nested Dirichlet process (NDP) (Rodríguez et al., 2008) is arguably the most famous Bayesian nonparametric prior to perform joint clustering of distributions and observations under mixture models. However, as pointed out by Camerlenghi et al. (2019) it suffers from a *degeneracy issue* that makes it unsuitable to face our species sampling problem. More precisely, it allows to naturally test for homogeneity of groups and to perform probabilistic clustering of groups since, contrary to the HDP case, a priori we have $\text{pr}(P_j \neq P_k) \in (0, 1)$, for any $j \neq k$. However, given that a single *species* (cluster of observations) is shared across groups j and k, i.e. $X_{j,i} = X_{k,l}$ for some $i, l \geq 1$, the species-populations P_j and P_k are almost surely equal. On the other hand, given that the two species-populations are not exactly equal they are independent and cannot share any species.

In order to overcome the restrictions not suitable for species sampling problems due to a DP prior exposed in Section 4.2, we first extend the hierarchical definition of the NDP to a composition of PYPs. However, also such

nested Pitman-Yor process (NPYP) suffers from the same *degeneracy issue* of the NDP. This will be overcome in Section 4.3, where we will introduce a novel prior for dependent species sampling processes that overcomes the issue combining the NPY and the HDP, taking the advantage of the two.

We say that $(P_1, \ldots, P_J)$ follows a NPYP distribution with vector of parameters $(\alpha, \gamma, \sigma, \theta, H)$, denoted $(P_1, \ldots, P_J) \sim \text{NPYP}(\alpha, \gamma, \sigma, \theta, H)$, if

$$P_j \mid Q \overset{\text{iid}}{\sim} Q \;\; j = 1, \ldots, J, \qquad Q \sim \text{PYP}(\alpha, \gamma; \text{PYP}(\sigma, \theta; H)). \tag{4.5}$$

In order to ease the understanding of the model we can rewrite the random distribution on the space of distributions Q exploiting the well-known stick-breaking representation of the Pitman-Yor process, so that

$$Q = \sum_{k \geq 1} \omega_k^* \delta_{P_k^*},$$

where the unique atoms P_k^* are random probabilities on the space of the observations and are i.i.d. samples from $\text{PYP}(\sigma, \theta; H)$, independent of the weights $(\omega_k^*)_{k \geq 1} \sim \text{GEM}(\alpha, \gamma)$. The discreteness of Q induces a probabilistic clustering of the groups since $\text{pr}(P_j = P_k) = \frac{1-\alpha}{\gamma+1} \in (0, 1)$. However, as for the NDP, given that a single atom is shared between the two distribution, such probability to degenerate to the exchangeable case is 1. Indeed, given that the two distributions P_j and P_k are different they are i.i.d. sampled from $\text{PYP}(\sigma, \theta; H)$ and thus their random atoms are i.i.d. sampled from a non-atomic distribution H and are almost surely different.

To overcome such issue of the NDP in mixture models (Camerlenghi et al., 2019) introduce a novel class of BNP priors named latent nested processes (LNPs). LNPs have the merit to be the first proposal to solve the degeneracy issue of the NDP. However, they are not suited for the study at hand, since computations become infeasible when there are more than two groups and in addition it forces the proportion of species, i.e. the weights, to be the same across groups.

Other proposals are available in the literature, exploiting hidden hierarchical Dirichlet process (HHDP) constructions for mixture models describe in Chapter 3. However, in addition to having a different focus, the theoretical results in Chapter 3 as well the proposed algorithm are not suited for the scenario we are considering, since they rely on the conjugacy and the finite dimensional approximations of the DP. See also Soriano and Ma (2019), Christensen and Ma (2020) and Beraha et al. (2020) for stimulating contributions to this literature. Note that, even if for practical reason we restrict ourselves to the case of composition of PYPs, the methodological arguments together with the algorithms developed in the present work can be easily adapted to a more general class of priors that arise from the composition of different Gibbs type priors, due to product form of their exchangeable partition probability function (EPPF).

4.3 Hidden hierarchical Pitman-Yor process

After having addressed the limitations of the HPYP and NPYP for the scopes at hand, we introduce a novel class of priors, called hidden hierarchical Pitman–Yor process (HHPYP), arising from composition of PYPs that overcomes the above mentioned issues. In particular, this construction is obtained combining the HPYP with the NPYP, as explained in Section 4.3.1, and allows for ties in the population distributions, without suffering from

the aforementioned *degeneracy* issue of the NPYP, thus making homogeneity testing of sub-groups effective, while simultaneously performing species sampling tasks borrowing information across populations.

4.3.1 Definition and basic properties

The HHPYP is obtained by taking a NPYP with discrete base measure distributed according to a PYP. This hierarchical construction, combined with the NDP, allows different populations P_j and P_k, $j \neq k$, to possibly share the same atoms, so that a tie in two observations in these groups does not imply $P_j = P_k$ with probability 1.

In formulae, we say that $(P_1, \ldots, P_J) \sim \text{HHPYP}(\alpha, \gamma, \sigma, \theta, \sigma_0, \theta_0; H)$ if

$$\begin{aligned} (P_1, \ldots, P_J) &\sim \text{NPYP}(\alpha, \gamma, \sigma, \theta; P_0^*) \\ P_0^* &\sim \text{PYP}(\sigma_0, \theta_0; H). \end{aligned} \tag{4.6}$$

For now on we assume that the common probability on the sample space H is non-atomic and for notational simplicity we just write $(P_1, \ldots, P_J) \sim \text{HHPYP}$. Furthermore, we assume the hyperparameters to be fixed, but in practice we can set a prior on them and all the results holds given the hyperparameters and it is straightforward to adapt the Gibbs sampler in Section 4.4 as for the usual species sampling under PYP prior in the exchangeable case.

It follows from (4.5) that we can alternatively characterize the P_j's to be i.i.d. sampled from $Q \sim \text{PYP}(\alpha, \gamma; \text{PYP}(\sigma, \theta; P_0^*))$, given P_0^*, which admits the representation

$$Q = \sum_{k \geq 1} \omega_k^* \delta_{P_k^*}, \tag{4.7}$$

where the weights $(\omega_k^*)_k \sim \text{GEM}(\alpha; \gamma)$ are independent from the distribution atoms. The unique underlying distributions $(P_k^*)_{k \geq 1}$ follow an infinite dimensional HPYP, that is

$$P_k^* \mid P_0^* \overset{\text{iid}}{\sim} \text{PYP}(\theta, \sigma; P_0^*) \ \ (k \geq 1), \quad P_0^* \sim \text{PYP}(\theta_0, \sigma_0; H). \tag{4.8}$$

The discreteness of Q allows to cluster the distributions. For instance, $\text{pr}(P_j = P_k) = \frac{1-\alpha}{\gamma+1} \in (0, 1)$, as for the NPYP. However, thanks to the discreteness of the common random base measure P_0^* the unique random distributions P_k^*'s are now dependent and share the same countable set of atoms allowing to share species across populations which is essential to overcome the aforementioned *degeneracy* issue.

In order to better understand the model, the role of the hyperparameters and the borrowing of strength we can derive the moments of the random probability measures $(P_1, \ldots, P_J) \sim \text{HHPYP}$ evaluated at an arbitrary measurable set A of the sample space $\mathbb{X}$. All the proofs are available in the Appendix. The expected value is $\mathbb{E}[P_j(A)] = H(A)$, as usual in species sampling processes, while the variance can be derived leveraging results on hierarchical models (Camerlenghi et al., 2019) and has the form

$$\text{Var}[P_j(A)] = \frac{H(A)[1 - H(A)]}{\theta_0 + 1} \left[(1 - \sigma_0) + (\theta_0 + \sigma_0) \frac{1 - \sigma}{\theta + 1} \right]. \tag{4.9}$$

We can also derive the expression for the correlation between P_j and P_k, $j \neq k$, which does not depend on

the specific set A, and thus it is often interpreted as a global measure of dependence between the random probabilities in Bayesian nonparametrics. It holds

$$\text{Cor}[P_j(A), P_{j'}(A)] = \frac{1-\alpha}{\gamma+1} + \frac{\gamma+\alpha}{\gamma+1}\frac{1-\sigma_0}{(1-\sigma_0)+(\theta_0+\sigma_0)\dfrac{1-\sigma}{\theta+1}}. \tag{4.10}$$

It is interesting to note the role played by the parameters α and γ, with the correlation decreasing as $\alpha \to 1$ or $\gamma \to \infty$: this is indeed consistent with the fact that in such scenarios we are decreasing the probabilities of homogeneity between the two populations. However, contrary to the NPYP (and its special case NDP), if $j \neq k$, P_j and P_k are not independent, but follow a bi-dimensional HPYP and we can control their dependence via the hyperparameters $(\sigma, \theta, \sigma_0, \theta_0)$ as for the well-known HPYP.

Finally, if the focus is predict future observations it is better to study the dependence directly in term of the observable random variables as de Finetti suggested. If the data matrix $\boldsymbol{X}$ is drawn from $(P_1, \ldots, P_J) \sim$ HHPYP, then

$$\text{Cor}(X_{j,i}, X_{k,l}) = \text{pr}(X_{j,i} = X_{k,l}) \tag{4.11}$$

$$= \begin{cases} \left[\left(\dfrac{1-\sigma}{\theta+1} + \dfrac{1-\sigma_0}{\theta_0+1}\dfrac{\theta+\sigma}{\theta+1}\right)(1-\alpha) + \dfrac{1-\sigma_0}{\theta_0+1}(\gamma+\alpha)\right](\gamma+1)^{-1} & \text{if } j \neq k \\ \left[1-\sigma + \dfrac{1-\sigma_0}{\theta_0+1}(\theta+\sigma)\right](\theta+1)^{-1} & \text{if } j = k. \end{cases} \tag{4.12}$$

Note that a priori correlation between observations, i.e. the probability that the observations belong to the same specie, arising from the same population is larger than the one between observations from different populations, which is an appealing feature from a modeling perspective. The fact that correlation between two observations coincides with the probability that they are equal is a very general result for species sampling models, both in the exchangeable and partially exchangeable cases. See the proof in the Appendix for further insights.

This hierarchical representations of general dependent species sampling processes points out that the dependence is controlled by the ties of the observations and the random partitions they induce. Thus, in order to understand the model and develop sampling schemes, we now study the random partitions structures of the distributions and populations induced by the ties.

A priori, the discreteness of Q induces a random partition $\Psi^{(J)}$ of $[J] = \{1, \ldots, J\}$ and thus a clustering of the distributions $P_1, \ldots, P_J$. More precisely, if $(P_1, \ldots, P_J) \sim$ HHPYP the probability law of $\Psi^{(J)}$ is characterized by the following EPPF, arising from the PYP,

$$\phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) = \frac{\prod_{r=1}^{R-1}(\gamma + r\,\alpha)}{(\gamma+1)_{J-1}} \prod_{r=1}^{R} (1-\alpha)_{m_r-1}, \tag{4.13}$$

where $(x)_J = x(x+1)\cdots(x+J-1)$ is the Jth ascending factorial, R is the random number of blocks of the partition of $[J]$ and m_r is the cardinality of the rth block in order of arrival of the unique P_j. Equation (4.13) immediately follows after recognizing that the underlying distributions $P_1^*, \ldots, P_R^*$ are almost surely different under the HHPYP, although they can share the same atoms.

Denoting with $\boldsymbol{S} = (S_1, \ldots, S_J)$ the cluster membership indicator vector of the J populations in the Chinese

restaurant process (CRP), the following Pölya urn scheme characterizes the distribution of $\boldsymbol{S} = (S_1, \ldots, S_J)$:

$$\text{pr}\big(S_{j+1} = s \mid \boldsymbol{S}^{-(j+)}\big) = \begin{cases} \frac{m_r^{-(j+)} - \alpha}{m_\cdot^{(-j+)} + j} & \text{if } s = S_r^{*-(j+)} \\ \frac{\gamma + \alpha R^{-(j+)}}{m_\cdot^{(-j+)} + j} & \text{if } s = \text{“new”}, \end{cases} \tag{4.14}$$

where we use the $\cdot$ symbol to indicate a summation over an index set, $(j+) = (j+1, \ldots, J)$ is the set of future populations not assigned to any restaurant yet, and $a^{-(b)}$ denotes the quantity a without considering the elements in b. We call $(S_r^* : r = 1, \ldots, R)$ the unique values of the restaurant assignment vector $\boldsymbol{S}$.

In addition, the discreteness of the P_j's induces a random partition of the observations $\boldsymbol{X}$ within and across populations. Calling D the overall number of unique values (number of species) in $\boldsymbol{X}$ and $\boldsymbol{n}_j = (n_{j,d} : d = 1, \ldots, D)$ the vector of cardinalities of the species observed in population j, $j = 1, \ldots, J$, the above mentioned partition structure of $\boldsymbol{X}$ is characterized by the pEPPF $\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J)$. In order to have a tractable form for it, in addition to the population assignment vector $\boldsymbol{S}$, we also make use of a further data augmentation, which corresponds to the usual table augmentation of the Chinese restaurant franchise (CRF) (see Teh (2006); Teh and Jordan (2010)). More precisely, exploiting that culinary metaphor, we introduce the variables $T_{j,i}$, $j = 1, \ldots, J$, $i = 1, \ldots, J_i$, representing the table at which observation i in population j sits and denote $\boldsymbol{T} = \{T_{j,i} : \ j = 1, \ldots, J, \ i = 1, \ldots, I_j\}$. Furthermore, we call $q_{r,t,d}$ the number of customers in restaurant r sitting at table t eating dish d. Marginalizing out the previous latent variables we obtain the following form for the pEPPF.

Theorem 4. *If $\boldsymbol{X}$ is drawn from $(P_1, \ldots, P_J) \sim \text{HHPYP}(\alpha, \gamma, \sigma, \theta, \sigma_0, \theta_0; H)$, then the random partition structure induced by the samples is characterized by the following pEPPF*

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) = \sum \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) \Phi_D^{(n)}(q_{1,\cdot,\cdot}, \ldots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0), \tag{4.15}$$

where the sum runs over all partitions of $[J]$, $\phi_R^{(J)}$ as in (4.13), and $\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \ldots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)$ is the pEPPF associated to an R-dimensional $\text{HPYP}(\sigma, \theta, \sigma_0, \theta_0; H)$.

Exploiting the aforementioned variable augmentation based on $\boldsymbol{T}$ and $\boldsymbol{S}$, and calling $X_1^*, \ldots, X_D^*$ the unique values in the sample $\boldsymbol{X}$, it follows from Bayes Theorem that the following urn scheme easily allows to sample from (4.6) in two steps:

(1) Assign the population to the different restaurant recursive from equation (4.14).

(2) Given the assignment of the populations to the restaurants via $\boldsymbol{S}$, sample the table assignments $\boldsymbol{T}$ and the observations values $\boldsymbol{X}$ recursively adapting the CRF (Teh, 2006) from

$\text{pr}(X_{j,i} = x, T_{j,i} = t \mid \boldsymbol{S}, \boldsymbol{X}^{-(j+i+)}, \boldsymbol{T}^{-(j+i+)}) =$

$$= \begin{cases} \frac{\theta_0 + D^{-(j+i+)}\sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(j+i+)}} \frac{\theta + \ell_{r,\cdot}^{-(j+i+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = \text{“new”} \text{ and } t = \text{“new”} \\ \frac{\omega_d^{-(j+i+)}}{\theta_0 + \ell_{\cdot,\cdot}^{-(j+i+)}} \frac{\theta + \ell_{r,\cdot}^{-(j+i+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = X_d^{*-(j+i+)} \text{ and } t = \text{“new”} \\ \frac{q_{r,t,d}^{-(j+i+)} - \sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = X_d^{*-(j+i+)} \text{ and } t = T_{r,d,l}^{*-(j+i+)}, \end{cases}$$

where $(j+i+) = \{(jl) : l \geq i\} \cup \{(kl) : k \geq j\}$ is the index set associated to the future random variables not

sampled yet, and $T^*_{r,d,l}$ denotes the value of the lth table in restaurant r serving dish d. Finally, $\ell_{r,d}$ represents the number of tables in restaurant r serving dish d. If we are interested not just in the clustering structure, but also on the specific value of the observations, we can sample the "new" values of the observations from the non-atomic base distribution H.

Notice that, contrary to the usual CRF characterizing the HPYP, a restaurant is not identified by a unique population, but different populations can be assigned to the same restaurant, thus sharing tables. On the other hand, if two populations are assigned to two different restaurants, they will not share any table. Since this urn scheme naturally extends the well-known CRF metaphor, with the additional property that a restaurant can be composed by more than one group, we call such a Pölya urn scheme hidden Chinese restaurant franchise (HCRF). Populations are clustered together when assigned to the same restaurant. In testing the homogeneity among different groups, one can then compute the posterior probability that two populations belong to the same cluster as discussed in the next section.

4.3.2 Population homogeneity testing

One of the main goals of the present work is to introduce a valuable model that, among usual inferential species sampling tasks, is able to assess which populations are homogeneous. Since the clustering is probabilistic, the key quantity of interest is the posterior probability of co-clustering for each couple of distributions P_j, P_k, $j \neq k$, namely $\text{pr}(P_j = P_k \mid \boldsymbol{X})$. These probabilities can be interpreted in terms of posterior evidence of homogeneity between the distributions P_j and P_k. Considering the case of $J = 2$ populations for ease of interpretation, and denoting $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ the vectors of the counts of the overall distinct D values in each of the two populations, the pEPPF characterizing the law of $\boldsymbol{X}$ can be written as

$$\Pi^{(n)}_D(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{1-\alpha}{\gamma+1}\Phi^{(n)}_D(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0) + \frac{\alpha+\gamma}{\gamma+1}\Phi^{(n)}_D(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0), \tag{4.16}$$

with $\Phi^{(n)}_D$ as in Theorem 4. As expected by the model specification (4.6), the pEPPF (4.16) can be seen as a convex combination of the probability laws of the random partitions induced by different HPYPs, the first composed by a single population with $\boldsymbol{n}_1 + \boldsymbol{n}_2$ vector of multiplicity, while the second formed by two distinct populations having multiplicity vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ respectively. From (4.16) one can easily derive the posterior probability to degenerate to the exchangeable case, that is of the event $\{P_1 = P_2\}$.

Proposition 4. *If $J = 2$ and $\boldsymbol{X}$ is sampled from $(P_1, P_2) \sim$ HHPYP, then the posterior probability of degeneracy is*

$$pr(P_1 = P_2 \mid \boldsymbol{X}) = \frac{(1-\alpha)\Phi^{(n)}_D(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}{(1-\alpha)\Phi^{(n)}_D(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0) + (\alpha+\gamma)\,\Phi^{(n)}_D(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}.$$

Notice that the HHPYP overcomes the degeneracy issue of the NDP allowing for the presence of shared species across populations, without implying to degenerate to exchangeability.

The above mentioned task is strictly related with hypothesis testing procedures. Indeed, assessing whether $P_1 = P_2$, can be rephrased as a test where

$$H_0 : \; S_1 = S_2 \quad vs. \quad H_1 : \; S_1 \neq S_2. \tag{4.17}$$

H_0 and H_1 specify two different models for the data matrix $\boldsymbol{X}$. The corresponding Bayes factor is then readily available and has the form

$$B_{01} = \frac{p(\boldsymbol{X} \mid H_0)}{p(\boldsymbol{X} \mid H_1)} = \frac{\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}{\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}.$$

For $J > 2$, the co-clustering posterior probabilities for each couple (j, k) can be easily computed via the marginal Gibbs sampler described in Section 4.4. It will be sufficient to count how many times out of the B Gibbs updates $S_j = S_k$ to get an MCMC estimate of $\text{pr}(P_j = P_k \mid \boldsymbol{X})$. Moreover, the testing procedure (4.17) can be straightforwardly extended to the generic null hypothesis

$$H_0 : \; S_{j_1} = S_{k_1}, \ldots, S_{j_C} = S_{k_C}, \text{ for some } \{j_1, \ldots j_C\}, \{k_1, \ldots k_C\} \subseteq [J],$$

with complementary alternative hypothesis H_1, with corresponding Bayes factor following from the specification of the summation in (4.16) to the cases specified by the null hypothesis and the alternative.

4.3.3 Inference on the number of species

Consistently with the above, let D be the overall random number of species (dishes) in the sample $\boldsymbol{X}$ of size $n = \sum_{j=1}^{J} I_j$, and define D_r the random number of species among the $q_{r,\cdot,\cdot}$ observations in the rth cluster of populations (restaurant). Call R the number of heterogeneous populations among the J populations. To keep the notation lighter, we suppress the dependence on n, J and $q_{r,\cdot,\cdot}$. The probabilistic behavior of D and R both on finite samples and when the overall numbers of observations n and populations J diverge is of utmost importance to deeper understand key properties of the proposed species sampling model.

First, note that $(T_{j,i} \mid S_j = r, P_r^*) \overset{\text{iid}}{\sim} P_r^*$, with $P_r^* \overset{\text{iid}}{\sim} \text{PYP}(\sigma, \theta, H)$, where H is a non atomic probability measure, so that, if we call L_r the number of distinct values in $\boldsymbol{T}_r = (T_{j,i} : S_j = r)$, $r = 1, \ldots, R$, these quantities are independent across restaurants.

We also denote by $D_{0,\ell}$ the random number of distinct values between ℓ exchangeable values generated from P_0^*. Notice that the distribution of R, L_r and $D_{0,\ell}$ can be derived via marginalization from the EPPF induced by a PYP with non-atomic base measure. More precisely,

$$\begin{aligned} p(R) &= \frac{1}{R!} \sum_{\boldsymbol{m} \in \mathcal{F}_R(J)} \binom{J}{m_1, \ldots, m_R} \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) \\ &= \frac{\prod_{r=1}^{R-1}(\gamma + r\,\alpha)}{(\gamma+1)_{J-1}} \frac{\mathscr{C}(J, R; \alpha)}{\alpha^R}, \end{aligned} \tag{4.18}$$

where $\mathcal{F}_R(J) = \{(m_1, \ldots, m_R) : m_r \geq 1, \sum_{r=1}^{R} m_R = J\}$. Here $\mathscr{C}(n, k; \sigma)$ represents the generalized factorial coefficient defined by $(\sigma t)_n = \sum_{k=1}^{n} \mathscr{C}(n, k; \sigma)(t)_k$ and computable as $\mathscr{C}(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-\sigma j)_n$ with the proviso $\mathscr{C}(0, 0; \sigma) = 1$ and $\mathscr{C}(n, 0; \sigma) = 1$ for any $n > 0$ and $\mathscr{C}(n, k; \sigma) = 0$ for any $k > n$. For an exhaustive review of the generalized factorial coefficients see Charalambides (2002).

Marginalizing out the corresponding EPPF we can also obtain:

$$p(D_{0,\ell}) = \frac{\prod_{d=1}^{D_{0,\ell}-1}(\theta_0 + d\,\sigma_0)}{(\theta_0+1)_{\ell-1}} \frac{\mathscr{C}(\ell, D_{0,\ell}; \sigma_0)}{\sigma_0^{D_{0,\ell}}}, \quad p(L_r) = \frac{\prod_{\ell=1}^{L_r-1}(\theta + \ell\,\sigma)}{(\theta_0+1)_{\ell-1}} \frac{\mathscr{C}(q_{r,\cdot,\cdot}, L_r); \sigma)}{\sigma^{L_r}}.$$

In the next Theorem we derive probability distribution of the overall number of species.

Theorem 5. *If the data matrix* $\boldsymbol{X}$ *is drawn from* $(P_1, \ldots, P_J) \sim$ HHPYP*, then*

$$\begin{aligned} p(D) &= \sum_{\boldsymbol{B} \in \rho(J)} \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) \sum_{L=D}^{n} pr(D_{0,L} = D)\, pr\left(\sum_{j=1}^{J} L_r = L\right) \\ &= \sum_{\boldsymbol{B} \in \rho(J)} \frac{\prod_{r=1}^{R-1}(\gamma + r\,\alpha)}{(\gamma+1)_{J-1}} \prod_{r=1}^{R} (1-\alpha)_{m_r - 1} \\ &\times \sum_{L=D}^{n} \frac{\prod_{d=1}^{D-1}(\theta_0 + d\,\sigma_0)}{(\theta_0+1)_{\ell-1}} \frac{\mathscr{C}(\ell, D; \sigma_0)}{\sigma_0^D} \frac{\prod_{\ell=1}^{L-1}(\theta + \ell\,\sigma)}{(\theta_0+1)_{\ell-1}} \frac{\mathscr{C}(q_{r,\cdot,\cdot}, L; \sigma)}{\sigma^L}, \end{aligned}$$

where $\rho(J)$ *is the space of the partitions of* $[J]$*.*

The distribution of the overall number of species D in Theorem 5 is quite involved. However, from such analytical formula we can derive a simple algorithm to sample from it after a variables augmentation.

From the composition structure points out in Theorem 5 we can also study the asymptotic behavior of the number of species as the sample size n diverges, which boils down to a simple analytical form. From now on, for an arbitrary function $f(n)$, we write $Y_n \asymp f(n)$ if the limit of $Y_n / f(n)$ as n diverges is almost surely a positive and finite random variable.

Theorem 6. *If the data matrix* $\boldsymbol{X}$ *is drawn from* $(P_1, \ldots, P_J) \sim$ HHPYP *and* D *is the overall number of distinct species in* J *populations of sample sizes* $I_1 = \ldots = I_J = I = n/J$*. Then*

$$D \asymp n^{\sigma\sigma_0}$$

as $n \to \infty$*.*

Notice that the HHPYP can be used also to discover the number of heterogeneous subpopulations R as the number of populations J grows. From (4.18) we have

$$R \asymp J^{\alpha},$$

as $j \to \infty$. That is the number of heterogeneous subpopulations follows a polynomial growth under model (4.6).

4.4 Marginal Gibbs sampler and predictive inference

Posterior inference can be efficiently performed thanks to the marginal Gibbs sampler described in the following section. The full conditionals for the augmented variables S_j and $\boldsymbol{T}_j$ have indeed a nice ratio expression, which is recovered exploiting Bayes theorem and the fact that, with such variable augmentation, the pEPPF admits a product form that simplifies between the numerator and the denominator. This results allow for interpretable and computationally tractable inference for all quantities of interest. These include the posterior distribution of the tables $\boldsymbol{T}$ and, more importantly, the posterior distribution of the vector of cluster assignments $\boldsymbol{S}$ and the predictive distribution of future observations. Such quantities can be used to perform population homogeneity

testing, and, for instance, to estimate the number of new species that are expected to be observed in an additional sample of $\boldsymbol{m} = (m_1, \ldots, m_J)$ observations.

4.4.1 Gibbs sampler

The proposed Gibbs sampler follows by extending the marginal Gibbs sampler for NDP mixture models in Zuanetti et al. (2018) to the species sampling framework presented in the present work. The main idea is that, after having set an initial configuration for the augmented variables $\boldsymbol{S}$ and $\boldsymbol{T}$, at each iteration one first updates the table assessment T_{ji} for each individual, and then updates the population cluster membership indicators S_j, $j = 1, \ldots, J$, via a Metropolis-Hastings within Gibbs step. Due to the fact that $\boldsymbol{T}_j$ must be coherent with S_j, the update of S_j is done jointly with an update of $\boldsymbol{T}_j$. The proposal distribution of the Metropolis step is such that it is easy to sample from and allows a fast evaluation of the acceptance probability. Performing homogeneity testing will then be immediate, as it will be sufficient to count the fraction of times two populations are clustered together out of the total number of iterations. In particular, the Gibbs sampler to perform posterior inference on the latent variables $\boldsymbol{S}$ and $\boldsymbol{T}$ is reported below.

`(0)` At $t = 0$ start from an initial configuration $\boldsymbol{S}$ and $\boldsymbol{T}$.

`(1)` At iteration $t \geq 1$

`(1.a)` With $X_{ji} = X_d^*$ sample latent variables T_{ji}, for $i = 1, \ldots, I_j$ and $j = 1, \ldots, J$ from

$$\text{pr}(T_{ji} = t \mid \boldsymbol{T}^{-(ji)}, \boldsymbol{X}, \boldsymbol{S}) \propto \begin{cases} q_{r,t,d}^{-(ji)} - \sigma & \text{if } t = T_{r,d,l}^{*-(ji)} \\ \frac{\omega_d^{-(ji)}}{\ell_{\cdot,\cdot}^{-(ji)} + \theta_0}(\theta + \ell_{r,\cdot}^{-(ji)}\sigma) & \text{if } t = \text{“new”}, \end{cases} \tag{4.19}$$

where $\omega_d^{-(ji)} = \ell_{\cdot,d}^{-(ji)} - \sigma_0$ if $\ell_{\cdot,d}^{-(ji)} > 0$ otherwise $\omega_d^{-(ji)} = 1$.

`(1.b)` When updating S_j, we will have to update the $\boldsymbol{T}_j$. This is done via the following efficient Metropolis-Hastings within Gibbs step. Call $Y = (S_j, \boldsymbol{T}_j)$ the vector of the current values for the jth population cluster assignment and the table assignments in there, the proposed new values $Y' = (S_j', \boldsymbol{T}_j{}')$ are sampled by the proposal distribution $q(\cdot \mid \cdot)$, which is defined hierarchically exploiting the results for the importance sampling density in (Maceachern et al., 1999):

$$q(Y' \mid Y) = p(S_j' \mid \boldsymbol{S}_{-j}) \prod_{i=1}^{I_j} p(T_{ji}' \mid \boldsymbol{T}_{-j}, \boldsymbol{T}_j'^{-(ji+)}, \boldsymbol{X}_j^{-(ji+)}, X_{ji}, S_j') \tag{4.20}$$

where $(ji+) = \{(jl) : l \geq i\}$ is the index set associated to the future random variables not yet sampled. Moreover, $p(S_j' \mid \boldsymbol{S}_{-j})$ is as in (4.14) with $(j+)$ replaced by (j) and $p(T_{ji}' \mid \boldsymbol{T}_{-j}, \boldsymbol{T}_j'^{-(ji+)}, \boldsymbol{X}_j^{-(ji+)}, X_{ji}, S_j')$ can be computed as in (4.19).

The proposed state Y' is accepted with probability $\min(1, A')$, where $A' = \frac{p(Y' \mid \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{X}) q(Y \mid Y')}{p(Y \mid \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{X}) q(Y' \mid Y)}$. The full conditional of Y can be expressed as

$$\begin{aligned} p(S_j, \boldsymbol{T}_j \mid \boldsymbol{X}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}) &= \frac{p(S_j, \boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{T}_{-j})}{p(\boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j})} \propto \\ &\propto p(S_j \mid \boldsymbol{S}_{-j}) p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j), \end{aligned} \tag{4.21}$$

so that

$$A' = \frac{p(\boldsymbol{T}'_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S'_j) \prod_{i=1}^{I_J} p(T_{ji} \mid \boldsymbol{X}^{-(ji+)}, X_{ji}\boldsymbol{T}_{-j}, \boldsymbol{T}_j^{-(ji+)}, \boldsymbol{S}_{-j}, S_j)}{p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j) \prod_{i=1}^{I_J} p(T'_{ji} \mid \boldsymbol{X}^{-(ji+)}, X_{ji}\boldsymbol{T}_{-j}, \boldsymbol{T}_j'^{-(ji+)}, \boldsymbol{S}_{-j}, S'_j)},$$

where the conditional distribution for $(\boldsymbol{T}_j, \boldsymbol{X}_j)$ has the form $p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}) = \prod_{i=1}^{I_j} p(T_{ji}, X_{ji} \mid \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}^{-(ji+)}, \boldsymbol{S})$. Thus,

$$A' = \prod_{i=1}^{I_j} \frac{p(X_{ji} \mid \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}_{-j}, \boldsymbol{T}_j'^{-(ji+)}, T'_{ji}, \boldsymbol{S}_{-j}, S'_j)}{p(X_{ji} \mid \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}_{-j}, \boldsymbol{T}_j^{-(ji+)}, T_{ji}, \boldsymbol{S}_{-j}, S_j)},$$

where

$$p(X_{ji} = x \mid \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}^{-(ji+)}, T_{ji} = t, \boldsymbol{S}) = \begin{cases} 1 & \text{if } t = T^*_{r,d,l} \text{ and } x = X_d^{*-(ji+)}, \\ \frac{\ell_{\cdot,d}^{-(ji+)} - \sigma_0}{\theta + \ell_{\cdot,\cdot}^{-(ji+)}} & \text{if } t = \text{“new”} \text{ and } x = X_d^{*-(ji+)}, \\ \frac{\theta + D^{-(ji+)}\sigma_0}{\theta + \ell_{\cdot,\cdot}^{-(ji+)}} & \text{if } t = \text{“new”} \text{ and } x = \text{“new”}. \end{cases} \tag{4.22}$$

4.4.2 Predictive distribution

Consider now the case where we want to make inference about an additional sample of $\boldsymbol{I}^{\text{“new”}} = (I_1^{\text{“new”}}, \ldots, I_J^{\text{“new”}})$ new observations, where m_j is the number of new observations in population j, for $j = 1, \ldots, J$. Le us denote $\boldsymbol{X}^{\text{new}} = \{X_{j,i}^{\text{new}} : \ j = 1, \ldots, J, \ i = 1, \ldots, I_j^{\text{“new”}}\}$ the values of such new observations and $\boldsymbol{T}^{\text{new}} = \{T_{j,i}^{\text{new}} : \ j = 1, \ldots, J, \ i = 1, \ldots, I_j^{\text{“new”}}\}$ the latent tables allocations in the HCRF metaphor.

The following urn scheme allows obtain sample $(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}})$ exploiting the output of the Gibbs sampler described in the previous section, since the sample can be obtained sequentially, exploiting the fact that $p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}) = \prod_{j=1}^{J} \prod_{i=1}^{m_j} p(X_{j,i}^{\text{new}}, T_{j,i}^{\text{new}} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{X}^{\text{new}-(j+i+)}, \boldsymbol{T}^{\text{new}-(j+i+)})$, where $\text{pr}(X_{j,i}^{\text{new}} = x, T_{j,i}^{\text{new}} = t \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{X}^{\text{new}-(j+i+)}, \boldsymbol{T}^{\text{new}-(j+i+)}) =$

$$= \begin{cases} \frac{\theta_0 + D^{-(j+i+)}\sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(j+i+)}} \frac{\theta + \ell_{r,\cdot}^{-(j+i+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = \text{“new”} \text{ and } t = \text{“new”} \\ \frac{\ell_{\cdot,d}^{-(j+i+)} - \sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(j+i+)}} \frac{\theta + \ell_{r,\cdot}^{-(j+i+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = X_d^{*-(j+i+)} \text{ and } t = \text{“new”} \\ \frac{q_{r,t,d}^{-(j+i+)} - \sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = X_d^{*-(j+i+)} \text{ and } t = T_{r,d,l}^{*-(j+i+)}, \end{cases}$$

being $(j + i+) = \{(jl) : l \geq i\} \cup \{(kl) : k \geq j\}$ the index set associated to the future random variables not yet sampled.

Thus, for each configuration $(\boldsymbol{S}, \boldsymbol{T})$ generated in the Gibbs sampler presented in Section 4.4.1, one can obtain a sample from $p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X})$, so that, after the burn-in period, samples from $p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{X})$ are obtained.

Appendix C

C.1. Proof of (4.9) and (4.10)

Proof. Note that $P_j \overset{\mathrm{d}}{=} P_1^*$.

$$\mathbb{E}[P_j(A)] = \mathbb{E}[P_1^*(A)] = H(A) \text{ since } P_1^* \text{ is a species sampling model}$$
$$\mathrm{Var}[P_j(A)] = \mathrm{Var}[P_1^*(A)]$$

Furthermore, we know that $\mathrm{Var}[P_0^*(A)] = H(A)[1-H(A)]\dfrac{1-\sigma_0}{\theta_0+1}$ and $\mathrm{Var}[P_1^*(A)] = \dfrac{H(A)[1-H(A)]}{\theta_0+1}\left[(1-\sigma_0)+(\theta_0+\sigma_0)\dfrac{1-\sigma}{\theta+1}\right]$, see Camerlenghi et al. (2019).

Moreover $\mathbb{E}[P_1^*(A)P_2^*(A)] = \mathbb{E}[\mathbb{E}[P_1^*(A) \mid P_0^*]\mathbb{E}[P_2^*(A) \mid P_0^*]] = \mathbb{E}[P_0^*(A)^2]$ and $\mathrm{pr}(P_j = P_{j'}) = \dfrac{1-\alpha}{\gamma+1}$ for $j \neq j'$. Thus,

$$\mathbb{E}[P_j(A)P_{j'}(A)] = \mathbb{E}[P_1(A)P_2(A) \mid P_1 = P_2]\mathrm{pr}(P_1 = P_2) + \mathbb{E}[P_1(A)P_2(A) \mid P_1 \neq P_2]\mathrm{pr}(P_1 \neq P_2) =$$
$$= \frac{1-\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)P_2^*(A)] = \frac{1-\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathbb{E}[P_0^*(A)^2].$$

From it we derive

$$\mathrm{Cov}[P_j(A), P_{j'}(A)] = \mathbb{E}[P_j(A)P_{j'}(A)] - H(A)^2 = \frac{1-\alpha}{\gamma+1}\mathrm{Var}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathrm{Var}[P_0^*(A)^2]$$

and

$$\mathrm{Cor}[P_j(A), P_{j'}(A)] = \frac{\mathrm{Cov}[P_j(A), P_{j'}(A)]}{\mathrm{Var}[P_1^*(A)]} = \frac{1-\alpha}{\gamma+1} + \frac{\gamma+\alpha}{\gamma+1}\frac{\mathrm{Var}[P_0^*(A)]}{\mathrm{Var}[P_1^*(A)]} =$$
$$= \frac{1-\alpha}{\gamma+1} + \frac{\gamma+\alpha}{\gamma+1}\frac{1-\sigma_0}{(1-\sigma_0)+(\theta_0+\sigma_0)\dfrac{1-\sigma}{\theta+1}} = \frac{1-\alpha+\dfrac{(\alpha+\gamma)(-1+\sigma_0)(1+\theta)}{-1+(-1+\sigma_0)\theta-\theta_0+\sigma(\sigma_0+\theta_0)}}{1+\gamma}$$

□

C.2. Proof of (4.11)

Proof. Note that $X_{j,i} \overset{\mathrm{d}}{=} X_l^*$. Thus,

$$\mathrm{Cov}[X_{j,i}, X_{j',i'}] = \mathbb{E}[\mathrm{Cov}(X_{j,i} = X_{j',i'} \mid \mathbf{1}(X_{j,i} = X_{j',i'}))]\mathrm{pr}(X_{j,i} = X_{j',i'}) + 0 = \mathrm{pr}(X_{j,i} = X_{j',i'})\mathrm{Var}(X_l^*)$$

Therefore $\mathrm{Cor}[X_{j,i}, X_{j',i'}] = \mathrm{pr}(X_{j,i} = X_{j',i'})$, where

$$\mathrm{pr}(X_{j,i'} = X_{j,i}) = \mathrm{pr}(X_{j,i'} = X_{j,i} \mid T_{j,i} = T_{j,i'})\mathrm{pr}(T_{j,i} = T_{j,i'}) + \mathrm{pr}(T_{j,i} \neq T_{j,i'})\mathrm{pr}(X_{j,i'} = X_{j,i} \mid T_{j,i} \neq T_{j,i'}) =$$
$$= \frac{1-\sigma}{\theta+1} + \frac{1-\sigma_0}{\theta_0+1}\frac{\theta+\sigma}{\theta+1}$$

and if $j \neq j'$

$$\begin{aligned}&\mathrm{pr}(X_{j,i'} = X_{j,i}) = \mathrm{pr}(X_{j,i} = X_{j',i'} \mid P_j = P_{j'})\mathrm{pr}(P_j = P_{j'}) + \mathrm{pr}(X_{j,i} = X_{j',i'} \mid P_j \neq P_{j'})\mathrm{pr}(P_j \neq P_{j'}) = \\ &= \left\{ \left[\frac{1-\sigma}{\theta+1} + \frac{1-\sigma_0}{\theta_0+1}\frac{\theta+\sigma}{\theta+1} \right](1-\alpha) + \frac{1-\sigma_0}{\theta_0+1}(\gamma+\alpha) \right\}(\gamma+1)^{-1}\end{aligned}$$

□

C.3. Proof of Theorem 4

Proof. In order to prove Theorem 4 first note that

Lemma 3. *The random partition structure induced by the samples $\boldsymbol{X}$ drawn from $(P_1, \ldots, P_J) \sim$ HHPYP given a particular partition of distributions $\Psi^{(J)} = \{B_1, \ldots, B_R\}$ is characterized by the pEPPF*

$$\Pi_D^{(n)}\left(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}\right) = \Phi_D^{(n)}\left(q_{1,\cdot,\cdot}, \ldots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0\right), \tag{4.23}$$

where $\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \ldots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)$ is the pEPPF associated to a R-dimensional HPYP$(\sigma, \sigma_0, \theta, \theta_0; H)$.

Indeed,

$$\begin{aligned}&\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}) = \\ &= \mathbb{E}\left[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} P_1^{n_{1,d}}(\mathrm{d}x_d) \ldots P_J^{n_{J,d}}(\mathrm{d}x_d) \;\middle|\; \Psi^{(J)} = \{B_1, \ldots, B_R\} \right] = \\ &= \mathbb{E}\left[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} P_1^{*q_{1,\cdot,d}}(\mathrm{d}x_d) \ldots P_R^{*q_{r,\cdot,d}}(\mathrm{d}x_d) \right] = \Phi_D^{(n)}(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \sigma, \theta, \sigma_0, \theta_0),\end{aligned}$$

where $\mathbb{X}_*^D = \mathbb{X}^D \setminus \{\boldsymbol{x} : x_i = x_j \text{ for some } i \neq j\}$ and $(P_1^*, \ldots, P_R^*) \sim$ HPYP$(\sigma, \sigma_0, \theta, \theta_0; H)$. Moreover, note that the R unique values between $(P_1, \ldots, P_J)$ are not necessary the first $(P_1^*, \ldots, P_R^*)$ but since $(P_k^*)_{k \geq 1}$ are exchangeable the third equality holds.

Therefore, applying Lemma 3

$$\begin{aligned}\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) &= \sum \mathrm{pr}(\Psi^{(J)} = \{B_1, \ldots, B_D\})\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \ldots, B_D\}) = \\ &= \sum \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma)\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \ldots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)\end{aligned}$$

□

C.4. Proof Proposition 4

Proof. In order to derive the posterior probability of degeneracy we rewrite the marginal likelihood as

$$p(\boldsymbol{X}) = \Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) \prod_{d=1}^{D} H(\mathrm{d}\boldsymbol{X}_d^*),$$

where $\{\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_D^*\}$ are the D unique values between $\boldsymbol{X}$ and $\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)$ is the pEPPF associated to the proposed model 4.16, that is

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \text{pr}(P_1 = P_2)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \text{pr}(P_1 \neq P_2)\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0),$$

Finally we prove the proposition by applying Bayes theorem

$$\begin{aligned}\text{pr}(P_1 = P_2 \mid \boldsymbol{X}) &= \frac{\text{pr}(P_1 = P_2)p(\boldsymbol{X} \mid P_1 = P_2)}{p(\boldsymbol{X})} \\ &= \frac{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0) + (\alpha + \gamma)\,\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}.\end{aligned}$$

□

C.5. Proof of Theorem 5

Proof. Note that applying Lemma 3 and Theorem 6 in (Camerlenghi et al., 2019) we have that

$$\text{pr}(D_n = D \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}) = \sum_{L=D}^{n} \text{pr}(D_{0,L} = D)\,\text{pr}\left(\sum_{j=1}^{J} L_{r,q_{r,\cdot,\cdot}} = L\right).$$

Then marginalizing out the population partition $\Psi^{(J)}$ we have

$$\text{pr}(D_n = D) = \sum_{\boldsymbol{B} \in \rho(J)} \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) \sum_{L=D}^{n} \text{pr}(D_{0,L} = D)\,\text{pr}\left(\sum_{j=1}^{J} L_{r,q_{r,\cdot,\cdot}} = L\right).$$

□

C.6. Proof of Theorem 6

Proof. Let $T(\boldsymbol{n}) \overset{\text{d}}{=} \sum_{r=1}^{R} L_{r,q_{r,\cdot,\cdot}} \leq D_n$, representing the number of tables in the franchise. The conditional independence arising from the hierarchical specification of the model (4.6) entails that $D_n = D_{0,T(\boldsymbol{n})}$ almost surely. Moreover, by the asymptotic of the number of species in the exchangeable case under a Pitman–Yor prior we have that for each $m_r = m_r(\Psi^{(J)}) \in \{0, \ldots, J\}$:

$$\frac{D_{0,I}}{I^{\sigma_0}} \xrightarrow{\text{a.s.}} C_0, \qquad \frac{L_{r,m_r I}}{I^{\sigma}} \xrightarrow{\text{a.s.}} C_r m_r^{\sigma},$$

as $I \to \infty$, where C_0 and C_r's are positive and finite random variables. Since $T(\boldsymbol{n}) = \sum_r^R L_{r,m_r I}$

$$\frac{T(\boldsymbol{n})}{I^{\sigma}} \xrightarrow{\text{a.s.}} \sum_{r=1}^{R} C_r m_r^{\sigma} = \eta(\Psi^{(J)}),$$

where $\eta = \eta(\Psi^{(J)})$ is a positive finite random variable. Thus,

$$\frac{D_{0,T(\boldsymbol{n})}}{D_{0,\eta I^\sigma}} = \frac{T(\boldsymbol{n})^{\sigma_0}}{(\eta I^\sigma)^{\sigma_0}} \frac{D_{0,T(\boldsymbol{n})}/T(\boldsymbol{n})^{\sigma_0}}{D_{0,\eta I^\sigma}/(\eta I^\sigma)^{\sigma_0}} \xrightarrow{\text{a.s.}} 1.$$

entailing

$$\frac{D_n}{I^{\sigma\sigma_0}} = \frac{D_{0,T(\boldsymbol{n})}}{D_{0,\eta I^\sigma}} \frac{D_{0,\eta I^\sigma}}{(I^\sigma)^{\sigma_0}} \xrightarrow{\text{a.s.}} C_0,$$

as $I \to \infty$. □

Chapter 5

Clustering consistency with Dirichlet process mixture modelling[1]

5.1 Introduction

Bayesian nonparametric methods have seen a huge development in the last decades, often standing out for flexibility and strong mathematical foundations; see Müller et al. (2018) and Ghosal and Van Der Vaart (2017) for recent stimulating accounts. The cornerstone of Bayesian nonparametrics is the model based on the Dirichlet process (Ferguson, 1973), which can be expressed as

$$X_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \qquad \tilde{P} \sim \text{DP}(\alpha, P_0). \tag{5.1}$$

where $\alpha > 0$ is the *concentration parameter* and P_0 is the *baseline distribution* over the sample space $(\mathsf{X}, \mathcal{X})$. Two main features that make this model particularly appealing are flexibility, which is assessed in terms of its topological support, and conjugacy.

The Dirichlet process is almost surely discrete and, if one wishes to model continuous data, it is useful to convolve it with a density kernel k. This yields the popular Dirichlet process mixture (Lo, 1984)

$$X_i \mid \theta_i \overset{\text{ind}}{\sim} k(\cdot|\theta_i), \quad \theta_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \quad \tilde{P} \sim \text{DP}(\alpha, P_0). \tag{5.2}$$

The model in (5.2) exhibits nice asymptotic properties in the context of density estimation: in many cases the posterior distribution of the unknown density concentrates at the true data-generating one at the minimax-optimal rate, up to a logarithmic factor (Ghosal et al., 1999; Ghosal and Van der Vaart, 2007). Such a model and many of its variants are widely used across scientific areas, thanks also to the availability of a wide variety of efficient computational methods to perform inference, see for instance Escobar and West (1995, 1998); Maceachern and Müller (1998); Neal (2000); Blei and Jordan (2006).

Thanks to the discreteness of the Dirichlet process, the latent parameters θ_i's exhibit ties with positive probability. Hence, the model in (5.2) is also routinely used to perform clustering since it partitions observations

[1] Joint work with Filippo Ascolani, Antonio Lijoi and Giacomo Zanella. Department of Decision Sciences, Bocconi University.

into groups based on whether their corresponding latent parameters θ_i coincide or not. The ubiquitous use of Dirichlet process mixtures for clustering motivates the interest in the asymptotic behavior of the posterior distribution of the underlying partition, and in particular on the inferred number of clusters (i.e. subpopulations), as the number of observations increases. Nguyen (2013) showed posterior consistency of the mixing distribution $\tilde{P}$ under general conditions. However this does not imply consistency for the number of clusters, due to the use of the Wasserstein distance. Miller and Harrison (2013) provided a negative result, showing that Dirichlet process mixtures are not consistent for the number of components when data are generated from a mixture with a single standard normal component, see also Miller and Harrison (2014) for extensions. These results, however, are derived under the assumption that the concentration parameter α is known and fixed. This is crucial, because the clustering behaviour of Dirichlet process mixtures is governed by the choice of α. Indeed, under (5.2), the prior probability of observing ties is purely a function of α, since $\text{pr}(\theta_i = \theta_j) = 1/(\alpha + 1)$. In order to have a more flexible distribution on the clustering of the data, in several applications one specifies a prior for α, leading to a mixture of Dirichlet processes in the sense of Antoniak (1974). Here we show that introducing such a prior has a major impact on the asymptotic behaviour of the number of clusters.

In this work we study the consistency of Dirichlet process mixtures under the commonly used assumption of assigning a prior π on the unknown concentration parameter. We show that Dirichlet process mixtures can be (and typically are) consistent for the number of clusters. We provide consistency results under fairly general conditions on π and for a moderately large class of kernels k, including Poisson, Uniform and Truncated Normal distributions. Following Miller and Harrison (2013), we focus on data-generating mixtures with a single component, although we expect our results to extend to the more general case of finite mixtures with multiple components. We stress that the framework we study is arguably closer to the way Dirichlet process mixtures are used in practice, compared to holding α fixed.

We note that studying an asymptotic regime where the data-generating truth is a mixture with a finite and fixed number of components entails some degree of model misspecification. Indeed Dirichlet process mixtures are nonparametric models with an a priori infinite number of components or, in other words, a number of clusters growing with the size of the dataset. Thus, our results can be interpreted as a form of robustness: despite assuming infinitely many components a priori, if the data-generating truth has finitely many the model can recover that by adapting appropriately the value of α. In particular we show that, when the data are generated from a mixture with one component, the posterior distribution of α converges to a point mass at 0 at a specific rate, which is crucial to ensure consistency. See Section 5.7 for more discussion and some related literature.

5.2 Dirichlet process mixtures and random partitions

Assigning a prior on the concentration parameter, we obtain a mixture model of the following form:

$$X_i|\theta_i \overset{\text{ind}}{\sim} k(\cdot|\theta_i), \quad \theta_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \quad \tilde{P} \mid \alpha \sim \text{DP}(\alpha, P_0), \quad \alpha \sim \pi. \tag{5.3}$$

Since we are interested in the distribution of the number of clusters, it is reasonable to rewrite (5.3) in terms of the distribution on partitions, related to the so-called Chinese restaurant process. For every pair of natural

numbers (n, s) such that $s \leq n$, denote with $\rho_s(n)$ the set of partitions of $\{1, \dots, n\}$ in s non empty subsets. Conditionally on α, the sequence $(\theta_i)_{i \geq 1}$ induces a prior distribution on the space of partitions of $\mathbb{N}$ that, for any $n \geq 2$, is characterized by

$$\text{pr}(A \mid \alpha) = \frac{\alpha^s}{\alpha^{(n)}} \prod_{j=1}^{s} (a_j - 1)!, \quad (A = \{A_1, \dots, A_s\} \in \rho_s(n), s \leq n), \tag{5.4}$$

where $\alpha^{(n)} = \alpha \cdots (\alpha + n - 1)$ is the ascending factorial and $a_j = |A_j|$. Conditional on the partition A, the distribution of the cluster-specific parameters $\hat{\theta}_{1:s} = (\hat{\theta}_1, \dots, \hat{\theta}_s)$ and of the associated observations $X_{1:n} = (X_1, \dots, X_n)$ is given by

$$\text{pr}(X_{1:n} \mid \hat{\theta}_{1:s}, A) = \prod_{j=1}^{s} \prod_{i \in A_j} k(X_i \mid \hat{\theta}_j), \quad \text{pr}(\hat{\theta}_{1:s} \mid A, \alpha) = \prod_{j=1}^{s} p_0(\hat{\theta}_j). \tag{5.5}$$

The number of clusters in a sample of size n is denoted by K_n and under (5.3) it has the following prior distribution

$$\text{pr}(K_n = s) = \int \sum_{A \in \rho_s(n)} \text{pr}(A \mid \alpha) \pi(\alpha) \, \mathrm{d}\alpha.$$

We are concerned in studying $\text{pr}(K_n = s \mid X_{1:n})$, so we are interested in the joint distribution of the vector $(X_{1:n}, K_n)$, given by

$$\text{pr}(X_{1:n} = x_{1:n}, K_n = s) = \sum_{A \in \rho_s(n)} \text{pr}(A) \prod_{j=1}^{s} m(x_{A_j}), \tag{5.6}$$

where $x_{1:n} = (x_1, \dots, x_n) \in \mathsf{X}^n$, where

$$m(x_{A_j}) = \int \prod_{i \in A_j} k(x_i \mid \theta) p_0(\theta) \mathrm{d}\theta.$$

is the marginal likelihood for the subset of observations identified by A_j, given that they are clustered together. Similarly to Miller and Harrison (2013), we study the asymptotic behavior of the posterior induced by model (5.3) when the observations are independent and identically distributed samples from a single-component mixture, that is we assume the following data generation mechanism:

$$X_i \overset{\text{iid}}{\sim} P, \quad (i = 1, 2, \dots), \tag{5.7}$$

where P is a fixed probability measure on X. We will let $P^{(n)}$ and $P^{(\infty)}$ be the product probability measures induced on X^n and X^∞ respectively, and denote (5.7) by $X_{1:\infty} \sim P^{(\infty)}$ for brevity. In the following we will consider P to be dominated by some measure, usually Lebesgue or counting one, and denote the resulting density by f. We say that model in (5.3) is *well-specified* for P if $k(\cdot|\theta) = f(\cdot)$ for some θ, that is if the data-generating distribution belongs to the family of kernels in (5.3).

Since data are generated from a single-component mixture, we say that posterior consistency for the number of clusters holds if $\text{pr}(K_n = 1 \mid X_{1:n}) \to 1$ as $n \to \infty$ in $P^{(\infty)}$-probability. Note that here the conditional probability $\text{pr}(K_n = 1 \mid X_{1:n})$ is defined with respect to the model in (5.3), while the convergence in probability is with respect to the data-generating process $X_{1:\infty} \sim P^{(\infty)}$. Since $\text{pr}(K_n = 1 \mid X_{1:n})$ lies between 0 and 1,

convergence in $P^{(\infty)}$-probability is equivalent to convergence in L^1 with respect to $P^{(\infty)}$ and thus we could equivalently define consistency in terms of L^1 convergence.

5.3 Consistency and random concentration parameter

Our proofs of consistency rely on the following lemma.

Lemma 4. *The convergence* $\mathrm{pr}(K_n = 1 \mid X_{1:n}) \to 1$ *as* $n \to \infty$ *in* $P^{(\infty)}$*-probability holds true if and only if one has, in* $P^{(\infty)}$*-probability,*

$$\sum_{s=2}^{n} \frac{\mathrm{pr}(K_n = s \mid X_{1:n})}{\mathrm{pr}(K_n = 1 \mid X_{1:n})} \to 0 \quad \text{as } n \to \infty\,. \tag{5.8}$$

Working with the ratios of conditional probabilities in (5.8) is beneficial, as the marginal distribution of $X_{1:n}$ involved in the definition of $\mathrm{pr}(K_n = 1 \mid X_{1:n})$ gets canceled and thus we can avoid studying it. Also, it is convenient to write such ratios of probabilities as follows: first, recall from (5.4) and (5.6) that

$$\mathrm{pr}(X_{1:n} = x_{1:n}, K_n = s) = \int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha) \mathrm{d}\alpha \sum_{A \in \rho_s(n)} \prod_{j=1}^{s} (a_j - 1)! m(x_{A_j}),$$

for any $s \geq 1$, which implies that

$$\frac{\mathrm{pr}(K_n = s \mid X_{1:n})}{\mathrm{pr}(K_n = 1 \mid X_{1:n})} = \underbrace{\frac{\int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha}{\int \frac{\alpha}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha}}_{C(n,s)} \underbrace{\sum_{A \in \rho_s(n)} \frac{\prod_{j=1}^{s} (a_j - 1)!}{(n-1)!} \frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}}_{R(n,s)}. \tag{5.9}$$

The decomposition of (5.9) into the factors $C(n, s)$ and $R(n, s)$ is useful to understand the role of the prior distribution over α, and to compare our results with the one of Miller and Harrison (2013, 2014). In particular the term $R(n, s)$ does not depend of the choice of prior π, and remains unchanged even if α is fixed. This is indeed the key term studied in Miller and Harrison (2013), where it is shown that, under some assumption, $\liminf R(n, s) > 0$ as $n \to \infty$ in $P^{(\infty)}$-probability for $s = 2$. On the contrary, the $C(n, s)$ incorporates information about α and its prior distribution. In the fixed α case, which can be thought as having a degenerate delta mass prior $\pi = \delta_\alpha$ for some $\alpha > 0$, the term $C(n, s)$ boils down to α^{s-1}, which is constant over n. This is sufficient for Miller and Harrison (2013) to deduce lack of consistency for fixed α, which in our context means that

$$\limsup \mathrm{pr}(K_n = 1 \mid X_{1:n}, \alpha) < 1 \tag{5.10}$$

as $n \to \infty$ in $P^{(\infty)}$-probability for any $\alpha > 0$.

However, once a non-degenerate prior π is employed, the term $C(n, s)$ depends on n and, as we show in the next section, converges to 0 as $n \to \infty$ under mild assumptions on π. Thus, $\liminf R(n, s) > 0$ is not anymore sufficient to establish whether consistency holds or not. Instead, one needs to compare the rate at which $C(n, s)$ goes to 0 with the behavior of $R(n, s)$, as done in the following sections. Note that further lower bounds for $R(n, s)$ for general values of s are given in Miller and Harrison (2014); Yang et al. (2019). However, once combined with $C(n, s)$, these are too rough to deduce either consistency or lack thereof. Therefore, we

need to exploit different techniques to derive the rate of $R(n, s)$.

Since $\text{pr}(K_n = 1 \mid X_{1:n}) = \int \text{pr}(K_n = 1 \mid X_{1:n}, \alpha)\pi(\alpha \mid X_{1:n})\,\text{d}\alpha$ and that by (5.10) $\limsup \text{pr}(K_n = 1 \mid X_{1:n}, \alpha) < 1$ for any $\alpha > 0$. This, however, does not imply that $\limsup \text{pr}(K_n = 1 \mid X_{1:n}) < 1$ because in this case the limit cannot be brought inside the integral. The main reason is that, in the asymptotic regime we are considering, the posterior distribution $\pi(\alpha \mid X_{1:n})$ concentrates around 0 as $n \to \infty$, see Proposition 6 below, and thus does not have a proper limit supported on $(0, \infty)$.

5.4 Asymptotic behaviour of the concentration parameter

We are now concerned with studying $C(n, s)$ in (5.9). We prove that for a large class of priors π, including the uniform and Gamma distributions, $C(n, s)$ goes to 0 with a logarithmic rate in n. The results of this section are unrelated to the choice of kernel k and data generating distribution f and thus can be useful to prove consistency, or lack thereof, for arbitrary Dirichlet process mixture models with random concentration parameter. In the next section we will combine them with the analysis of $R(n, s)$ for specific choices of k and f to deduce consistency results.

In order to facilitate the intuition, the term $C(n, s)$ can be interpreted as the $(s-1)$-moment of α conditional on all the n observations being clustered together. Indeed, under (5.3) it holds

$$\pi(\alpha \mid K_n = 1) \propto \frac{\alpha}{\alpha^{(n)}}\pi(\alpha).$$

and thus $C(n, s) = \int \alpha^{s-1}\pi(\alpha \mid K_n = 1)\,d\alpha = E(\alpha^{s-1} \mid K_n = 1)$.

We will derive asymptotic results under the following assumptions:

- *A*1. *Absolute continuity*: the prior of α admits a density π with respect to the Lebesgue measure;
- *A*2. *Polynomial behaviour around the origin*: $\exists\, \epsilon$, δ, β such that $\forall \alpha \in (0, \epsilon)$ it holds $\frac{1}{\delta}\alpha^{\beta} \leq \pi(\alpha) \leq \delta\alpha^{\beta}$;
- *A*3. *Subfactorial moments*: $\exists\, D, \nu > 0$ such that $\int \alpha^s \pi(\alpha)\,d\alpha < D\rho^{-s}\Gamma(\nu+s+1)$ for every s and for sufficiently large ρ.

The first two assumptions are sufficient to study the behaviour of $C(n, s)$, as explained in the next Lemma. The third requirement, instead, will be useful specifically for the consistency purposes: the minimum value of ρ needed depends on the problem at hand, that is on the specific choice of $k(\cdot)$ in (5.3) and P_0 in (5.7).

Proposition 5. *Suppose π satisfies assumptions A*1 *and A*2*. Then there exist $F, G > 0$ such that for any $s < n$ it holds*

$$F\,\frac{\Gamma\{s+\beta+1, \epsilon\log(n)\}}{\{\log(n)+1\}^s} \leq C(n, s+1) \leq \frac{Gs}{\epsilon^s}E(\alpha^s)\frac{\Gamma\{s+\beta+1, \epsilon\log(n)\}}{\log\{n/(1+\epsilon)\}^s},$$

where $\Gamma(x, y)$ is the lower incomplete Gamma function and $E(\alpha^s) = \int \alpha^s \pi(\alpha)\,d\alpha$.

Thus, for fixed s, $C(n, s)$ decreases logarithmically. Coherently with our intuition, by looking at (5.9), the addition of a prior helps the model recovering the correct number of clusters, in this case by favouring a smaller number of clusters. The outlined assumptions are satisfied by common families of distributions, as displayed in the next lemma.

Lemma 5. *The following choices of $\pi(\cdot)$ satisfy assumptions A1, A2 and A3:*

- *Any distribution with bounded support that satisfies assumptions A1 and A2, such as the uniform distribution over $(0, c)$, with $c > 0$;*
- *The Generalized Gamma distribution with density proportional to $\alpha^{d-1}e^{-\left(\frac{\alpha}{a}\right)^p}$, provided that $p > 1$;*
- *The Gamma distribution with shape ν and rate ρ, provided that ρ is high enough.*

5.5 Consistency results for specific examples

In this section we analyze specific choices for the data-generating distribution P and kernel k, while in the next section we provide a general result. Notice that Miller and Harrison (2014) prove that, with a fixed α, consistency does not hold for any of the cases illustrated in this section. Thus, our results suggest that placing a prior on the concentration parameter and learning it from the observed data is critical to achieve consistency for Dirichlet process mixture models.

The following results are established by combining upper bounds on $C(n, s)$ and $R(n, s)$ to prove the convergence in (5.8). Often, instead of proving directly convergence in probability of (5.8), we will prove convergence in L^1, that is a sufficient condition. In this way we will avoid the study of the specific partition at hand. The following Lemma shows how the problem simplifies in this case.

Lemma 6. *Assume $(X_1, \dots, X_n)$ are exchangeable random variables. Then*

$$E\left\{\sum_{A\in\rho_s(n)} \frac{\prod_{j=1}^s (a_j-1)!}{(n-1)!}\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})}\right\} = \sum_{\boldsymbol{a}\in\mathcal{I}_s(n)} \frac{n}{s!\prod_{j=1}^s a_j} E\left\{\frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right\},$$

where the sum runs over $\mathcal{I}_s(n) = \{\boldsymbol{a} \in \{1, \dots, n\}^s : \sum_{j=1}^s a_j = n\}$ and $A^{\boldsymbol{a}}$ is an arbitrary partition in $\rho_s(n)$ such that $|A_j^{\boldsymbol{a}}| = a_j$ for $j = 1, \dots, s$.

5.5.1 Gaussian mixtures

We start with a simple case to illustrate the impact of the random concentration parameter on consistency. Indeed, we specialize the model in (5.3) to the case of Gaussian kernels and assume constant data, equal to some fixed real number θ^*. More precisely, set

$$f = \delta_{\theta^*}, \quad k(\cdot|\theta) = \mathrm{N}(\theta, 1), \quad p_0 = \mathrm{N}(0, 1). \tag{5.11}$$

Unlike all other examples we will consider below, this case is not well-specified (as $k(\cdot|\theta) \neq f(\cdot)$ for every θ), which makes the definition of true or data-generating number of clusters more delicate. Nonetheless, being an example with constant data, one would hope the posterior of the number of cluster to concentrate on one cluster. However, even in such limiting case with constant data, Miller and Harrison (2013) show that under (5.3) with fixed concentration parameter $\mathrm{pr}(K_n = 1|X_{1:n})$ does not go to 1 as n diverges. The following theorem shows that assuming a prior on α changes the asymptotic posterior behaviour of K_n.

Theorem 7. *Consider (f, k, p_0) as in (5.11) and assume $\pi(\cdot)$ satisfies assumptions A1–A2 and A3 with $\rho > 16$. Then,*

$$pr(K_n = 1 \mid X_1, \dots, X_n) \to 1$$

$P^{(\infty)}$-almost surely as $n \to \infty$.

5.5.2 Poisson case

In case of discrete observations, a popular choice for the kernels in (5.3) is given by Poisson densities. Consider then the following well-specified setting:

$$f = \text{Po}(\theta^*), \quad k(\cdot|\theta) = \text{Po}(\theta), \quad p_0 = \text{Exp}(1), \tag{5.12}$$

where $\theta^* > 0$ is the true expected value. The following lemma exploits the Poisson-Gamma conjugacy and provides an upper bound to the expectations involved in Lemma 6.

Lemma 7. *Consider (f, k, p_0) as in (5.12). Then*

$$E\left\{\frac{\prod_{j=1}^{s} m(X_{A_j^a})}{m(X_{1:n})}\right\} \leq e^{\theta^*} \frac{n+1}{\prod_{j=1}^{s}(a_j+1)}.$$

We have the following consistency result.

Theorem 8. *Consider f, k and p_0 as in (5.12) and assume that π satisfies assumptions A1, A2 and A3 (with $\rho > 14$). Then*

$$pr(K_n = 1 \mid X_1, \dots, X_n) \to 1$$

as $n \to \infty$ in $P^{(\infty)}$-probability.

5.5.3 Uniform case

We now consider a well-specified Dirichlet process mixtures of uniform distributions:

$$f = \text{Unif}(\theta^* - c, \theta^* + c), \quad k(\cdot|\theta) = \text{Unif}(\theta - c, \theta + c), \quad p_0 = \text{Unif}(\theta^* - c, \theta^* + c), \tag{5.13}$$

where $\theta^* \in \mathbb{R}$ is a fixed location parameter and $c > 0$. The choice of p_0 is made for ease of computations, in the next section a much more general framework will be considered. Here the marginal distribution is available and given by the following lemma.

Lemma 8. *Consider k and p_0 as in (5.13). Then it holds*

$$m(x_{1:n}) = \frac{2c - \{\max(x_{1:n}, \theta^*) - \min(x_{1:n}, \theta^*)\}}{(2c)^{n+1}}, \qquad (x_{1:n} \in [\theta^* - c, \theta^* + c]^n).$$

With a suitable application of Hölder inequality it is possible to prove consistency through Lemma 6.

Theorem 9. *Consider f, k and p_0 as in (5.13), and assume π satisfies assumptions A1, A2 and A3 (with $\rho \geq 38$). Then*

$$pr(K_n = 1 \mid X_1, \ldots, X_n) \to 1$$

as $n \to \infty$ in $P^{(\infty)}$-probability.

5.6 General consistency result for location families with bounded support

In this section we provide a general consistency result for location families with bounded support, which generalizes the uniform case of Section 5.5.3. We consider kernels of the form

$$k(x \mid \theta) = g(x - \theta) \qquad (x \in \mathbb{R})$$

where $c > 0$ and $\theta \in \mathbb{R}$ is a location parameter. Here g is a density function on the real line satisfying the following assumptions:

*B*1. g is strictly positive on some interval $[a, b]$ and 0 elsewhere;

*B*2. g is differentiable with bounded derivative in (a, b);

*B*3. The base measure P_0 is absolutely continuous with respect to the Lebesgue measure, with bounded density p_0.

The above assumptions essentially require that the kernel is a location-family distribution with positive density on a bounded support. Under the above assumption we have the following consistency result.

Theorem 10. *Suppose k and p_0 satisfy assumptions B1-B3; and $\pi(\cdot)$ satisfies assumptions A1–A3. Then, for any $f = k(\cdot|\theta^*)$ with θ^* belonging to the interior support of p_0, we have*

$$pr(K_n = 1 \mid X_1, \ldots, X_n) \to 1$$

as $n \to \infty$ in $P^{(\infty)}$-probability.

The class is fairly general and it includes, as relevant special cases, the previously studied uniform distribution and the truncated Gaussian distribution among others.

Finally, we note that all the previous consistency results are related to another result of general interest, which is that the posterior distribution of the concentration parameter goes to zero in the asymptotic regime we are considering.

Proposition 6. *Under any of the setting in Theorem 7, 8, 9 and 10, we have*

$$\pi(\alpha \mid X_1, \ldots, X_n) \to \delta_0$$

weakly as $n \to \infty$ in $P^{(\infty)}$-probability.

5.7 Discussion

There are many avenues to extend our results, such as considering broader settings with a general number of components for the data-generating truth (including infinite mixtures) and different mixture kernels. Some or the results and tools we introduced here may prove useful to accomplish such tasks. Another interesting question worth studying is whether consistency can also be attained estimating the concentration parameter through maximization of the marginal likelihood, in an empirical Bayes fashion (Liu, 1996; McAuliffe et al., 2006). In this work we preferred to focus on the fully Bayesian approach because it is arguably the approach most commonly employed by practitioners using Dirichlet process mixtures. Moreover, the empirical Bayes estimator of α may not be well defined on $(0, \infty)$ because the marginal likelihood can easily have its maximum at both 0 or infinity, thus raising theoretical and practical issues.

We note that the asymptotic analysis of the posterior distribution of the number of clusters for Dirichlet process mixtures has recently attracted considerable theoretical interest (Yang et al., 2019; Ohn and Lin, 2020), and has motivated various methodological developments (Miller and Harrison, 2018; Zeng and Duan, 2020). Ohn and Lin (2020) showed that, if α is sent deterministically to 0 at appropriate rates as $n \to \infty$, the posterior distribution of the number of clusters concentrates on finite values when data are generated from a finite mixture, which is a necessary condition for consistency. Such results are similar in spirit to ours, although we consider the substantially different setting where α is learned through a prior, which is arguably more natural in a Bayesian framework. Finally, our results also provide an answer, at least partially, to the question of Yang et al. (2019): *"there exists a natural way to correct the problem instead of truncating the number of clusters?"*, by showing that placing a prior on α can be sufficient to recover consistency.

Appendix D

D.1. Proof of Lemma 4

Proof. By construction it holds

$$\mathrm{pr}(K_n = 1 \mid X_{1:n}) = 1 - \sum_{s=2}^{n} \mathrm{pr}(K_n = s \mid X_{1:n} = x_{1:n}).$$

Dividing by the left hand side and rearranging we get

$$\mathrm{pr}(K_n = 1 \mid X_{1:n}) = \left(1 + \sum_{s=2}^{n} \frac{\mathrm{pr}(K_n = s \mid X_{1:n} = x_{1:n})}{\mathrm{pr}(K_n = 1 \mid X_{1:n} = x_{1:n})} \right)^{-1}.$$

The result follows immediately. □

D.2. Proof of Proposition 5

By assumptions $A1$ and $A2$ there exist $\epsilon, \delta, \beta > 0$ such that

$$\frac{1}{\delta^2} \frac{\int_0^\epsilon \frac{\alpha^{s+\beta+1}}{\alpha^{(n)}} \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{\beta+1}}{\alpha^{(n)}} \mathrm{d}\alpha} \leq \frac{\int_0^\epsilon \frac{\alpha^{s+1}}{\alpha^{(n)}} \pi(\alpha) \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha}{\alpha^{(n)}} \pi(\alpha) \mathrm{d}\alpha} \leq \delta^2 \frac{\int_0^\epsilon \frac{\alpha^{s+\beta+1}}{\alpha^{(n)}} \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{\beta+1}}{\alpha^{(n)}} \mathrm{d}\alpha}. \tag{5.14}$$

Notice that, if assumption $A2$ holds for $\epsilon' \geq 1$, it holds automatically for $\epsilon < 1$. Thus, without loss of generality, we will assume $\epsilon < 1$. Thus, the main object of interest will be

$$E_n\{\alpha^s\} = \int_0^\epsilon \alpha^s p_n(\alpha)\mathrm{d}\alpha,$$

where E_n denotes the expected value with respect to the probability distribution with density

$$p_n(\alpha) = \frac{f_n(\alpha)}{\int_0^\epsilon f_n(x)\,\mathrm{d}x}, \quad f_n(x) = \frac{x^{\beta+1}}{x^{(n)}}\,\mathbb{1}_{(0,\epsilon)}(x), \tag{5.15}$$

where $\mathbb{1}_A$ stands for the indicator function of set A. We now provide some lemmas that will be useful to prove Proposition 1.

Lemma 9. *Let f and g be two pdf's on $\mathbb{R}$ such that $g(x)/f(x)$ is non-decreasing in x. Then $\int h(x)f(x)\mathrm{d}x \leq \int h(x)g(x)\mathrm{d}x$ for any non-decreasing $h : \mathbb{R} \to \mathbb{R}$.*

Proof. Let $X \sim f$ and $Y \sim g$. Since $g(x)/f(x)$ is non-decreasing we have $g(x_0)f(x_1) \leq g(x_1)f(x_0)$ for any $x_0 < x_1$. Thus we have

$$F_Y(x_1)f(x_1) = \int_{-\infty}^{x_1} g(x_0)f(x_1)\mathrm{d}x_0 \leq \int_{-\infty}^{x_1} g(x_1)f(x_0)\mathrm{d}x_0 = F_X(x_1)g(x_1)$$

and

$$\{1 - F_X(x_0)\}g(x_0) = \int_{x_0}^{\infty} g(x_0)f(x_1)\mathrm{d}x_1 \leq \int_{x_0}^{\infty} g(x_1)f(x_0)\mathrm{d}x_1 = \{1 - F_Y(x_0)\}f(x_0).$$

It follows

$$\frac{F_Y(x)}{F_X(x)} \leq \frac{g(x)}{f(x)} \leq \frac{1 - F_Y(x)}{1 - F_X(x)},$$

for every $x \in \mathbb{R}$, which implies

$$\frac{F_Y(x)}{1 - F_Y(x)} \leq \frac{F_X(x)}{1 - F_X(x)}.$$

Thus, Y stochastically dominates X, i.e. the corresponding cdf's satisfy $F_Y(x) \leq F_X(x)$ for every $x \in \mathbb{R}$, which implies that $E\{h(X)\} \leq E\{h(Y)\}$ for any non-decreasing h. □

Lemma 10. *Under assumptions $A1$ and $A2$, for any $n > s \geq 1$ it holds*

$$\frac{\Gamma[s+\beta+1, \epsilon\{\log(n)+1\}]}{\delta^2\Gamma[\beta+1, \epsilon\{\log(n)+1\}]}\{\log(n)+1\}^{-s} \leq \frac{\int_0^\epsilon \frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} \leq \frac{\delta^2\Gamma\{s+\beta+1, \epsilon\log(n)\}}{\Gamma\{\beta+1, \epsilon\log(n)\}}\log\{n/(1+\epsilon)\}^{-s},$$

where $\Gamma(x, y)$ is the lower incomplete Gamma function and $\epsilon, \delta, \beta > 0$ are such that for every $\alpha \in (0, \epsilon)$ it holds $\frac{1}{\delta}\alpha^\beta \leq \pi(\alpha) \leq \delta\alpha^\beta$.

Proof. By (5.14) it suffices to find suitable bounds of $E_n\{\alpha^s\}$. For the upper inequality we apply Lemma 9 with $f = p_n$, $g(\alpha) \propto (cn)^{-\alpha}\alpha^\beta\mathbb{1}_{(\alpha\in[0,\epsilon])}$ with $c = (1+\epsilon)^{-1}$ and $h(\alpha) = \alpha^s$. To verify that $g(\alpha)/p_n(\alpha)$ is

non-decreasing for $\alpha \in (0, \epsilon]$ we compute

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log\left\{\frac{g(\alpha)}{p_n(\alpha)}\right\} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\left\{-\alpha \log(cn) + \sum_{i=1}^{n-1} \log(\alpha + i)\right\} = -\log\left(\frac{n}{1+\epsilon}\right) + \sum_{i=1}^{n-1} \frac{1}{\alpha + i}$$
$$\geq -\log\left(\frac{n+\epsilon}{1+\epsilon}\right) + \sum_{i=1}^{n-1} \frac{1}{i+\epsilon} \geq 0,$$

where the last inequality follows by a standard property of the harmonic series: $\int_1^k \frac{1}{x+\epsilon}\, dx < \sum_{i=1}^{k-1} \frac{1}{i+\epsilon}$ for any $k > 1$. Thus, since $h(\alpha) = \alpha^s$ is non-decreasing in α it follows by Lemma 9 that

$$E_n\{\alpha^s\} \leq \frac{\int_0^\epsilon \alpha^{s+\beta}(cn)^{-\alpha}\mathrm{d}\alpha}{\int_0^\epsilon \alpha^{\beta}(cn)^{-\alpha}\,\mathrm{d}\alpha} = \frac{\log(cn)^{-s}\int_0^{\epsilon\log(cn)} z^{s+\beta}e^{-z}\mathrm{d}z}{\int_0^{\epsilon\log(cn)} z^{\beta}e^{-z}\,\mathrm{d}z} =$$
$$= \frac{\log(cn)^{-s}\Gamma\{s+\beta+1, \epsilon\log(cn)\}}{\Gamma\{\beta+1, \epsilon\log(cn)\}}.$$

For the lower bound we apply Lemma 9 with $f(\alpha) \propto (en)^{-\alpha}\alpha^{\beta}\mathbb{1}_{(\alpha\in[0,\epsilon])}$, $g(\alpha) = p_n(\alpha)$ and $h(\alpha) = \alpha^s$. To verify that $p_n(\alpha)/f(\alpha)$ is non-decreasing for $\alpha \in (0, \epsilon]$ we compute

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log\left\{\frac{p_n(\alpha)}{f(\alpha)}\right\} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\left[-\sum_{i=1}^{n-1} \log(\alpha + i) + \alpha\{\log(n) + 1\}\right] = -\sum_{i=1}^{n-1} \frac{1}{\alpha + i} + \log(n) + 1$$
$$\geq -\sum_{i=1}^{n-1} \frac{1}{i} + \log(n) + 1 \geq 0,$$

where the last inequality follows by a standard property of the harmonic series: $\sum_{i=1}^{k} \frac{1}{i} \leq \log(k) + 1$ for any $k \geq 1$. Thus, since $h(\alpha) = \alpha^s$ is non-decreasing in α it follows by Lemma 9 that

$$E_n\{\alpha^s\} \geq \frac{\int_0^\epsilon \alpha^{s+\beta}(en)^{-\alpha}\mathrm{d}\alpha}{\int_0^\epsilon \alpha^{\beta}(en)^{-\alpha}\,\mathrm{d}\alpha} = \frac{\log(en)^{-s}\int_0^{\epsilon\log(en)} z^{s+\beta}e^{-z}\mathrm{d}z}{\int_0^{\epsilon\log(en)} z^{\beta}e^{-z}\,\mathrm{d}z} =$$
$$= \frac{\log(en)^{-s}\Gamma\{s+\beta+1, \epsilon\log(en)\}}{\Gamma\{\beta+1, \epsilon\log(en)\}}.$$

Combining the bounds with (5.14) we obtain the desired results. □

Lemma 11. *For any $\epsilon > 0$, there exists $M > 0$ such that, for any $n \geq 1$, it holds*

$$M \int_0^\epsilon \frac{\alpha}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha \geq \int_\epsilon^\infty \frac{\alpha}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha\,.$$

Proof. Define $p = \frac{\int_\epsilon^\infty \alpha\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\frac{\epsilon}{2}} \alpha\pi(\alpha)\,\mathrm{d}\alpha}$. Then

$$\int_0^\epsilon \frac{\alpha}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha - \int_\epsilon^\infty \frac{\alpha}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha = \int_0^\epsilon \frac{\alpha}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha - \int_0^{\frac{\epsilon}{2}} p\frac{\alpha}{\epsilon^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha$$
$$\geq \int_0^{\frac{\epsilon}{2}} \frac{\alpha}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha - \int_0^{\frac{\epsilon}{2}} p\frac{\alpha}{\epsilon^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha.$$

Choose m such that $\left(\frac{\epsilon}{2}\right)^{(m)} < \frac{\epsilon^{(m)}}{p}$, which is always possible because $\left(\epsilon^{(m)}\right)^{-1}\left(\frac{\epsilon}{2}\right)^{(m)} \to 0$ as $m \to \infty$. Thus

$$\int_0^{\epsilon} \frac{\alpha}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha \geq \int_{\epsilon}^{\infty} \frac{\alpha}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha, \quad n \geq m\,,$$

and it suffices to set $M = \max\{P, 1\}$ with

$$P = \max_{1\leq i\leq m}\left\{\frac{\int_{\epsilon}^{\infty}\frac{\alpha}{\alpha^{(i)}}\,\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\frac{\alpha}{\alpha^{(i)}}\,\pi(\alpha)\,\mathrm{d}\alpha}\right\}.$$

□

of Proposition 5. We first prove the upper bound. We have

$$C(n, s+1) \leq \frac{\int_0^{\infty}\frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\frac{\alpha}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} = \frac{\int_0^{\epsilon}\frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\frac{\alpha}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} + \frac{\int_0^{\epsilon}\frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\frac{\alpha}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}\frac{\int_{\epsilon}^{\infty}\frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}.$$

Moreover, it holds

$$\frac{\int_{\epsilon}^{\infty}\frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} \leq \frac{\int_{\epsilon}^{\infty}\alpha^{s}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\alpha^{s}\pi(\alpha)\,\mathrm{d}\alpha} \leq \delta\frac{\int_{\epsilon}^{\infty}\alpha^{s}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\alpha^{s+\beta}\,\mathrm{d}\alpha} \leq \delta\, E\{\alpha^s\}\frac{s+\beta+1}{\epsilon^{s+\beta+1}},$$

where the first inequality follows since $\alpha^{(n)} \geq \epsilon^{(n)}$ for $\alpha \in (\epsilon, \infty)$ and $\alpha^{(n)} \leq \epsilon^{(n)}$ for $\alpha \in (0, \epsilon)$, while the second one follows from assumption $A2$. Moreover E stands for the expected value with respect to π. Thus from Lemma 10 it holds

$$C(n, s+1) \leq \frac{\delta^2\left\{1 + E\{\alpha^s\}\frac{s+\beta+1}{\epsilon^{s+\beta+1}}\right\}\Gamma\{s+\beta+1, \epsilon\log(n)\}}{\Gamma\{\beta+1, \epsilon\log(n)\}}\log\{n/(1+\epsilon)\}^{-s}.$$

Then choose $G = \frac{4\delta^2}{\epsilon^{\beta+1}\Gamma(\beta+1, \epsilon\log 2)}$. For the lower bound, apply Lemma 10 and Lemma 11 to get

$$C(n, s+1) \geq \frac{1}{M+1}\frac{\int_0^{\epsilon}\frac{\alpha^{s+1}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\epsilon}\frac{\alpha}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} \geq \frac{1}{M+1}\frac{\Gamma[s+\beta+1, \epsilon\{\log(n)+1\}]}{\delta^2\Gamma[\beta+1, \epsilon\{\log(n)+1\}]}\{\log(n)+1\}^{-s}.$$

Then choose $F = \frac{1}{(M+1)\delta^2\Gamma(\beta+1)}$. □

The following corollary of Proposition 5 will be useful in the proof of Theorem 7 below.

Corollary 5. *Suppose π satisfies assumptions A1 and A2. Then there exists $G > 0$ such that for any $s < n$ and $n \geq 4$ it holds*

$$C(n, s+1) \leq \frac{G\Gamma(2+\beta)2^s s}{\epsilon}E(\alpha^s)\log\{n/(1+\epsilon)\}^{-1},$$

Proof. By Proposition 5 we have

$$C(n, s+1) \leq \frac{Gs}{\epsilon^s}E(\alpha^s)\frac{\Gamma\{s+\beta+1, \epsilon\log(n)\}}{\log\{n/(1+\epsilon)\}^s}.$$

Note that

$$\frac{\Gamma\{s+\beta+1, \epsilon\log(n)\}}{\epsilon^s\log^s\{n/(1+\epsilon)\}} \leq \frac{\Gamma(2+\beta)}{\epsilon}\left[\frac{\log(n)}{\log\{n/(1+\epsilon)\}}\right]^{s-1}\log\{n/(1+\epsilon)\}^{-1}.$$

Moreover, since $\epsilon < 1$, we have $\log\{n/(1+\epsilon)\} \geq \frac{1}{2}\log(n)$ for any $n \geq 4$. Combining the inequalities above we obtain the desired result. □

D.3. Proof of Lemma 5

Proof. Assumptions $A1$ and $A2$ are immediately satisfied in all three cases discussed in the statement of the lemma. We thus focus on proving that $A3$ is satisfied, considering each of the three cases separately. Suppose first that the support of the density π is contained in $[0, c]$ with $c > 0$. Then

$$\int_0^\infty \alpha^s \pi(\alpha)\, \mathrm{d}\alpha \leq c^s\,.$$

Thus in this case assumption $A3$ is satisfied for any $\rho > 0$ because $c^s < D\rho^{-s}\Gamma(s+1)$ with $D = \max_{s\in\mathbb{N}} \frac{(c\rho)^s}{\Gamma(s+1)}$ for any $\rho > 0$. Suppose now the prior is given by a Generalized Gamma distribution, so that

$$\int_0^\infty \alpha^s \pi(\alpha)\, \mathrm{d}\alpha = \frac{p}{a^d \Gamma\left(\frac{d}{p}\right)} \int_0^\infty \alpha^{d+s-1} e^{-\left(\frac{\alpha}{a}\right)^p} \mathrm{d}\alpha\,.$$

The condition $p > 1$ implies that, for every fixed $\rho > 0$ and $a > 0$, there exists $k > 0$ such that $\rho\alpha \leq \left(\frac{\alpha}{a}\right)^p$ for any $\alpha \geq k$. Thus

$$\begin{aligned}\int_0^\infty \alpha^{d+s-1} e^{-\left(\frac{\alpha}{a}\right)^p} \mathrm{d}\alpha &\leq \int_0^k \alpha^{s+d-1} e^{-\left(\frac{\alpha}{a}\right)^p} \mathrm{d}\alpha + \int_k^\infty \alpha^{s+d-1} e^{-\rho\alpha}\, \mathrm{d}\alpha \\ &\leq k^{s+d-1} e^{-\left(\frac{k}{a}\right)^p} + \rho^{-d-s}\Gamma(s+d).\end{aligned}$$

Also,

$$\begin{aligned}\int_0^\infty \alpha^s \pi(\alpha)\, \mathrm{d}\alpha \leq & \frac{p}{a^d \Gamma\left(\frac{d}{p}\right)} \Gamma(s+d) \left[\frac{k^{s+d-1} e^{-\left(\frac{k}{a}\right)^p}}{\Gamma(s+d)} + \rho^{-d-s}\right] \leq \\ &\leq D\rho^{-s}\Gamma(s+d),\end{aligned}$$

with $D = \max_{s\in\mathbb{N}} \frac{p}{a^d \Gamma\left(\frac{d}{p}\right)} \left[\frac{k^{s+d-1} e^{-\left(\frac{k}{a}\right)^p} \rho^s}{\Gamma(s+d)} + \rho^{-d}\right]$, so that also in this case assumption $A3$ is satisfied for any $\rho > 0$. Finally, in the case of Gamma distribution we get

$$\int_0^\infty \alpha^s \pi(\alpha)\, \mathrm{d}\alpha = \frac{\Gamma(\nu+s)}{\Gamma(\nu)} \rho^{-s},$$

and assumption $A3$ holds with ρ high enough, as desired. □

D.4. Proof of Lemma 6

Proof. Consider $R(n,s)$ as in (5.9). Taking the expectation with respect to the data generating distribution we have

$$\begin{aligned} E\{R(n,s)\} &= \sum_{A\in\rho_s(n)} \frac{\prod_{j=1}^s (a_j-1)!}{(n-1)!} E\left\{\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})}\right\} \\ &= \sum_{\boldsymbol{a}\in\mathcal{I}_s(n)} \binom{n}{a_1\cdots a_j} \frac{\prod_{j=1}^s (a_j-1)!}{s!(n-1)!} E\left\{\frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right\} \\ &= \sum_{\boldsymbol{a}\in\mathcal{I}_s(n)} \frac{n}{s!\prod_{j=1}^s a_j} E\left\{\frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right\}. \end{aligned}$$

□

D.5. Proof of Theorem 7

In order to prove Theorem 7, we first need the following results.

Lemma 12. *Let k and p_0 be as in (5.11) and $x_1 = \cdots = x_n = \theta^*$ for some $\theta^* \in \mathbb{R}$. Then*

$$\frac{\prod_{j=1}^s m(x_{A_j})}{m(x_{1:n})} = \left\{\frac{n+1}{\prod_{j=1}^s (a_j+1)}\right\}^{1/2} \exp\left\{\frac{\theta^{*2}}{2}\left(-\frac{n^2}{n+1} + \sum_{j=1}^s \frac{a_j^2}{a_j+1}\right)\right\} < \left(\frac{n}{\prod_{j=1}^s a_j}\right)^{1/2},$$

for any $s = 1,\ldots,n$ and any partition $A = \{A_1,\ldots,A_s\} \in \rho_s(n)$.

Proof. The equality follows after writing down the marginal likelihood of x_{A_j} as

$$m(x_{A_j}) = (a_j+1)^{-1/2} p_0(\theta^*)^{a_j} \exp\left\{\frac{\theta^{*2}}{2}\frac{a_j^2}{a_j+1}\right\},$$

and then computing the resulting expression for $m(x_{1:n})^{-1}\prod_{j=1}^s m(x_{A_j})$. The inequality follows from

$$\frac{n+1}{\prod_{j=1}^s (a_j+1)} \le \frac{n}{\prod_{j=1}^s a_j},$$

and

$$\begin{aligned} -\frac{n^2}{n+1} + \sum_{j=1}^s \frac{a_j^2}{a_j+1} &= n - \frac{n^2}{n+1} + \sum_{j=1}^s \left(\frac{a_j^2}{a_j+1} - a_j\right) = \frac{n}{n+1} - \sum_{j=1}^s \frac{a_j}{a_j+1} = \\ &= \sum_{j=1}^s a_j \left(\frac{1}{n+1} - \frac{1}{a_j+1}\right) \le 0. \end{aligned}$$

□

Lemma 13. *For any $p > 1$ and for any integers $s \ge 2$ and $n \ge s$ it holds*

$$\sum_{\boldsymbol{a}\in\mathcal{I}_s(n)} \left(\frac{n}{\prod_{j=1}^s a_j}\right)^p < C_p^{s-1},$$

where the sum runs over $\mathcal{F}_s(n) = \{\boldsymbol{a} \in \{1,\dots,n\}^s : \sum a_i = n\}$ *and* $C_p = 2^p\zeta(p)$, *with* $\zeta(p) = \sum_{a=1}^{\infty} \frac{1}{a^p} < \infty$.

Proof. We prove the result by induction. Consider the base case $s = 2$. By the strict convexity of $x \mapsto x^p$ for $p > 1$ we have

$$\sum_{\boldsymbol{a}\in\mathcal{F}_2(n)} \left(\frac{n}{a_1 a_2}\right)^p = \sum_{a=1}^{n-1} \left\{\frac{n}{a(n-a)}\right\}^p = 2^p \sum_{a=1}^{n-1} \left(\frac{1}{2}\frac{1}{a} + \frac{1}{2}\frac{1}{n-a}\right)^p < 2^p \sum_{a=1}^{n-1} \frac{1}{a^p} < C_p,$$

for any $n \geq 2$. For the induction step, assume that for some $s \geq 3$ we have

$$\sum_{\boldsymbol{a}\in\mathcal{F}_{s-1}(n)} \left(\frac{n}{\prod_{j=1}^{s-1} a_j}\right)^2 < C_p^{s-2}$$

for all $n \geq s-1$. Then

$$\begin{aligned}
\sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \left(\frac{n}{\prod_{j=1}^{s} a_j}\right)^p &= \sum_{a_s=1}^{n-s+1} \sum_{(a_1,\dots,a_{s-1})\in\mathcal{F}_{s-1}(n-a_s)} \left(\frac{n}{\prod_{j=1}^{s} a_j}\right)^p \\
&= \sum_{a_s=1}^{n-s+1} \left\{\frac{n}{(n-a_s)a_s}\right\}^p \sum_{(a_1,\dots,a_{s-1})\in\mathcal{F}_{s-1}(n-a_s)} \left(\frac{n-a_s}{\prod_{j=1}^{s-1} a_j}\right)^p \\
&\leq C_p^{s-2} \sum_{a_s=1}^{n-s+1} \left\{\frac{n}{(n-a_s)a_s}\right\}^p < C_p^{s-1},
\end{aligned}$$

and thus the thesis follows by induction. □

In the following we will drop the subscript in C_p when the value of p is clear from the context, thus denoting $C = C_p$.

of Theorem 7. First we study $R(n,s)$ as defined in (5.9). Since all the observations are almost surely equal, we have

$$R(n,s) = \sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \frac{n}{s!\prod_{j=1}^{s} a_j} \frac{\prod_{j=1}^{s} m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}.$$

Thus, applying Lemma 12 and then Lemma 13 with $p = 3/2$, the constant $C = 2^{\frac{3}{2}}\zeta\left(\frac{3}{2}\right) < 8$ is such that

$$R(n,s) < \frac{1}{s!} \sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \left(\frac{n}{\prod_{j=1}^{s} a_j}\right)^{3/2} < \frac{C^{s-1}}{s!}.$$

From Corollary 5 we have

$$C(n,s+1) \leq \frac{G\Gamma(2+\beta)2^s s}{\epsilon} E(\alpha^s)\log\{n/(1+\epsilon)\}^{-1}, \quad n \geq 4\,. \tag{5.16}$$

Thus, combining the inequalities above with (5.9) and assumption $A3$ we have

$$\begin{aligned}\sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n=s+1|X_{1:n})}{\mathrm{pr}(K_n=1|X_{1:n})} &= \sum_{s=1}^{n-1} C(n,s+1)R(n,s+1)\\ &\leq \frac{DG\Gamma(2+\beta)}{\epsilon\log\{n/(1+\epsilon)\}}\underbrace{\sum_{s=1}^{n-1}\frac{s(2C)^s\rho^{-s}\Gamma(\nu+s+1)}{(s+1)!}}_{<\infty}\to 0 \qquad \text{as } n\to\infty\,,\end{aligned} \tag{5.17}$$

where the finiteness follows from $\rho > 16 > 2C$. Then we conclude applying a variation of Lemma 4 with equalities and limits in probability replaced by almost sure equalities and limits (the proof of Lemma 4 extends trivially to that case). □

D.6. Proof of Lemma 7 and Theorem 8

of Lemma 7. First note that under the setting of (5.3) with (k, p_0) as in (5.12) we have

$$\begin{aligned}m(x_{1:n}) &= \int_0^\infty \prod_{i=1}^n k(x_i\mid\theta)e^{-\theta}\,\mathrm{d}\theta = \int_0^\infty \frac{\theta^{\sum_{i=1}^n x_i}e^{-(n+1)\theta}}{\prod_{i=1}^n x_i!}\,\mathrm{d}\theta\\ &= \frac{\Gamma(\sum_{i=1}^n x_i+1)}{\prod_{i=1}^n x_i!}(n+1)^{-\left(1+\sum_{i=1}^n x_i\right)}.\end{aligned}$$

Thus, for any integer $s \leq n-1$ and $\{A_1,\ldots,A_s\}\in\rho_s(n)$ we want to study

$$E\left\{\frac{\prod_{j=1}^s m(X_{A_j^a})}{m(X_{1:n})}\right\} = \frac{n+1}{\prod_{j=1}^s(a_j+1)}E\left\{\frac{\prod_{j=1}^s\Gamma\left(Z_j+1\right)}{\Gamma\left(\sum_{j=1}^s Z_j+1\right)}\prod_{j=1}^s\left(\frac{n+1}{a_j+1}\right)^{Z_j}\right\},$$

where $Z_j=\sum_{i\in A_j^a}X_i$. Since $Z_j\sim\mathrm{Po}(\theta^* a_j)$ independently for $j=1,\ldots,s$, by the integral representation of the multivariate Beta function and Tonelli's Theorem we get

$$\begin{aligned}E\left\{\frac{\prod_{j=1}^s\Gamma\left(Z_j+1\right)}{\Gamma\left(\sum_{j=1}^s Z_j+1\right)}\prod_{j=1}^s\left(\frac{n+1}{a_j+1}\right)^{Z_j}\right\} &= E\left\{\int_{\Delta_{s-1}}\prod_{j=1}^s t_j^{Z_j}\left(\frac{n+1}{a_j+1}\right)^{Z_j}\right\}\mathrm{d}t_j\\ &= \int_{\Delta_{s-1}}\prod_{j=1}^s E\left\{\left(t_j\frac{n+1}{a_j+1}\right)^{Z_j}\right\}\mathrm{d}t_j\,,\end{aligned}$$

where Δ_s is the $(s-1)$-dimensional probability simplex. By the expression for the probability-generating function of Poisson distributions and simple manipulations we have

$$\begin{aligned}E\left\{\left(t_j\frac{n+1}{a_j+1}\right)^{Z_j}\right\} &= \exp\left\{\theta^* t_j\frac{a_j}{a_j+1}(n+1)-\theta^* a_j\right\}\\ &\leq \exp\left\{\theta^* t_j(n+1)-\theta^* a_j\right\}.\end{aligned}$$

Then

$$E\left\{\frac{\prod_{j=1}^{s}\Gamma(Z_j+1)}{\Gamma(\sum_{j=1}^{s}Z_j+1)}\prod_{j=1}^{s}\left(\frac{n+1}{a_j+1}\right)^{Z_j}\right\} \leq e^{-\theta^* n}\int_{\Delta_{s-1}}\prod_{j=1}^{s}e^{\theta^* t_j(n+1)}\,\mathrm{d}t_j \leq e^{\theta^*}\int_{\Delta_{s-1}}\mathrm{d}t_j \leq e^{\theta^*}.$$

□

D.7. Proof of Theorem 8

Proof. By Lemma 6 and 7 we have

$$E\{R(n,s)\} \leq \frac{e^{\theta^*}}{s!}\sum_{\boldsymbol{a}\in\mathcal{F}_s(n)}\left(\frac{n}{\prod_{j=1}^{s}a_j}\right)^2,$$

with $R(n,s)$ defined as in (5.9). Thus, Lemma 13 with $p=2$ implies that

$$E\{R(n,s)\} \leq e^{\theta^*}\frac{C^{s-1}}{s!}.$$

with $C = 4\zeta(2) < 7$. From Corollary 5 we have

$$C(n,s+1) \leq \frac{G\Gamma(2+\beta)2^s s}{\epsilon}E(\alpha^s)\log\{n/(1+\epsilon)\}^{-1}, \quad n \geq 4\,.$$

Thus, combining the inequalities above with (5.9) and assumption $A3$ we have

$$\begin{aligned} E\left\{\sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n=s+1|X_{1:n})}{\mathrm{pr}(K_n=1|X_{1:n})}\right\} &= \sum_{s=1}^{n-1}C(n,s+1)E\{R(n,s+1)\} \\ &\leq \frac{DGe^{\theta^*}\Gamma(2+\beta)}{\epsilon\log\{n/(1+\epsilon)\}}\underbrace{\sum_{s=1}^{n-1}\frac{s(2C)^s\rho^{-s}\Gamma(\nu+s+1)}{(s+1)!}}_{<\infty} \to 0 \qquad \text{as } n\to\infty\,, \end{aligned}$$

where finiteness follows from $\rho > 14 > 2C$. This implies that

$$\sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n=s+1|X_{1:n})}{\mathrm{pr}(K_n=1|X_{1:n})} \to 0$$

in L^1 and thus in $P^{(\infty)}$-probability as $n\to\infty$. □

D.8. Proof of Lemma 8 and Theorem 9

of Lemma 8. Note that $x_i \in (\theta - c, \theta + c)$ for all $i \in \{1, \dots, n\}$ if and only if $\theta \in (\max(x_{1:n}) - c, \min(x_{1:n}) + c)$. Thus

$$\begin{aligned} m(x_{1:n}) &= \frac{1}{(2c)^{n+1}} \int_\Theta \prod_{i=1}^n \mathbb{1}_{(\theta-c,\theta+c)}(x_i) \mathbb{1}_{(\theta^*-c,\theta^*+c)}(\theta) \mathrm{d}\theta \\ &= \frac{1}{(2c)^{n+1}} \int_\Theta \mathbb{1}_{(\max(x_{1:n})-c,\min(x_{1:n})+c)}(\theta) \mathbb{1}_{(\theta^*-c,\theta^*+c)}(\theta) \mathrm{d}\theta \\ &= \frac{2c - \{\max(x_{1:n}, \theta^*) - \min(x_{1:n}, \theta^*)\}}{(2c)^{n+1}} . \end{aligned}$$

□

Define $\mathrm{Range}(X_{1:n}) = \max(X_{1:n}) - \min(X_{1:n})$. Thus, Lemma 8 has an important corollary, that is stated after a technical lemma.

Lemma 14. *Let $A \subset \{1, \dots, n\}$ such that $|A| = a$, Then it holds:*

$$\frac{2c - \{\max(X_A, \theta^*) - \min(X_A, \theta^*)\}}{(2c)^{a+1}} \leq \frac{2c - \mathrm{Range}(X_A)}{(2c)^{a+1}}.$$

Proof. The result follows immediately from $\max(X_A, \theta^*) \geq \max(X_A)$ and $\min(X_A, \theta^*) \leq \min(X_A)$. □

Corollary 6. *In the setting of (5.3) with (f, k, p_0) as in (5.13), define $\Omega_n = \{x \in X^\infty \mid \max(x_{1:n}) \geq \theta^* \text{ and } \min(x_{1:n}) \leq \theta^*\}$. Then*

$$\frac{\prod_{j=1}^{s+1} m(X_{A_j})}{m(X_{1:n})} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \leq \frac{\prod_{j=1}^{s+1} \{2c - \mathrm{Range}(X_{A_j})\}}{(2c)^s \{2c - \mathrm{Range}(X_{1:n})\}}. \tag{5.18}$$

Proof. As regards the numerator, apply firstly Lemma 8 and then Lemma 14 to get

$$m(X_{A_j}) = \frac{2c - \{\max(X_{A_j}, \theta^*) - \min(X_{A_j}, \theta^*)\}}{(2c)^{a_j+1}} \leq \frac{2c - \mathrm{Range}(X_{A_j})}{(2c)^{a_j+1}}, \quad j = 1, \dots, s+1 .$$

Apply Lemma 8 to $m(x_{1:n})$ for any $x \in \Omega_n$, to get

$$\begin{aligned} m(X_{1:n}) \mathbb{1}_{\Omega_n}(X_{1:\infty}) &= \frac{2c - \{\max(X_{1:n}, \theta^*) - \min(X_{1:n}, \theta^*)\}}{(2c)^{n+1}} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \\ &= \frac{2c - \{\max(X_{1:n}) - \min(X_{1:n})\}}{(2c)^{n+1}} \mathbb{1}_{\Omega_n}(X_{1:\infty}), \end{aligned}$$

as desired. □

The lemma below shows that, in order to prove Theorem 9, it is sufficient to show $\mathbb{1}_{\Omega_n}(X_{1:\infty}) \sum_{s=1}^{n-1} \frac{\mathrm{pr}(K_n = s+1 | X_{1:n})}{\mathrm{pr}(K_n = 1 | X_{1:n})} \to 0$ in $P^{(\infty)}$-probability.

Lemma 15. *Consider f as in (5.13) and define $\Omega_n = \{x \in X^\infty \mid \max(x_{1:n}) \geq \theta^* \text{ and } \min(x_{1:n}) \leq \theta^*\}$. Let $\{Y_n\}$ be a sequence of positive random variables. Thus, $Y_n \mathbb{1}_{\Omega_n}(X_{1:\infty}) \to 0$ in $P^{(\infty)}$-probability implies $Y_n \to 0$ in $P^{(\infty)}$-probability.*

Proof. First of all, by definition of f we have

$$\max(X_{1:n}) \to \theta^* + c, \quad \min(X_{1:n}) \to \theta^* - c$$

almost surely with respect to $P^{(\infty)}$ as $n \to \infty$. Then $P^{(\infty)}(\Omega_n) \to 1$, as $n \to \infty$, by definition of Ω_n. Thus, fix $\epsilon > 0$ and notice that

$$P^{(\infty)}\left(Y_n > \epsilon\right) = P^{(\infty)}\left(\{Y_n > \epsilon\} \cap \Omega_n\right) + P^{(\infty)}\left(\{Y_n > \epsilon\} \cap \Omega_n^c\right).$$

The first term on the right hand side goes to 0, since $Y_n \mathbb{1}_{\Omega_n}(X_{1:\infty}) \to 0$ in $P^{(\infty)}$-probability, while the second vanishes because $P^{(\infty)}(\Omega_n^c) \to 0$, both as $n \to \infty$. □

Combining Corollary 6 and Lemma 15 we are ready to prove Theorem 9.

of Theorem 9. From Corollary 6 we have

$$\frac{\prod_{j=1}^{s+1} m(X_{A_j})}{m(X_{1:n})} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \leq \frac{\prod_{j=1}^{s+1}\{2c - \text{Range}(X_{A_j})\}}{(2c)^s\{2c - \text{Range}(X_{1:n})\}}.$$

Note that $\{2c - \text{Range}(X_{A_j})\}/(2c) \sim \text{Beta}(2, a_j - 1)$ independently for $j = 1, \ldots, s$. Moreover, recall that if $Z \sim \text{Beta}(\alpha, \beta)$ then for $p > -\alpha$:

$$E(Z^p) = \frac{\Gamma(\alpha + p)\Gamma(\alpha + \beta)}{\Gamma(\alpha + p + \beta)\Gamma(\alpha)}.$$

Thus, by Hölder's inequality (with exponents 3 and 3/2) we get

$$\begin{aligned} E\left\{\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\right\} &\leq E\left[\prod_{j=1}^{s} m(X_{A_j})^3\right]^{1/3} E\left[m(X_{1:n})^{-3/2}\right]^{2/3} \\ &= \left\{\frac{\Gamma(5)}{\Gamma(2)}\right\}^{s/3}\left\{\frac{\Gamma(1/2)}{\Gamma(2)}\right\}^{2/3}\left\{\prod_{j=1}^{s}\frac{\Gamma(1+a_j)}{\Gamma(a_j+4)}\right\}^{1/3}\left\{\frac{\Gamma(1+n)}{\Gamma(n-1/2)}\right\}^{2/3}. \end{aligned}$$

By the recursive definition of Gamma function and recalling that $\Gamma(1/2) = \pi^{1/2}$, the upper bound above becomes

$$\begin{aligned} E\left\{\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\right\} &\leq 24^{s/3}\pi^{1/3}\left\{\prod_{j=1}^{s}\frac{\Gamma(1+a_j)}{\Gamma(a_j+4)}\right\}^{1/3}\left\{\frac{\Gamma(1+n)}{\Gamma(n-1/2)}\right\}^{2/3} \\ &= 24^{s/3}\pi^{1/3}\left\{\prod_{j=1}^{s}\frac{1}{(a_j+3)(a_j+2)(a_j+1)}\right\}^{1/3}\left\{\frac{(n-1/2)\Gamma(1+n)}{\Gamma(n+1/2)}\right\}^{2/3}. \end{aligned}$$

Moreover, exploiting again the recursive definition of the Gamma function, Gautschi's Inequality, i.e. $\frac{\Gamma(1+n)}{\Gamma(n+1/2)} \leq (n+1)^{1/2}$, and $(n+1)/(a_j+1) < n/a_j$, we have

$$E\left\{\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\right\} \leq 24^{s/3}K\left\{\prod_{j=1}^{s}\frac{(n+1)^3}{(a_j+1)^3}\right\}^{1/3} \leq 24^{s/3}K\left(\frac{n^3}{\prod_{j=1}^{s} a_i^3}\right)^{1/3} = 24^{s/3}K\frac{n}{\prod_{j=1}^{s} a_j}.$$

Thus, applying Lemma 13 with $p = 2$ and $C = 4\zeta(2) < 7$ we get

$$E\{R(n,s)\} \leq \frac{24^{s/3}K}{s!} \sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \left(\frac{n}{\prod_{j=1}^s a_j}\right)^2 < \frac{C^{s-1}24^{s/3}K}{s!}.$$

With a reasoning similar to the proof of Theorem 8, noticing that $\rho \geq 38 > 24^{1/3} \times 2C$, we get

$$\mathbb{1}_{\Omega_n}(X_{1:\infty}) \sum_{s=1}^{n-1} \frac{\mathrm{pr}(K_n = s+1|X_{1:n})}{\mathrm{pr}(K_n = 1|X_{1:n})} \to 0$$

in $P^{(\infty)}$-probability. Lemma 15 with $Y_n = \sum_{s=1}^{n-1} \frac{\mathrm{pr}(K_n=s+1|X_{1:n})}{\mathrm{pr}(K_n=1|X_{1:n})}$ concludes the proof. □

D.9. Proof of Theorem 10

Through a linear rescaling, without loss of generality we may assume $[a,b] = [-c,c]$ and $\theta^* = 0$. In the following we denote by $X_{(r)}$ the $r-$th order statistic of the vector $X_{1:n}$.

We assume to be in the well-specified framework and we rewrite the assumptions $B1$, $B2$, $B3$ and 0 belonging to the support of P_0 as

$C1$. $\exists\, m, M$ such that $0 < m \leq g(x) \leq M < \infty$ for any $x \in [-c,c]$;

$C2$. g is differentiable on $(-c,c)$ and $\exists\, R$ such that $|\frac{g'(x)}{g(x)}| \leq R < \infty$ for any $x \in (-c,c)$;

$C3$. $\exists\, U > 0$ such that $h(y) = p_0(y) + p_0(-y) \leq U$ for any $y \in [0,2c]$;

$C4$. $\exists\, L > 0$ such that $p_0(\theta) \geq L$ for any θ in a neighborhood of 0.

We first study the following

$$\begin{aligned}\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})} &= \frac{\int_{\mathbb{R}^s} \prod_{j=1}^s \prod_{i\in A_j} g(X_i - \theta_j)\, p_0(\theta_j)\mathrm{d}\theta_j}{\int_{\mathbb{R}} \prod_{i=1}^n g(X_i - \theta)\, p_0(\theta)\mathrm{d}\theta} \\ &= \frac{\int_{\mathbb{R}^s} \prod_{j=1}^s \prod_{i\in A_j} \frac{g(X_i-\theta_j)}{g(X_i)} p_0(\theta_j)\, \mathrm{d}\theta_j}{\int_{\mathbb{R}} \prod_{i=1}^n \frac{g(X_i-\theta)}{g(X_i)} p_0(\theta)\, \mathrm{d}\theta}. \end{aligned} \tag{5.19}$$

The next lemma provides a bound for the numerator in (5.19).

Lemma 16. *Under $X_{1:\infty} \sim P^{(\infty)}$, we have*

$$E\left\{\int_{\mathbb{R}^s} \prod_{j=1}^s \prod_{i\in A_j} \frac{g(X_i-\theta_j)}{g(X_i)} p_0(\theta_j)\, \mathrm{d}\theta_j\right\} \leq \left(\frac{U}{m}\right)^s \prod_{j=1}^s \frac{1}{a_j+1}.$$

with m defined in $C1$.

Proof. Taking the expectation under $P^{(\infty)}$ we have

$$E\left\{\int_{\mathbb{R}^s} \prod_{j=1}^s \prod_{i\in A_j} \frac{g(X_i-\theta_j)}{g(X_i)} p_0(\theta_j)\, \mathrm{d}\theta_j\right\} = \int_{\mathbb{R}^s} \int_{[-c,c]^n} \prod_{j=1}^s \prod_{i\in A_j} g(x_i - \theta_j) p_0(\theta_j)\, \mathrm{d}x_i\, \mathrm{d}\theta_j, \tag{5.20}$$

by Tonelli's Theorem. By the change of variables $z = x - \theta_j$, we have

$$\int_{-c}^{c} g(x-\theta_j)\mathbb{1}_{[\theta_j-c,\theta_j+c]}(x)\,\mathrm{d}x = \int_{-c-\theta_j}^{c-\theta_j} g(z)\mathbb{1}_{[-c,c]}(z)\,\mathrm{d}z.$$

If $\theta_j > 0$, then

$$\begin{aligned}\int_{-c-\theta_j}^{c-\theta_j} g(z)\mathbb{1}_{[-c,c]}(z)\,\mathrm{d}z &= \mathbb{1}_{[0,2c]}(\theta_j)\int_{-c}^{c-\theta_j} g(z)\,\mathrm{d}z \\ &= \mathbb{1}_{[0,2c]}(\theta_j)\left(1-\int_{c-\theta_j}^{c} g(z)\,\mathrm{d}z\right) \leq \mathbb{1}_{[0,2c]}(|\theta_j|)\,(1-m|\theta_j|)\,.\end{aligned}$$

Similarly, if $\theta_j < 0$ we get

$$\begin{aligned}\int_{-c-\theta_j}^{c-\theta_j} g(z)\mathbb{1}_{[-c,c]}(z)\,\mathrm{d}z &= \mathbb{1}_{[-2c,0]}(\theta_j)\int_{-c-\theta_j}^{c} g(z)\,\mathrm{d}z \\ &= \mathbb{1}_{[-2c,0]}(\theta_j)\left(1-\int_{-c}^{-c-\theta_j} g(z)\,\mathrm{d}z\right) \leq \mathbb{1}_{[0,2c]}(|\theta_j|)\,(1-m|\theta_j|)\,.\end{aligned}$$

Thus, we proved

$$\int_{-c}^{c} g(x-\theta_j)\mathbb{1}_{[\theta_j-c,\theta_j+c]}(x)\,\mathrm{d}x \leq \mathbb{1}_{[0,2c]}(|\theta_j|)\,(1-m|\theta_j|)\,, \quad j=1,\ldots,s,$$

that implies

$$\prod_{j=1}^{s}\prod_{i\in A_j}\int_{-c}^{c} g(x-\theta_j)\mathbb{1}_{[\theta_j-c,\theta_j+c]}(x)\,\mathrm{d}x \leq \prod_{j=1}^{s}\mathbb{1}_{[0,2c]}(|\theta_j|)\,(1-m|\theta_j|)\,.$$

Considering h defined in $C3$, we have

$$\int_{\mathbb{R}}\mathbb{1}_{[0,2c]}(|\theta_j|)\,(1-m|\theta_j|)\,p_0(\theta_j)\,\mathrm{d}\theta_j = \int_0^{2c}(1-m|\theta_j|)\,h(\theta_j)\,\mathrm{d}\theta_j, \quad j=1,\ldots,s.$$

Directly from (5.20) we get

$$\begin{aligned}E\left\{\int_{\mathbb{R}^s}\prod_{j=1}^{s}\prod_{i\in A_j}\frac{g(X_i-\theta_j)}{g(X_i)}p_0(\theta_j)\,\mathrm{d}\theta_j\right\} &= \int_{\mathbb{R}^s}\int_{[-c,c]^n}\prod_{j=1}^{s}\prod_{i\in A_j} g(x_i-\theta_j)p_0(\theta_j)\,\mathrm{d}x_i\,\mathrm{d}\theta_j \\ &\leq \prod_{j=1}^{s}\int_0^{2c}(1-m|\theta_j|)\,h(\theta_j)\,\mathrm{d}\theta_j.\end{aligned} \tag{5.21}$$

With U as defined in $C3$, we have

$$\int_0^{2c}(1-my)^{a_j}h(y)\,\mathrm{d}y \leq U\int_0^{2c}(1-my)^{a_j}\,\mathrm{d}y.$$

Now consider the change of variables $u = 1-my$ and compute

$$\int_0^{2c}(1-my)^{a_j}\,\mathrm{d}y = \frac{1}{m}\int_{1-2mc}^{1}u^{a_j}\,\mathrm{d}u = \frac{1-(1-2mc)^{a_j+1}}{m(a_j+1)} \leq \frac{1}{m(a_j+1)}.$$

Finally, through (5.21), we have

$$E\left\{\int_{\mathbb{R}^s}\prod_{j=1}^{s}\prod_{i\in A_j}\frac{g(X_i-\theta_j)}{g(X_i)}p_0(\theta_j)\,\mathrm{d}\theta_j\right\}\leq\prod_{j=1}^{s}\int_0^{2c}(1-m|\theta_j|)\,h(\theta_j)\,\mathrm{d}\theta_j$$
$$\leq\left(\frac{U}{m}\right)^s\prod_{j=1}^{s}\frac{1}{a_j+1},$$

as desired. □

As regards the denominator of (5.19), we start by studying the asymptotic behaviour of the maximum.

Lemma 17. *Let $X_{(n)}=\max(X_{1:n})$. Then*

$$Y_n=\min\left[1,n\sqrt{\log n}\{c-X_{(n)}\}\right]\to 1$$

in $P^{(\infty)}$-probability as $n\to\infty$.

Proof. We have to prove that $\forall\epsilon>0$

$$\mathrm{pr}\left(|1-Y_n|>\epsilon\right)\to 0$$

as $n\to\infty$, where pr is evaluated with respect to $P^{(\infty)}$. By definition we have

$$\mathrm{pr}(1-Y_n>\epsilon)=\mathrm{pr}\left[n\sqrt{\log n}\{c-X_{(n)}\}\leq 1-\epsilon\right]=\mathrm{pr}\left[X_{(n)}\geq c-\frac{1-\epsilon}{n\sqrt{\log n}}\right]$$
$$=1-\left\{1-\int_{c-\frac{1-\epsilon}{n\sqrt{\log n}}}^{c}g(x)\,\mathrm{d}x\right\}^n.$$

Thus, by $C1$ we have that $\int_{c-\frac{1-\epsilon}{n\sqrt{\log n}}}^{c}g(x)\,\mathrm{d}x\leq\frac{M(1-\epsilon)}{n\sqrt{\log n}}$, so that

$$\mathrm{pr}(1-Y_n>\epsilon)\leq 1-\left(1-\frac{M(1-\epsilon)}{n\sqrt{\log n}}\right)^n=1-e^{-\frac{M(1-\epsilon)}{\sqrt{\log n}}+n\,\mathrm{o}\left(\frac{1}{n\sqrt{\log n}}\right)}\to 0,$$

by the Taylor expansion of the logarithmic function. □

Lemma 18. *For any $x_{1:n}\in[-c,c]^n$ it holds*

$$\prod_{i=1}^{n}\frac{g(x_i-\theta)}{g(x_i)}\geq e^{-R}\mathbb{1}_{[0,\frac{1}{n}]}(|\theta|)\mathbb{1}_{[x_{(n)}-c,x_{(1)}+c]}(\theta).$$

with R defined in $C2$.

Proof. Define $p(x):=\log g(x)$, with $x\in[-c,c]$, so that $p'(x)=\frac{g'(x)}{g(x)}$. By C2 and the Fundamental Theorem of Calculus

$$|p(y)-p(x)|=\left|\int_x^y p'(t)\,\mathrm{d}t\right|\leq\int_x^y\left|\frac{g'(t)}{g(t)}\right|\mathrm{d}t\leq R|y-x|,\quad -c<x\leq y<c.$$

Thus, we have

$$\frac{g(x-\theta)}{g(x)}=e^{p(x-\theta)-p(x)}=e^{-(p(x)-p(x-\theta))}\geq e^{-R|\theta|},\quad x\in[-c,c].$$

Finally we get

$$\prod_{i=1}^{n} \frac{g(x_i-\theta)}{g(x_i)} \geq e^{-Rn|\theta|}\mathbb{1}_{[x_{(n)}-c,x_{(1)}+c]}(\theta) \geq e^{-Rn|\theta|}\mathbb{1}_{[0,\frac{1}{n}]}(|\theta|)\mathbb{1}_{[x_{(n)}-c,x_{(1)}+c]}(\theta)$$
$$\geq e^{-R}\mathbb{1}_{[0,\frac{1}{n}]}(|\theta|)\mathbb{1}_{[x_{(n)}-c,x_{(1)}+c]}(\theta).$$

□

Lemma 19. *There exists $N \in \mathbb{N}$ and $K > 0$ such that for all $n \geq N$ and $x_{1:n} \in [-c, c]^n$ it holds*

$$\int_{\mathbb{R}} \prod_{i=1}^{n} \frac{g(x_i-\theta)}{g(x_i)} p_0(\theta)\,\mathrm{d}\theta \geq \frac{KY_n}{n\sqrt{\log n}},$$

with Y_n defined as in Lemma 17.

Proof. Notice that, by $C4$, there exists $N \in \mathbb{N}$ such that $p_0(\theta) \geq L$ for any $\theta \in \left[-\frac{1}{N}, 0\right]$. Thus, applying Lemma 18 and considering $n \geq N$, we get

$$\int_{\mathbb{R}} \prod_{i=1}^{n} \frac{g(x_i-\theta)}{g(x_i)} p_0(\theta)\,\mathrm{d}\theta \geq e^{-R} \int_{\mathbb{R}} \mathbb{1}_{[0,\frac{1}{n}]}(|\theta|)\mathbb{1}_{[x_{(n)}-c,x_{(1)}+c]}(\theta)\, p_0(\theta)\,\mathrm{d}\theta$$
$$\geq e^{-R} \int_{-\frac{1}{n}}^{0} \mathbb{1}_{\{x_{(n)} \leq \theta + c\}}\, p_0(\theta)\,\mathrm{d}\theta \geq L e^{-R} \min\left\{\frac{1}{n}, c - X_{(n)}\right\}.$$

with L defined in $C4$. Thus, multiplying both the numerator and the denominator by $n \log n$, with $n \geq N$, we have

$$\int_{\mathbb{R}} \prod_{i=1}^{n} \frac{g(x_i-\theta)}{g(x_i)} p_0(\theta)\,\mathrm{d}\theta \geq 2Le^{-R} \min\left\{\frac{1}{n}, c - X_{(n)}\right\}$$
$$\geq \frac{K \min\left[1, n\sqrt{\log n}\{c - X_{(n)}\}\right]}{n\sqrt{\log n}} = \frac{KY_n}{n\sqrt{\log n}},$$

with $K = 2Le^{-R}$ and $Y_n = \min\left[1, n\sqrt{\log n}\{c - X_{(n)}\}\right]$. □

D.10. Proof of Theorem 10

We start with a technical Lemma.

Lemma 20. *Define $Y_n = \min\left[1, n\sqrt{\log n}\{c - X_{(n)}\}\right]$ and let Z_n be such that*

$$Z_n \mathbb{1}_{(1/2,1]}(Y_n) \to 0$$

in $P^{(\infty)}$-probability as $n \to \infty$. Then $Z_n \to 0$ in $P^{(\infty)}$-probability as $n \to \infty$.

Proof. By assumption $P^{(\infty)}\left(\mathbb{1}_{(1/2,1]}(Y_n)Z_n > \epsilon\right) \to 0$ for any $\epsilon > 0$, while $P^{(\infty)}\left(Y_n > 1/2\right) \to 1$ by Lemma 17, both as $n \to \infty$. Thus, we have

$$P^{(\infty)}\left(Z_n > \epsilon\right) \leq P^{(\infty)}\left(\{Z_n > \epsilon\} \cap \{Y_n > 1/2\}\right) + P^{(\infty)}\left(Y_n \leq 1/2\right) \to 0$$

as $n \to \infty$. □

Lemma 21. *Consider the setting of (5.3) with (f, k, p_0) as in Theorem 10. Moreover, assume $\pi(\alpha)$ satisfies assumptions A1, A2 and A3. Then, under $X_{1:\infty} \sim P^{(\infty)}$ we have*

$$E\left\{\mathbb{1}_{[1/2,1]}(Y_n)\sum_{s=1}^{n-1}\frac{pr(K_n = s+1 \,|\, X_{1:n})}{pr(K_n = 1 \,|\, X_{1:n})}\right\} \to 0$$

as $n \to \infty$.

Proof. By Lemma 19, for n large enough we have

$$\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})} \leq \frac{n\sqrt{\log n}}{KY_n}\int_{\mathbb{R}^s}\prod_{j=1}^{s}\prod_{i\in A_j}\frac{g(X_i-\theta_j)}{g(X_i)}p_0(\theta_j)\,\mathrm{d}\theta_j, \tag{5.22}$$

and thus

$$\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\mathbb{1}_{[1/2,1]}(Y_n) \leq \frac{2n\sqrt{\log n}}{K}\int_{\mathbb{R}^s}\prod_{j=1}^{s}\prod_{i\in A_j}\frac{g(X_i-\theta_i)}{g(X_i)}p_0(\theta_j)\,\mathrm{d}\theta_j. \tag{5.23}$$

By Lemma 16 we get

$$E\left\{\int_{\mathbb{R}^s}\prod_{j=1}^{s}\prod_{i\in A_j}\frac{g(X_i-\theta_j)}{g(X_i)}p_0(\theta_j)\,\mathrm{d}\theta_j\right\} \leq \left(\frac{U}{m}\right)^s\prod_{j=1}^{s}\frac{1}{a_j+1},$$

and finally

$$E\left\{\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\mathbb{1}_{[1/2,1]}(Y_n)\right\} \leq \frac{2(U/m)^s\sqrt{\log n}}{K}\frac{n}{\prod_{j=1}^{s}(a_j+1)}.$$

Recalling the definition of $R(n,s)$ in (5.9) and applying Lemma 6, we have

$$\begin{aligned} E\left\{\mathbb{1}_{[1/2,1]}(Y_n)R(n,s)\right\} &= \sum_{\boldsymbol{a}\in\mathcal{F}_s(n)}\frac{n}{s!\prod_{j=1}^{s}a_j}E\left\{\frac{\prod_{j=1}^{s} m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\mathbb{1}_{[1/2,1]}(Y_n)\right\} \\ &\leq \frac{2}{K}\frac{(U/m)^s\sqrt{\log n}}{s!}\sum_{\boldsymbol{a}\in\mathcal{F}_s(n)}\left\{\frac{n}{\prod_{j=1}^{s}(a_j+1)}\right\}^2. \end{aligned}$$

Lemma 13 with $p=2$ implies that there exists a positive constant C such that

$$E\left\{\mathbb{1}_{[1/2,1]}(Y_n)R(n,s)\right\} \leq \frac{2}{K}\frac{C^{s-1}(U/m)^s\sqrt{\log n}}{s!}.$$

Moreover, from Corollary 5 and $A3$, we have

$$\begin{aligned} C(n,s+1) &\leq \frac{G\Gamma(2+\beta)2^s s}{\epsilon}E(\alpha^s)\log\{n/(1+\epsilon)\}^{-1} \\ &\leq \frac{DG\Gamma(2+\beta)2^s s}{\epsilon}\rho^{-s}\Gamma(\nu+s+1)\log\{n/(1+\epsilon)\}^{-1}, \quad n \geq 4\,. \end{aligned}$$

Thus, combining the inequalities above with (5.9), we have

$$E\left\{\mathbb{1}_{[1/2,1]}(Y_n)\sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n=s+1|X_{1:n})}{\mathrm{pr}(K_n=1|X_{1:n})}\right\}=\sum_{s=1}^{n-1}C(n,s+1)E\{\mathbb{1}_{[1/2,1]}(Y_n)R(n,s+1)\}$$
$$\leq \frac{2(U/m)DG\Gamma(2+\beta)\sqrt{\log n}}{K\epsilon\log\{n/(1+\epsilon)\}}\underbrace{\sum_{s=1}^{n-1}\frac{s(2CU/m)^s\rho^{-s}\Gamma(\nu+s+1)}{(s+1)!}}_{<\infty}\to 0 \qquad \text{as } n\to\infty,$$

as $n\to\infty$, where finiteness follows by taking ρ sufficiently large. □

of Theorem 10. By Lemma 21 it holds

$$\mathbb{1}_{[1/2,1]}(Y_n)\sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n=s+1\,|\,X_{1:n})}{\mathrm{pr}(K_n=1\,|\,X_{1:n})}\to 0$$

in $P^{(\infty)}$–probability as $n\to\infty$. The desired result then follows from Lemma 20 with $Z_n=\sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n=s+1\,|\,X_{1:n})}{\mathrm{pr}(K_n=1\,|\,X_{1:n})}$. □

D.11. Proof of Proposition 6

Proof. Under (5.3), for any $\epsilon>0$ we have

$$\mathrm{pr}(\alpha<\epsilon\mid X_1,\ldots,X_n)=\sum_{s=1}^{n}\mathrm{pr}(\alpha<\epsilon\mid K_n=s)\ \mathrm{pr}(K_n=s\mid X_1,\ldots,X_n)=$$
$$\geq \mathrm{pr}(\alpha<\epsilon\mid K_n=1)\ \mathrm{pr}(K_n=1\mid X_1,\ldots,X_n).$$

By Theorem 10, $\mathrm{pr}(K_n=1\mid X_1,\ldots,X_n)\to 1$ in $P^{(\infty)}$-probability as $n\to\infty$. Moreover, by Proposition 5 with $s=2$ we get

$$E(\alpha\mid K_n=1)\to 0,$$

as $n\to\infty$. It follows $\mathrm{pr}(\alpha<\epsilon\mid K_n=1)\to 1$ in $P^{(\infty)}$-probability as $n\to\infty$, as desired. □

Bibliography

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

Andrieu, C. and A. Doucet (2002). Particle filtering for partially observed Gaussian state space models. *Journal of the Royal Statistical Society: Series B 64*, 827–836.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*(6), 1152–1174.

Arellano-Valle, R. B. and A. Azzalini (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics 33*, 561–574.

Argiento, R., A. Cremaschi, and M. Vannucci (2020). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association 115*(529), 318–333.

Arnold, B. C. and R. J. Beaver (2000). Hidden truncation models. *Sankhyā: Series A 62*, 23–35.

Arnold, B. C., R. J. Beaver, A. Azzalini, N. Balakrishnan, A. Bhaumik, D. Dey, C. Cuadras, and J. M. Sarabia (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test 11*, 7–54.

Atkins, A., M. Niranjan, and E. Gerding (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science 4*, 120–137.

Azzalini, A. and A. Bacchieri (2010). A prospective combination of phase II and phase III in drug development. *Metron 68*, 347–369.

Azzalini, A. and A. Capitanio (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B 61*, 579–602.

Azzalini, A. and A. Capitanio (2014). *The Skew-normal and Related Families*. Cambridge University Press.

Azzalini, A. and A. Dalla Valle (1996). The multivariate skew-normal distribution. *Biometrika 83*, 715–726.

Bassetti, F., R. Casarin, and L. Rossini (2020). Hierarchical species sampling models. *Bayesian Analalysis 15*(3), 809–838.

Beraha, M., A. Guglielmi, and F. A. Quintana (2020). The semi-hierarchical Dirichlet Process and its application to clustering homogeneous distributions. *Prepint at arXiv: 2005.10287*.

Blei, D. M. and M. I. Jordan (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis 1*(1), 121–143.

Botev, Z. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B 79*, 125–148.

Bunge, J. and M. Fitzpatrick (1993). Estimating the number of species: A review. *Journal of the American Statistical Association 88*(421), 364–373.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodríguez (2019). Latent nested nonparametric priors (with discussion). *Bayesian Analysis 14*(4), 1303–1356.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *The Annals of Statistics 47*(1), 67–92.

Camerlenghi, F., A. Lijoi, and I. Prünster (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis 156*, 18–28.

Camerlenghi, F., A. Lijoi, and I. Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics 45*(4), 1062–1091.

Canale, A., R. Corradin, and B. Nipoti (2019). Importance conditional sampling for Bayesian nonparametric mixtures. *Preprint arXiv: 1906.08147*.

Carlin, B. P., N. G. Polson, and D. S. Stoffer (1992). A Monte Carlo approach to non-normal and nonlinear state-space modeling. *Journal of the American Statistical Association 87*, 493–500.

Chao, A. (1981). On estimating the probability of discovering a new species. *Annals of Statistics 9*(6), 1339–1342.

Chao, A. and S. M. Lee (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association 87*(417), 210–217.

Charalambides, C. A. (2002). *Enumerative Combinatorics*. Chapman and Hall/CRC.

Chib, S. and E. Greenberg (1998). Analysis of multivariate probit models. *Biometrika 85*, 347–361.

Chopin, N. and J. Ridgway (2017). Leave Pima indians alone: Binary regression as a benchmark for Bayesian computation. *Statistical Science 32*, 64–87.

Christensen, J. and L. Ma (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(1), 127–153.

Cifarelli, D. M. and E. Regazzini (1978). Problemi statistici non parametrici in condizioni di scambiabilita parziale e impiego di medie associative. Technical report, Quaderni Istituto Matematica Finanziaria dell'Universita di Torino.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (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.

Doucet, A., N. De Freitas, and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Doucet, A., S. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10*, 197–208.

Doucet, A. and A. M. Johansen (2009). A tutorial on particle filtering and smoothing: fifteen years later. *Handbook of Nonlinear Filtering 12*, 656–704.

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics*, pp. 223–273. Cambridge University Press.

Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika 106*, 765–779.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Efron, B. and T. Ronald (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika 63*(3), 435–447.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association 89*(425), 268–277.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Escobar, M. D. and M. West (1998). Computing nonparametric hierarchical models. In *Practical nonparametric and semiparametric Bayesian statistics*, pp. 1–22. Springer, New York, NY.

Ewens, W. J. (1990). Population genetics theory - the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177–227. Dordrecht: Springer.

Favaro, S., A. Lijoi, R. H. Mena, and I. Prünster (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.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*(2), 209–230.

de Finetti, Bruno (1938). Sur la condition d'equivalence partielle. *Actualitès Scientifiques et Industrielles 739*, 5–18.

Foti, N. J. and S. A. Williamson (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Trans. Pattern Anal. Mach. Intell. 37*(2), 359–371.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics 2*, 1360–1383.

Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics 27*(1), 143–158.

Ghosal, S. and A. W. Van der Vaart (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics 35*(2), 697–723.

Ghosal, S. and A. W. Van Der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

González-Farías, G., A. Domínguez-Molina, and A. K. Gupta (2004). Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference 126*, 521–534.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40*(3-4), 237–264.

Good, I. J. and G. H. Toulmin (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika 43*(1-2), 45–63.

Gordon, N. J., D. J. Salmond, and A. F. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F-radar and Signal Processing 140*, 107–113.

Gupta, A. K., M. A. Aziz, and W. Ning (2013). On some properties of the unified skew-normal distribution. *Journal of Statistical Theory and Practice 7*, 480–495.

Gupta, A. K., G. González-Farías, and J. A. Domínguez-Molina (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis 89*, 181–190.

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

Ishwaran, H. and M. Zarepour (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics 30*(2), 269–283.

Johndrow, J. E., A. Smith, N. Pillai, and D. B. Dunson (2019). Mcmc for imbalanced categorical data. *Journal of the American Statistical Association 114*, 1394–1403.

Julier, S. J. and J. K. Uhlmann (1997). New extension of the Kalman filter to nonlinear systems. In *Proceedings SPIE 3068, Signal Processing, Sensor Fusion, and Target Recognition*, pp. 182–194.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*. Springer.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*, 35–45.

Kara, Y., M. A. Boyacioglu, and Ö. K. Baykan (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert Systems with Applications 38*, 5311–5319.

Keane, M. P. and K. I. Wolpin (2009). Empirical applications of discrete choice dynamic programming models. *Review of Economic Dynamics 12*, 1–22.

Kim, K.-j. and I. Han (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications 19*, 125–132.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics 5*, 1–25.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika 94*(4), 769–786.

Lijoi, A., R. H. Mena, and I. Prünster (2008). A Bayesian nonparametric approach for comparing clustering structures in EST libraries. *Journal of Computational Biology 15*(10), 1315–1327.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Stat. 24*(3), 911–930.

Liu, J. S. and R. Chen (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association 93*, 1032–1044.

Liu, X., M. J. Daniels, and B. Marcus (2009). Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. *Journal of the American Statistical Association 104*, 429–438.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics 12*(1), 351–357.

MacDonald, I. L. and W. Zucchini (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. CRC.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, The Ohio State University.

Maceachern, S. N., M. Clyde, and J. S. Liu (1999). Sequential importance sampling for nonparametric Bayes models: the next generation. *Canadian Journal of Statistics 27*(2), 251–267.

Maceachern, S. N. and P. Müller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*(2), 223–238.

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *Journal of the American Statistical Association 99*(468), 1108–1118.

Mao, C. X. and B. G. Lindsay (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika 89*(3), 669–681.

McAuliffe, J. D., D. M. Blei, and M. I. Jordan (2006, jan). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Stat. Comput. 16*(1), 5–14.

Meilă, M. (2007). Comparing clusterings-an information based distance. *Journal of Multivariate Analysis 98*(5), 873–895.

Miller, J. W. and M. T. Harrison (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in neural information processing systems*, pp. 199–206.

Miller, J. W. and M. T. Harrison (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research 15*, 3333–3370.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356.

Muliere, P. and L. Tardella (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics 26*(2), 283–297.

Müller, P., F. Quintana, and G. L. Rosner (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics 20*(1), 260–278.

Müller, P., F. A. Quintana, and G. L. Page (2018). Nonparametric Bayesian inference in applications. *Statistical Methods & Applications 27*(2), 175–206.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics 41*(1), 370–400.

Ohn, I. and L. Lin (2020). Optimal bayesian estimation of gaussian mixtures with growing number of components. *Preprint at arXiv: 2007.09284*.

Page, G. L. and F. A. Quintana (2016). Spatial product partition models. *Bayesian Analysis 11*(1), 265–298.

Page, G. L. and F. A. Quintana (2018). Calibrating covariate informed product partition models. *Statistics and Computing 28*(5), 1009–1031.

Pakman, A. and L. Paninski (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics 23*, 518–542.

Petris, G., S. Petrone, and P. Campagnoli (2009). *Dynamic Linear Models with R*. Springer.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series 30*, 245–267.

Pitman, J. (2006). *Combinatorial stochastic processes*. Springer.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association 94*, 590–599.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*(483), 483–1131.

Samuels, S. M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics 14*(4), 373–383.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*(2), 639–650.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*, 115–131.

Soriano, J. and L. Ma (2019). Mixture modeling on related samples by ψ-stick breaking and kernel perturbation. *Bayesian Analysis 14*(1), 161–180.

Soyer, R. and M. Sung (2013). Bayesian dynamic probit models for the analysis of longitudinal data. *Computational Statistics & Data Analysis 68*, 388–398.

Stockwell, D. R. B. and A. T. Peterson (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling 148*(1), 1–13.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, Morristown, NJ, USA, pp. 985–992. Association for Computational Linguistics.

Teh, Y. W. and M. I. Jordan (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian nonparametrics*, Chapter 5, pp. 158–207. Cambridge University Press.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Uhlmann, J. K. (1992). Algorithms for multiple-target tracking. *American Scientist 80*, 128–141.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer Science & Business Media.

Wade, S. and Z. Ghahramani (2018). Bayesian cluster analysis: point estimation and credible balls. *Bayesian Analysis 13*(2), 559–626.

West, M. and J. Harrison (2006). *Bayesian Forecasting and Dynamic Models*. Springer Science & Business Media.

Yang, C.-Y., N. Ho, and M. I. Jordan (2019). Posterior distribution for the number of clusters in Dirichlet process mixture models. *Preprint at arXiv: 1905.09959*.

Zeng, C. and L. L. Duan (2020). Quasi-Bernoulli stick-breaking: infinite mixture with cluster consistency. *Preprint at arXiv: 2008.09938*.

Zuanetti, D. A., P. Müller, Y. Zhu, S. Yang, and Y. Ji (2018). Clustering distributions with the marginalized nested Dirichlet process. *Biometrics 74*(2), 584–594.