Università Commerciale "Luigi Bocconi"
PhD School

PhD program in: Statistics
Cycle: XXXI
Disciplinary Field: secs-s/01

Finite-dimensional nonparametric priors: theory and applications

Advisor: Igor Prünster
Co-Advisor: Antonio Lijoi

PhD Thesis by: Tommaso Rigon
ID Number: 3005213

Academic Year 2019/2020

THESIS DECLARATION

I, the undersigned

Surname: Rigon
Name: Tommaso
ID Number: 3005213
Thesis title: Finite-dimensional nonparametric priors: theory and applications
PhD in: Statistics
Cycle: XXXI

DECLARES

Under my responsibility:

1) that, according to Italian Republic Presidential Decree no. 445, 28th December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove true, all benefits included in this declaration and those of the temporary "embargo" are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30th April 1999, to keep a copy of the thesis on deposit at the "Biblioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary "embargo" protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

3) that I will submit the thesis online in unalterable format to Bocconi University, that, by means of "Institutional Research Information System (IRIS)", will permits online consultation of the complete text (except in cases of temporary "embargo");

4) the thesis is protected by the regulations governing copyright (Italian law no. 633, 22nd April 1941 and subsequent modifications). The exception is the right of Università Commerciale "Luigi Bocconi" to reproduce the same, quoting the source, for research and teaching purposes;

5) that the copy of the thesis submitted online is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is absolved from

any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal, and shall be exempt from any requests or claims from third parties;

7) that the thesis is not subject to "embargo", i.e. that it is not the result of work included in the regulations governing industrial property; it was not written as part of a project financed by public or private bodies with restrictions on the diffusion of the results; is not subject to patent or protection registrations.

Date: November 25, 2019

Tommaso Rigon

Contents

Acknowledgements

There are several people that I want to thank at the end of this journey. It is hard to imagine how my PhD would have been without them.

In first place, thanks to Antonio and Igor. I have learned a lot from them on how to do research. I am very grateful for the uncountable and fruitful discussions we had about Bayesian nonparametrics and Statistics in general. Beside, they were also real mentors. They introduced me to the academic world and they patiently showed me how to sail in these uncharted waters, either in their offices or in front of a good beer.

Thanks to Daniele, who has been a friend, a scientific collaborator, and an academic big brother. With his example, he encouraged me to strive for the best and to not rest on my laurels. Arguing with him is still one of my favorite activities.

Thanks also to Bruno, who does not like the spotlight but is always in the rearguard to motivate, to spur me into new projects, and to suggest the right direction.

A special thank goes also to the fellows of the XXXI cycle: Andrea, Giovanni, and Hristo. It has been a hard life, but we also had a lot of fun. Speaking of which, I would like to thank all the aficionados of Bar Bruto, whose names have been omitted to preserve their ongoing PhDs and careers. Thanks also to Marta, my favorite lunchmate, whose critical views have often smoothed my rigid opinions. A big thank also to all the exceptional students, researchers, and professors I had the chance to meet in these years, and in particular Giacomo, Sonia, and Raffaella, among those in the Bocconi crew.

Finally, thanks to my family, who constantly supported me during these years. And to Caterina, whose joyful and indispensable presence has made everything easier.

Abstract

The investigation of flexible classes of discrete prior has been an active research line in Bayesian statistics. Several contributions were devoted to the study of nonparametric priors, including the Dirichlet process, the Pitman–Yor process and normalized random measures with independent increments (NRMI). In contrast, only few finite-dimensional discrete priors are known, and even less come with sufficient theoretical guarantees. In this thesis we aim at filling this gap by presenting several novel general classes of parametric priors closely connected to well-known infinite-dimensional processes, which are recovered as limiting case. A priori and posteriori properties are extensively studied. For instance, we determine explicit expressions for the induced random partition, the associated urn schemes and the posterior distributions. Furthermore, we exploit finite-dimensional approximations to facilitate posterior computations in complex models beyond the exchangeability framework. Our theoretical and computational findings are employed in a variety of real statistical problems, covering toxicological, sociological, and marketing applications.

Chapter 1

Introduction

1.1 Discrete dependence structures

The statistical investigation of *discrete random structures* has been a very lively area of research in recent years. Bayesian nonparametric (BNP) discrete priors have found wide applicability in numerous settings that include, among others, flexible density estimation, model-based clustering, density regression, functional data analysis, and hidden Markov models (Hjort et al., 2010). The literature on discrete nonparametric priors flourished after the seminal paper of Ferguson (1973), in which the Dirichlet process (DP) was introduced. Well-known limitations of the DP have fostered the research of novel discrete nonparametric priors, which are nowadays well established inferential tools. Among them we recall the Pitman–Yor (PY) process (Ishwaran & James, 2001; Pitman & Yor, 1997), the normalized inverse-Gaussian process (Lijoi et al., 2005), the normalized generalized gamma process (Lijoi et al., 2007) and the very general classes of Gibbs-type priors (Gnedin & Pitman, 2005; De Blasi et al., 2015), and of homogeneous normalized random measures with independent increments (NRMIS) (Regazzini et al., 2003).

These priors have been employed to address predictive inference, with species sampling data, and density estimation, through hierarchical mixture models. In such a setting, the underlying assumption is that the observations $\theta_1, \ldots, \theta_n$ are drawn from an *exchangeable* sequence of random elements $(\theta_i)_{i \geqslant 1}$. More formally, let Θ be the sample space, which is assumed to be Polish, and let $\mathscr{B}(\Theta)$ denote its Borel σ-algebra. Moreover, $\mathscr{P}_\Theta$ stands for the space of probability measures on Θ. Then, the celebrated de Finetti's theorem guarantees the existence of a random probability measure conditionally on which the Θ-valued random variables are independent and identically distributed (iid), namely for any $n \geqslant 1$

$$\begin{aligned} (\theta_1, \ldots, \theta_n \mid \tilde{p}) &\overset{\text{iid}}{\sim} \tilde{p}, \\ \tilde{p} &\sim Q, \end{aligned} \tag{1.1}$$

where $\tilde{p}$ is a random probability measure, having either a parametric or a nonparametric form, which in turns follows the law of Q on the space $\mathscr{P}_\Theta$, the *prior* distribution in Bayesian inference. The PY process and the class of homogeneous NRMIs are instances of discrete prior laws Q, namely random probability measures of the form

$$\tilde{p}^{(\infty)} = \sum_{h=1}^{\infty} \xi_h \delta_{\tilde{\phi}_h}, \tag{1.2}$$

where the sequence of random Θ-valued locations $(\tilde{\phi}_h)_{h\geqslant 1}$ and the random weights $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots)$ are independent. Furthermore, the $\tilde{\phi}_h$'s are iid draws from a probability measure P, which is often assumed to be *diffuse*, that is $P(\{\theta\}) = 0$ for any $\theta \in \Theta$. Allowing the baseline measure P to have atoms is far from being inconsequential. Indeed, in such a case the random probability measure $\tilde{p}$ can not be regarded as a species sampling model and therefore the classical theoretical framework (Pitman, 1996) does not apply. For example, Carlton (2002) stressed that the posterior distribution of a Pitman–Yor process with a purely atomic baseline measure was, at the time, still unknown. This important theoretical gap was recently addressed e.g. by Canale et al. (2017); Camerlenghi et al. (2018). Such a setup entails challenging technical hurdles when it comes to determining distributional properties of interest for Bayesian inference.

When investigating covariate-dependent data $\{(\theta_{xi})_{i\geqslant 1} : x \in \mathbb{X}\}$ in a Bayesian framework, with $\mathbb{X}$ being the covariate space, the standard assumption of exchangeability is not appropriate since it amounts to considering the data as being homogeneous. The covariate $x \in \mathbb{X}$ is actually a source of heterogeneity that one has to take into account and a different symmetry condition among the data should be specified. The case $\mathbb{X} = \{1, \dots, L\}$, corresponding to a finite covariate space, identifies data that are recorded under L different, though related, experimental conditions. In view of this, a natural dependence structure is implied by *partial exchangeability*, according to which exchangeability holds true within each of the L separate groups each of $n^{(l)}$ observations, for $l = 1, \dots, L$, but not across them. Then, the array of Θ-valued random elements $\{(\theta_{li})_{i\geqslant 1} : l = 1, \dots, L\}$ is partially exchangeable if and only if for any $i_l = 1, \dots, n^{(l)}$ and $l = 1, \dots, L$

$$\begin{gathered}(\theta_{1i_1}, \dots, \theta_{Li_L}) \mid (\tilde{p}_1, \dots, \tilde{p}_L) \overset{\text{iid}}{\sim} \tilde{p}_1 \times \cdots \times \tilde{p}_L, \\ (\tilde{p}_1, \dots, \tilde{p}_L) \sim Q_L,\end{gathered} \tag{1.3}$$

for some probability measure Q_L on the product space $\mathscr{P}_\Theta^L$. Hence, conditionally on the vector $(\tilde{p}_1, \dots, \tilde{p}_L)$, the θ_{li}'s are independent and identically distributed within, but only independent across groups. The measure Q_L plays the role of prior distribution and in addition governs the *dependence across groups*.

An early proposal for Q_L appeared in Cifarelli & Regazzini (1978), but the decisive boost to the literature came after the seminal paper of MacEachern (1999). In this thesis we will rely on a hierarchical construction of Q_L and assume that the elements of the collection $\{\tilde{p}_1, \ldots, \tilde{p}_L\}$ are conditionally iid, given another discrete random probability measure $\tilde{p}_0$, such that

$$\begin{aligned} \left(\tilde{p}_l^{(\infty)} \mid \tilde{p}_0^{(\infty)}\right) &= \sum_{h=1}^{\infty} \xi_{lh}\delta_{\tilde{\phi}_{lh}}, \qquad \left(\tilde{\phi}_{lh} \mid \tilde{p}_0^{(\infty)}\right) \overset{\text{iid}}{\sim} \tilde{p}_0^{(\infty)}, \qquad l = 1, \ldots, L; \quad h \geqslant 1, \\ \tilde{p}_0^{(\infty)} &= \sum_{h=1}^{\infty} \xi_{0h}\delta_{\tilde{\phi}_{0h}}, \qquad \tilde{\phi}_{0h} \overset{\text{iid}}{\sim} P, \qquad h \geqslant 1, \end{aligned} \tag{1.4}$$

where P is some diffuse probability measure on Θ. Note that in view of this specification, one marginally has $\mathbb{E}(\tilde{p}_l \mid \tilde{p}_0) = \tilde{p}_0$ for each $l = 1, \ldots, L$. Thus, dependence across groups in (1.3) is induced by considering an exchangeable collection $\{\tilde{p}_1^{(\infty)}, \ldots, \tilde{p}_L^{(\infty)}\}$ of random probability measures. Note that the baseline distribution $\tilde{p}_0^{(\infty)}$ is almost surely purely atomic, implying that specification (1.4) entails similar technical difficulties that arises in the exchangeable model (1.2) when the baseline distribution is not diffuse. Such a model, when the $\tilde{p}_l^{(\infty)}$'s and $\tilde{p}_0^{(\infty)}$ are Dirichlet processes, has been proposed in Teh et al. (2006) and takes on the name of *hierarchical Dirichlet process* (HDP). The HDP has been successfully applied, e.g., to topic modeling (Teh et al., 2006), speaker diarization (Fox et al., 2011) and the analysis of fMRI data (Zhang et al., 2016). For a stimulating account on its use in several modeling and applied frameworks see Teh & Jordan (2010). An extension to the wider class of *normalized random measures* was proposed in Camerlenghi et al. (2019), which further provides a systematic investigation of the most relevant distributional properties for Bayesian inference. The achievement of these results heavily benefits from the nice probabilistic structure of the completely random measures (CRMs) that are used to define the underlying random probability measures. It is worth recalling that other examples of CRM-based priors Q_L are available in the literature, the most recent examples being Lijoi et al. (2014a,b), Lijoi & Nipoti (2014) and Griffin & Leisen (2017).

Exchangeable and the partially exchangeable settings constitutes a crucial building block for the construction of more complex models in which latent quantities, rather than the raw data, are assumed to be (partially) exchangeable. Hence, the theoretical investigation of this framework is motivated by applications well-beyond models (1.1) and (1.3). However, in some cases one might be interested in regression models where the entire distribution of a response variable is unknown and changes with a general vector of predictors $\boldsymbol{x} \in \mathbb{X} \subseteq \mathbb{R}^p$. Indeed, the increased flexibility provided by these procedures allows improvements in inference and prediction compared to classical regression frameworks, as seen in applications (e.g. Dunson & Park, 2008; Griffin & Steel,

2011; Wade et al., 2014). Mixture models based on (1.2), such as the Dirichlet process mixture of Lo (1984), have key computational benefits (e.g. Escobar & West, 1995; Neal, 2000), and provides a consistent strategy for density estimation (e.g. Ghosal et al., 1999; Tokdar, 2006; Ghosal & Van Der Vaart, 2007). This has motivated different generalizations of (1.2) by allowing the random mixing measure $\tilde{p}_{\boldsymbol{x}}$ to change with $\boldsymbol{x} \in \mathbb{X} \subseteq \mathbb{R}^p$ (MacEachern, 1999, 2000). Popular representations consider predictor-independent mixing weights ξ_h as in (1.2), and incorporate changes with $\boldsymbol{x} \in \mathbb{X}$ in the atoms $\tilde{\phi}_{\boldsymbol{x}h}$; see for instance De Iorio et al. (2004); Gelfand et al. (2005); De la Cruz-Mesía et al. (2007). As noted in MacEachern (2000) the predictor-independent assumption for the mixing weights might have limited flexibility in practice. This has motivated more general formulations in which also the weights $\xi_{\boldsymbol{x}h} = \xi_h(\boldsymbol{x})$ vary with the predictors. Relevant examples include the order-based dependent Dirichlet process (Griffin & Steel, 2006), the kernel stick-breaking process (Dunson & Park, 2008), and the infinite mixture model with predictor-dependent weights (Antoniano-Villalobos et al., 2014). In this thesis we will rely on a specific predictor-dependent formulation for $\tilde{p}_{\boldsymbol{x}}$ called *logit stick-breaking process* (LSBP), defined for any $\boldsymbol{x} \in \mathbb{X} \subseteq \mathbb{R}^p$ as

$$\tilde{p}_{\boldsymbol{x}} = \sum_{h=1}^{\infty} \xi_h(\boldsymbol{x}) \delta_{\tilde{\phi}_h}, \qquad \xi_h(\boldsymbol{x}) = \nu_h(\boldsymbol{x}) \prod_{l=1}^{H-1} \{1 - \xi_l(\boldsymbol{x})\}, \qquad h \geqslant 2, \tag{1.5}$$

with $\xi_1(\boldsymbol{x}) = \nu_1(\boldsymbol{x})$ and $\tilde{\phi}_h \overset{\text{iid}}{\sim} P$, where each stick-breaking weight $\nu_h(\boldsymbol{x}) \in (0,1)$ relates to a function $\eta_h(\boldsymbol{x}) \in \mathbb{R}$ of the covariates through the logit link. Such a formulation is closely related to the probit stick-breaking prior (PSBP) of Rodriguez & Dunson (2011) and it has been employed in Ren et al. (2011) for image segmentation.

1.2 Main contributions of the thesis

The remarkable advances in the BNP literature—outlined in the previous section—have not been paralleled by a similar wealth of proposals in the *finite-dimensional* setting, namely priors characterized by finitely many parameters. This includes discrete law of the form

$$\tilde{p}^{(H)} = \sum_{h=1}^{H} \pi_h \delta_{\tilde{\theta}_h}, \qquad \tilde{\theta}_h \overset{\text{iid}}{\sim} P, \qquad H \geqslant 1, \tag{1.6}$$

where the positive weights $\sum_{h=1}^{H} \pi_h = 1$ almost surely and the collections $\{\tilde{\theta}_1, \ldots, \tilde{\theta}_H\}$ and $\{\pi_1, \ldots, \pi_H\}$ are independent. Indeed, only few alternatives to (1.6) are known beyond Dirichlet-like structures and most of them outside the Bayesian realm. For example, there is an interesting piece of literature focusing on compositional data and spurred by the

pioneering work in Aitchison (1985) on general classes of distributions on the simplex, but none of them has been actually used as a prior distribution for modeling either the data or some latent feature. Indeed, the lack of deep theoretical results has prevented the development of more flexible classes of priors, as well as simple sampling algorithms that may facilitate their usage in applications.

A classical and popular prior choice for the weights $(\pi_1, \dots, \pi_H)$ in equation (1.6) is the symmetric $\textsc{dirichlet}(c/H, \dots, c/H)$ distribution. For instance Malsiner-Walli et al. (2016) suggest its usage for sparse finite mixture models as a way to circumvent the issue of selecting the number of mixture components, on the ground of asymptotic results presented by Rousseau & Mengersen (2011). The symmetric Dirichlet specification above, when directly used for exchangeable data $(\theta_1, \dots, \theta_n \mid \tilde{p}^{(H)}) \overset{\text{iid}}{\sim} \tilde{p}^{(H)}$, is often referred to as *Dirichlet multinomial model*, finite-dimensional Dirichlet process, or Fisher process. See for instance Kingman (1975); Ishwaran & Zarepour (2000, 2002) for further discussions. In addition, it is well-known that for H large enough the Dirichlet multinomial might be regarded as an approximation of the DP, and the implications of such a usage are detailed for instance in Green & Richardson (2001) or Ishwaran & Zarepour (2000).

Given the amount of interesting properties characterizing the symmetric Dirichlet, it is natural to ask whether there exist some equivalent and tractable formulations for the Pitman–Yor process and for homogeneous NRMIs. As we shall see, the answer is positive. Specifically, in Chapter 2 and Chapter 3 we introduce and investigate novel classes of finite-dimensional discrete priors that naturally generalize the Dirichlet multinomial and whose limits are the aforementioned nonparametric priors. We drop the Dirichlet specification since it displays serious drawbacks and limitations that are well-known in the literature. For example, it is very sensitive to the choice of its hyperparameter, hence requiring a careful calibration of the total mass parameter c. Moreover, the underlying clustering structure induced by the Dirichlet distribution is somehow restrictive, since it depends only on one parameter, thus calling for more flexible specifications. This has some effects on the structure of the associated system of predictive distributions. In the nonparametric case, these aspects are surveyed and illustrated in De Blasi et al. (2015).

Motivated by similar considerations, in Chapter 4 we present a novel *enriched* finite-dimensional discrete prior whose limit is the enriched functional Dirichlet process of Scarpa & Dunson (2014). The emphasis of such a contribution is on Bayesian functional clustering and this allow us to illustrate the practical advantages that a finite-dimensional specification might have in business applications.

The Dirichlet multinomial process has been exploited also as a *computational tool* for approximating the DP. However, other approximations exist. For instance, Muliere & Tardella (1998) rely on a truncation of the stick-breaking representation of the DP. Such

an idea was popularized and extended to the Pitman–Yor case by Ishwaran & James (2001). Following a similar line of reasoning, the contributions of Arbel & Prünster (2017) and Arbel et al. (2018) discuss truncation-based approximations for homogeneous NRMIs and for the Pitman–Yor process, respectively. The excellent performance of the above methods motivated us to develop similar approximate forms also for models (1.4) and (1.5). Specifically, in Chapter 5 and Chapter 6 we present novel computational strategies based on finite-dimensional prior for general hierarchical processes and for the LSBP, respectively. In both cases these methods comes with theoretical guarantees as we formally quantify the discrepancy between the finite- and the infinite-dimensional processes.

As was made clear, the core concept which motivates and unifies the contributions of this thesis is the notion of finite-dimensional nonparametric priors. The only exception to this general scheme is the work of Chapter 7, in which we provide theoretical justifications for a widely used variational Bayes approach for logistic regression. These theoretical advances will play a central role in the derivation of variational strategies for the LSBP, which are discussed in Chapter 6.

Before providing a concise account of each specific contribution, we owe a comment about the "finite-dimensional nonparametric" terminology. Indeed, within the BNP framework, the term "nonparametric" usually refers to the fact that the support of the prior distribution is infinite-dimensional. In this thesis, we embrace a different perspective and rely on classes of finite-dimensional priors whose flexibility can be increased at will, eventually converging to some well-defined infinite-dimensional prior. This argument is not completely new (Green & Richardson, 2001; Miller & Harrison, 2018) and leads to a broader definition of BNP, a perspective which seems to be supported in the review paper "*Bayesian nonparametric inference - why and how*" by Müller & Mitra (2013), who state in the conclusion:

> «*We started out by defining* BNP *as probability models for infinite-dimensional random quantities like curves or densities. It might be more fittingly called "massively parametric Bayes". The label nonparametric has been used because inference under* BNP *models often looks similar to (genuinely) nonparametric classical inference.*»

The finite-dimensional priors we propose in this thesis well fit this broad definition of "nonparametric", being highly flexible, massively parametric, and well-defined at the limit. We shall stress, however, that a large amount of parameters does not automatically define a "nonparametric" prior. Indeed, a strong prerequisite of any nonparametric procedure is the possibility of increasing the model flexibility at will, a requirement that is not necessarily satisfied by general massively parametric proposals.

1.3 Summary of the specific contributions

Each of the following sections corresponds to the homonymous chapter of the thesis and summarizes its main findings. The notation through these Chapters is largely consistent—and strongly consistent within the same Chapter—meaning that in some cases the same symbol has been used along the thesis to denote different but conceptually similar quantities. For example, the vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\intercal$ always denote regression coefficients through the thesis, although they will be appear e.g. in a convex mixture regression model in Chapter 2 and in a logistic regression model in Chapter 7. Therefore, to avoid confusions, each quantity is either recalled or re-defined within each Chapter.

1.3.1 On a finite-dimensional Pitman–Yor process

In Chapter 2 we aim at studying a novel finite-dimensional random probability measure in the form of equation (1.6) which we term *Pitman–Yor multinomial* process. We show that such a prior may be seen as a finite-dimensional analogue of the Pitman–Yor process and naturally generalizes the Dirichlet distribution on the simplex. Besides yielding a considerable degree of modeling flexibility, it preserves analytical and computational tractability. In first place, the Pitman–Yor multinomial process may serve as a very effective tool for computational purposes in a nonparametric setting. A popular class of Markov chain Monte Carlo algorithms for nonparametric mixture models, usually referred to as blocked Gibbs sampler, relies on the truncation of a stick-breaking representation of the mixing Pitman–Yor process (Ishwaran & James, 2001). If $(\nu_j)_{j\geqslant 1}$ is a sequence of independent random variables with $\nu_j \sim \text{Beta}(1-\sigma, c+j\sigma)$, $\sigma \in [0,1)$, $c > -\sigma$, and P a probability measure defined over Θ, then a prior on the space of density functions is the distribution of

$$\int_\Theta \mathcal{K}(y;\theta)\tilde{p}^{(\infty)}(d\theta), \qquad \tilde{p}^{(\infty)} = \sum_{h=1}^{\infty} \xi_h \, \delta_{\tilde{\phi}_h}, \qquad \tilde{\phi}_h \overset{\text{iid}}{\sim} P, \tag{1.7}$$

where $\xi_1 = \nu_1$, $\xi_h = \nu_h \prod_{j=1}^{h-1}(1-\nu_j)$ for $h \geqslant 2$ and where $\mathcal{K}$ is some transition kernel such that $\int_{\mathbb{R}} \mathcal{K}(y;\theta)\,dy = 1$ for any $\theta \in \Theta$. When it comes to evaluating Bayesian inferences, the infinite series defining $\tilde{p}^{(\infty)}$ in (1.7) cannot be computed and one conveniently relies on a suitable finite-dimensional approximation obtained by truncating $\tilde{p}_\infty$ at some level H, i.e.

$$\tilde{p}_{\text{tr}}^{(H)} = \sum_{h=1}^{H-1} \xi_h \, \delta_{\tilde{\phi}_h} + (1 - |\boldsymbol{\xi}^{(H-1)}|) \, \delta_{\tilde{\phi}_H}, \tag{1.8}$$

where $\boldsymbol{\xi}^{(H-1)} = (\xi_1, \ldots, \xi_{H-1})$ and $|\boldsymbol{\xi}^{(H-1)}| = \xi_1 + \cdots + \xi_{H-1}$. This approach has some limitations since one can hardly identify marginal probabilistic structures of interest such as, for example, the law of the induced exchangeable random partition, the probability distribution of the number of clusters or the prediction rule associated to (1.8). The results that are displayed in Chapter 2 will successfully address the above issues by relying on the Pitman–Yor multinomial process, which stands as an alternative finite-dimensional approximation of $\tilde{p}^{(\infty)}$. It will be shown that especially for non-informative specifications of $\tilde{p}^{(\infty)}$, namely those corresponding to values of $\sigma > 1/2$, the Pitman–Yor multinomial process is a more accurate approximation of $\tilde{p}^{(\infty)}$ compared to the truncated stick-breaking representation (1.8). In addition, the distributional results we achieve allow for a straightforward implementation of novel generalised Blackwell–MacQueen sampling schemes for evaluating point estimates, as well as conditional algorithms for uncertainty quantification. When $\sigma = 0$, the finite-dimensional model we propose clearly boils down to the Dirichlet multinomial with H atoms.

On top of its computational relevance in Bayesian nonparametrics, the Pitman–Yor multinomial process has important applications in finite mixture modeling. In this setting the value H represents a conservative upper bound for the number of mixture components. As discussed in Section 1.2, when $\sigma = 0$, such an approach find asymptotic justifications in Rousseau & Mengersen (2011). Hence, the Pitman–Yor multinomial process stands as a natural generalisation of such a method and it translates the advantages of the Pitman–Yor process into the finite-dimensional settings. This additional flexibility permits a much finer control of the underlying random partition, and in particular allows for more robust specification of the cluster distribution, which is typically very informative in the Dirichlet setting (Lijoi et al., 2007).

The impact of our proposal, and of the related distributional results, will be displayed by considering a covariate-dependent mixture. It will be assumed that the data $Y_1, \ldots, Y_n$ are such that

$$(Y_i \mid \tilde{p}_{x_i}) \overset{\text{ind}}{\sim} \int_\Theta \mathcal{K}(y; \theta) \tilde{p}_{x_i}(d\theta), \qquad i = 1, \ldots, n, \tag{1.9}$$

where $x_i \geqslant 0$ is a covariate associated to Y_i and

$$\tilde{p}_x = \{1 - f(x)\}\tilde{p}^{(H)} + f(x)\delta_{\tilde{\theta}_\infty}, \qquad x \geqslant 0, \tag{1.10}$$

is modeled as a convex linear combination of a discrete random measure $\tilde{p}^{(H)}$ on Θ, and a random point mass at $\tilde{\theta}_\infty \in \Theta$. This modeling framework is analogous to the one proposed in Canale et al. (2018), who engage in a toxicological study originally conducted by Longnecker et al. (2001) and later discussed also in Dunson & Park (2008).

The aim of these investigations was to assess the relationship between the DDE persistent metabolite of the pesticide DDT, and the risk of premature delivery. Hence, in (1.9) the DDE and the gestational age at delivery for the ith woman in the study are the covariate $x_i \geqslant 0$ and the response Y_i, respectively. While Canale et al. (2018) rely on Dirichlet-like priors in various modeling steps, e.g. for the law of $\tilde{p}$, here we leverage on the Pitman–Yor multinomial process. The smooth transition from $\tilde{p}^{(H)}$ to the point mass at $\tilde{\theta}_\infty$ is regulated by a nondecreasing bounded function $f(x) \in [0, 1]$ defined for $x \geqslant 0$, which equals zero in the origin, i.e. $f(0) = 0$. The Pitman–Yor multinomial process might be used also for the semi-parametric estimation of $f(x)$. Indeed, in quantitative risk assessment one customarily assumes that the $f(x)$ can be expressed as a linear combination of pre-specified basis functions

$$f(x) = \sum_{m=1}^{M} \mathcal{B}_m(x)\beta_m, \qquad x \geqslant 0, \tag{1.11}$$

where $\mathcal{B}_1(x), \ldots, \mathcal{B}_M(x)$ are nondecreasing and such that $\mathcal{B}_m \in [0, 1]$ for $m = 1, \ldots, M$. Under this choice, the constraints on $f(x)$ are automatically satisfied if $0 \leqslant \beta_m \leqslant 1$ and $\sum_{m=1}^{M} \beta_m = 1$. Hence, a symmetric Dirichlet distribution, corresponding to the weights of a Dirichlet multinomial random measure, might be adopted as prior choice for the parameters $\beta_1, \ldots, \beta_M$. However, question remains on the choice of its hyperparameters, which might affect the estimate of $f(x)$ if not suitably calibrated. To overcome this issue, we replace the symmetric Dirichlet with the more flexible ratio-stable distribution, the law associated to the weights of a Pitman–Yor multinomial process, which is shown to provide robust and reliable inferential results even under miscalibrated choices of the hyperparameters.

1.3.2 Finite-dimensional normalized random measures

In Chapter 3 we move beyond the Pitman–Yor multinomial case and we study a much broader class of finite-dimensional random probability measures having form (1.6) which we term *normalized infinitely divisible multinomial* (NIDM) processes. These priors are the finite-dimensional analogue of the class of homogeneous NRMIs. Our theoretical results are general, but particular emphasis will be given to the special case called *normalized generalized gamma multinomial process*.

As an illustrative application, we consider the INVALSI 2016-2017 dataset, a national examination conducted in Italy. Specifically, we aim at measuring the teaching competencies of a set of schools by taking into account the socio-demographic characteristics of its students. Great effort has been made by the INVALSI institution to provide reliable quantifications of the effect of each school on the test performance. Indeed, such an

indicator is nationally relevant especially for the development and the evaluation of educational policies. We address this problem via semi-parametric modeling with nonparametric school-specific random effects, which will be interpreted as a proxy of the added-value of the school. Stated in more general terms, let $Y_1, \dots, Y_n$ be a sequence of $\mathbb{Y}$–valued random elements (they will be schools' effects in our motivating application), and let $\mathcal{K} : \mathbb{Y} \times \Theta \to \mathbb{R}_+$ be a transition kernel such that $y \mapsto \mathcal{K}(y;\theta)$ is a density function on $\mathbb{Y}$, for any $\theta \in \Theta$. Then, conditionally on a random probability measure $\tilde{p}^{(H)}$, we suppose that

$$(Y_1, \dots, Y_n \mid \tilde{p}^{(H)}) \overset{\text{iid}}{\sim} \int_\Theta \mathcal{K}(y;\theta)\tilde{p}^{(H)}(d\theta),$$

with $\tilde{p}^{(H)}$ being a NIDM process of the form (1.6). The proposed NIDM processes allow for a finer control of the underlying clustering mechanism and for a robustification of the estimation process. We will provide both theoretical and empirical evidence in support of this claim, in line to what was already noticed in the infinite-dimensional case (Lijoi et al., 2007).

The study of this novel class of priors will benefit from its connection with homogeneous NMRIS, whose theory has achieved remarkable advances in the recent years (Lijoi & Prünster, 2010), and such a connection suggests several practical advantages. As a by-product of our investigation, we note that by virtue of their close relationship to homogeneous NRMI, one might employ NIDM as approximations of their infinite-dimensional counterpart. Besides the theoretical interest that a result of this type may give rise to, it is also very helpful from a practical standpoint since it helps lightening computational bottlenecks. Indeed, posterior inference for NRMIS might involve Ferguson & Klass (1972) representations, hence requiring numerical and analytical approximations. In contrast, the posterior structure of several NIDMS can be computed exactly. Such a gain, however, does not come for free since the probabilistic structure of our model yields some challenging technical hurdles when it comes to determining distributional properties of interest for Bayesian inference. These difficulties parallel those that arise when one uses a discrete base measure for a nonparametric prior process. See, e.g., Canale et al. (2017) for an example in the Pitman–Yor case.

1.3.3 Functional clustering via finite-dimensional enriched priors

There is an increasingly rich literature about BNP models for clustering functional observations. However, most of the recent proposals rely on infinite-dimensional characterizations that might lead to overly complex cluster solutions. Motivated by an application in e-commerce, In Chapter 4 we propose a novel finite-dimensional dicrete

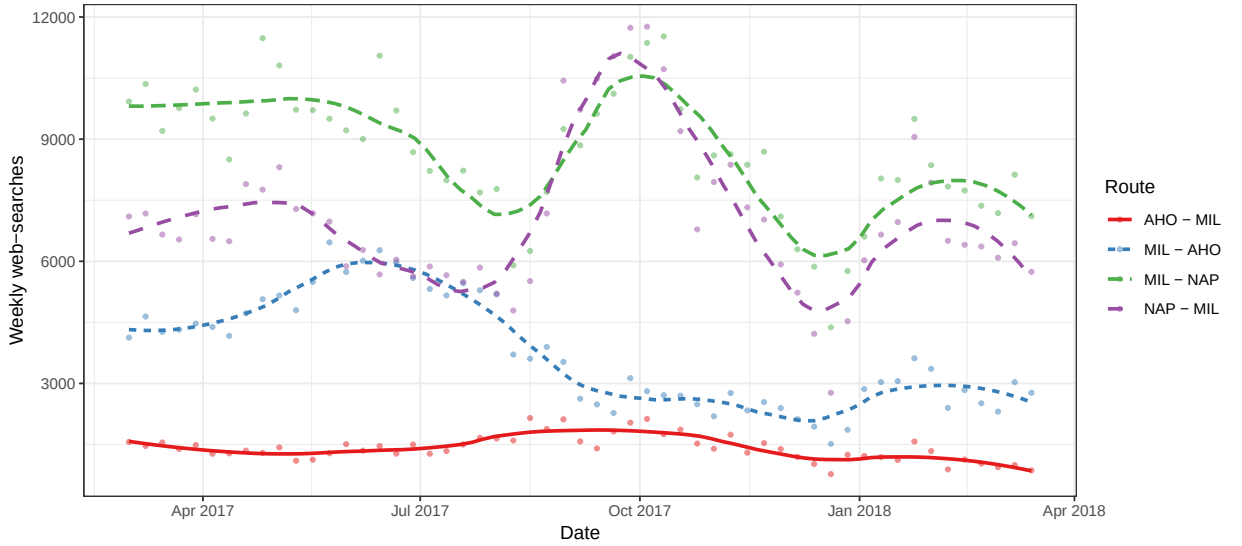


Figure 1.1: Number of the web searches on an Italian website in the period between March 2017 and March 2018. The origin and the destination of each route are coded as follows: `MIL` = Milan, `NAP` = Naples, `AHO` = Alghero. Smoothed trajectories are obtained using a nonparametrics `loess` estimate.

prior that we term *enriched Dirichlet multinomial process*. Our proposal accommodates the incorporation of functional constraints while bounding the model complexity.

In our motivating application, a private company selling flight tickets is interested in understanding the preferences and the needs of its customers, to implement effective marketing strategies and to provide tailored solutions to its clients. In this specific industry, a major goal is to assess the interests of customers towards each flight route, which represents the functional unit in our analysis. The involved number of flight routes is quite large and therefore route-specific marketing actions are practically unfeasible, since they would require massive human interventions. A possible solution is to consider groups (clusters) of similar routes to allow the development of cluster-specific policies which have an impact on homogeneous segments of the market. Such a strategy is highly effective as long as the number of clusters is *limited*.

The entries of the dataset at our disposal are the number of times that each route has been searched on the company's website, comprising a collection of weekly counts for each flight route. These longitudinal measurements are characterized by relevant temporal patters that can be exploited to produce a finer partition of the market, compared to approaches based on static indicators. This is immediately evident from Figure 1.1, where the smoothed trajectories of two different routes are depicted. From a modeling perspective, we are given a collection of functional observations—one for each flight route—and we aim at partitioning them into groups. Let us assume that the route-specific measurements $Y_i(t)$ can be regarded as error-prone realizations of

unknown functions $f_i(t)$, for each route $i = 1, \dots, n$, and time value $t \in \mathbb{R}_+$, that is

$$Y_i(t) = f_i(t) + \epsilon_i(t), \qquad i = 1, \dots, n, \tag{1.12}$$

with $\epsilon_i(t)$ denoting a random noise term, independent over flight routes and time. The additive specification (1.12) customarily serves as starting point in functional data analysis (Ramsay & Silverman, 2005). A natural way to group different functions in (1.12) is through Bayesian mixtures. Functional clustering via finite mixtures have been provably effective in applications (e.g. Heard et al., 2006), but question remains on the choice of mixture components, i.e. the number of clusters. A possible solution is to rely on BNP priors, and one may follow Bigelow & Dunson (2009) who proposed a spline formulation for each f_i together with a DP for the associated regression coefficients. Similarly, Ray & Mallick (2006) adopted the DP in conjunction with wavelets. The resulting process is called functional Dirichlet process (FDP). In Dunson & Park (2008) such a model has been employed for joint modeling of functional observations with a response variable, whereas in Petrone et al. (2009) a hybrid FDP is proposed, allowing realizations of $f_i(t)$ to share atoms in different local regions.

Although the latter methods enable flexible clustering and they are excellent tools for density estimation, their practical usage might be limited here. Indeed, the employment of a model with an unbounded number of groups might undermine the original goal, namely providing small dimensional summaries of flight routes. Furthermore, all the above models seem to rely too much on data while ignoring accumulated knowledge from past analyses. For example, it is known that some flight routes are characterized by a strong cyclical component, e.g. the one depicted in Figure 1.1, and one may want to include this aspect in the model. The latter remark motivated Scarpa & Dunson (2009) to propose a contaminated FDP accounting for parametric functional specifications.

To overcome all the above limitations we propose an enriched functional Dirichlet multinomial process (E-FDMP), which has a *bounded complexity* in terms of number of clusters and can easily incorporate prior knowledge about functional shapes. We will show that the proposed model converges to the enriched class of functional Dirichlet processes (E-FDP) presented in Scarpa & Dunson (2014), when the number of clusters is allowed to be infinite, while being reminiscent of the enriched Dirichlet process of Wade et al. (2011). Specifically, the underlying clustering mechanism can be described in terms of a two-step enriched urn-scheme, extending the well-know Blackwell & MacQueen (1973) Pólya urn. Such a theoretical development clarifies the interpretation of the involved random partition and it is helpful in the practical specification of the hyperparameters.

1.3.4 Computational advances for hierarchical models

Hierarchical normalized discrete random measures in equation (1.4) identify a general class of priors that is suited to flexibly learn how the distribution of a response variable changes across groups of observations. Although current theory on hierarchies of non-parametric priors yields all relevant tools for drawing posterior inference (Camerlenghi et al., 2019), their implementation comes at a high computational cost, especially when one has to deal with the analysis of large datasets.

In Chapter 5 we fill this gap by proposing a finite-dimensional approximation for a general class of hierarchical processes, which leads to an efficient *conditional* Gibbs sampling algorithm. Most of the current algorithms for posterior inference with hierarchical processes are of *marginal* type, that is they rely on the marginalization of the random probability measures $(\tilde{p}_1^{(\infty)}, \ldots, \tilde{p}_L^{(\infty)})$. While having some computational advantages, this rules out the possibility of obtaining complex posterior functionals of the vector $(\tilde{p}_1^{(\infty)}, \ldots, \tilde{p}_L^{(\infty)})$, which are often of interest in several applied contexts such as, for example, credible intervals. To overcome this difficulty, we propose a simple and efficient *conditional* Gibbs sampler for a wide class of hierarchical discrete random probability measures that includes the HDP as a special case. The actual implementation of the algorithm is eased by an a priori approximation of the infinite-dimensional process, based on a deterministic truncation of the random probability measure $\tilde{p}_0$. We provide theoretical support for such a truncation, borrowing ideas from the arguments of Muliere & Tardella (1998), Ishwaran & James (2001), and Arbel et al. (2018) within the exchangeable setting.

It is finally worth noting that building upon model (1.4) and, then, truncating to the Hth term, one can obtain the building block of a mixture model for partially exchangeable data that is discussed in detail Chapter 5. Most notably, such a model also has some connections with the Latent Dirichlet Allocation (LDA) of Blei et al. (2003), of which our proposal is a generalization. In fact, we work with a wider class of distributions compared to the Dirichlet distribution used in LDA. Additionally, while in LDA dependence among mixing distributions is induced through an approximate empirical Bayes procedure that determines the numerical value of certain hyperparameters of the model, here our full Bayesian analysis makes use of suitable prior laws for all the parameters and hyperparameters of the model. Finally, as for the choice of H, that is the number of latent topics in the terminology of topic modeling, in LDA it is selected so that it minimizes some out-of-sample goodness-of-fit metric. On the other hand, we choose H in order to achieve a satisfactory approximation of the infinite-dimensional process; the actual number of latent topics is elegantly and effectively regulated by the prior. We stress that our model is not confined to topic modeling with categorical data: indeed, they may

cope with observations taking values in general Polish spaces, thus allowing for a much broader applicability.

1.3.5 Computational advances for logit stick-breaking priors

The formulations for Bayesian density regression recalled in Section 1.2 are very flexible covariate-dependent nonparametric priors. However, this comes at a computational cost. In particular, the availability of simple algorithms for tractable posterior inference is limited by the specific construction of these representations.

The above issue motivates alternative formulations which preserve theoretical properties, but facilitate tractable posterior computation under a broader variety of algorithms. In Chapter 6 we aim to address this goal via a LSBP prior, that has been defined in equation (1.5). The proposed formulation is closely related to the probit stick-breaking prior (PSBP) of Rodriguez & Dunson (2011). Indeed, as we will discuss in Chapter 6, both LSBP and PSBP are characterized by a continuation-ratio representation (Tutz, 1991), which allows to express the underlying clustering assignment in terms of independent and sequential binary regressions. This representation has key computational benefits and has been exploited by Rodriguez & Dunson (2011) to derive a Markov chain Monte Carlo (MCMC) algorithm for posterior inference. However, while the MCMC for PSBP relies on the truncated Gaussian data augmentation for probit regression (Albert & Chib, 1993), the one for LSBP exploits the recent Pólya-gamma data augmentation for logistic regression (Polson et al., 2013), which might improve mixing compared to the PSBP, especially in imbalanced situations (Johndrow et al., 2018). As we will clarify in Chapter 6, these imbalanced settings can also occur in our case, since the binary regressions are associated to latent clustering allocations.

Besides developing tractable Gibbs sampling methods, we further derive alternative computational routines which address the scalability and mixing issues of MCMC in high-dimensional studies. Specifically, in Chapter 6 we illustrate a tractable expectation-maximization (EM) routine for point estimation, and a simple variational Bayes (VB) algorithm for scalable inference. Both strategies leverage again the sequential representation of the LSBP and the associated Pólya-gamma data augmentation. Note that a VB routine for LSBP is also presented in Ren et al. (2011), but it is based on the bound of Jaakkola & Jordan (2000). As a consequence of the recent theoretical findings in Durante & Rigon (2019), which are broadly summarized in Chapter 7, it can be shown that our approach is intimately related to the one of Ren et al. (2011), although being developed by means of seemingly unrelated strategies. Finally, while tractable algorithms such as EM or VB could be possibly obtained also for PSBP, we are not aware of any actual

discussion or implementation. Indeed, the analytical derivations might be slightly more complex in the PSBP case compared to the LSBP.

We shall emphasize that the overarching focus of our contribution is not on developing a novel methodological framework for Bayesian density regression, but on deriving a broad set of routine-use computational strategies under a suitable and tractable representation. To our knowledge this goal remains partially unaddressed, but represents a fundamental condition to facilitate routine implementation of Bayesian density regression by practitioners. The three proposed algorithms are empirically compared using a real data toxicology study, previously considered in Dunson & Park (2008) as well as in Chapter 2.

1.3.6 Conditionally conjugate variational Bayes for logistic models

Chapter 7 represents an exception to the common thread of finite-dimensional nonparametric priors underlying this thesis. Nonetheless, the theoretical advances contained there are key for a deeper understanding of VB approaches for logit-based models, like those developed in Chapter 6 for the LSBP.

Variational Bayes (VB) is a common strategy for approximate Bayesian inference, but simple methods are only available for specific classes of models including, in particular, representations having conditionally conjugate constructions within an exponential family. Models with logit components are an apparently notable exception to this class, due to the absence of conjugacy between the logistic likelihood and the Gaussian priors for the coefficients in the linear predictor. To facilitate approximate inference within this widely used class of models, Jaakkola & Jordan (2000) proposed a simple variational approach which relies on a family of tangent quadratic lower bounds of logistic log-likelihoods, thus restoring conjugacy between these approximate bounds and the Gaussian priors.

This strategy is still implemented successfully, but less attempts have been made to formally understand the reasons underlying its excellent performance. Following a review on VB for logistic models, in Chapter 7 we cover this gap by providing a formal connection between the above bound and a recent Pólya-gamma data augmentation for logistic regression. Such a result places the computational methods associated with the aforementioned bounds within the framework of variational inference for conditionally conjugate exponential family models, thereby allowing recent advances for this class to be inherited also by the methods relying on Jaakkola & Jordan (2000), such as the LSBP prior of Chapter 6.

1.4 Related work and future directions

Before providing a concise account of possible future directions, it worth mentioning some related works on finite-dimensional nonparametric priors and logit-based models that have not been included in the thesis because of space constraints but also to provide a more coherent and homogeneous treatment of the topic. In first place, the contribution of Rigon, Durante & Torelli (2019) covers both nonparametric random effects via Dirichlet multinomial process priors and logistic regressions via Pólya-gamma data augmentations. Extensions of the ideas presented in Chapter 7, including stochastic variational inference and EM strategies for logistic regression models, are instead discussed in Durante & Rigon (2019). In addition, the contributions of Chapter 6 and Chapter 7 fostered the research of novel algorithms for covariate-dependent latent class analysis, which are discussed in Durante, Canale & Rigon (2019). Finally, modeling strategies for functional observations—covered in the contribution of Chapter 4—might have sensible applications in neuroscience, especially for the analysis of FMRI data. Indeed, the development of tailored models for FMRI data is an active research line, as testified by the contribution of Caponera, Denti, Rigon, Sottosanti & Gelfand (2018).

Several generalizations and developments of the work developed in this thesis can be envisioned. In first place, the enriched Dirichlet multinomial process presented in Chapter 4 might be readily combined with the Pitman–Yor multinomial and NIDM processes introduced in Chapter 2 and Chapter 3. Such a generalization would allow for an even finer control of the partition mechanism. While Gibbs sampling methods in such a setting would be straightforward to implement, variational Bayes strategies would not be trivial and therefore worthwhile of future research. Another possible usage of the Pitman–Yor multinomial process is within the framework of hierarchical processes, where it might be employed in place of the truncated stick-breaking representation in Chapter 5. Further applications of such a process could be within the context of hidden Markov models for speaker diarization, hence generalizing the model of Fox et al. (2011).

The contributions of this thesis will hopefully foster further research about finite-dimensional discrete priors given that, so far, they have been extremely useful in a wide variety of statistical problems.

Chapter 2

On a finite-dimensional Pitman–Yor process

2.1 Summary

The chapter is organized as follows. In Section 2.2 we define the Pitman–Yor multinomial process and we provide different characterizations. In Section 2.3 we study its properties and in particular we derive closed form expressions for the law of the random partition, the distribution of the number of clusters, and the associated urn schemes. In Section 2.4 we characterize its posterior distribution, which can be regarded as quasi-conjugate, paralleling the terminology used for the Pitman–Yor process. A sampling algorithm which allows to draw independent posterior values is proposed. Finally, in Section 2.5, we show that the Pitman–Yor multinomial process can be regarded as a weak-limit approximation of the Pitman–Yor and we discuss its advantages over to the truncated stick-breaking representation. In Section 2.6 we conduct a simulation study to assess the empirical performance of the proposed prior. In Section 2.7 the Pitman–Yor multinomial is employed for convex mixture regression modeling, and its practical gains over the Dirichlet multinomial are emphasized.

2.2 The Pitman–Yor multinomial process

The Pitman–Yor multinomial process is built upon the Pitman–Yor (Perman et al., 1992), also known as the two-parameter Poisson–Dirichlet process, briefly recalled in (1.7). Notice that the probability distribution P of the atoms $\tilde{\phi}_h$ is sometimes termed the baseline measure and is such that $\mathbb{E}\{\tilde{p}^{(\infty)}(A)\} = P(A)$ for any measurable subset A of Θ. We will henceforth use the notation $\tilde{p}^{(\infty)} \sim \text{PY}(\sigma, c; P)$ and P is typically chosen to be diffuse, i.e. $P(\{\theta\}) = 0$ for any $\theta \in \Theta$. The Pitman–Yor multinomial process corresponds to the case where P is replaced by some discrete random probabiity measure with finitely many support points, as the following definition clarifies.

Definition 2.1. A discrete random probability measure $\tilde{p}^{(H)}$ is a Pitman–Yor multinomial process if it admits the hierarchical representation

$$(\tilde{p}^{(H)} \mid \tilde{p}_0^{(H)}) \sim \textsc{py}(\sigma, c; \tilde{p}_0^{(H)}), \qquad \tilde{p}_0^{(H)} = \frac{1}{H} \sum_{h=1}^{H} \delta_{\tilde{\theta}_h}, \qquad H \geqslant 1, \tag{2.1}$$

where $\tilde{\theta}_h$ are independent and identically distributed Θ-valued random variables with common distribution P. We will write $\tilde{p}^{(H)} \sim \textsc{pym}(\sigma, c; P)$.

When $\sigma = 0$, the random probability measure $\tilde{p}_H$ in Definition 2.1 reduces to the Dirichlet multinomial process, which indeed admits such a hierarchical representation (Ishwaran & Zarepour, 2000). Though $\tilde{p}^{(H)}$ is finite-dimensional, one can give an alternative and equivalent definition in terms of the infinite-dimensional counterpart $\tilde{p}^{(\infty)} \sim \textsc{py}(\sigma, c; P)$. Indeed, for any finite and measurable partition $B_1, \ldots, B_d$ of Θ the vector $\{\tilde{p}^{(\infty)}(B_1), \ldots, \tilde{p}^{(\infty)}(B_{d-1})\}$ identifies a probability distribution on the simplex known as ratio-stable (Carlton, 2002), so that

$$\{\tilde{p}^{(\infty)}(B_1), \ldots, \tilde{p}^{(\infty)}(B_{d-1})\} \sim \textsc{rs}\{\sigma, c; P(B_1), \ldots, P(B_d)\},$$

where we agree that $\tilde{p}(B_i) = 0$ almost surely if $P(B_i) = 0$, for any $i = 1, \ldots, d$. Moment formulae for ratio-stable laws can be found in Carlton (2002). Unsurprisingly, the weights of a Pitman–Yor multinomial process follow a ratio-stable distribution, as summarized in the following proposition, whose proof is straightforward.

Proposition 2.1. *A Pitman–Yor multinomial process* $\tilde{p}^{(H)} \sim \textsc{pym}(\sigma, c; P)$ *admits the following marginal representation*

$$\tilde{p}^{(H)} \overset{d}{=} \sum_{h=1}^{H} \pi_h \delta_{\tilde{\theta}_h}, \qquad (\pi_1, \ldots, \pi_{H-1}) \sim \textsc{rs}(\sigma, c; 1/H, \ldots, 1/H), \qquad \tilde{\theta}_h \overset{iid}{\sim} P. \tag{2.2}$$

The density function of the weights $(\pi_1, \ldots, \pi_{H-1})$ is generally not available in closed form, besides some special cases. When $\sigma = 0$, corresponding to the Dirichlet multinomial process, the distribution of the weights is that of a symmetric Dirichlet distribution with parameters $(c/H, \ldots, c/H)$. When $\sigma = 1/2$ the density function is available in closed form, and it was firstly obtained by Carlton (2002). The lack of a closed form expression of the density function is not a concern for Bayesian inference, since a ratio-stable distribution can be sampled for any admissible value of σ and c both a priori and a posteriori, as detailed henceforth. The proposed algorithm will arise from a hierarchical representation of ratio-stable distributions in terms of tempered-stable and gamma random variables, which can be easily simulated. To this end, we briefly recall that a positive random

variable J is *tempered-stable* if for some $c > 0$, $\sigma \in (0,1)$ and $\kappa \geqslant 0$, its Laplace transform is

$$\mathbb{E}(e^{-\lambda J}) = \exp\left[-c\{(\lambda+\kappa)^\sigma - \kappa^\sigma\}\right], \qquad \lambda > 0,$$

and we shall use the notation $J \sim \text{TS}(c,\sigma,\kappa)$. Note that such a random variable can be efficiently sampled, for instance by means of the algorithm of Ridout (2009). When $\sigma = 1/2$, the random variable $J \sim \text{TS}(1/H, 1/2, \kappa)$ has inverse Gaussian distribution, while setting $\kappa = 0$ leads to the positive-stable distribution. The main result we will rely on for computational purposes is the following.

Proposition 2.2. *Let* $\tilde{p}^{(H)} \sim \text{PYM}(\sigma, c; P)$ *with* $\sigma \in (0,1)$ *and* $c \geqslant 0$. *Then the weights of* $\tilde{p}^{(H)}$ *in* (2.2) *admit the representation* $(\pi_1,\ldots,\pi_H) \stackrel{d}{=} (J_1/\sum_{h=1}^H J_h, \ldots, J_H/\sum_{h=1}^H J_h)$ *and*

$$(J_h \mid U) \overset{\text{iid}}{\sim} \text{TS}(1/H, \sigma, U), \qquad U^\sigma \sim \text{GA}(c/\sigma, 1),$$

where we agree that $U = 0$ *almost surely if* $c = 0$.

Although the above hierarchical representation holds only for $c \geqslant 0$, one can make it fully general through the following argument. For any c one can conveniently represent the distribution of weights in Proposition 2.1 as

$$(\pi_1,\ldots,\pi_H) \stackrel{d}{=} W(\zeta_1,\ldots,\zeta_H) + (1-W)(\pi_1^*,\ldots,\pi_H^*),$$

where the random variable $W \sim \text{BETA}(1-\sigma, c+\sigma)$, and the random vectors $(\zeta_1,\ldots,\zeta_H) \sim \text{MULTINOM}(1/H,\ldots,1/H)$ and $(\pi_1^*,\ldots,\pi_{H-1}^*) \sim \text{RS}(\sigma, c+\sigma; 1/H,\ldots,1/H)$ are mutually independent. See Carlton (2002). Since $c+\sigma$ is positive, the simulation of $(\pi_1^*,\ldots,\pi_H^*)$ can be addressed by means of Proposition 2.2 and this allows sampling of $(\pi_1,\ldots,\pi_H)$ for any $c > -\sigma$.

Remark 2.1. The simulation of each J_h in Proposition 2.2 might be the source of numerical issues, for values of σ close to 0. This occurs because the distribution of U, then, places mass on very large values that might cause overflows. However, such a problem can be easily circumvented by considering the rescaled random variables $\tilde{J}_h = J_h/\{(\sigma/c)^{1/\sigma}\}$ whose distribution is $(\tilde{J}_h \mid \tilde{U}) \sim \text{TS}\{c/(\sigma H), \sigma, \tilde{U}\}$ with $\tilde{U}^\sigma \sim \text{GA}(c/\sigma, c/\sigma)$. This leads to more stable algorithms because $\mathbb{E}(\tilde{U}^\sigma) = 1$. The rescaling constant cancels in the normalization and therefore one has equivalently that $(\pi_1,\ldots,\pi_H) \stackrel{d}{=} (\tilde{J}_1/\sum_{h=1}^H \tilde{J}_h, \ldots, \tilde{J}_H/\sum_{h=1}^H \tilde{J}_h)$.

The hierarchical representation of Proposition 2.2 is a useful practical tool for simulating ratio-stable random vectors. However, a further and extremely useful characterization of the random variables $J_1,\ldots,J_H$ is available. Specifically, we show that

the law of $J_1, \ldots, J_H$ can be obtained via polynomial tilting of a collection of independent and identically distributed positive-stable random variables. Such a construction is reminiscent of the change of measure formula given in Pitman & Yor (1997), for the infinite-dimensional case. The connection is of great theoretical importance for the derivation of posterior quantities, as detailed in the Appendix.

Proposition 2.3. *Let* $\tilde{p}^{(H)} \sim \textsc{pym}(\sigma, c; P)$ *with* $\sigma \in (0,1)$ *and* $c > -\sigma$. *Then the vector of jumps* $(J_1, \ldots, J_H)$ *identifying the weights of* $\tilde{p}^{(H)}$ *in Proposition 2.2 is such that*

$$\mathbb{E}\left\{\exp\left(-\sum_{h=1}^{H}\lambda_h J_h\right)\right\} = \frac{\Gamma(c+1)}{\Gamma(c/\sigma+1)}\mathbb{E}\left\{\left(\sum_{h=1}^{H} J_h^{(\sigma)}\right)^{-c}\exp\left(-\sum_{h=1}^{H}\lambda_h J_h^{(\sigma)}\right)\right\},$$

for any $\lambda_1, \ldots, \lambda_H > 0$, *where* $J_h^{(\sigma)} \overset{\text{iid}}{\sim} \textsc{ts}(1/H, \sigma, 0)$.

2.3 Distributional properties

The Pitman–Yor multinomial process is almost surely discrete. This implies that a sample of n random elements $(\theta_1, \ldots, \theta_n \mid \tilde{p}^{(H)}) \overset{\text{iid}}{\sim} \tilde{p}^{(H)}$, with $\tilde{p}^{(H)} \sim \textsc{pym}(\sigma, c; P)$, will display ties with positive probability. If $K_{n,H} = k \leqslant \min\{n, H\}$ is the number of distinct values, say $\theta_1^*, \ldots, \theta_k^*$ in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, we let $n_1, \ldots, n_k$ denote their respective frequencies, so that $\sum_{j=1}^{k} n_j = n$. This induces a random partition $\Psi_{n,H}$ of $[n] = \{1, \ldots, n\}$ into k sets $C_1, \ldots, C_k$ such that i and j are in the same set when $\theta_i = \theta_j$. The probability distribution of such a random partition is the so-called *exchangeable partition probability function*, which is defined by

$$\Pi_H(n_1, \ldots, n_k) = \mathbb{P}(\Psi_{n,H} = \{C_1, \ldots, C_k\}) = \sum_{i_1 \neq \cdots \neq i_k} \mathbb{E}\left(\prod_{j=1}^{k} \pi_{i_j}^{n_j}\right),$$

where the vector $(n_1, \ldots, n_k)$ of positive integers is such that $n_j = \sharp C_j$ and $\sum_{j=1}^{k} n_j = n$ and the sum runs over all the positive and distinct integers $(i_1, \ldots, i_k)$ in $\{1, \ldots, H\}$. As discussed in Pitman (1996), when P is diffuse the exchangeable partition probability function characterizes the underlying random probability measure and yields, as a by-product, the system of predictive distributions. We briefly recall these in the infinite-dimensional case, namely when $(\phi_1, \ldots, \phi_n \mid \tilde{p}^{(\infty)}) \overset{\text{iid}}{\sim} \tilde{p}^{(\infty)}$, with $\tilde{p}^{(\infty)} \sim \textsc{py}(\sigma, c; P)$ and P is diffuse. The exchangeable partition probability function is

$$\Pi_\infty(n_1, \ldots, n_k) = \frac{\prod_{j=1}^{k-1}(c + j\sigma)}{(c+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}, \tag{2.3}$$

where $(a)_n = a(a+1)\cdots(a+n-1)$ for any real a and integer $n \geqslant 1$ is the ascending factorial, with $(a)_0 = 1$. Moreover, if one conditions on $\phi = (\phi_1, \ldots, \phi_n)$ featuring k distinct values $\phi_1^*, \ldots, \phi_k^*$, the predictive distribution of ϕ_{n+1} is

$$\mathbb{P}(\phi_{n+1} \in A \mid \phi) = \frac{c + k\sigma}{c+n} P(A) + \frac{1}{c+n} \sum_{j=1}^{k} (n_j - \sigma) \delta_{\phi_j^*}(A). \tag{2.4}$$

The following result provides the finite-dimensional counterpart to (2.3) and is expressed in terms of the generalized factorial coefficients (Charalambides, 2002), defined as

$$\mathscr{C}(n, k; \sigma) := \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n. \tag{2.5}$$

Henceforth, we shall further assume that P is diffuse and $\sigma \in (0, 1)$, so that the well-known Dirichlet case might be obtained by taking the limit as $\sigma \to 0$.

Theorem 2.1. *The exchangeable partition probability function induced by a Pitman–Yor multinomial process* $\tilde{p}^{(H)} \sim \textsc{pym}(\sigma, c; P)$ *is*

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} \frac{1}{(c+1)_{n-1}} \sum_{\boldsymbol{\ell}} \frac{\Gamma(c/\sigma + |\boldsymbol{\ell}|)}{\sigma \Gamma(c/\sigma + 1)} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{H^{\ell_j}},$$

where the sum runs over all the vectors $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ *such that* $\ell_j \in \{1, \ldots, n_j\}$ *and* $|\boldsymbol{\ell}| = \ell_1 + \cdots + \ell_k$.

Based on this result, one may determine the system of predictive distributions corresponding to the Pitman–Yor multinomial process and the related urn-scheme. This admits a tractable form if one conditions on $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ that will act as latent variables, thus simplifying computations. Firstly, it can be easily noted that

$$\mathbb{P}(\ell_1 = l_1, \ldots, \ell_k = l_k \mid \boldsymbol{\theta}) \propto \Gamma(c/\sigma + |\boldsymbol{l}|) \prod_{j=1}^{k} \frac{\mathscr{C}(n_j, l_j; \sigma)}{H^{l_j}}, \tag{2.6}$$

and this is concentrated on $\boldsymbol{l} = (l_1, \ldots, l_k)$ such that $l_j \in \{1, \ldots, n_j\}$. These latent random variables can be interpreted in terms of the multiroom Chinese restaurant metaphor—as described in Chapter 3 of this thesis—but we do not pursue the discussion here. An efficient algorithm for sampling independent values from (2.6) is available and presented in Section 2.4. This is very useful since it enables the Monte Carlo approximation of its expectation.

Theorem 2.2. *Let* $(\theta_1, \ldots, \theta_n \mid \tilde{p}^{(H)}) \overset{\text{iid}}{\sim} \tilde{p}^{(H)}$ *and* $\tilde{p}^{(H)} \sim \text{PYM}(\sigma, c; P)$. *If* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ *displays* k *distinct values* $\theta_1^*, \ldots, \theta_k^*$ *with frequencies* $n_1, \ldots, n_k$, *then*

$$\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}) = \left(1 - \frac{k}{H}\right)\left(\frac{c + |\bar{\boldsymbol{\ell}}|\sigma}{c+n}\right) P(A) + \sum_{j=1}^{k} \left(\frac{1}{H}\frac{c + |\bar{\boldsymbol{\ell}}|\sigma}{c+n} + \frac{n_j - \bar{\ell}_j\sigma}{c+n}\right) \delta_{\theta_j^*}(A), \tag{2.7}$$

having set $\bar{\boldsymbol{\ell}} = (\bar{\ell}_1, \ldots, \bar{\ell}_k) = \mathbb{E}(\boldsymbol{\ell} \mid \boldsymbol{\theta})$, $|\bar{\boldsymbol{\ell}}| = \bar{\ell}_1 + \cdots + \bar{\ell}_k$ *and* $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ *is the vector of integer-valued random variables whose distribution is described in* (2.6).

The well-known predictive distribution of the Dirichlet multinomial process is recovered as particular case of Theorem 2.2 after setting $\sigma = 0$. In this special case, the predictive law (2.7) does not depend on the conditional expectations of the underlying latent variables. Moreover, as $H \to \infty$ one can easily see that $\bar{\ell}_j \to 1$ and $|\bar{\boldsymbol{\ell}}| \to k$ in probability. This unsurprisingly implies that, as H increases, the predictive distributions in (2.4) and (2.7) get closer.

The closed form expression of $\Pi_H(n_1, \ldots, n_k)$ in Theorem 2.2 is essential for determining the probability distribution of $K_{n,H}$, the number of random sets. Beside its theoretical relevance, the law of $K_{n,H}$ is often of great importance in applications, e.g. for mixture modeling or for Bayesian clustering. Compared to the Dirichlet multinomial special case, the Pitman–Yor multinomial process induces a richer parametrization for $K_{n,H}$, hence increasing the model flexibility. The Dirichlet multinomial case is recovered as $\sigma \to 0$ in the following theorem.

Theorem 2.3. *If* $(\theta_1, \ldots, \theta_n \mid \tilde{p}^{(H)}) \overset{\text{iid}}{\sim} \tilde{p}^{(H)}$ *and* $\tilde{p}^{(H)} \sim \text{PYM}(\sigma, c; P)$, *the probability distribution of the number of distinct values* $K_{n,H}$ *in* $\boldsymbol{\theta}$ *equals*

$$\mathbb{P}(K_{n,H} = k) = \frac{H!}{(H-k)!}\frac{1}{(c+1)_{n-1}} \sum_{\ell=k}^{n} \frac{1}{H^\ell}\frac{\Gamma(c/\sigma + \ell)}{\Gamma(c/\sigma + 1)} \mathscr{S}(\ell, k)\mathscr{C}(n, \ell; \sigma),$$

for any $k \leqslant \min\{H, n\}$, *where* $\mathscr{S}(\ell, k) = 1/k! \sum_{r=0}^{k} (-1)^{k-r} k! \{\ell!(k-\ell)!\} r^\ell$ *is the Stirling number of the second kind.*

The parameter c controls the location of $K_{n,H}$, while σ regulates both the location and the variability. As suggested by Figure 2.1, the choice $\sigma = 0$ leads to very informative prior distributions, implying that the choice of the location parameter c in the Dirichlet multinomial process is tricky and heavily influences the inferential results. The additional parameter σ of the Pitman–Yor multinomial process allows to circumvent this difficulty, without the need of a hyperprior distribution on c. To illustrate this phenomenon, we depict in Figure 2.2 the distribution of $K_{n,H}$ for different choices of (σ, c), keeping fixed its expectation. As the stable parameter σ increases, the law of $K_{n,H}$ becomes less

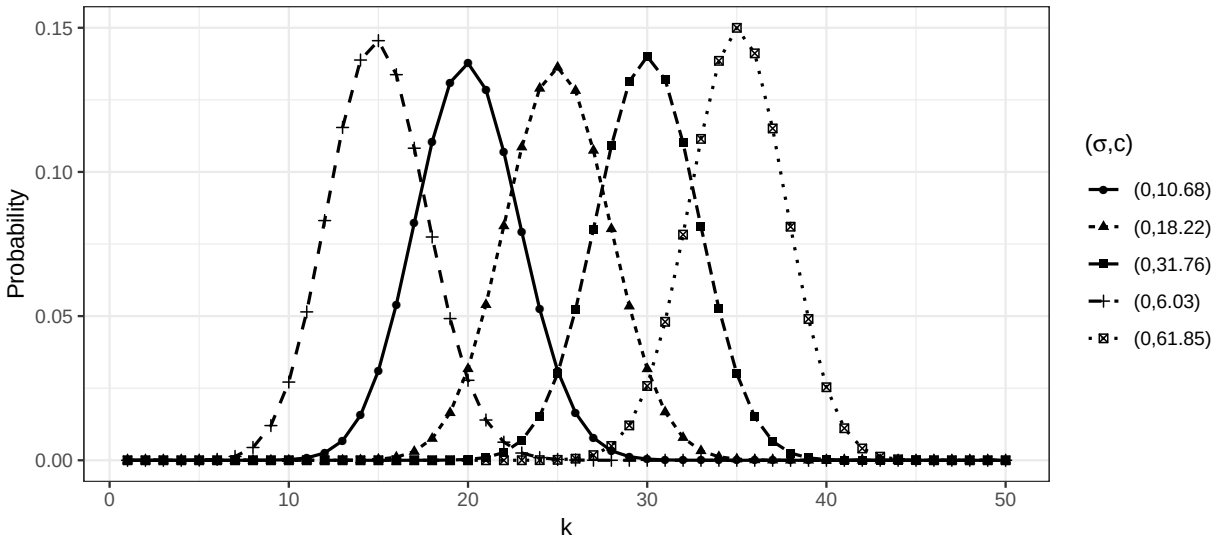


Figure 2.1: Distribution of the number of clusters $\mathbb{P}(K_{n,H} = k)$ in the Dirichlet case ($\sigma = 0$), when $n = 100$, $H = 50$, and for various choices of the location parameter c.

informative. This is further reflected in a higher degree of flexibility and robustness of the Pitman–Yor multinomial process compared to the Dirichlet. This will be empirically confirmed in the convex mixture regression application of Section 2.6. See also De Blasi et al. (2015) and Canale & Prünster (2017) for further discussions on the robustness issue.

2.4 Posterior distribution and latent variables sampling

The usage of the Pitman–Yor multinomial process in applications is greatly facilitated by the availability of its posterior distribution. Indeed, the posterior law of a Pitman–Yor multinomial, conditionally on the set of latent variables (2.6), is available in closed form and it can be written as a linear combination of Dirichlet and ratio-stable distributions. Such a representation parallels the quasi-conjugate posterior characterization of the Pitman–Yor process (Lijoi et al., 2008). When the sample $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ displays $k < H$ distinct values $\theta_1^*, \dots, \theta_k^*$, we let $\bar{\theta}_{k+1}, \dots, \bar{\theta}_H$ represent the point masses in $\tilde{p}^{(H)}$ that are not included in $\boldsymbol{\theta}$, up to a permutation.

Theorem 2.4. *Let* $(\theta_1, \dots, \theta_n \mid \tilde{p}^{(H)}) \overset{\text{iid}}{\sim} \tilde{p}^{(H)}$ *and* $\tilde{p}^{(H)} \sim \text{PYM}(\sigma, c; P)$ *with* P *diffuse. Moreover, let* $\boldsymbol{\ell} = (\ell_1, \dots, \ell_k)$ *be a collection of random variables having distribution* (2.6)*. Then, the posterior distribution of* $\tilde{p}^{(H)}$ *conditional on* $\boldsymbol{\theta}$ *and* $\boldsymbol{\ell}$ *is*

$$(\tilde{p}^{(H)} \mid \boldsymbol{\theta}, \boldsymbol{\ell}) \overset{d}{=} \sum_{j=1}^{k} (W_j + W_{k+1} R_j) \delta_{\theta_j^*} + W_{k+1} \sum_{j=k+1}^{H} R_j \delta_{\bar{\theta}_j},$$

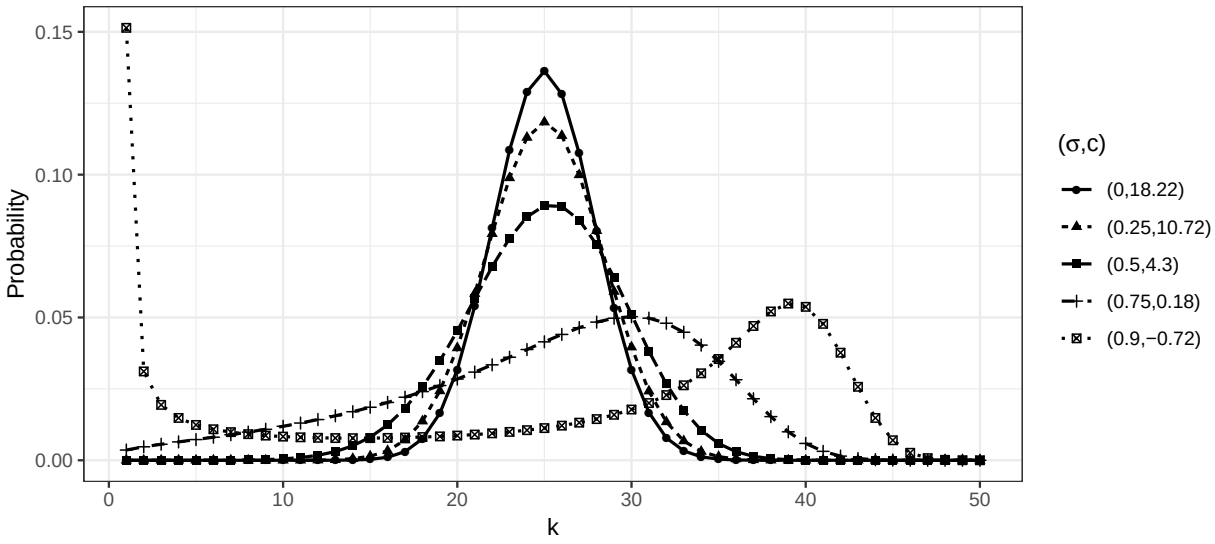


Figure 2.2: Distribution of the number of clusters $\mathbb{P}(K_{n,H} = k)$ in the Pitman–Yor multinomial case when $n = 100$, $H = 50$, and for various choices of (σ, c) so that the expected value $\mathbb{E}(K_{n,H}) = 25$ is fixed.

where $\bar{\theta}_{k+1}, \dots, \bar{\theta}_H$ *are independent and identically distributed from* P. *Moreover,*

$$(W_1, \dots, W_k \mid \boldsymbol{\theta}, \boldsymbol{\ell}) \sim \textsc{dir}(n_1 - \ell_1\sigma, \dots, n_k - \ell_k\sigma, c + |\boldsymbol{\ell}|\sigma),$$

has Dirichlet distribution and it is independent on $(R_1, \dots, R_H \mid \boldsymbol{\theta}, \boldsymbol{\ell})$ *which follows a ratio-stable distribution with updated parameters*

$$(R_1, \dots, R_{H-1} \mid \boldsymbol{\theta}, \boldsymbol{\ell}) \sim \textsc{rs}(\sigma, c + |\boldsymbol{\ell}|\sigma; 1/H, \dots, 1/H).$$

The ratio-stable distribution appearing in Theorem 2.4 is such that $c + |\boldsymbol{\ell}|\sigma > 0$ almost surely for any σ and c, implying that the hierarchical representation of $(R_1, \dots, R_H \mid \boldsymbol{\theta}, \boldsymbol{\ell})$ in terms of tempered-stable random variables, as for Proposition 2.1, can be always exploited directly. It is easy to check that as $\sigma \to 0$ the posterior distribution of the Dirichlet multinomial is recovered, while also being independent on $\boldsymbol{\ell}$.

Therefore, provided that we can simulate independent values from (2.6), we can obtain independent posterior samples for $\tilde{p}^{(H)}$ without the need of Markov chain Monte Carlo. To this end, note that the law of $\boldsymbol{\ell}$ in equation (2.6) is discrete with finite support, meaning that in principle one could directly sample from it. However, standard strategies are computationally feasible only in very simple cases, because the number of support points rapidly increases with n and $(n_1, \dots, n_k)$. We address this issue with a data-augmentation step. Indeed, by expanding over the gamma integral in (2.6), we recognize that, conditionally on a latent variable V, the discrete random variables $\boldsymbol{\ell}$

become independent and therefore much easier to simulate. Specifically, we have

$$\mathbb{P}(\ell_1 = l_1, \dots, \ell_k = l_k \mid \boldsymbol{\theta}, V) \propto \prod_{j=1}^{k} \left(\frac{V}{H}\right)^{l_j} \mathscr{C}(n_j, l_j; \sigma),$$

where V is a positive random variable on $(0, \infty)$ having density, conditional on $\boldsymbol{\theta}$, given by

$$p(v) \propto e^{-v} v^{c/\sigma - 1} \prod_{j=1}^{k} \sum_{\ell_j=1}^{n_j} \left(\frac{v}{H}\right)^{\ell_j} \mathscr{C}(n_j, \ell_j; \sigma).$$

A draw from the above density can be simulated using acceptance-rejection strategies. We obtained good empirical performance with the classic ratio-of-uniform acceptance-rejection algorithm applied on the logarithmic scale $\log V$; see e.g. Devroye (1986). We remark that the generalized factorial coefficients appearing in the above distributions should not be computed directly from their definition, but exploiting instead the recursive relationship $\mathscr{C}(n+1, k; \sigma) = \mathscr{C}(n, k; \sigma)(n - k\sigma) + \sigma\mathscr{C}(n, k-1; \sigma)$, with initial conditions $\mathscr{C}(0, 0; \sigma) = 1$, $\mathscr{C}(n, 0; \sigma) = 0$ for $n > 0$ and $\mathscr{C}(n, k; \sigma) = 0$ for $k > n$.

Remark 2.2. For the sake of the exposition, we described the posterior distribution of the random probability measure $\tilde{p}^{(H)}$. However, Theorem 2.4 leads, with the obvious modifications, to the posterior distribution of the random probabilities $(\pi_1, \dots, \pi_H)$ under multinomial sampling. In such a setting, the frequencies $n_1, \dots, n_k$ correspond to the k occupied cells in a multinomial distribution having probability vector $(\pi_1, \dots, \pi_H)$ and a ratio-stable prior. Then the posterior of $(\pi_1, \dots, \pi_H)$ will coincide with the weights of $\tilde{p}^{(H)}$ in Theorem 2.4.

2.5 Weak limit representation of the Pitman–Yor process

In this section we draw a sharp connection between the Pitman–Yor multinomial process and the infinite-dimensional Pitman–Yor, which is recovered as limiting case when $H \to \infty$. This formal relationship sheds some further light on the interpretation of (σ, c), while motivating the usage of $\tilde{p}^{(H)} \sim \textsc{pym}(\sigma, c; P)$ as an approximation of the infinite-dimensional process $\tilde{p}^{(\infty)} \sim \textsc{py}(\sigma, c; P)$. Our next theorem relies on the notion of weak convergence for random measures; see e.g. Daley & Vere-Jones (2008). Weak convergence implies convergence in distribution also of continuous and bounded functionals, hence including finite dimensional distributions.

Theorem 2.5. *Let* $\tilde{p}^{(H)} \sim \textsc{pym}(\sigma, c; P)$ *and* $\tilde{p}^{(\infty)} \sim \textsc{py}(\sigma, c; P)$. *Then the law of the process* $\tilde{p}^{(H)}$ *weakly converges to the law of* $\tilde{p}^{(\infty)}$ *as* $H \to \infty$. *We will write* $\tilde{p}^{(H)} \xrightarrow{\text{wd}} \tilde{p}^{(\infty)}$.

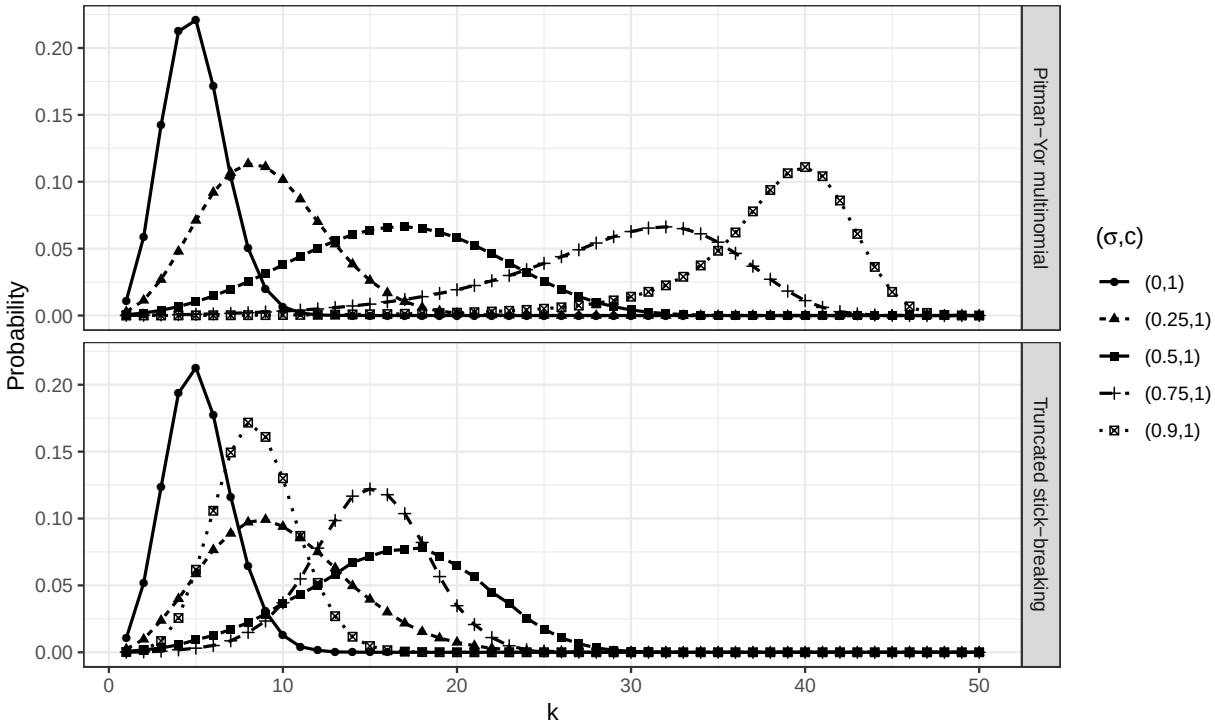


Figure 2.3: Distribution of the number of clusters $\mathbb{P}(K_{n,H} = k)$ in the Pitman–Yor multinomial case (upper plot), and in the truncated stick-breaking case (lower plot), when $n = 100$, $H = 50$, and $c = 1$, for various choices of σ. The distribution of the truncated stick-breaking is obtained averaging over 10^4 Monte Carlo simulations.

On the light of the above theorem, one might want to compare the Pitman–Yor multinomial process with the truncated stick-breaking representation $\tilde{p}_{tr}^{(H)}$ in (1.8), whose use in mixture modeling has become increasingly popular after the work of Ishwaran & James (2001). We devote the remaining of this section to qualitative and formal comparisons between these two weak limit representations.

In first place, it should be acknowledged that the truncated stick-breaking construction lacks a deep theoretical understanding as one cannot rely on results analogous to the ones that we have displayed in the previous sections on the Pitman–Yor multinomial process. Specifically, the exchangeable partition probability function, the associated predictive schemes and the distribution of the number of clusters are not available in closed form. Hence, the usage of $\tilde{p}_{tr}^{(H)}$ as prior law can be motivated only when H is large enough, so that one can consider sampled trajectories of $\tilde{p}_{tr}^{(H)}$, conditional on the data, as reasonable approximations of the realizations of the posterior infinite-dimensional process. In contrast, the Pitman–Yor multinomial can be studied and used even for small values of H, regardless its closeness to the limit case.

In terms of quality of the approximation, there is rather striking argument in favor of the Pitman–Yor multinomial process. It is well-known that the truncation level H

required to be reasonably close to $\tilde{p}^{(\infty)}$ might be exceptionally large, especially when the stable parameter σ approaches 1. In practice, one would typically choose the largest truncation level H which maintains computations feasible. However, this might lead to very poor approximations of the infinite process if the truncated stick-breaking $\tilde{p}^{(H)}_{tr}$ were employed. As an illustration, consider the following example: suppose we are given a sample of $n = 100$ observations and a conservative truncation level $H = 50$ is selected. Then, one might expect that a higher value of σ implies on average an increased number of clusters, paralleling the behavior of the Pitman–Yor process. Unfortunately, this is not the case when the truncated stick-breaking prior is employed, as shown in Figure 2.3. Indeed, the distribution of $K_{n,H}$ increases at first but then decreases as a function of σ and a similar mechanism would hold also for the parameter c. Broadly speaking, this occurs because of the stick-breaking truncation: large values of either σ or c push, on average, the mass of $\tilde{p}^{(H)}_{tr}$ towards the last atom, eventually making $\tilde{p}^{(H)}_{tr}$ collapse to a single random mass. This is a strongly undesirable behavior which has no modeling justification, and furthermore it undermines one of the most appealing property of the Pitman–Yor, namely the ability of controlling the variability of the cluster distribution. On the other hand, the Pitman–Yor multinomial process preserves the peculiar characteristics of the Pitman–Yor process, as shown in Figure 2.3, while still being computationally tractable.

We now conduct a formal comparison between $\tilde{p}^{(H)}$ and $\tilde{p}^{(H)}_{tr}$ within the context of mixture modeling. If the data $(Y_1, \ldots, Y_n \mid \tilde{p}^{(\infty)}) \overset{iid}{\sim} \int_\Theta \mathcal{K}(y;\theta)\tilde{p}^{(\infty)}(d\theta)$ as in (1.7), then the corresponding marginal density is

$$m^{(\infty)}(\boldsymbol{Y}) = \mathbb{E}\left\{\prod_{i=1}^{n}\int_\Theta \mathcal{K}(Y_i;\theta)\tilde{p}^{(\infty)}(d\theta)\right\}, \tag{2.8}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and the expected value is taken with respect to the prior law of $\tilde{p}^{(\infty)}$. Similarly, we define the marginal densities $m^{(H)}$ and $m^{(H)}_{tr}$ as in (2.8), having replaced $\tilde{p}^{(\infty)}$ with the approximations $\tilde{p}^{(H)}$ and $\tilde{p}^{(H)}_{tr}$, respectively. Upper-bounds of the total variation distance between these marginal densities were obtained by Ishwaran & James (2001) in the truncated Pitman–Yor case and Ishwaran & Zarepour (2000, 2002) in the Dirichlet multinomial case. When $\sigma = 0$, the the total variation distance between $m^{(\infty)}$ and $m^{(H)}_{tr}$ vanishes exponentially fast. On the basis of this result Ishwaran & Zarepour (2000) argued that the truncated stick-breaking representation $\tilde{p}^{(H)}_{tr}$ might constitute a better approximation than $\tilde{p}^{(H)}$ in the Dirichlet case. However, the aforementioned exponential decay does not occur for general values of σ and furthermore the quality of the truncated stick-breaking approximation deteriorates as σ increases. These aspects are clarified in the following proposition.

Proposition 2.4. *Let* $m^{(\infty)}$, $m^{(H)}$ *and* $m_{tr}^{(H)}$ *be the marginal densities defined in* (2.8). *If* $\sigma \in (0,1)$ *and* P *is diffuse then*

$$d_{TV}\left\{m_{tr}^{(H)}, m^{(\infty)}\right\} \leqslant 2\left[1 - \left\{1 - \frac{\left(\frac{c}{\sigma}+1\right)_{H-1}}{\left(\frac{c}{\sigma}+\frac{1}{\sigma}\right)_{H-1}}\right\}^{n}\right] = \mathcal{O}(H^{-\frac{1}{\sigma}+1}),\ H \to \infty.$$

If $\Psi_{n,H}$ *and* $\Psi_{n,\infty}$ *are the random partitions associated to* $\tilde{p}^{(H)}$ *and* $\tilde{p}^{(\infty)}$ *respectively, then*

$$d_{TV}\left\{m^{(H)}, m^{(\infty)}\right\} \leqslant d_{TV}(\Psi_{n,H}, \Psi_{n,\infty}) = \mathcal{O}\left(\frac{1}{H}\right), \quad H \to \infty.$$

The total variation distance $d_{TV}(\Psi_{n,H}, \Psi_{n,\infty})$ can be obtained explicitly, although the actual computation could be cumbersome, since it requires the summation over the space of the partitions of $[n]$. We remark that the proportionality constants relative to the above convergence rates are known and they are reported in the Appendix; they are omitted for the sake of the exposition.

The convergence rates of Proposition 2.4 provide some guidance about the advantages of both the approximations. When $\sigma > 1/2$ the convergence rate of $d_{TV}(\Psi_{n,H}, \Psi_{n,\infty})$ is linear regardless the value of σ. In contrast, the upper-bound in the truncated stick-breaking case displays slower converge rates as σ increases, and it is not anymore exponential when $\sigma > 0$. This fact, together with the qualitative findings illustrated in Figure 2.3, suggests that the Pitman–Yor multinomial prior might be preferable especially when σ is large. When $\sigma < 1/2$ the truncated stick-breaking approximation might behave better than the Pitman–Yor multinomial in terms of convergence rates, but the unappealing behavior of $\tilde{p}_{tr}^{(H)}$ highlighted in Figure 2.3 might still occur.

2.6 Simulation study

The additional flexibility provided by the Pitman–Yor multinomial prior is empirically illustrated on a simulated dataset. To ease our discussion, we focus on a simplified version of the model in equations (1.9) and (1.10), which arises when the transition function $f(x) = 0$ is set to zero almost surely for any $x \geqslant 0$. Therefore, in this Section we shall assume that observations $Y_1, \ldots, Y_n$ are conditionally independent realizations from a mixture model

$$(Y_1, \ldots, Y_n \mid \tilde{p}^{(H)}) \overset{iid}{\sim} \sum_{h=1}^{H} \pi_h \mathcal{K}(y; \tilde{\theta}_h),$$

where the random weights and the random locations have Pitman–Yor multinomial distribution. The simulated dataset consists on an independent sample of $n = 300$

observations from a mixture of Gaussians. Specifically, the data generating process is

$$\frac{1}{4}\mathcal{N}(y;-2,0.2^2)+\frac{1}{8}\mathcal{N}(y;-1,0.2^2)+\frac{1}{4}\mathcal{N}(y;0,0.2^2)+\frac{1}{8}\mathcal{N}(y;1,0.2^2)+\frac{1}{4}\mathcal{N}(y;2,0.2^2),$$

where $\mathcal{N}(y;\mu,\sigma^2)$ denotes the density function of a Gaussian distribution with mean μ and variance σ^2. The true number of mixture components is 5 and we would like to infer it from the data. This is a classical problem in Bayesian mixture modeling, which was addressed for instance in Richardson & Green (1997) by placing a prior distribution on H. In contrast, we rely on the approach advocated by Malsiner-Walli et al. (2016), which has foundations on the asymptotic results of Rousseau & Mengersen (2011). In such a setting H is assumed to be large enough, meaning that it should be interpreted as an upper bound for the true number of components. The optimal number of clusters is inferred by inspecting the posterior distribution of $K_{n,H}$, the random number of distinct values. While such an approach is appealing because of its simplicity, in the Dirichlet multinomial case the results may depend on the choice of the parameter c, as discussed for instance by Ishwaran & Zarepour (2000). The Pitman–Yor multinomial prior addresses this difficulty and allows for a more robust specification without the need of a hyperprior for c. This is achieved by simply enlarging the prior variability of $K_{n,H}$, which is indeed quite low in the Dirichlet case. The simulation study we conduct serves as an empirical confirmation of this aspect, which was theoretically investigated a priori in Section 2.3.

We let the kernel function $\mathcal{K}(y;\theta)$ to be a Gaussian density having mean μ and variance σ^2, with $\theta=(\mu,\sigma^2)$. We choose conditionally conjugate priors for the atoms $\tilde{\theta}_h=(\tilde{\theta}_{1h},\tilde{\theta}_{2h})$ for $h=1,\dots,H$, and in particular we assume independent gamma priors for the precisions $\tilde{\theta}_{2h}^{-1}\overset{\text{iid}}{\sim}\text{GA}(a_\sigma,b_\sigma)$ and independent Gaussian priors for the locations $\tilde{\theta}_{11},\dots,\tilde{\theta}_{1H}$, with mean μ_μ and variance σ_μ^2. We set the hyperparameters consistently with the data generating process to make the prior distributions centered on the true values but still relatively vague. More precisely, we set $\mu_\mu=0$ and $\sigma_\mu^2=1000$, whereas we let $a_\sigma=2.5$ and $b_\sigma=0.1$.

We let $H=20$, a fairly conservative upper bound for the true number of mixture components, which is 5 in our simulation study. We consider four different prior specifications for the ratio-stable parameters σ and c, as summarized in Table 2.1. In two of these scenarios, the stable parameter σ is set to 0, corresponding to the Dirichlet multinomial, which we aim at comparing with the Pitman–Yor prior. As evidenced in Table 2.1, two hyperparameters settings are well calibrated, in the sense that the expected values of the number of clusters $K_{n,H}$ are both close to 5 a priori. Under these calibrated choices, we expect that the model is able to correctly recover the correct number of clusters a posteriori. However, in real applications one does not know the true number

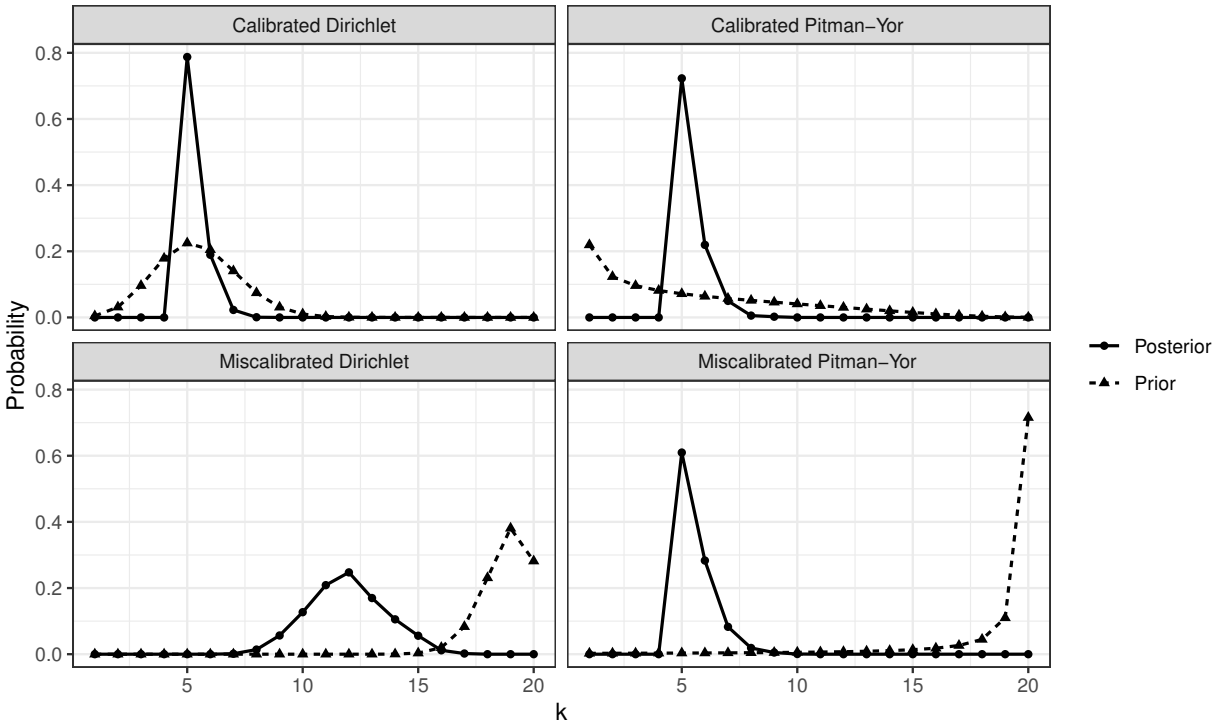


Figure 2.4: Prior and posterior distributions of the number of clusters $\mathbb{P}(K_{n,H} = k)$ corresponding to the four scenarios described in Table 2.1. The a posteriori distributions are obtained averaging over 5000 Markov chain Monte Carlo samples.

	Calibrated DM	Calibrated PYM	Miscalibrated DM	Miscalibrated PYM
c	1	-0.18	20	-0.02
σ	0	0.40	0	0.80
$\mathbb{E}(K_{n,H})$	5.42	5.42	18.81	18.81

Table 2.1: Hyperparameter settings for the simulation study. DM denotes the Dirichlet multinomial.

of components and therefore she might inadvertently adopt a miscalibrated prior for the data at hand. This scenario is mimicked by considering hyperparameters that lead to a priori expectations for $K_{n,H}$ close to the upper bound, leading to an "overfitted" mixture model.

Posterior samples for $K_{n,H}$ in each scenario can be obtained via Markov chain Monte Carlo through classical Gibbs sampling schemes for finite mixture models, and leveraging Theorem 2.4 for the step involving the full-conditional distribution of $(\pi_1, \dots, \pi_H)$. We run the algorithm for 7000 iterations holding out the first 2000 as burn-in period. From the results in Figure 2.4 it is evident that both the calibrated choices (Dirichlet and Pitman–Yor) are able to recover the correct number of clusters, which is unsurprising given that the prior law was already concentrated around the true value. Conversely,

in the overfitted scenarios the differences are marked: under the Dirichlet multinomial specification the distribution of $K_{n,H}$ struggles to deviate from the prior, whereas in the Pitman–Yor multinomial case the posterior law of $K_{n,H}$ correctly recover the true number of mixture components. This behavior motivates the usage of the Pitman–Yor multinomial to robustify mixture modeling, with applications beyond the convex mixture regression modeling presented in this chapter.

2.7 Convex mixture regression modeling

The Dichlorodiphenyldichloroethylene (DDE) is a persistent metabolite of the pesticide DDT, and it is measured in the maternal serum during the third trimester of pregnancy. Although the DDT has been widely used against malaria-transmitting mosquitoes, its presence has been linked to preterm birth, a major contributor to infant mortality (Longnecker et al., 2001). Hence, there is strong interest in relating the dose level of the DDE to the corresponding risk of premature delivery. In Longnecker et al. (2001) the gestational age at delivery is dichotomized using a 37-week cut-off, following standard practice in reproductive epidemiology. Although this might simplify the modeling, it arguably leads to a loss of information (Dunson & Park, 2008; Canale et al., 2018). The convex mixture regression model outlined in Section 2.1 provides the basis for a flexible and interpretable method for quantitative risk assessment.

Recall that the observations $Y_1, \ldots, Y_n$ for $n = 2312$ represent the gestational ages at delivery (weeks) and they are independent draws from the mixture model of equation (1.9) in Chapter 1. The covariate-dependent random probability measure of equation (1.10) is employed as mixing measure. To improve flexibility and robustness, we leverage on the Pitman–Yor multinomial specification for two different components of the model. Specifically, the discrete random probability measure $\tilde{p}$ in the convex mixture representation (1.10) will follow a Pitman–Yor multinomial and we assume a ratio-stable distribution for the weights $(\beta_1, \ldots, \beta_M)$ in equation (1.11). To summarize, the gestational ages at delivery $Y_1, \ldots, Y_n$ are conditionally independent draws from the mixture density

$$\{1 - f(x)\} \sum_{h=1}^{H} \pi_h \mathcal{K}(y; \tilde{\theta}_h) + f(x) \mathcal{K}(y; \tilde{\theta}_\infty), \qquad x \geqslant 0, \tag{2.9}$$

where $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$ are independent and identically distributed random variables from P and where the mixing weights $(\pi_1, \ldots, \pi_H)$ are such that

$$(\pi_1, \ldots, \pi_{H-1}) \sim \text{RS}(\sigma, c; 1/H, \ldots, 1/H),$$

	Low Variance Dir.	Low Variance RS	High Variance Dir.	High Variance RS
c	1	0	1	0
σ	0	0.3	0	0.3
c_β	20	1.1	2	-0.7
σ_β	0	0.9	0	0.9
$\mathrm{sd}(\beta_m)$	0.07	0.07	0.17	0.17

Table 2.2: Hyperparameter settings for the convex mixture regression model.

for $\sigma \in [0, 1)$ and $c > -\sigma$. Moreover, recall from equation (1.11) in Chapter 1 that $f(x) = \sum_{m=1}^{M} \mathcal{B}_m(x)\beta_m$, then we specify

$$(\beta_1, \ldots, \beta_{M-1}) \sim \mathrm{RS}(\sigma_\beta, c_\beta; 1/M, \ldots, 1/M),$$

with $\sigma_\beta \in [0, 1)$ and $c_\beta > -\sigma_\beta$. Although several alternatives are available for the shape-constrained basis functions $\mathcal{B}_1(x), \ldots, \mathcal{B}_M(x)$, a tractable default choice is the I-splines basis (Ramsay, 1988), with the knots placed on the empirical quantiles of the DDE. This is slightly different from the approach of Canale et al. (2018), who consider I-splines with equally spaced knots. Moreover, we set $\mathcal{B}_M(x) = 0$ to allow asymptotes in $f(x)$.

Compared to more complex covariate-dependent approaches, the convex mixture regression model is appealing for quantitative risk assessment because of its intuitive interpretation. Specifically, when there is no exposure to DDE ($x = 0$), observations are drawn from a mixture model directed by $\tilde{p}^{(H)}$ because $f(0) = 0$. Conversely, at high exposure levels ($x \to \infty$), observations smoothly shift towards a more adverse health profile, represented by $\tilde{\theta}_\infty$. Such a transition is regulated by the function $f(x)$, which has an explicit interpretation. Let $\tilde{F}_x(y)$ be the cumulative distribution function associated to the density in equation (2.9). A common risk assessment measure is the additional risk function $\tilde{F}_x(a) - \tilde{F}_0(a)$, which is evaluated in some fixed clinical threshold a. Hence, one can show that $f(x) \propto \tilde{F}_x(a) - \tilde{F}_0(a)$, implying that $f(x)$ constitutes a standardized measure of risk which does not depend on the chosen threshold a. Because of this property, the transition function $f(x)$ is of great inferential interest.

We estimate the convex mixture regression model under different hyperparameters settings, which are reported in Table 2.2. Moreover, we let $H = M = 10$, to parallel the choices of Canale et al. (2018). Since the true data generating mechanism is unknown, there is no clear notion of calibrated model. Hence, to emphasize the differences between Dirichlet and Pitman–Yor priors, we consider two different variability levels for the vector $(\beta_1, \ldots, \beta_M)$ appearing in the specification of $f(x)$. The variance of each weight β_m can be obtained from the formula $\mathrm{var}(\beta_m) = H^{-1}(1 - H)^{-1}(1 - \sigma_\beta)/(c_\beta + 1)$; see Carlton

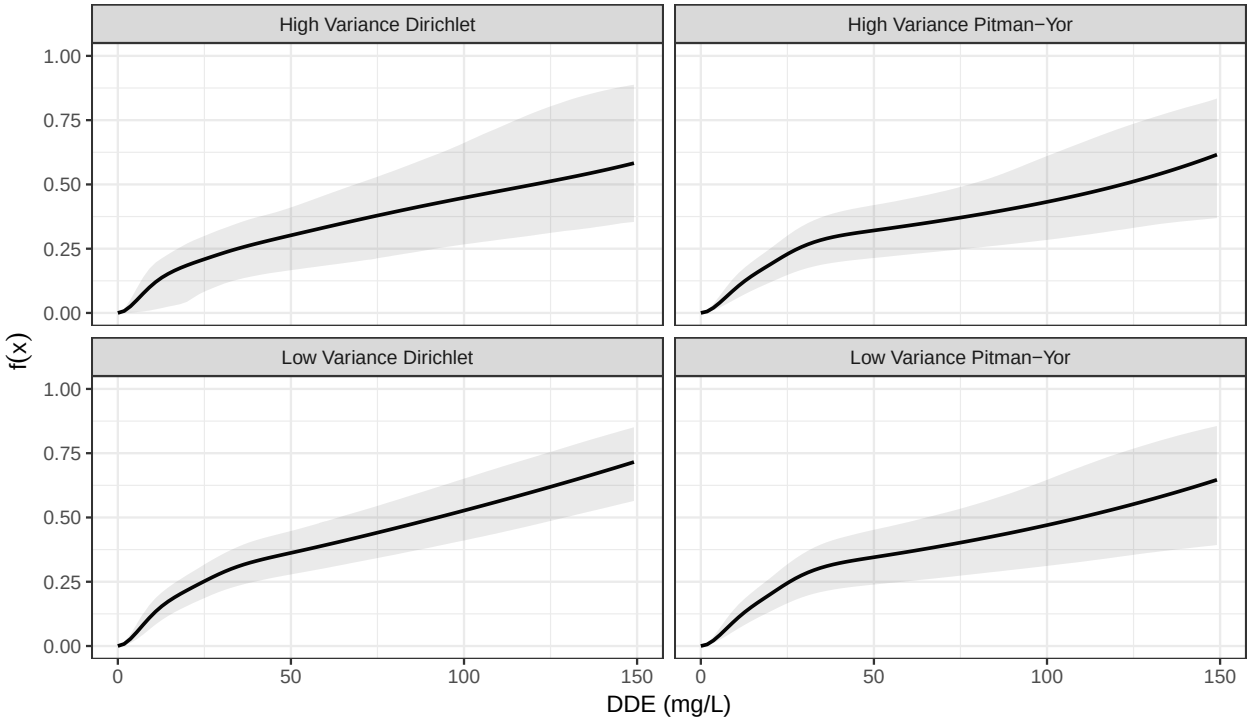


Figure 2.5: Posterior summaries of the functions $f(x)$ in the four scenarios described in Table 2.2. The solid lines correspond to the posterior means, whereas the shaded areas denote 95% pointwise credible intervals.

(2002). In the low variance scenarios, the prior is quite concentrated around its prior expectation, and vice versa in the high variance ones.

Consistently with Canale et al. (2018), let the kernel function $\mathcal{K}(y;\theta)$ in equation (1.9) be a Gaussian density having mean μ and variance σ^2, with $\theta = (\mu, \tau)$. Under this choice, the prior distributions for each atom $\tilde{\theta}_h = (\tilde{\theta}_{1h}, \tilde{\theta}_{2h})$ for $h = 1, \ldots, H$, in equation (2.9) and for the adverse health profile atom $\tilde{\theta}_\infty = (\tilde{\theta}_{1\infty}, \tilde{\theta}_{2\infty})$ can be chosen conditionally conjugate. In first place, we assume independent gamma priors for the precisions $\tilde{\theta}_{2h}^{-1} \overset{\text{iid}}{\sim} \text{GA}(a_\sigma, b_\sigma)$, independently also on $\tilde{\theta}_{2\infty} \sim \text{GA}(a_\sigma, b_\sigma)$. Moreover, we specify independent truncated Gaussian distributions for the locations $\tilde{\theta}_{11}, \ldots, \tilde{\theta}_{1H}$ and $\tilde{\theta}_{1\infty}$ with mean μ_μ and variance σ_μ^2. The truncations are imposed to met an adversity health profile property, namely that

$$\tilde{\theta}_{1\infty} < \tilde{\theta}_{1h}, \qquad h = 1, \ldots, H,$$

almost surely. Broadly speaking, such a constraint enforces large values of the `DDE` ($x \to \infty$) to be associated on average with a greater risk of premature birth. The extent of such a risk will be inferred from the data. We set the hyperparameters consistently

with previous works, so that $a_\sigma = b_\sigma = 2$ and $\sigma_\mu^2 = 10$, whereas we set $\mu_\mu = 39.27$, the arithmetic mean of the observed gestational ages at delivery (weeks).

Markov chain Monte Carlo is required for approximating the posterior distribution of a convex mixture regression model, and this can be accomplished via Gibbs sampling. Our algorithm closely resembles the one of Canale et al. (2018), with straightforward adjustments in the updates of the vectors $(\pi_1, \ldots, \pi_H)$ and $(\beta_1, \ldots, \beta_M)$, which are modified according to Theorem 2.4. We run the Gibbs sampling algorithm for 60000 iterations, having discarded the first 10000 samples as burn in period. The estimated curves $f(x)$, under the four scenarios of Table 2.2, are depicted in Figure 2.5 together with their credible intervals. In the low variance scenario the Dirichlet multinomial prior slightly overestimate the function $f(x)$ compared to the other estimates and to the results in Canale et al. (2018), while also underestimating its variability. Conversely, the Pitman–Yor multinomial prior, while having the same variability a priori of the Dirichlet, it recovers a posteriori essentially the same variability level provided in the high variance scenarios. This effect is due to the robustness property of the Pitman–Yor multinomial prior, which was discussed in Section 2.3 and empirically demonstrated in the simulation study of Section 2.6.

2.8 Appendix

Throughout the Appendix, we will make extensive use of an alternative construction of the Pitman–Yor process based on completely random measures. Refer to Lijoi & Prünster (2010) for a review on nonparametric priors using completely random measures as a unifying concept. Any homogeneous and almost surely finite completely random measures without fixed points of discontinuity is characterized by the Laplace functional

$$\mathbb{E}\left\{e^{-\int_\Theta f(\theta)\tilde{\mu}(d\theta)}\right\} = \exp\left[-\int_{\mathbb{R}_+\times\Theta}\left\{1-e^{-sf(\theta)}\right\}\rho(s)dscP(d\theta)\right],$$

for any $f : \Theta \to \mathbb{R}_+$ and with $\rho(s)dscP(d\theta)$ the Lévy intensity function associated to $\tilde{\mu}$. The σ-stable process (Kingman, 1975) is identified by setting $c = 1$, $\rho(s) = \sigma s^{-1-\sigma}/\Gamma(1-\sigma)$, for some $\sigma \in (0,1)$, and letting $\mathbb{P}_\sigma$ denote its probability distribution. Let $\mathbb{P}_{\sigma,c}$ be another probability measure which is absolutely continuous with respect to $\mathbb{P}_\sigma$ and such that

$$\frac{d\mathbb{P}_{\sigma,c}}{d\mathbb{P}_\sigma}(p) = \frac{\Gamma(c+1)}{\Gamma(c/\sigma+1)}p^{-c}(\Theta). \tag{2.10}$$

The resulting random measure $\tilde{\mu}_{\sigma,c}$ with distribution $\mathbb{P}_{\sigma,c}$ is almost surely discrete while not completely random. Clearly when $c = 0$ then $\tilde{\mu}_\sigma = \tilde{\mu}_{\sigma,0}$ is a σ-stable

completely random measure. Moreover, $\tilde{p}^{(\infty)} = \tilde{\mu}_{\sigma,c}/\tilde{\mu}_{\sigma,c}(\Theta)$ is a Pitman–Yor process $\tilde{p}^{(\infty)} \sim \text{PY}(\sigma, c; P)$.

Proofs of Propositions 2.1,2.2 and 2.3

Because of the almost sure discreteness of $\tilde{p}_0^{(H)}$, one can equivalently write the Pitman–Yor multinomial process as

$$\tilde{p}^{(H)} = \sum_{h=1}^{\infty} \xi_h \delta_{\tilde{\phi}_h} = \sum_{h=1}^{H} \pi_h \delta_{\tilde{\theta}_h}, \qquad \pi_h = \sum_{j:\tilde{\phi}_j = \tilde{\theta}_h} \xi_j,$$

where $\tilde{\theta}_h$ are independent and identically distributed draws from P, while $\tilde{\phi}_h$ are, conditionally on $\tilde{p}_0^{(H)}$, independent and identically distributed draws from $\tilde{p}_0^{(H)}$. The distribution of $(\pi_1, \ldots, \pi_{H-1})$ is obtained after noting that

$$(\pi_1, \ldots, \pi_{H-1} \mid \tilde{p}_0^{(H)}) = \{\tilde{p}^{(H)}(\{\tilde{\theta}_1\}), \ldots, \tilde{p}^{(H)}(\{\tilde{\theta}_{H-1}\}) \mid \tilde{p}_0^{(H)}\} \sim \text{RS}(\sigma, c; 1/H, \ldots, 1/H),$$

since $\tilde{p}_0^{(H)}(\{\tilde{\theta}_h\}) = 1/H$ for $h = 1, \ldots, H$. However, this implies that $(\pi_1, \ldots, \pi_{H-1})$ is independent on $(\tilde{\theta}_1, \ldots, \tilde{\theta}_H)$, thus proving Proposition 2.1. Now assume $\sigma \in (0, 1)$ and $c > 0$. By exploiting the change of measure formula (2.10), we can represent the Pitman–Yor multinomial process as $(\tilde{p}^{(H)} \mid \tilde{p}_0^{(H)}) = (\tilde{\mu}_{\sigma,c}/\tilde{\mu}_{\sigma,c}(\Theta) \mid \tilde{p}_0^{(H)})$, where the Laplace functional of $(\tilde{\mu}_{\sigma,c} \mid \tilde{p}_0^{(H)})$ for any measurable function $f : \Theta \to \mathbb{R}_+$ is

$$\begin{aligned}
\mathbb{E}\left\{e^{-\int_\Theta f(\theta)\tilde{\mu}_{\sigma,c}(d\theta)} \mid \tilde{p}_0^{(H)}\right\} &= \frac{\Gamma(c+1)}{\Gamma(c/\sigma+1)}\mathbb{E}\left\{\tilde{\mu}_\sigma(\Theta)^{-c} e^{-\int_\Theta f(\theta)\tilde{\mu}_\sigma(d\theta)} \mid \tilde{p}_0^{(H)}\right\} \\
&= \frac{c}{\Gamma(c/\sigma+1)}\int_0^\infty u^{c-1}\mathbb{E}\left\{e^{-u\tilde{\mu}_\sigma(\Theta)-\int_\Theta f(x)\tilde{\mu}_\sigma(dx)} \mid \tilde{p}_0^{(H)}\right\} du \\
&= \frac{c}{\Gamma(c/\sigma+1)}\int_0^\infty u^{c-1} e^{-u^\sigma}\mathbb{E}\left\{e^{-\int_\Theta f(\theta)\tilde{\mu}_\sigma^{(u)}(d\theta)} \mid \tilde{p}_0^{(H)}\right\} du,
\end{aligned} \tag{2.11}$$

where $(\tilde{\mu}_\sigma^{(u)} \mid \tilde{p}_0^{(H)})$ is a completely random measure with tilted Lévy intensity $\rho^{(u)}(s) = \sigma/\Gamma(1-\sigma)s^{-1-\sigma}e^{-us}$ and baseline probability measure $\tilde{p}_0^{(H)}$, which identifies a generalized gamma process, and whose finite-dimensional distribution are tempered-stables random variables. Hence, we can write

$$(\pi_1, \ldots, \pi_{H-1} \mid U, \tilde{p}_0^{(H)}) \overset{d}{=} \left(\frac{J_1}{\sum_{h=1}^H J_h}, \ldots, \frac{J_H}{\sum_{h=1}^H J_h} \mid U, \tilde{p}_0^{(H)}\right),$$

with $U^\sigma \sim \text{GA}(c/\sigma, 1)$ and $(J_h \mid U, \tilde{p}_0^{(H)}) = (\tilde{\mu}_\sigma^{(u)}(\{\tilde{\theta}_h\}) \mid U, \tilde{p}_0^{(H)}) \sim \text{TS}(1/H, \sigma, U)$ independently and identically distributed for $h = 1, \ldots, H$, which concludes the proof for

$c > 0$. As before, the dependence on $\tilde{p}_0^{(H)}$ can be dropped because $\tilde{p}_0^{(H)}(\{\tilde{\theta}_h\}) = 1/H$ for $h = 1, \ldots, H$, which implies the independence between $J_1, \ldots, J_H$ and $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$. Clearly when $c = 0$ the Laplace functional of $(\tilde{\mu}_\sigma \mid \tilde{p}_0^{(H)})$ is already that of a σ-stable completely random measure, and the result immediately follows. The proof of Proposition 2.3 is a direct consequence of the first equality in equation (2.11).

Proof of Theorem 2.1

The exchangeable partition probability function by definition is

$$\Pi_H(n_1, \ldots, n_k) = \sum_{i_1 \neq \cdots \neq i_k} \mathbb{E}\left(\prod_{j=1}^k \pi_{i_j}^{n_j}\right) = \frac{H!}{(H-k)!}\mathbb{E}\left(\prod_{j=1}^k \pi_j^{n_j}\right),$$

where the sum runs over all the vectors $(i_1, \ldots, i_k)$ of distinct positive integers such that $i_j \in \{1, \ldots, H\}$, whereas the second equality is a consequence of the symmetricity of the law of the weights. Moreover, recalling the change of measure (2.10), the above expected value con be expressed as

$$\begin{aligned}\mathbb{E}\left(\prod_{j=1}^k \pi_j^{n_j}\right) &= \mathbb{E}\left\{\prod_{j=1}^k \frac{J_j^{n_j}}{(\sum_{h=1}^H J_h)^{n_j}}\right\} = \frac{\Gamma(c+1)}{\Gamma(c/\sigma+1)}\mathbb{E}\left\{\tilde{\mu}_\sigma(\Theta)^{-c-n}\prod_{j=1}^k \tilde{\mu}_\sigma(\{\tilde{\theta}_j\})^{n_j}\right\} \\ &= \frac{1}{\Gamma(c/\sigma+1)}\frac{1}{(c+1)_{n-1}}\int_0^\infty u^{c+n-1}e^{-u^\sigma}\prod_{j=1}^k \mathcal{V}_{n_j,H}(u)du,\end{aligned}$$

where each $\mathcal{V}_{m,H}(u)$ is defined, for every $m \geqslant 1$ and $u > 0$, as follow

$$\mathcal{V}_{m,H}(u) = \left\{(-1)^m \frac{\partial^m}{\partial u^m} e^{-u^\sigma/H}\right\} e^{u^\sigma/H} = \sum_{\ell=1}^m H^{-\ell}\mathscr{C}(m, \ell; \sigma)u^{-m+\ell\sigma}. \tag{2.12}$$

The last equality may be proved with some combinatorial manipulation; its derivation entails similar steps as those in the supplementary material of Camerlenghi et al. (2019). Hence, one has that

$$\begin{aligned}\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!}\frac{1}{\Gamma(c/\sigma+1)}\frac{1}{(c+1)_{n-1}} \\ \times \sum_{\boldsymbol{\ell}}\int_0^\infty u^{c+|\boldsymbol{\ell}|\sigma}e^{-u^\sigma}du\prod_{j=1}^k H^{-\ell_j}\mathscr{C}(n_j, \ell_j; \sigma),\end{aligned}$$

where the sum runs over $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ for $\ell_j \in \{1, \ldots, n_j\}$. The change of variable $v = u^\sigma$ in the above integral yields the desired result.

Proof of Theorem 2.2

The predictive distribution is easily obtained from the exchangeable partition probability function, since

$$\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}) \propto \Pi_H(n_1, \ldots, n_k, 1) P(A) + \sum_{j=1}^{k} \Pi_H(n_1, \ldots, n_j + 1, \ldots, n_k) \delta_{\theta_j^*}(A).$$

The above coefficients can be both expressed as a mixture over $\boldsymbol{\ell}$. Let $\mathscr{V}_{n,k} = \prod_{j=1}^{k-1}(c + j\sigma)/(c+1)_{n-1}$, then from Theorem 2.1

$$\Pi_H(n_1, \ldots, n_k, 1) = \frac{H!}{(H-k-1)!} \sum_{\boldsymbol{\ell}} \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|+1}}{H^{|\boldsymbol{\ell}|+1}} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{\sigma^{\ell_j}}.$$

Moreover, exploiting the recursive relationship $\mathscr{C}(n_j + 1, \ell_j; \sigma) = \mathscr{C}(n_j, \ell_j; \sigma)(n_j - \ell_j\sigma) + \sigma\mathscr{C}(n_j, \ell_j - 1; \sigma)$ (Charalambides, 2002), and after some algebraic manipulations can express the term $\Pi_H(n_1, \ldots, n_j + 1, \ldots, n_k)$ as follows

$$\frac{H!}{(H-k)!} \sum_{\boldsymbol{\ell}} \left\{ \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|+1}}{H^{|\boldsymbol{\ell}|+1}} \prod_{i=1}^{k} \frac{\mathscr{C}(n_i, \ell_i; \sigma)}{\sigma^{\ell_i}} + \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|}}{H^{|\boldsymbol{\ell}|}} (n_j - \ell_j\sigma) \prod_{i=1}^{k} \frac{\mathscr{C}(n_i, \ell_i; \sigma)}{\sigma^{\ell_i}} \right\}.$$

Then, by augmenting over the set of random variables $(\ell_1, \ldots, \ell_k)$ and after normalization, one obtains

$$\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}, \boldsymbol{\ell}) = \left(1 - \frac{k}{H}\right) \left(\frac{c + |\boldsymbol{\ell}|\sigma}{c+n}\right) P(A) + \sum_{j=1}^{k} \left(\frac{1}{H} \frac{c + |\boldsymbol{\ell}|\sigma}{c+n} + \frac{n_j - \ell_j\sigma}{c+n}\right) \delta_{\theta_j^*}(A),$$

and the desired predictive distribution follows after taking the expectation with respect to (2.6).

Proof of Theorem 2.3

For any $k \leqslant \min\{H, n\}$ the probability of the number of clusters can be expressed in terms of the exchangeable partition probability function

$$\begin{aligned}
\mathbb{P}(K_{n,H} = k) &= \frac{1}{k!} \sum_{\boldsymbol{n}^{(k)} \in \Delta_n} \binom{n}{n_1, \ldots, n_k} \Pi_H(n_1, \ldots, n_k) \\
&= \frac{H!}{(H-k)!} \sum_{s=k}^{n} \frac{\mathscr{V}_{n,s}}{\sigma^s H^s} \sum_{\boldsymbol{n}^{(k)} \in \Delta_n} \sum_{\boldsymbol{\ell} \in \Delta_s(\boldsymbol{n}^{(k)})} \frac{1}{k!} \binom{n}{n_1, \ldots, n_k} \prod_{j=1}^{k} \mathscr{C}(n_j, \ell_j; \sigma),
\end{aligned}$$

where the first sum $\boldsymbol{n}^{(k)} \in \Delta_n$ runs over all the positive integers $\boldsymbol{n}^{(k)} = (n_1, \ldots, n_k)$ such that $|\boldsymbol{n}^{(k)}| = n$, and where $\boldsymbol{\ell} \in \Delta_s(\boldsymbol{n}^{(k)})$ denotes the sum over all the integers $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ such that $\ell_j \in \{1, \ldots, n_j\}$ and $|\boldsymbol{\ell}| = s$. Then, by interchanging the order of the summation and exploiting well-known combinatorial identities, we obtain

$$\begin{aligned}\mathbb{P}(K_{n,H} = k) &= \frac{H!}{(H-k)!} \sum_{s=k}^{n} \frac{\mathscr{V}_{n,s}}{\sigma^s H^s} \sum_{\boldsymbol{\ell} \in \Delta_s} \frac{1}{k!} \sum_{\boldsymbol{n}^{(k)} \in \Delta_n(\boldsymbol{\ell})} \binom{n}{n_1, \ldots, n_k} \prod_{j=1}^{k} \mathscr{C}(n_j, \ell_j; \sigma) \\ &= \frac{H!}{(H-k)!} \sum_{s=k}^{n} \frac{\mathscr{V}_{n,s}}{\sigma^s H^s} \mathscr{S}(s, k) \mathscr{C}(n, s; \sigma),\end{aligned}$$

where $\boldsymbol{n}^{(k)} \in \Delta_n(\boldsymbol{\ell})$ denotes the summation over all the integers $(n_1, \ldots, n_k)$ such that $n_j \in \{\ell_j, \ldots, n\}$ and $|\boldsymbol{n}^{(k)}| = n$, whereas $\boldsymbol{\ell} \in \Delta_s$ denotes the summation over all the integers $\boldsymbol{\ell}$ such that $|\boldsymbol{\ell}| = s$. Then the result follows after some algebra.

Proof of Theorem 2.4

We first state, without proof, the following technical lemma concerning the posterior distribution of $\tilde{p}_0^{(H)}$. The proof is based on basic properties of species sampling models and it is given in the Appendix of Chapter 3.

Lemma 2.1. *Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be a draw from an exchangeable sequence directed by a Pitman–Yor multinomial process and let P be diffuse. Then the posterior distribution of $\tilde{p}_0^{(H)}$ is*

$$(\tilde{p}_0^{(H)} \mid \boldsymbol{\theta}) \stackrel{d}{=} \frac{1}{H} \left(\sum_{j=1}^{k} \delta_{\theta_j^*} + \sum_{j=k+1}^{H} \delta_{\tilde{\theta}_j} \right),$$

where $\tilde{\theta}_{k+1}, \ldots, \tilde{\theta}_h$ are independent and identically distributed draws from P.

Because of the symmetry of the weights, we can assume without loss of generality that the distinct values $\theta_1^*, \ldots, \theta_k^*$ are associated to the first k random weights $\pi_1, \ldots, \pi_k$ of $\tilde{p}^{(H)}$. Recalling representation (2.10), for any function $f : \Theta \to \mathbb{R}_+$, the Laplace functional of $(\tilde{\mu}_{\sigma,c} \mid \tilde{p}_0^{(H)})$ given the observations is

$$\mathbb{E}\left\{ e^{-\tilde{\mu}_{\sigma,c}(f)} \mid \boldsymbol{\theta}, \tilde{p}_0^{(H)} \right\} = \frac{\mathbb{E}\left\{ e^{-\int_\Theta f(\theta) \tilde{\mu}_{\sigma,c}(d\theta)} \prod_{j=1}^{k} \pi_j^{n_j} \mid \tilde{p}_0^{(H)} \right\}}{\mathbb{E}\left(\prod_{j=1}^{k} \pi_j^{n_j} \mid \tilde{p}_0^{(H)} \right)},$$

where $\tilde{\mu}_{\sigma,c}(f) = \int_\Theta f(\theta)\tilde{\mu}_{\sigma,c}(d\theta)$. Hence, following the same steps as for Theorem 2.1, the above Laplace functional may be written as

$$\mathbb{E}\left\{e^{-\tilde{\mu}_{\sigma,c}(f)} \mid \boldsymbol{\theta}, \tilde{p}_0^{(H)}\right\} = \\ = \frac{\int_0^\infty u^{c+n-1} e^{-\frac{1}{H}\sum_{j=1}^k \{f(\theta_j^*)+u\}^\sigma} e^{-\frac{1}{H}\sum_{j=k+1}^H \{f(\tilde{\theta}_j)+u\}^\sigma} \prod_{j=1}^k \mathcal{V}_{n_j,H}\{f(\theta_j^*)+u\}du}{\int_0^\infty u^{c+n-1} e^{-u^\sigma} \prod_{j=1}^k \mathcal{V}_{n_j,H}(u)du}.$$

where each $\mathcal{V}_{m,H}(u)$ is defined as in (2.12). Hence, by augmenting the above Laplace functional over the set of latent variables $\boldsymbol{\ell}$ with distribution function (2.6), we obtain that

$$\mathbb{E}\left\{e^{-\tilde{\mu}_{\sigma,c}(f)} \mid \boldsymbol{\theta}, \boldsymbol{\ell}, \tilde{p}_0^{(H)}\right\} \propto \\ \propto \int_0^\infty u^{c+|\boldsymbol{\ell}|\sigma-1} e^{-\frac{1}{H}\sum_{j=k+1}^H \{f(\tilde{\theta}_j)+u\}^\sigma} e^{-\frac{1}{H}\sum_{j=1}^k \{f(\theta_j^*)+u\}^\sigma} \prod_{j=1}^k \left\{1+\frac{f(\theta_j^*)}{u}\right\}^{-n_j+\ell_j\sigma} du.$$

Hence, after normalization, the Laplace functional equals

$$\mathbb{E}\left\{e^{-\tilde{\mu}_{\sigma,c}(f)} \mid \boldsymbol{\theta}, \boldsymbol{\ell}, \tilde{p}_0^{(H)}\right\} = \int_0^\infty \prod_{j=k+1}^H \mathbb{E}\left\{e^{-f(\tilde{\theta}_j)J_j^*}\right\} \prod_{j=1}^k \mathbb{E}\left\{e^{-f(\theta_j^*)(J_j^*+I_j)}\right\} p^{(\infty)}(u)du,$$

where $p^{(\infty)}(u) = \sigma/\Gamma(c/\sigma+|\boldsymbol{\ell}|)u^{c+|\boldsymbol{\ell}|\sigma-1}e^{-u^\sigma}\mathbb{1}_{(0,\infty)}(u)$ is a density function and the corresponding random variable, say U, is such that $U^\sigma \sim \text{GA}(c/\sigma+|\boldsymbol{\ell}|,1)$. Hence, conditionally on U and by marginalizing over $\tilde{p}_0^{(H)}$ as for Lemma 2.1, we get the following posterior representation for the unormalized Pitman–Yor multinomial process

$$(\tilde{\mu}_{\sigma,c} \mid \boldsymbol{\theta}, \boldsymbol{\ell}, U) \stackrel{d}{=} \sum_{j=1}^k (J_j^* + I_j)\delta_{\theta_j^*} + \sum_{j=k+1}^H J_j^*\delta_{\tilde{\theta}_j}, \tag{2.13}$$

where $\tilde{\theta}_j$ are independent and identically distributed random variables from P and in addition

$$(J_h^* \mid \boldsymbol{\theta}, \boldsymbol{\ell}, U) \overset{\text{iid}}{\sim} \text{TS}(1/H, \sigma, U), \qquad h = 1, \ldots, H,$$

whereas

$$(I_j \mid \boldsymbol{\theta}, \boldsymbol{\ell}, U) \overset{\text{ind}}{\sim} \text{GA}(n_j - \ell_j\sigma, U), \qquad j = 1, \ldots, k.$$

Equation (2.13) already leads to a posterior representation of $\tilde{p}^{(H)}$. The normalization and the subsequent marginalization with respect to the random variable U leads to the

final representation. In first place, set

$$W_j = \frac{I_j}{\sum_{j'=1}^{k} I_{j'} + \sum_{h=1}^{H} J_h^*}, \qquad j = 1, \dots, k, \qquad W_{k+1} = \frac{\sum_{j=1}^{k} J_j^*}{\sum_{j=1}^{k} I_j + \sum_{h=1}^{H} J_h^*},$$

whereas set $R_h = J_h^* / \sum_{h'=1}^{H} J_{h'}^*$ for $h = 1, \dots, H$. Note that the distribution of the vector od random variables $(W_1, \dots, W_k, W_{k+1})$ can be represented as follow

$$(\tilde{W}_1(1 - W_{k+1}), \dots, \tilde{W}_k(1 - W_{k+1}), W_{k+1}), \quad \tilde{W}_j = \frac{I_j}{\sum_{j'=1}^{k} I_{j'}}, \quad j = 1, \dots, k,$$

and therefore the vector $(\tilde{W}_1, \dots, \tilde{W}_k)$, given $\boldsymbol{\theta}$ and $\boldsymbol{\ell}$, has Dirichlet distribution and it is independent on $(R_1, \dots, R_H)$, on W_{k+1} and on U . Indeed, the random variable U cancels almost surely in the ratio $I_j / \sum_{j'=1}^{k} I_{j'}$, whereas the independence on W_{k+1} is a consequence of the independence of the Dirichlet distribution with its own total mass $\sum_{j'=1}^{k} I_{j'}$. Because of Proposition 2.1, we recognize that $(R_1, \dots, R_H \mid \boldsymbol{\theta}, \boldsymbol{\ell})$ follows a ratio-stable distribution. We now prove that, given $\boldsymbol{\theta}$ and $\boldsymbol{\ell}$, the random vector $(R_1, \dots, R_H)$ is independent on W_{k+1}, which in turns is shown to be beta distributed.

Let $p(s_1, \dots, s_H, \iota, u)$ be the density function associated to the random variables $(J_1^*, \dots, J_H^*)$, $\sum_{j=1}^{k} I_j$ and U, respectively, given the observations $\boldsymbol{\theta}$ and the latent variables $\boldsymbol{\ell}$. Representation (2.13) implies that such a density factorizes as

$$p(s_1, \dots, s_H, \iota, u) = p^{(\infty)}(u) p_u(\iota) \prod_{h=1}^{H} p_u^{(\sigma)}(s_h),$$

where $p^{(\infty)}(u)$ is defined as before, where $p_u(\iota)$ is the density of a gamma random variable, and where each $p_u^{(\sigma)}(s_h)$ represents the density of a tempered stable distribution. Now consider the change of variable $r_h = s_h/(s_1 + \cdots + s_H)$ for $h = 1, \dots, H-1$, $s = s_1 + \cdots + s_H$, $w = (s_1 + \cdots + s_H)/\{\iota + (s_1 + \cdots + s_H)\}$ and $u = u$. The resulting density is given by

$$\begin{aligned} p(r_1, \dots, r_{H-1}, s, w, u) &= s^H w^{-2} p^{(\infty)}(u) p_u\{s(1-w)w^{-1}\} \prod_{h=1}^{H} p_u^{(\sigma)}(s r_h), \\ &\propto u^{n+c-1} s^{n+H-|\boldsymbol{\ell}|\sigma-1} \frac{(1-w)^{n-|\boldsymbol{\ell}|\sigma-1}}{w^{n-|\boldsymbol{\ell}|\sigma+1}} e^{-u\{s(1-w)w^{-1}\} - u^\sigma} \prod_{h=1}^{H} p_u^{(\sigma)}(s r_h), \\ &\propto u^{n+c-1} e^{-us/w} s^{n+H-|\boldsymbol{\ell}|\sigma-1} \frac{(1-w)^{n-|\boldsymbol{\ell}|\sigma-1}}{w^{n-|\boldsymbol{\ell}|\sigma+1}} \prod_{h=1}^{H} p^{(\sigma)}(s r_h), \end{aligned}$$

where $r_H = 1 - (r_1 + \cdots + r_{H-1})$ and where it has been exploited the relationship $p_u^{(\sigma)}(sr_h) = e^{u^\sigma/H} e^{-usr_h} p^{(\sigma)}(sr_h)$, with $p^{(\sigma)}(sr_h)$ denoting the density of a positive stable distribution. Then, the integration over u and s leads to $p(r_1, \ldots, r_{H-1}, w) = \int_0^\infty \int_0^\infty p(r_1, \ldots, r_{H-1}, s, w, u) du ds$ and

$$p(r_1, \ldots, r_{H-1}, w) \propto w^{c+|\ell|\sigma-1}(1-w)^{n-|\ell|\sigma-1} \int_0^\infty s^{H-|\ell|\sigma-c-1} \prod_{h=1}^{H} p^{(\sigma)}(sr_h) ds.$$

This concludes the proof since $(W_{k+1} \mid \boldsymbol{\theta}, \boldsymbol{\ell}) \sim \text{BETA}(c + |\ell|\sigma, n - |\ell|\sigma)$ and it is independent on $(R_1, \ldots, R_H \mid \boldsymbol{\theta}, \boldsymbol{\ell})$.

Proof of Theorem 2.5

The proof relies on the convergence of the exchangeable partition probability function in Theorem 2.1 to that of the Pitman–Yor process, which can be easily checked. Indeed, for any collection of measurable subsets $A_1, \ldots, A_n$ of Θ it holds

$$\begin{aligned}\mathbb{P}(\theta_1 \in A_1, \ldots, \theta_n \in A_n) &= \sum_{\Psi} \Pi_H(n_1, \ldots, n_k) \prod_{j=1}^{k} P(\cap_{i \in C_j} A_i) \\ &\to \sum_{\Psi} \Pi_\infty(n_1, \ldots, n_k) \prod_{j=1}^{k} P(\cap_{i \in C_j} A_i) = \mathbb{P}(\phi_1 \in A_1, \ldots, \phi_n \in A_n), \quad H \to \infty,\end{aligned}$$

with $(\theta_n)_{n \geqslant 1}$ and $(\phi_n)_{n \geqslant 1}$ being two exchangeable sequences directed by $\tilde{p}_H$ and $\tilde{p}_\infty$, respectively, and where the sum runs over the space of partitions with $\Psi = \{C_1, \ldots, C_k\}$. Using de Finetti representation theorem, the latter is equivalent to

$$E\left\{\prod_{i=1}^{n} \tilde{p}_H(A_i)\right\} \to E\left\{\prod_{i=1}^{n} \tilde{p}_\infty(A_i)\right\}, \qquad H \to \infty.$$

Since the above convergence holds for any collection of sets $A_1, \ldots, A_n$ and any $n \geqslant 1$, we can write equivalently

$$E\left\{\prod_{j=1}^{k} \tilde{p}_H(B_j)^{n_j}\right\} \to E\left\{\prod_{j=1}^{k} \tilde{p}_\infty(B_j)^{n_j}\right\}, \qquad H \to \infty, \tag{2.14}$$

for any measurable collection $B_1, \ldots, B_k$. The random vector $\{\tilde{p}_\infty(B_1), \ldots, \tilde{p}_\infty(B_k)\}$ is positive and bounded, thus being fully determined by its cross-moments. Hence, the vector version of Theorem 30.2 in Billingsley (1995) ensures that convergence of the

cross-moments (2.14) implies the convergence in distribution, namely

$$\{\tilde{p}^{(H)}(B_1),\dots,\tilde{p}^{(H)}(B_k)\} \to \{\tilde{p}^{(\infty)}(B_1),\dots,\tilde{p}^{(\infty)}(B_k)\}, \quad H \to \infty,$$

in the sense of weak convergence. The weak convergence of the whole process is then guaranteed by Theorem 11.1.VII in Daley & Vere-Jones (2008).

Proof of Proposition 2.4

From Theorem 1 and Theorem 2 in Ishwaran & James (2001) it follows that

$$d_{TV}\left\{m_{tr}^{(H)}, m^{(\infty)}\right\} \leqslant 2\left[1-\left\{1-\frac{\left(\frac{c}{\sigma}+1\right)_{H-1}}{\left(\frac{c}{\sigma}+\frac{1}{\sigma}\right)_{H-1}}\right\}^{n}\right].$$

We will say that two sequences of real numbers $(a_h)_{h\geqslant 1}$ and $(b_h)_{h\geqslant 1}$ are asymptotically equivalent, written $a_h \approx b_h$, if $\lim_{h\to\infty} a_h/b_h = 1$. The order of convergence stated in Proposition 2.4 is a direct consequence of the standard properties of the Gamma function. Indeed, note that

$$\frac{\left(\frac{c}{\sigma}+1\right)_{H-1}}{\left(\frac{c}{\sigma}+\frac{1}{\sigma}\right)_{H-1}} \approx \frac{\Gamma\left(\frac{c}{\sigma}+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{c}{\sigma}+1\right)} H^{-\frac{1}{\sigma}+1}, \quad H \to \infty,$$

from which it follows that

$$2\left[1-\left\{1-\frac{\left(\frac{c}{\sigma}+1\right)_{H-1}}{\left(\frac{c}{\sigma}+\frac{1}{\sigma}\right)_{H-1}}\right\}^{n}\right] \approx 2n\frac{\Gamma\left(\frac{c}{\sigma}+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{c}{\sigma}+1\right)} H^{-\frac{1}{\sigma}+1} = \mathcal{O}(H^{-\frac{1}{\sigma}+1}), \quad H \to \infty.$$

Moreover, following the same steps as for Theorem 4 in Ishwaran & Zarepour (2002) we get

$$\begin{aligned} d_{TV}\left\{m^{(H)}, m^{(\infty)}\right\} &\leqslant d_{TV}(\Psi_{n,H}, \Psi_{n,\infty}) \\ &\leqslant \frac{1}{2}\sum_{\Psi} |\Pi_H(n_1,\dots,n_k) - \Pi_\infty(n_1,\dots,n_k)| \\ &\leqslant \frac{1}{2}\sum_{\Psi} \Pi_\infty(n_1,\dots,n_k)\left|1-\frac{\Pi_H(n_1,\dots,n_k)}{\Pi_\infty(n_1,\dots,n_k)}\right|, \end{aligned}$$

where the sum runs over all the partitions Ψ of $[n]$ with cardinalities $n_1,\dots,n_k$. From the proof of Theorem 2.1 and Theorem 2.2 it follows that

$$\frac{\Pi_H(n_1,\dots,n_k)}{\Pi_\infty(n_1,\dots,n_k)} = \frac{H!}{H^k(H-k)!}\sum_{\ell}\frac{\mathscr{V}_{n,|\ell|}}{\mathscr{V}_{n,k}}\frac{1}{H^{|\ell|-k}}\prod_{j=1}^{k}\frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\sigma^{\ell_j}}\frac{1}{(1-\sigma)_{n_j-1}},$$

from which we obtain

$$\lim_{H\to\infty} H\left|1-\frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_\infty(n_1,\ldots,n_k)}\right| = \sum_{j=1}^{k} \frac{\mathscr{V}_{n,k+1}}{\mathscr{V}_{n,k}} \frac{\mathscr{C}(n_j,2;\sigma)}{\sigma^2} \frac{1}{(1-\sigma)_{n_j-1}},$$

since the term not vanishing in the summation over $\boldsymbol{\ell}$ are those for which $\ell_1+\cdots+\ell_k = k+1$, meaning that each $\ell_j = 1$ but one equal to 2, and recalling also that $\mathscr{C}(n_j,1;\sigma) = \sigma(1-\sigma)_{n_j-1}$. Hence, one get

$$d_{TV}(\Psi_{n,H},\Psi_{n,\infty}) \approx \frac{1}{2H}\sum_{\Psi} \Pi_\infty(n_1,\ldots,n_k) \sum_{j=1}^{k} \frac{\mathscr{V}_{n,k+1}}{\mathscr{V}_{n,k}} \frac{\mathscr{C}(n_j,2;\sigma)}{\sigma^2} \frac{1}{(1-\sigma)_{n_j-1}} = \mathcal{O}\left(\frac{1}{H}\right),$$

which concludes the proof.

Chapter 3

Finite-dimensional normalized random measures

3.1 Summary

The chapter is organized as follows. In Section 3.2 we review some background material about completely random measures and homogeneous normalized random measures. In Section 3.3 we define NIDM processes and we discuss several *a priori* properties, such as the law of the random partition they induce, and the weak convergence to NRMIS. In Section 3.4, we discuss generalized urn schemes and posterior characterizations, with a particular emphasis on the normalized generalized gamma (NGG) multinomial process. In Section 3.5 we employ the NGG multinomial prior for the analysis of a real dataset, highlighting practical advantages over existing methods.

3.2 Homogeneous normalized random measures

As discussed in Chapter 1, NIDM processes are closely related to homogeneous NRMIS, whose definition is therefore briefly recalled here. To this end, we also recall the definition of a noteworthy class of random measures.

Let Θ be a separable and complete metric space and let $\mathscr{B}(\Theta)$ be its Borel σ-field. We will denote with $\mathcal{M}_\Theta$ the space of boundedly finite measures on $\{\Theta, \mathscr{B}(\Theta)\}$ and with $\mathscr{M}_\Theta$ the corresponding σ-algebra. For technical details on the construction of $(\mathcal{M}_\Theta, \mathscr{M}_\Theta)$ one may refer to Daley & Vere-Jones (2008). A random element, say $\tilde{\mu}^{(\infty)}$ defined on some probability space and taking values in $(\mathcal{M}_\Theta, \mathscr{M}_\Theta)$ such that $\tilde{\mu}^{(\infty)}(B_1), \ldots, \tilde{\mu}^{(\infty)}(B_d)$ are mutually independent random variables for any choice of pairwise disjoint $B_1, \ldots, B_d$ in $\mathscr{B}(\Theta)$, and for any $d \geqslant 2$, is a *completely random measure* (CRM). As exhaustively discussed in Kingman (1967), a CRM with no fixed points of discontinuity and no deterministic drift can be represented as $\tilde{\mu}^{(\infty)} = \sum_{h=1}^{\infty} J_h \delta_{\tilde{\phi}_h}$ and is characterized by the Laplace functional

transform

$$\mathbb{E}\left(e^{-\int_\Theta f(\theta)\tilde{\mu}^{(\infty)}(d\theta)}\right) = \exp\left\{-\int_{\Theta\times\mathbb{R}_+}\left(1-e^{-sf(\theta)}\right)\mathcal{L}(ds,d\theta)\right\}, \tag{3.1}$$

where $f:\Theta\to\mathbb{R}_+$ is a measurable function such that $\int_\Theta f(\theta)\tilde{\mu}^{(\infty)}(d\theta)<\infty$ almost surely, whereas the measure $\mathcal{L}$ on $\mathbb{R}_+\times\Theta$, termed Lévy measure, or intensity, characterizes the CRM and is such that $\int_{\mathbb{R}_+\times A}\min\{1,s\}\mathcal{L}(ds,d\theta)<\infty$ for any $A\in\mathscr{B}(\Theta)$. In the following, we consider only *homogeneous* CRMs, which amounts to having a Lévy intensity of the form

$$\mathcal{L}(ds,d\theta) = \rho(s)ds\; cP(d\theta),$$

where P is a probability measure over $\{\Theta,\mathscr{B}(\Theta)\}$ and $c\in\mathbb{R}_+$ is a positive constant. We will use the notation $\tilde{\mu}^{(\infty)}\sim \text{CRM}(c,\rho;P)$. If one additionally has that $\int_{\mathbb{R}_+}\rho(s)\,ds=\infty$, then $0<\tilde{\mu}^{(\infty)}(\Theta)<\infty$ almost surely, and a homogeneous NRMI is defined as

$$\tilde{p}^{(\infty)} = \sum_{h=1}^{\infty}(J_h/\bar{J})\delta_{\tilde{\phi}_h},$$

where $\bar{J}=\sum_{h=1}^{\infty}J_h=\tilde{\mu}^{(\infty)}(\Theta)$, the $\tilde{\phi}_h$'s are iid draws from P, independently also from the jumps $(J_h)_{h\geqslant 1}$. We will write $\tilde{p}^{(\infty)}\sim\text{NRMI}(c,\rho;P)$ to denote such a random probability measure. Several relevant nonparametric priors are NRMIs and a noteworthy class, which is the object of investigation in the present chapter, arises when

$$\rho(s) = \frac{1}{\Gamma(1-\sigma)}\,s^{-1-\sigma}\,e^{-\kappa s}, \tag{3.2}$$

whose additional parameters $0\leqslant\sigma<1$ and $\kappa\geqslant 0$ are such that at least one of them is positive (Brix, 1999). The resulting NRMI is often referred to as *normalized generalized gamma* (NGG) process, and it includes some well–known nonparametric priors as special cases. See Lijoi et al. (2007). For example, if we set $\sigma=0$ and $\kappa=1$ we recover the Dirichlet process, whereas for $\sigma=\kappa=\frac{1}{2}$ one obtains the normalized inverse–Gaussian process introduced in Lijoi et al. (2005). Finally, with $\kappa=0$ we get the normalized σ-stable process (Kingman, 1975).

Both NRMIs and the novel class of NIDM processes are discrete random probability measures. Thus, as discussed in Chapter 2, when their law identifies the de Finetti measure of the exchangeable sequence of Θ–valued random elements $(\theta_n)_{n\geqslant 1}$, there will be ties with positive probability, namely $\mathbb{P}[\theta_i=\theta_{i'}]>0$ for any $i\neq i'$. Hence, an n-sample $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_n)$ will display $K_n=k\leqslant n$ distinct values, say $\theta_1^*,\ldots,\theta_k^*$, with respective frequencies $n_1,\ldots,n_k$, so that $\sum_{j=1}^k n_j=n$. This amount to saying that $\boldsymbol{\theta}$ induces a random partition Ψ_n of $[n]=\{1,\ldots,n\}$ into k sets $C_1,\ldots,C_k$ such that $i\in C_j$

if and only if $\theta_i = \theta_j^*$. As discussed in Pitman (1996) and previously in Chapter 2, this clustering mechanism is regulated by a symmetric function called *exchangeable partition probability function* (EPPF), whose definition was given in Chapter 2 and it is recalled here

$$\Pi(n_1, \ldots, n_k) = \mathbb{P}(\Psi_n = \{C_1, \ldots, C_k\}) = \sum_{i_1 \neq \cdots \neq i_k} \mathbb{E}\left(\prod_{j=1}^{k} \pi_{i_j}^{n_j}\right), \tag{3.3}$$

where the vector $\boldsymbol{n}_k = (n_1, \ldots, n_k)$ of positive integers is such that $n_j = \sharp C_j$ and $\sum_{j=1}^{k} n_j = n$ and the sum runs over all the positive and distinct integers $(i_1, \ldots, i_k)$. The EPPF is an extremely useful tool and it has a simple interpretation: it is the probability of recording a specific partition induced by $\boldsymbol{\theta}$ into k distinct groups, each represented by respective distinct values $\theta_1^*, \ldots, \theta_k^*$, with vector of corresponding frequencies $\boldsymbol{n}^{(k)} = (n_1, \ldots, n_k)$. The availability of the EPPF yields, as a by-product, the system of predictive distributions. Indeed, the conditional probabilities that θ_{n+1} displays a new value generated from the diffuse base measure P or coincides with the previously observed θ_j^*, given $\boldsymbol{\theta}$, are

$$w_0(\boldsymbol{n}^{(k)}) = \frac{\Pi(n_1, \ldots, n_k, 1)}{\Pi(n_1, \ldots, n_k)}, \qquad w_j(\boldsymbol{n}^{(k)}) = \frac{\Pi(n_1, \ldots, n_j + 1, \ldots, n_k)}{\Pi(n_1, \ldots, n_k)}, \quad j = 1, \ldots, k.$$

and, hence, for any $A \in \mathscr{B}(\Theta)$

$$\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}) = w_0(\boldsymbol{n}^{(k)}) P(A) + \sum_{j=1}^{k} w_j(\boldsymbol{n}^{(k)}) \delta_{\theta_j^*}(A).$$

For homogeneous NRMIs with diffuse baseline measure P the expression of the EPPF will be denoted with $\Pi_\infty(n_1, \ldots, n_k)$, and it is known as it can be expressed in terms of the underlying parameters (c, ρ). Indeed, if $\tau_m(u) := \int_{\mathbb{R}_+} s^m e^{-us} \rho(s)\, ds$, for any integer $m \geqslant 1$, then

$$\Pi_\infty(n_1, \ldots, n_k) = \frac{c^k}{\Gamma(n)} \int_{\mathbb{R}_+} u^{n-1}\, e^{-c\psi(u)} \prod_{j=1}^{k} \tau_{n_j}(u)\, du, \tag{3.4}$$

where the function $\psi(u) = \int_{\mathbb{R}_+} (1 - e^{-us}) \rho(s) ds$ is termed *Laplace exponent*. See, e.g., James et al. (2009). In the NGG special case, one finds out that

$$\Pi_\infty(n_1, \ldots, n_k) = \mathscr{V}_{n,k} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1},$$

with $(a)_n = a(a+1)\cdots(a+n-1)$, for any real a and integer $n \geqslant 1$, being the ascending factorial, and with $(a)_0 = 1$. If we set $\bar{c} = c\kappa^\sigma/\sigma$, and with $\Gamma(x; a) = \int_x^\infty s^{a-1} e^{-s} ds$ for

any $a > 0$ being the incomplete gamma function, one further knows that

$$\mathscr{V}_{n,k} := \frac{e^{\bar{c}}\sigma^{k-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \bar{c}^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \bar{c}\right). \tag{3.5}$$

Thus, a NGG random measure is a Gibbs type prior (Gnedin & Pitman, 2005; De Blasi et al., 2015), and the only one also being a NRMI (Lijoi et al., 2008). As a final remark, note that the parametrization (c, κ, σ) is redundant, since the above EPPF only depends on the parameter vector $(\bar{c}, \sigma)$, meaning that one can fix either c or κ without loss of generality.

3.3 Normalized infinitely divisible multinomial processes

In order to define a more flexible class of finite-dimensional priors that overcomes the limitations of the standard Dirichlet multinomial model, we do rely on a normalization procedure similar to the one that yields NRMIS. In doing so, we make use of finitely many jumps whose distribution has Laplace transform available in closed form. For this reason we make use of random measures with finitely many support points and with masses having an infinitely divisible distribution.

3.3.1 NIDM processes

The main building block of the new class of processes that we are proposing is a collection of independent and infinitely divisible random variables. In particular, we will deal with *finite* and *strictly positive* infinitely divisible random variables *without drift*, say J, whose probability distribution has Laplace transform that can be expressed as

$$\mathbb{E}\left(e^{-\lambda J}\right) = \exp\{-c\psi(\lambda)\} = \exp\left(-c\int_{\mathbb{R}_+}(1-e^{-\lambda s})\rho(s)\,ds\right), \tag{3.6}$$

for any $\lambda > 0$ and some positive constant $0 < c < \infty$. See Sato (1999) for details. The function $\psi(\lambda)$ is the Laplace exponent, whereas ρ is any non–negative measurable function such that $\int_{\mathbb{R}_+} \min\{1, s\}\rho(s)\,ds < \infty$ and $\int_{\mathbb{R}_+} \rho(s)\,ds = \infty$. Note that these conditions coincide with those involved in the construction of homogeneous NRMIS. A random variable J having Laplace transform (3.6) will be denoted $J \sim \text{ID}(c, \rho)$.

Definition 3.1. A measurable function $\tilde{\mu}^{(H)}$ defined on some probability space and taking values in $(\mathcal{M}_\Theta, \mathscr{M}_\Theta)$ such that

$$\tilde{\mu}^{(H)} = \sum_{h=1}^{H} J_h \delta_{\tilde{\theta}_h}, \qquad J_h \overset{\text{iid}}{\sim} \text{ID}\left(\frac{c}{H}, \rho\right), \qquad \tilde{\theta}_h \overset{\text{iid}}{\sim} P, \tag{3.7}$$

where $0 < c < \infty$, with $H < \infty$, and where P is a probability measure defined over $\{\Theta, \mathscr{B}(\Theta)\}$, is a *multinomial* random measure with *infinitely divisible* jumps. We will use the notation $\tilde{\mu}^{(H)} \sim \text{IDM}(c, \rho; P)$.

It is important to stress that IDM's are not completely random, because the random variables $\tilde{\mu}^{(H)}(B_1), \ldots, \tilde{\mu}^{(H)}(B_d)$ are not mutually independent random variables for any choice of pairwise disjoint sets $B_1, \ldots, B_d$ in $\mathscr{B}(\Theta)$ and for any $d \geqslant 2$. Although this fact might appear as counter-intuitive at a first glance, it can be understood with the following example. Let $A \in \mathscr{B}(\Theta)$ be such that $0 < P(A) < 1$, and note that $\mathbb{P}(\tilde{\mu}^{(H)}(A) = 0) = \{1 - P(A)\}^H > 0$. On the other hand, $\mathbb{P}(\tilde{\mu}^{(H)}(A) = 0 \mid \tilde{\mu}^{(H)}(\Theta \setminus A) = 0) = 0$ since $\mathbb{P}(\tilde{\mu}^{(H)}(\Theta) > 0) = 1$. Hence, independence between $\tilde{\mu}^{(H)}(A)$ and $\tilde{\mu}^{(H)}(\Theta \setminus A)$ does not hold true. Nonetheless, the Laplace functional transform of a IDM is available and equals

$$\mathbb{E}\left(e^{-\int_\Theta f(\theta)\tilde{\mu}^{(H)}(d\theta)}\right) = \left\{\int_\Theta \exp\left(-\frac{c}{H}\int_{\mathbb{R}_+}\left(1 - e^{-sf(\theta)}\right)\rho(s)ds\right) P(d\theta)\right\}^H, \tag{3.8}$$

where $f : \Theta \to \mathbb{R}_+$ is any measurable function such that $\int_\Theta f(\theta)\tilde{\mu}^{(H)}(d\theta) < \infty$ almost surely. Now set $f(\theta) = \lambda \mathbb{1}_A(\theta)$ with $\lambda > 0$, $A \in \mathscr{B}(\Theta)$ and $\mathbb{1}_A$ denoting the indicator function of A. Then the Laplace transform of $\tilde{\mu}^{(H)}(A)$ is

$$\mathbb{E}\left(e^{-\lambda\tilde{\mu}^{(H)}(A)}\right) = \left[1 - P(A) + P(A)\exp\left\{-\frac{c}{H}\psi(\lambda)\right\}\right]^H, \tag{3.9}$$

where ψ is the Laplace exponent defined in (3.6). It is apparent that $\tilde{\mu}^{(H)}(A)$ equals in distribution a binomial compound random element, namely

$$(\tilde{\mu}^{(H)}(A) \mid \tilde{m}) \stackrel{d}{=} \sum_{j=0}^{\tilde{m}} J_j, \qquad \tilde{m} \sim \text{BINOMIAL}\{H, P(A)\},$$

where, conditional on $\tilde{m}$, the random variables $J_1, \ldots, J_{\tilde{m}}$ are iid from $\text{ID}(c/H, \rho)$ and J_0 is a point mass at zero: this provides an interesting and alternative representation of the random variable $\tilde{\mu}^{(H)}(A)$.

The random measure displayed in Definition 3.1 is the main tool for defining normalized infinitely divisible multinomial processes. In this respect, the construction is reminiscent of the normalized infinitely divisible family of distributions in Favaro et al. (2011), which is based on the normalization of infinitely divisible random variables. If $J_i \stackrel{\text{ind}}{\sim} \text{ID}(c_i, \rho)$, for $i = 1, \ldots, d$ and with $c_i > 0$, the random vector

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{d-1}) = \left(\frac{J_1}{\sum_{i=1}^d J_i}, \ldots, \frac{J_{d-1}}{\sum_{i=1}^d J_i}\right),$$

is a *normalized infinitely divisible* (NID) random vector and will be denoted as $\boldsymbol{\pi} \sim \text{NID}(c_1, \ldots, c_d; \rho)$. The case $c_i = 0$ is not explicitly considered here, since it would imply $\pi_i = 0$ and the distribution of $\boldsymbol{\pi}$ would degenerate on a lower dimensional simplex.

Definition 3.2. Let $\tilde{\mu}^{(H)} \sim \text{IDM}(c, \rho; P)$ as in Definition 3.1. If

$$\tilde{p}^{(H)} = \frac{\tilde{\mu}^{(H)}}{\tilde{\mu}^{(H)}(\Theta)} = \sum_{h=1}^{H} \pi_h \delta_{\tilde{\theta}_h},$$

where $(\pi_1, \ldots, \pi_H) = (J_1/\sum_{h=1}^{H} J_h, \ldots, J_H/\sum_{h=1}^{H} J_h)$, then $\tilde{p}^{(H)}$ is a *normalized infinitely divisible multinomial process*. We will use the notation $\tilde{p}^{(H)} \sim \text{NIDM}(c, \rho; P)$.

From such a definition, it is apparent that $\tilde{\mu}(\Theta) \sim \text{ID}(c, \rho)$, thus ensuring that the above normalization is well–defined. As for the probability weights assigned to the points $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$, one has

$$(\pi_1, \ldots, \pi_{H-1}) \sim \text{NID}\left(\frac{c}{H}, \ldots, \frac{c}{H}; \rho\right).$$

Note that a NIDM process is also a proper species sampling model (1.6) with $H < \infty$, when P is diffuse. If we set $\rho(s) = s^{-1}e^{-s}$ we recover the Dirichlet multinomial process, whose weights $(\pi_1, \ldots, \pi_{H-1}) \sim \text{DIRICHLET}(c/H, \ldots, c/H)$. Henceforth, we plan to use ρ as in (3.2) and obtain a novel and tractable class of NIDM processes, which we will term NGG multinomial process.

It is useful to point out that NIDM processes display a hierarchical structure. Indeed, it holds

$$(\tilde{\mu}^{(H)} \mid \tilde{p}_{0,H}) \sim \text{CRM}(c, \rho; \tilde{p}_0^{(H)}), \qquad \tilde{p}_0^{(H)} = \frac{1}{H}\sum_{h=1}^{H} \delta_{\tilde{\theta}_h},$$

with $\tilde{\theta}_h \overset{\text{iid}}{\sim} P$, then $\tilde{\mu}^{(H)} \sim \text{IDM}(c, \rho; P)$. Hence, if we pick $\tilde{p}^{(H)} = \tilde{\mu}^{(H)}/\tilde{\mu}^{(H)}(\Theta)$ one has that $\tilde{p}^{(H)} \sim \text{NIDM}(c, \rho; P)$ and in view of the previous remark, it can be described through the following hierarchical model

$$(\tilde{p}^{(H)} \mid \tilde{p}_0^{(H)}) \sim \text{NRMI}(c, \rho; \tilde{p}_0^{(H)}), \qquad \tilde{p}_0^{(H)} = \frac{1}{H}\sum_{h=1}^{H} \delta_{\tilde{\theta}_h}, \tag{3.10}$$

In words, any NIDM can be represented as a hierarchical process with a NRMI having a discrete baseline measure at the bottom of the hierarchy. Note that when P is diffuse, the law of $\tilde{p}_0^{(H)}$ is that of a specific Gibbs type prior, arising when the discount parameter goes to $-\infty$; see Gnedin & Pitman (2005) for details. Furthermore, this representation relates

any NIDM process to hierarchical constructions like those presented in the contribution of Camerlenghi et al. (2018).

In view of (3.10), it is instructive to compare the finite-dimensional distributions of a NIDM $\tilde{p}^{(H)}$ with those of a $\tilde{p}^{(\infty)} \sim \text{NRMI}(c, \rho; P)$. In particular, it follows that, for any finite partition $\{B_1, \ldots, B_d\}$ of Θ into $\mathscr{B}(\Theta)$–sets, the distribution of $\{\tilde{p}^{(H)}(B_1), \ldots, \tilde{p}^{(H)}(B_{d-1})\}$ can be expressed as a mixture of a NID distribution with multinomial weights, motivating the NIDM denomination. More precisely,

$$\{\tilde{p}^{(H)}(B_1), \ldots, \tilde{p}^{(H)}(B_{d-1})\} \mid (\tilde{m}_1, \ldots, \tilde{m}_d) \sim \text{NID}\left(c\,\frac{\tilde{m}_1}{H}, \ldots, c\,\frac{\tilde{m}_d}{H};\ \rho\right),$$
$$(\tilde{m}_1, \ldots, \tilde{m}_d) \sim \text{MULTINOMIAL}\left[H, \{P(B_1), \ldots, P(B_d)\}\right].$$

On the other hand, since $\tilde{p}^{(\infty)}(B_i) = J_i / \sum_{j=1}^{d} J_j$, where $J_i = \tilde{\mu}^{(\infty)}(B_i) \sim \text{ID}(cP(B_i); \rho)$, one has

$$\{\tilde{p}^{(\infty)}(B_1), \ldots, \tilde{p}^{(\infty)}(B_{d-1}))\} \sim \text{NID}\left(cP(B_1), \ldots, cP(B_d); \rho\right).$$

An application of the strong law of large numbers yields $(c\tilde{m}_1/H, \ldots, c\tilde{m}_d/H) \xrightarrow{\text{a.s.}} \{cP(B_1), \ldots, cP(B_d)\}$ as $H \to \infty$. This provides a heuristic argument for arguing that NIDM processes and homogeneous NRMIs are closely related when H is large. As it will be more formally and rigorously discussed in the next section, NIDM processes weakly converge to the corresponding homogeneous NRMI as $H \to \infty$.

Finally, note that the moments of a NIDM process can be obtained in closed form. Here we confine the attention to the first two moments, which have simple analytical forms. Define $\mathcal{I}(c, \rho) = c \int_{\mathbb{R}_+} u e^{-c\psi(u)} \tau_2(u) du$.

Proposition 3.1. *Let* $\tilde{p}^{(H)} \sim \text{NIDM}(c, \rho; P)$ *be a* NIDM *process. Moreover, let* $A, A_1, A_2 \in \mathscr{B}(\Theta)$ *and set* $C := A_1 \cap A_2$. *Then*

$$\mathbb{E}\{\tilde{p}^{(H)}(A)\} = P(A),$$
$$\text{Var}\{\tilde{p}^{(H)}(A)\} = P(A)\{1 - P(A)\}\left\{\mathcal{I}(c, \rho) + \frac{1 - \mathcal{I}(c, \rho)}{H}\right\},$$
$$\text{Cov}\{\tilde{p}^{(H)}(A_1), \tilde{p}^{(H)}(A_2)\} = \{P(C) - P(A_1)P(A_2)\}\left\{\mathcal{I}(c, \rho) + \frac{1 - \mathcal{I}(c, \rho)}{H}\right\}.$$

Unsurprisingly, when $H \to \infty$ the moments of a NIDM process converge to those of a NRMI, as one would expect from the previous discussion.

3.3.2 Weak convergence of NIDM processes

The previous discussion suggests that, when H is large enough, a NIDM approaches a nonparametric prior that corresponds to a homogeneous NRMI. To make this more

precise, endow $\mathcal{M}_\Theta$ with the topology of *vague convergence* (see e.g. Chap. 4, Kallenberg, 2017) and use the notation $\mu_n \xrightarrow{\text{vd}} \mu$ to identify a sequence $(\tilde{\mu}_n)_{n\geqslant 1}$ of boundedly finite measures on Θ that vaguely converges to μ. We will also make use of the concise notation $\tilde{\mu}(f) = \int_\Theta f(\theta)\tilde{\mu}(d\theta)$. The main result of this Section concerns the convergence of IDM random measures to homogeneous CRMs, and it is summarized in the following theorem.

Theorem 3.1. *Let* $\tilde{\mu}^{(H)} \sim \text{IDM}(c, \rho; P)$ *and* $\tilde{\mu}^{(\infty)} \sim \text{CRM}(c, \rho; P)$. *Then as* $H \to \infty$

(1) $\tilde{\mu}^{(H)} \xrightarrow{\text{vd}} \tilde{\mu}^{(\infty)}$,

(2) $\mathbb{E}\left(e^{-\tilde{\mu}^{(H)}(f)}\right) \to \mathbb{E}\left(e^{-\tilde{\mu}^{(\infty)}(f)}\right)$ *for any positive and measurable function* $f : \Theta \to \mathbb{R}_+$ *such that* $\int_\Theta \psi\{f(\theta)\}P(d\theta) < \infty$.

Note that the two statement are not equivalent, because the former can be deduced from the latter, but not viceversa. Note that if f in the above theorem is integrable with respect to P then the condition $\int_\Theta \psi\{f(\theta)\}P(d\theta) < \infty$ is satisfied.

Now recall that $\tilde{p}^{(H)} \sim \text{NIDM}(c, \rho; P)$ and that $\tilde{p}^{(\infty)} \sim \text{NRMI}(c, \rho; P)$. An important implication of Theorem 3.1 is the convergence of the finite-dimensional distributions of $\tilde{p}^{(H)}$ to those of $\tilde{p}^{(\infty)}$. Indeed, substituting in Theorem 3.1 the simple function $f(\theta) = \sum_{i=1}^d \lambda_i \mathbb{1}_{A_i}(\theta)$, for any collection of sets $A_1, \ldots, A_d \in \mathscr{B}(\Theta)$ and positive constants $\lambda_1, \ldots, \lambda_d > 0$, as a consequence of the continuous mapping theorem one has that

$$\{\tilde{p}^{(H)}(A_1), \ldots, \tilde{p}^{(H)}(A_d)\} \xrightarrow{\text{d}} \{\tilde{p}^{(\infty)}(A_1), \ldots, \tilde{p}^{(\infty)}(A_d)\}, \quad \text{as } H \to \infty.$$

When working with random probability measures, the convergence of the finite-dimensional distributions suffices to guarantee the weak convergence of the whole process, which will be indicated with the $\xrightarrow{\text{wd}}$ notation; see Kallenberg (Theorem 4.11, 2017).

Corollary 3.1. *Let* $\tilde{p}^{(H)} \sim \text{NIDM}(c, \rho; P)$ *and let* $\tilde{p}^{(\infty)} \sim \text{NRMI}(c, \rho; P)$. *Then* $\tilde{p}^{(H)} \xrightarrow{\text{wd}} \tilde{p}^{(\infty)}$ *as* $H \to \infty$.

The above statement implies the convergence in distribution of general functionals $\tilde{p}^{(H)}(f) \xrightarrow{\text{d}} \tilde{p}^{(\infty)}(f)$ as $H \to \infty$ when f is a continuous and bounded function. In the Dirichlet multinomial process case, related results were previously obtained in Kingman (1975), Muliere & Secchi (1996) and Green & Richardson (2001).

Remark 3.1. Although our focus here is on random probability measures, we note that Theorem 3.1 could be useful also when one needs to approximate any homogeneous CRM by means of a finite-dimensional IDM process. This might be the case, for instance, in Bayesian nonparametric survival analysis where CRMs are used to define mixture hazard rate functions or cumulative hazards; see e.g. Lijoi & Prünster (2010) for a review.

3.3.3 Random partitions and number of clusters

In this section we study the clustering mechanism underlying a NIDM process that amounts to determining the EPPF, namely the probability distribution of the induced exchangeable partition: this will be denoted by Π_H and it is defined through (3.3), with $\tilde{p}$ being replaced by $\tilde{p}^{(H)}$. To be more specific, it will be shown that the EPPF of any NIDM process is a finite mixture of the partition functions arising in the infinite-dimensional setting. Before stating the theorem, let us introduce some further quantity of interest. Define for any $m \geqslant 1$

$$\mathcal{V}_{m,H}(u) := \left((-1)^m \frac{\partial^m}{\partial u^m} e^{-\frac{c}{H}\psi(u)}\right) e^{\frac{c}{H}\psi(u)} = \frac{c}{H}\Delta_{m,H}(u),$$

and set $\mathcal{V}_{0,H} := 1$, where $\psi(u)$ is the Laplace exponent defined as in (3.6). Moreover, for any vector $\boldsymbol{x} \in \mathbb{R}^p$, we let $|\boldsymbol{x}| = \sum_{i=1}^p x_i$.

Theorem 3.2. *Let $(\theta_n)_{n\geqslant 1}$ be an exchangeable sequence directed by a* NIDM *process prior with diffuse baseline P. Then, the associated* EPPF *when $k \leqslant \min\{n, H\}$ is*

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{H^k(H-k)!}\frac{c^k}{\Gamma(n)}\int_{\mathbb{R}_+} u^{n-1} e^{-c\psi(u)} \prod_{j=1}^k \Delta_{n_j,H}(u)du,$$

Moreover, if Π_∞ is the EPPF *of the corresponding homogeneous* NRMI *in* (3.3)*, one has*

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{H^k(H-k)!}\sum_{\ell}\frac{1}{H^{|\ell|-k}}\prod_{j=1}^k\frac{1}{\ell_j!}\sum_{e_j}\binom{n_j}{e_{j1}, \ldots, e_{j\ell_j}} \times \Pi_\infty(e_{11}, \ldots, e_{1\ell_1}, \ldots, e_{k1}, \ldots, e_{k\ell_k}), \tag{3.11}$$

where the first sum runs over all vectors $\ell = (\ell_1, \ldots, \ell_k)$ such that $\ell_j \in \{1, \ldots, n_j\}$, and the jth of the k sums runs over $e_j = (e_{j1}, \ldots, e_{j\ell_j})$ such that $e_{jr} \geqslant 1$ and with $|e_j| = n_j$.

This mixture representation in (3.11) is reminiscent of the one in Camerlenghi et al. (2018) for hierarchical NRMIs, and indeed NIDM processes can be represented in a hierarchical fashion, as for equation (3.10). Hence, peculiar properties of infinite-dimensional NRMIs will be inherited by NIDM processes for any choice of H. Furthermore, Corollary 3.1 suggests that one might get $\lim_{H\to\infty}\Pi_H(n_1, \ldots, n_k) = \Pi_\infty(n_1, \ldots, n_k)$, that is, that the EPPF associated to a NIDM process converges to the one associated to a homogeneous NRMI. This is indeed the case, and it can shown directly through the representation of Π_∞ in (3.4) and by noting that $\lim_{H\to\infty}\Delta_{m,H}(u) = \tau_m(u)$ for any $u > 0$. Working along these lines, one can identify bounds for the ratio between Π_H and Π_∞ and these may be used to assess their closeness.

Theorem 3.3. *For any* $k \leqslant \min\{n, H\}$ *one has*

$$\frac{H!}{H^k(H-k)!} \leqslant \frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_\infty(n_1,\ldots,n_k)} \leqslant \int_{\mathbb{R}_+} \prod_{j=1}^{k} \frac{\Delta_{n_j,H}(u)}{\tau_{n_j}(u)} p^{(\infty)}(u)du,$$

where $p^{(\infty)}(u) \propto u^{n-1}e^{-c\psi(u)} \prod_{j=1}^{k} \tau_{n_j}(u)\mathbb{1}_{(0,\infty)}(u)$ *is a density function.*

Note that $p^{(\infty)}(u)$ is the density function of a latent random variable that is used in James et al. (2009) to provide posterior characterizations of NRMIs. Both bounds converge to 1 as $H \to \infty$. As a simple application of Theorem 3.3, one might obtain bounds also for the predictive distributions, by exploiting their relationship with the EPPF. For NGG multinomial processes and, a fortiori, in the Dirichlet multinomial case, the EPPF and the related bounds can be computed explicitly. This is illustrated in the following examples.

Example 3.1 (Dirichlet multinomial process). Let the NIDM process be characterized by the intensity function $\rho(s) = s^{-1}e^{-s}$. On the basis of Theorem 3.2, one has for any $k \leqslant \min\{n, H\}$

$$\Pi_H(n_1,\ldots,n_k) = \frac{H!}{(H-k)!}\frac{1}{(c)_n}\prod_{j=1}^{k}\left(\frac{c}{H}\right)_{n_j}.$$

This EPPF can be found, e.g., in Green & Richardson (2001). Note that Π_H identifies the building block of a Gibbs–type prior with $\sigma < 0$: indeed any such prior would arise from a mixture of Dirichlet multinomial distributions, with respect to H. See, e.g., De Blasi et al. (2015). Straight application of Theorem 3.3 yields

$$\frac{H!}{H^k(H-k)!} \leqslant \frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_\infty(n_1,\ldots,n_k)} \leqslant \prod_{j=1}^{k}\frac{(1+c/H)_{n_j-1}}{(n_j-1)!}.$$

This makes clear that Π_H and Π_∞ are close when either H is large, as it is natural to expect on the basis of Corollary 3.1, or the total mass parameter c is small. Note that this is the same bound obtained in the Appendix of Ishwaran & Zarepour (2002) by means of different techniques, and thus Theorem 3.3 can be seen as a generalization of their result to NIDM processes. Details on these derivations are given in the Appendix.

Example 3.2 (NGG multinomial process). Let the NIDM process be characterized by the generalized gamma intensity function $\rho(s)$ given in (3.2). On the basis of Theorem 3.2, one has for any $k \leqslant H$

$$\Pi_H(n_1,\ldots,n_k) = \frac{H!}{(H-k)!}\sum_{\ell}\frac{\mathscr{V}_{n,|\ell|}}{H^{|\ell|}}\prod_{j=1}^{k}\frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\sigma^{\ell_j}},$$

where $\mathscr{C}(n,k;\sigma)$ are the generalized factorial coefficients, defined as in equation (2.5) of Chapter 2, and where $\mathscr{V}_{n,k}$ are the coefficients defined in (3.5). Hence, the EPPF of a NGG multinomial process has a simpler form compared to the general equation (3.11), because it only depends on the integers $\ell_1,\dots,\ell_k$. This also enables the practical evaluation of Theorem (3.3), yielding to

$$\frac{H!}{H^k(H-k)!} \leqslant \frac{\Pi_H(n_1,\dots,n_k)}{\Pi_\infty(n_1,\dots,n_k)} \leqslant \sum_{\boldsymbol{\ell}} \frac{\mathscr{V}_{n,|\boldsymbol{\ell}|}}{\mathscr{V}_{n,k}} \left(\frac{c}{\sigma H}\right)^{|\boldsymbol{\ell}|-k} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\mathscr{C}(n_j,1;\sigma)},$$

where $\mathscr{C}(n_j,1;\sigma) = \sigma(1-\sigma)_{n_j-1}$. Details on these derivations are given in the Appendix.

The availability of Π_H naturally leads one to address the problem of determining the distribution of the number of partition sets $K_{n,H}$, that is, the law of the number of distinct values observed in a sample $\boldsymbol{\theta}$ under a NIDM process prior. As one might expect, $K_{n,H}$ converges to the number of partition sets $K_{n,\infty}$, namely the number of distinct values generated by an exchangeable n-sample from homogeneous NRMI, when $H \to \infty$. Another interesting connection between $K_{n,H}$ and $K_{n,\infty}$ is formalized in the following theorem.

Theorem 3.4. *For any* $k \leqslant \min\{H,n\}$

$$\mathbb{P}(K_{n,H}=k) = \frac{H!}{H^k(H-k)!} \sum_{\ell=0}^{n-k} \frac{1}{H^\ell} \mathscr{S}(\ell+k,k)\mathbb{P}(K_{n,\infty}=\ell+k),$$

where $\mathscr{S}(\ell,k) = \frac{1}{k!}\sum_{r=0}^{k}(-1)^{k-r}\binom{k}{r}r^\ell$ *is the Stirling number of the second kind for* $\ell,k \geqslant 0$. *Moreover, the expected value of* $K_{n,H}$ *is given by*

$$\mathbb{E}(K_{n,H}) = H - H\mathbb{E}\left\{\left(\sum_{h=1}^{H-1}\pi_h\right)^n\right\} = H - H\mathbb{E}\left\{\left(1-\frac{1}{H}\right)^{K_{n,\infty}}\right\}.$$

From Theorem 3.4 it can be easily seen that $\mathbb{P}(K_{n,H}=k) \to \mathbb{P}(K_{n,\infty}=k)$ for all k as $H \to \infty$, since $\mathscr{S}(k,k)=1$. Hence, $K_{n,H} \xrightarrow{d} K_{n,\infty}$. Additionally, we have that $\mathbb{E}(K_{n,H}) \to \mathbb{E}(K_{n,\infty})$ as $H \to \infty$, and the following asymptotic expansion holds

$$\frac{\mathbb{E}(K_{n,H})}{\mathbb{E}(K_{n,\infty})} = 1 - \frac{1}{2H}\left(\frac{\mathbb{E}(K_{n,\infty}^2)}{\mathbb{E}(K_{n,\infty})} - 1\right) + \mathcal{O}\left(\frac{1}{H^2}\right), \qquad \text{as } H \to \infty.$$

Thus, the convergence of the expected value to the infinite case occurs at the linear rate $\mathcal{O}(1/H)$. We expanded the above ratio up the the second order, to gain further understanding about the speed at which $\mathbb{E}(K_{n,H})$ approaches its limit: broadly speaking, quick convergence occurs whenever $\mathbb{E}(K_{n,\infty}^2) \approx \mathbb{E}(K_{n,\infty})$, which is the case when $\mathbb{E}(K_{n,\infty})$

is relatively small. On the other hand, one needs a large value of H when trying to approximate an infinite-dimensional process having a high number of expected clusters in a sample of size n. This is in line with the discussion of Example 3.1.

The actual evaluation of the probability distribution of $K_{n,H}$ in Theorem 3.4 might be cumbersome due to the presence of the Stirling numbers. Thus, in cases where it is more convenient to rely on the probability distribution of $K_{n\infty}$ it may be interesting to provide simple bounds for the ratio $\mathbb{P}(K_{n,H} = k)/\mathbb{P}(K_{n,\infty} = k)$. This is achieved in the next Theorem.

Theorem 3.5. *For any* $k \leqslant \min\{H, n-1\}$

$$\frac{H!}{H^k(H-k)!} \leqslant \frac{\mathbb{P}(K_{n,H}=k)}{\mathbb{P}(K_{n,\infty}=k)} \leqslant \frac{H!}{H^k(H-k)!}\left\{1+\frac{1}{2}\sum_{\ell=1}^{n-k}\left(\frac{k}{H}\right)^{\ell}\binom{\ell+k}{k}\frac{\mathbb{P}(K_{n,\infty}=\ell+k)}{\mathbb{P}(K_{n,\infty}=k)}\right\},$$

whereas when $k = n = H$, *it holds* $\mathbb{P}(K_{n,H}=n)/\mathbb{P}(K_{n,\infty}=n) = H^{-n}H!/(H-n)!$.

Interestingly, the lower bound in the above theorem does not depend on the specific NIDM process, and actually coincide with the one obtained by Ishwaran & Zarepour (2002) in the special case of the Dirichlet multinomial NIDM. Instead, the upper bound can be lower than 1, and therefore it is usually tighter than the one already known for the Dirichlet prior. Hence, besides being a generalization to all NIDM processes, Theorem 3.5 also yields an improvement over existing bounds.

Example 3.3 (Dirichlet multinomial process, cont'd). A straightforward application of Theorem 3.4 yields

$$\mathbb{P}(K_{n,H}=k) = \frac{H!}{(H-k)!}\frac{(-1)^k}{(c)_n}\mathscr{C}(n,k;-c/H), \qquad k=1,\dots,\min\{H,n\}.$$

This simple form is due to the Gibbs-type structure of the Dirichlet multinomial process, and indeed the above formula might be deduced from Gnedin & Pitman (2005). In the relevant particular case $c = H$, that is, when the weights of the NIDM process are uniformly distributed, the following simplification occurs

$$\mathbb{P}(K_{n,H}=k) = \frac{n!}{(H)_n}\binom{H}{k}\binom{n-1}{k-1},$$

for $k = 1,\dots,\min\{H,n\}$, a particular instance of hypergeometric distribution. As for the expected value of $K_{n,H}$, for any value of c we have the following simple formula

$$\mathbb{E}(K_{n,H}) = H - (H-1)\frac{(c+1-c/H)_{n-1}}{(c+1)_{n-1}}.$$

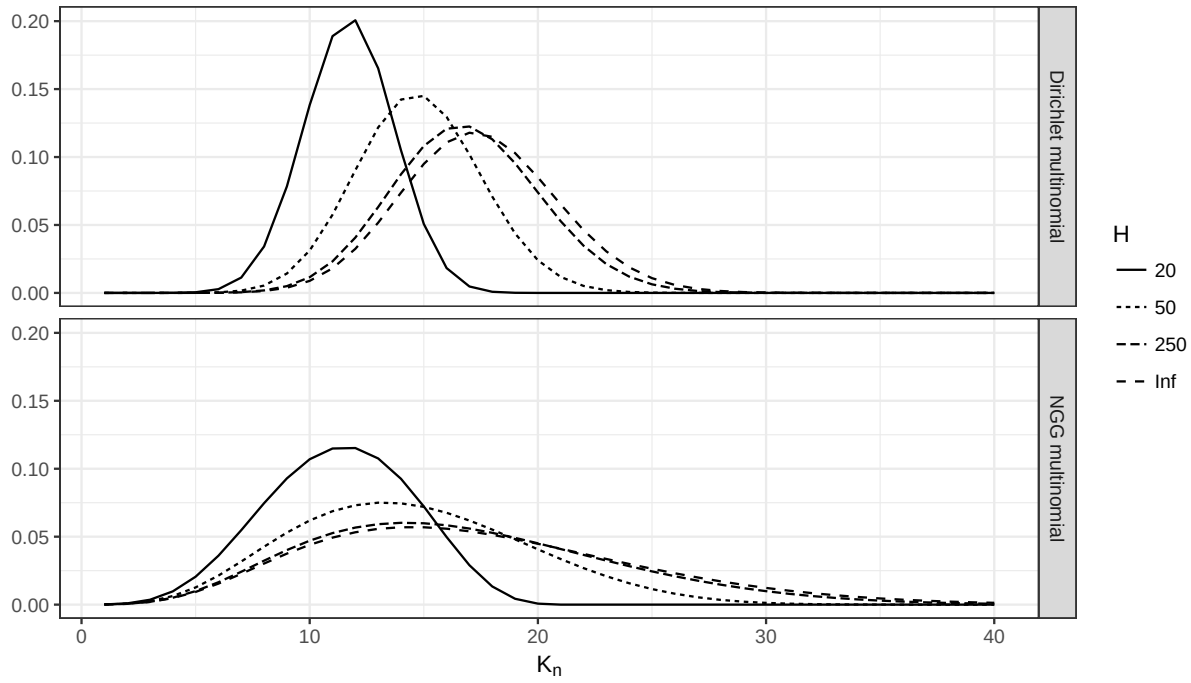


Figure 3.1: Distribution of $K_{100,H}$ with $n = 100$, under a Dirichlet multinomial process ($c = 5.87$), and a NGG multinomial process ($c = 1/2, \kappa = 1, \sigma = 1/2$), for different levels of $H \in \{20, 50, 250, \infty\}$. The distribution is depicted only for the values in the interval $[1, 40]$ for graphical reasons.

Example 3.4 (NGG multinomial process, cont'd). Direct application of Theorem 3.4 yields

$$\mathbb{P}(K_{n,H} = k) = \frac{H!}{(\sigma H)^k (H-k)!} \sum_{\ell=0}^{n-k} \frac{\mathscr{V}_{n,\ell+k}}{H^\ell} \mathscr{S}(\ell+k, k) \frac{\mathscr{C}(n, \ell+k; \sigma)}{\sigma^\ell},$$

for any $k = 1, \ldots, \min\{H, n\}$, since $\mathbb{P}(K_{n,\infty} = k) = \mathscr{V}_{n,k} \sigma^{-k} \mathscr{C}(n, k; \sigma)$. Furthermore, the expected value of $K_{n,H}$ is given by

$$\mathbb{E}(K_{n,H}) = H - H \sum_{\ell=1}^{n} \left(1 - \frac{1}{H}\right)^\ell \mathscr{V}_{n,\ell} \frac{\mathscr{C}(n, \ell; \sigma)}{\sigma^\ell}.$$

To illustrate the clustering mechanism, in Figure 3.1 we depicted the distribution of the random variable $K_{100,H}$ under both a Dirichlet multinomial process and a NGG multinomial process, for different values of H. To make these prior choices comparable, we have set the hyperparameters c and $(\bar{c}, \sigma)$ such that the exepcted number of clusters for the corresponding infinite-dimensional NRMIs $\mathbb{E}(K_{100,\infty})$ is the same. As highlighted in Figure 3.1, the distribution of $K_{100,H}$ under the NGG multinomial prior is "flatter", i.e. less informative, compared to the one induced by Dirichlet multinomial prior, for any of the values of H being considered. We note that the variance of $K_{n,H}$ in the NGG prior can be tuned through the parameter σ. When $\sigma \to 0$ the Dirichlet multinomial

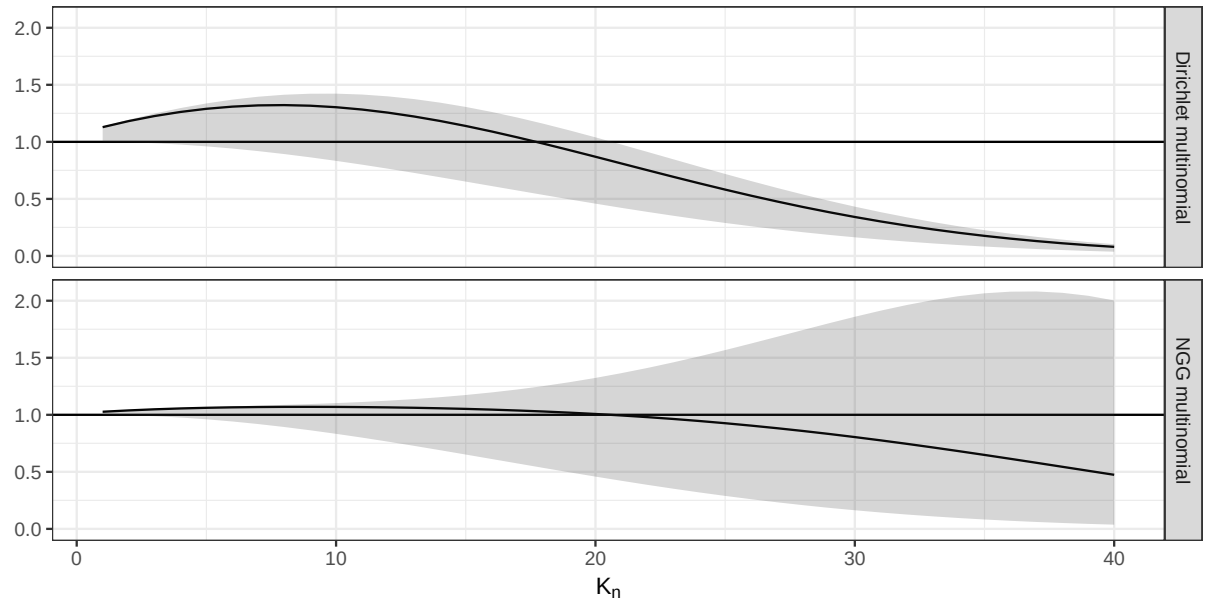


Figure 3.2: Exact value (solid lines) and bounds (shaded areas) for the ratio $\mathbb{P}(K_{100,250} = k)/\mathbb{P}(K_{100,\infty} = k)$ under a Dirichlet multinomial process ($c = 5.87$), and a NGG multinomial process ($c = 1/2, \kappa = 1, \sigma = 1/2$), for different values of k. The ratio is depicted only for the values in the interval $[1, 40]$ for graphical reasons.

case is recovered and, as such, it identifies a limiting scenario. If $H \to \infty$, this effect of σ was extensively discussed in Lijoi et al. (2007) and it comes to no surprise it is reflected also in the finite H case, in view of Theorem 3.4. More generally, Theorems 3.2-3.4 confirm, altogether, that NIDM processes inherit several structural properties of their infinite-dimensional counterpart even when H is finite. Nonetheless, we remark that if one is interested in approximating the infinite-dimensional NRMI prior, a relatively large value of H is typically required. For instance, Figure 3.1 suggests that we should set at least $H = 250$ to suitably approximate both the Dirichlet and the NGG processes with their NIDM counterparts, in this specific setting. Finally, using the same hyperparameter settings as before, we depict in Figure 3.2 the bounds as for Theorem 3.5, together with the exact value, when $H = 250$. Note that the bounds become less informative essentially in those values $K_{100,250} = k$ having negligible probabilities, i.e. in the tails of the distribution.

3.4 Posterior characterizations

We complete here the picture of the distributional properties of NIDM processes by determining their posterior distribution. While for prediction the EPPF of Theorem 3.2 might suffice, inference on non-linear functionals of $\tilde{p}^{(H)}$, such as quantiles or credible intervals, relies on the posterior distribution of $\tilde{p}^{(H)}$ given a sample $\boldsymbol{\theta}$. To be more precise

we will refer to a framework where

$$(\theta_1, \ldots, \theta_n \mid \tilde{p}^{(H)}) \overset{\text{iid}}{\sim} \tilde{p}^{(H)}, \qquad \tilde{p}^{(H)} \sim \text{NIDM}(c, \rho; P). \tag{3.12}$$

We will display the predictive distribution for θ_{n+1}, given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and will, then, provide two representations of the posterior distribution of $\tilde{p}^{(H)}$ one of which is effectively described in terms of the the multiroom Chinese restaurant metaphor introduced in Camerlenghi et al. (2019). While our results are general, particular emphasis is given to the NGG multinomial process and, despite the lack of conjugacy, we are able to devise efficient algorithms for exact (iid) sampling of the posterior. In contrast with most widely known samplers for homogeneous NRMIs, which require truncating certain series representations, the posterior laws of NIDM processes can be sampled exactly.

3.4.1 Predictive distributions and posterior laws

Since the EPPF Π_H can be evaluated through Theorem 3.2, it is straightforward to compute the predictive distributions. To this end, for any $n \geqslant 1$ we define a positive random variable U whose density function on $\mathbb{R}_+$, conditional on the sample $\boldsymbol{\theta}$, equals

$$p^{(H)}(u) \propto u^{n-1} e^{-c\psi(u)} \prod_{j=1}^{k} \Delta_{n_j,H}(u). \tag{3.13}$$

The normalizing constant of the above density is finite and it essentially identifies the EPPF of a NIDM process. The density function $p^{(H)}(u)$ parallels the one, say $p^{(\infty)}(u)$, of the latent variable appearing in the posterior representation for NRMIs. See James et al. (2009). Note actually that $p^{(H)}(u)$ converges pointwise to $p^{(\infty)}(u)$ as $H \to \infty$ since $\Delta_{m,H}(u) \to \tau_m(u)$ for any $m \geqslant 1$ and $u > 0$.

Corollary 3.2. *Let $\theta_1, \ldots, \theta_n$ be as in (3.12), and such that they admit k distinct values $\theta_1^*, \ldots, \theta_k^*$ with θ_j^* having frequency n_j, then for any $A \in \mathscr{B}(\Theta)$*

$$\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}) = w_0(\boldsymbol{n}^{(k)}) P(A) + \sum_{j=1}^{k} w_j(\boldsymbol{n}^{(k)}) \delta_{\theta_j^*}(A),$$

where $\boldsymbol{n}^{(k)} = (n_1, \ldots, n_k)$, $w_0(\boldsymbol{n}^{(k)}) = \left(1 - \frac{k}{H}\right) \frac{c}{n} \int_{\mathbb{R}_+} u \Delta_{1,H}(u) p^{(H)}(u) du$ and

$$w_j(\boldsymbol{n}^{(k)}) = \frac{1}{n} \int_{\mathbb{R}_+} u \frac{\Delta_{n_j+1,H}(u)}{\Delta_{n_j,H}(u)} p^{(H)}(u) du, \qquad j = 1, \ldots, k.$$

This result is reminiscent of the predictive distributions obtained for homogeneous NRMIs, and indeed similar sampling strategies can be borrowed from that context. For example, note that conditionally on the latent variable U one has

$$\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}, U) \propto \left(1 - \frac{k}{H}\right) c\Delta_{1,H}(U)P(A) + \sum_{j=1}^{k} \frac{\Delta_{n_j+1,H}(U)}{\Delta_{n_j,H}(U)} \delta_{\theta_j^*}(A).$$

Hence, one can devise a generalized Pólya-urn scheme by first drawing U from its density $p^{(H)}(u)$ and then sampling from the predictive distribution using the above formula. The terms $\Delta_{m,H}$ might be expensive to compute in practice, mainly because of their combinatorial nature. However, this issue can be attenuated by relying on the recursive definition of $\Delta_{m,H}$ provided in James et al. (2006). Furthermore, in the fairly general NGG multinomial case, the weights $\Delta_{m,H}$ have an explicit formula, and this allow the implementation of an exact sampling algorithm for U; see Appendix 3.6.

Remark 3.2. If one is only interested in obtaining an exchangeable draw for $\boldsymbol{\theta}$, a direct strategy consists in simulating $\tilde{p}^{(H)}$ and then, conditionally on it, sampling iid observations from $\tilde{p}^{(H)}$, according to (3.12). Indeed, any ID random variable whose Laplace exponent is available in closed form can be sampled, for example through the general algorithm of Ridout (2009), and this enables the simulation of NIDM processes.

We now provide a posterior characterization for the law of the random measure $\tilde{\mu}^{(H)}$ given $\boldsymbol{\theta}$: the posterior distribution of $\tilde{p}^{(H)}$ can, then, be recovered by normalization. To ease notation, the posterior law is expressed conditionally on the latent variable U, whose density is $p^{(H)}(u)$. Moreover, when the sample $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ displays $k < H$ distinct values $\theta_1^*, \ldots, \theta_k^*$, we let $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ represent the point masses in $\tilde{p}^{(H)}$ that are not included in $\boldsymbol{\theta}$, up to a permutation.

Theorem 3.6. *Let $\theta_1, \ldots, \theta_n$ be as in (3.12) and, conditionally on $\boldsymbol{\theta}$, let U be a random variable with density $p^{(H)}(u)$ as in (3.13). Then*

$$(\tilde{\mu}^{(H)} \mid \boldsymbol{\theta}, U) \stackrel{d}{=} \sum_{j=k+1}^{H} J_j^* \delta_{\bar{\theta}_j} + \sum_{j=1}^{k} (J_j^* + I_j)\delta_{\theta_j^*}, \tag{3.14}$$

where $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ are iid draws from P and

$$(J_h^* \mid U) \stackrel{\text{iid}}{\sim} \text{ID}\left(\frac{c}{H}, \rho^*\right), \qquad \rho^*(s) = e^{-Us}\rho(s), \qquad h = 1, \ldots, H.$$

Finally, the jumps $I_1, \ldots, I_k$ are independent and nonnegative random variables characterized by

$$\mathbb{E}\left(e^{-\lambda I_j} \,\middle|\, \boldsymbol{\theta}, U\right) = \frac{\Delta_{n_j,H}(\lambda + U)}{\Delta_{n_j,H}(U)}, \qquad j = 1, \ldots, k.$$

This representation is closely related to the posterior representation for homogeneous NRMIs, which can be recovered as $H \to \infty$. Indeed, the first term in (3.14) converges to a CRM with the exponentially tilted Lévy intensity ρ^*, as a consequence of Theorem 3.1. On the other hand, $J_j^* \xrightarrow{d} 0$ and $\mathbb{E}\left(e^{-\lambda I_j} \mid \boldsymbol{\theta}, U\right) \to \tau_{n_j}(\lambda + U)/\tau_{n_j}(U)$ for any $j = 1, \ldots, k$: hence, the second term on the right–hand–side of (3.14) converges to the fixed jumps component of the posterior representation of NRMIs. See James et al. (2009). Interestingly, a structural property is shared by NRMIs and NIDMs: conditionally on a latent variable U the posterior law is a mixture of: (i) a component with a tilted intensity and (ii) a collection of independent jumps corresponding to the distinct values $\theta_1^*, \ldots, \theta_k^*$ in the sample $\boldsymbol{\theta}$. However, it must be stressed that for NIDMs the tilted component vanishes as soon as $k = H$ distinct values are recorded in the sample, and the posterior distribution will coincide with the law of a measure with jumps at fixed locations identified by the distinct values $\theta_1^*, \ldots, \theta_H^*$.

Example 3.5 (Dirichlet multinomial process, cont'd). Note that $\Delta_{m,H}(\lambda + u)/\Delta_{m,H}(u) = \tau_m(\lambda + u)/\tau_m(u)$ for any $m \geqslant 1$ and H, implying that the random variables in Theorem 3.6 have a simple form

$$(J_j^* \mid \boldsymbol{\theta}, U) \overset{\text{iid}}{\sim} \text{GAMMA}\left(\frac{c}{H}, 1 + U\right), \qquad (I_j \mid \boldsymbol{\theta}, U) \overset{\text{ind}}{\sim} \text{GAMMA}(n_j, 1 + U),$$

for $j = 1, \ldots, H$, where we agree that $I_j = 0$ a.s. for any $j > k$. Then, after normalization we get

$$(\tilde{p}^{(H)} \mid \boldsymbol{\theta}, U) \overset{d}{=} \sum_{j=k+1}^{H} \frac{J_j^*}{\sum_{h=1}^{H}(J_h^* + I_h)}\delta_{\tilde{\theta}_j} + \sum_{j=1}^{k} \frac{J_j^* + I_j}{\sum_{h=1}^{H}(J_h^* + I_h)}\delta_{\theta_j^*},$$

which can be shown not to depend on the latent variable U. Also, the above weights have Dirichlet distribution with parameters $(n_1 + c/H, \ldots, n_k + c/H, c/H, \ldots, c/H)$. Finally, it is easy to check that Corollary 3.2 can be specialized to obtain the well-known predictive distributions

$$\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}) = \left(1 - \frac{k}{H}\right)\frac{c}{c+n}P(A) + \sum_{j=1}^{k}\frac{n_j + c/H}{c+n}\delta_{\theta_j^*}(A).$$

Example 3.6 (NGG multinomial process, cont'd). Note that for any $m \geqslant 1$ one has

$$\Delta_{m,H}(u) = \sum_{\ell=1}^{m} \varsigma_{m,\ell,H}(u) = \sum_{\ell=1}^{m}\left(\frac{c}{H}\right)^{\ell-1}\frac{\mathscr{C}(m,\ell;\sigma)}{\sigma^\ell}(\kappa + u)^{-m+\ell\sigma}.$$

Moreover, the random variables $J_1^*, \ldots, J_H^*$ of Theorem 3.6 are *conditionally conjugate*, because $\rho^*(s) = e^{-us}\rho(s) = \frac{1}{\Gamma(1-\sigma)}s^{-1-\sigma}e^{-(\kappa+u)s}$ identifies an updated generalized

gamma process. The distribution of each $J_1^*, \dots, J_H^*$ is known as *tempered stable* and there exists several methods for drawing samples from it; see e.g. Ridout (2009) and references therein. Furthermore, the random variables $I_1, \dots, I_k$ given U have the following mixture densities

$$p_{I_j}(w \mid u) = \sum_{\ell=1}^{n_j} \frac{\varsigma_{n_j,\ell,H}(u)}{\Delta_{n_j,H}(u)} \textsc{gamma}(w; n_j - \ell\sigma, \kappa + u), \tag{3.15}$$

for $j = 1, \dots, k$, where $\textsc{gamma}(w; a, b)$ denotes the density function of a gamma random variable. Finally, some algebra yields

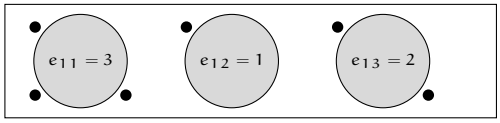
$$\begin{aligned} \mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}, U) \propto \left(1 - \frac{k}{H}\right) c(\kappa + U)^{\sigma} P(A) \\ + \sum_{j=1}^{k} \left[\frac{1}{H} c(\kappa + U)^{\sigma} + \sum_{\ell_j=1}^{n_j} \frac{\varsigma_{n_j,\ell_j,H}(U)}{\Delta_{n_j,H}(U)} (n_j - \ell_j\sigma) \right] \delta_{\theta_j^*}(A). \end{aligned} \tag{3.16}$$

for any $A \in \mathscr{B}(\Theta)$.

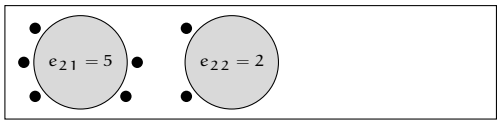
Remark 3.3. To enable posterior inference through random sampling it suffices to simulate iid U values from $p^{(H)}(u)$ and, then, make use of the above posterior representation. Although $p^{(H)}(u)$ is known up to a normalizing constant, we can nonetheless draw samples by acceptance–rejection algorithms. The simulation of the limiting density $p^{(\infty)}(u)$ was typically addressed via Markov Chain Monte Carlo (Lijoi et al., 2007). However, this further complication can be avoided in the NGG setting, given the availability of efficient algorithms for exact sampling, which are discussed in Appendix 3.6 both for $p^{(H)}(u)$ and $p^{(\infty)}(u)$. As such, these algorithms might be useful beyond their application to NIDMs.

3.4.2 Multiroom Chinese restaurant metaphor

To gain further insights about structural properties of NIDM processes, we now describe a data-augmentation based on the hierarchical representations (3.10)-(3.11). To this purpose, it is worth recalling the so-called *multiroom Chinese restaurant* metaphor coined by Camerlenghi et al. (2019), which can be adapted to NIDM processes. Suppose that there exists a restaurant which serves a *finite* number of dishes H, corresponding to iid draws from the diffuse P. Each restaurant has infinitely many rooms, and each room contains infinitely many tables and is associated to a single dish out of the H available from the menu. The first customer seats in one of the tables of the first room and selects a dish. The nth customer can either select a dish previously chosen by the other $n-1$ customers or she can choose a new dish. In the former case, she will be seated in the room serving the dish of choice and she may be seated either at a new table or at an



Room 1, associated to the 1st dish.



Room 2, associated to the 2nd dish.

Figure 3.3: The multiroom chinese restaurant metaphor: circles represent tables and bullets represent customers. The number of tables for these two rooms are $(\ell_1, \ell_2) = (3, 2)$ so that $|\boldsymbol{\ell}| = \ell_1 + \ell_2 = 5$. The number of customers eating the first dish (i. e. first room) is $n_1 = \sum_{r=1}^{\ell_1} e_{1r} = 6$, while the number of customers eating the second dish (i.e. second room) are $n_2 = \sum_{r=1}^{\ell_2} e_{2r} = 7$.

existing one. If a new dish is chosen, she will sit in a new room and at a new table. An illustration of this generative scheme is depicted in Figure 3.3.

Recalling the notation used so far, the entries of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ represent the dishes eaten by the n customers of the restaurant, whereas the labels identifying the single tables and their respective frequencies tables are unobservable and, hence, *latent quantities*. Specifically, we consider the latent random variables $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \ldots, \mathcal{T}_n)$, which can be thought of as the label of the table where each customer is seated. Recall that k distinct dishes are served at the restaurant, that is, there are $\theta_1^*, \ldots, \theta_k^*$ distinct values having frequencies $n_1, \ldots, n_k$, meaning that the total number of customers seating in room j or, equivalently, eating dish j, corresponds to the frequency n_j. Finally, each $\ell_j \in \{1, \ldots, n_j\}$ represents the number of tables in room j where customers are seated, while each e_{jr} denotes the number of customers seating at table r in room j, so that $\sum_{j=1}^{k} \sum_{r=1}^{\ell_j} e_{jr} = \sum_{j=1}^{k} n_j = n$. When $H \to \infty$, the probability of observing a new table where the same dish is being served tends to zero, implying that each room will have only one table: in formula, one has the following convergences $\ell_j \xrightarrow{d} 1$ for any $j = 1, \ldots, k$ implying that $|\boldsymbol{\ell}| \xrightarrow{d} k$. We denote the $|\boldsymbol{\ell}|$ distinct labels of the tables with $\mathcal{T}_{j1}^*, \ldots, \mathcal{T}_{j\ell_j}^*$ having frequencies e_{jr} for $r = 1, \ldots, \ell_j$ and $j = 1, \ldots, k$. Thus, the joint augmented model for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \ldots, \mathcal{T}_n)$ follows immediately from equation (3.11). Indeed, one has the following

Corollary 3.3. *The joint probability distribution of* $(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}, \Psi_{n,H})$*, where* $\Psi_{n,H}$ *is the partition of* $[n]$ *induced by* $\boldsymbol{\theta}$ *through* (3.12)*, is*

$$\prod_{j=1}^{k} P(d\theta_j^*) \prod_{r=1}^{\ell_j} P(d\mathcal{T}_{jr}^*) \times \left[\frac{H!}{(H-k)!} \frac{1}{H^{|\boldsymbol{\ell}|}} \Pi_\infty(e_{11}, \ldots, e_{1\ell_1}, \ldots, e_{k1}, \ldots, e_{k\ell_k}) \right]. \tag{3.17}$$

The baseline distribution of $\mathcal{T}$ is set equal to P for simplicity, but any other diffuse probability measure would obviously work. Indeed, only the clustering mechanism implied by the table configuration is relevant to our purposes, and not the actual labels. Equation (3.17) is hence clear: conditionally on the table configuration induced by $\mathcal{T}$, the predictive distribution for θ_{n+1} can be easily obtained. In turns, given the previously observed values $\boldsymbol{\theta}$, the table configuration can be drawn efficiently through a Gibbs sampler. For the sake of the exposition, we do not attempt the full derivation of these conditional distributions, but the interest reader may refer to Camerlenghi et al. (2019) for a detailed discussion, though in a different setting.

Example 3.7 (NGG multinomial process, cont'd). Conditionally on the latent random variables $\mathcal{T}$, the predictive probabilities can be readily obtained from equation (3.17), so that

$$\begin{aligned}\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}, \mathcal{T}) &= \\ &= \left(1 - \frac{k}{H}\right) \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|+1}}{\mathscr{V}_{n,|\boldsymbol{\ell}|}} P(A) + \sum_{j=1}^{k} \left[\frac{1}{H} \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|+1}}{\mathscr{V}_{n,|\boldsymbol{\ell}|}} + \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|}}{\mathscr{V}_{n,|\boldsymbol{\ell}|}} (n_j - \ell_j \sigma)\right] \delta_{\theta_j^*}(A).\end{aligned}$$

Hence, a relevant simplification occurs when considering NGG multinomial processes. In particular, the above conditional distribution depends on the table configuration only through the number of distinct tables $\ell_1, \ldots, \ell_k$ rather than the table-specific frequencies e_{jr}. This is a major computational advantage, since we only need to sample k latent variables rather than n, as in general NIDM processes. Moreover, note that the above conditional law is intimately related to (3.16), having expanded over the table configuration and marginalizing over the latent variable U.

In the following, we expand the posterior characterization of Theorem 3.6 by conditioning also on the table configuration. The random variable U, conditionally on $(\boldsymbol{\theta}, \mathcal{T})$, is a nonnegative latent variable whose density function is given by

$$p^{(\infty)}(u) \propto u^{n-1} e^{-c\psi(u)} \prod_{j=1}^{k} \prod_{r=1}^{\ell_j} \tau_{e_{jr}}(u). \tag{3.18}$$

Thus, conditionally also on $\mathcal{T}$, the latent variable U has the same structure of that involved in the posterior derivation of NRMIS. Similar simplifications occurs also for the fixed jump component, as summarized in the next corollary.

Corollary 3.4. *Let $\theta_1, \ldots, \theta_n$ be a draw from an exchangeable sequence directed by $\tilde{p}^{(H)} \sim$ NIDM$(c, \rho; P)$, with P diffuse, as in (3.12). Moreover, let the conditional distribution of U, given*

$(\boldsymbol{\theta}, \mathcal{T})$, *have density function* $p^{(\infty)}(u)$ *defined as in* (3.18). *Then,*

$$(\tilde{\mu}^{(H)} \mid \boldsymbol{\theta}, \mathcal{T}, U) \stackrel{d}{=} \sum_{j=k+1}^{H} J_j^* \delta_{\bar{\theta}_j} + \sum_{j=1}^{k} (J_j^* + \sum_{r=1}^{\ell_j} I_{jr}) \delta_{\theta_j^*}(A),$$

where $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ *are iid draws from* P, *and*

$$(J_h^* \mid U) \overset{\text{iid}}{\sim} \text{ID}\left(\frac{c}{H}, \rho^*\right), \qquad \rho^*(s) = e^{-Us}\rho(s), \qquad h = 1, \ldots, H.$$

Moreover, the jumps I_{jr} *are independent and nonnegative random variables having density*

$$p_{jr}(s \mid \boldsymbol{\theta}, \mathcal{T}, U) \propto e^{-sU} s^{e_{jr}} \rho(s), \qquad r = 1, \ldots, \ell_j, \quad j = 1, \ldots, k.$$

Hence, conditionally on the table configuration, the posterior structure of $\tilde{p}^{(H)}$ closely resemble the one of homogeneous NRMIs, for any finite value of H. Specifically, the distribution of the random variables I_{jr} have the same kernel of those involved in NRMIs.

Example 3.8 (NGG multinomial process, cont'd). Specializing Corollary 3.4 we obtain that

$$(\sum_{r=1}^{\ell_j} I_{jr} \mid \boldsymbol{\theta}, \mathcal{T}, U) \sim \text{GAMMA}(n_j - \ell_j \sigma, \kappa + U), \qquad j = 1, \ldots, k,$$

which depends on $\mathcal{T}$ only through the number of tables $\ell_1, \ldots, \ell_k$. Note that this representation is essentially the augmentation of equation (3.15) with respect to the number of tables. Note that when $\sigma = 0$, the posterior law $\tilde{p}^{(H)}$ becomes independent on the table configuration.

3.5 The INVALSI dataset

We consider a publicly available dataset gathered by INVALSI, which is an institute for the assessment of the Italian education system. In particular, the 2016-2017 dataset we are going to analyze is part of a national examination program conducted in Italy with the aim of "carrying out periodic and systematic checks on knowledge and skills of students", as declared in the official documentation of the INVALSI statistical service[1]. A great effort was put by the INVALSI in order to quantify the *added-value* of a school, based on these data. The Bayesian framework constitutes a natural choice when trying to combine multiple sources of information, i.e. the schools. This can be accomplished through hierarchical nonparametric models, which enable flexible borrowing of information between different

[1]Such documentation is available, only in Italian, at: https://invalsi-serviziostatistico.cineca.it

institutions (see e.g. Dunson, 2010). A broad and systematic socio-demographical analysis is beyond the aims of this chapter, and we hence limit ourselves in the presentation of novel modeling strategies based on NIDM priors.

In view of our discussion, we confine the analysis to data related to 8th grade students from schools in the cities of Padova and Bolzano: more specifically we focus on those questions related to the comprehension of the Italian language. Having omitted few observations for which covariates were not available, the resulting dataset comprises a total of $N = 9808$ observations (students), belonging to 100 educational institutions. The INVALSI test is composed by 45 questions and the performance of each student might be well summarized by the proportions of correct answers. To ease the modeling process we take a logistic transformation of the original proportions, and we define the score S_{ij} for the ith student in the jth school as

$$S_{ij} := \text{logit}\left(\frac{\#\text{ of correct answers, } i\text{th student } j\text{th school} + 1/2}{\#\text{ of questions} + 1}\right), \quad i = 1, \dots, N_j,$$

and $j = 1, \dots, 100$, where N_j denotes the number of students in the jth school. In the above ratio, we added a small correction to the original proportions to avoid boundary issues. Such a transformation maps the original scores into $\mathbb{R}$, and therefore it is more amenable for classical linear modeling with Gaussian errors. Consistent with this, we model the scores as follows

$$S_{ij} = \mu_j + \boldsymbol{x}_{ij}^\intercal \boldsymbol{\beta} + \epsilon_{ij}, \qquad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

for $i = 1, \dots, N_j$ and $j = 1, \dots, 100$, where $\boldsymbol{x}_{ij} = (1, x_{ij1}, \dots, x_{ijp})^\intercal$ is a vector of student-specific covariates which are associated to a $p + 1$ dimensional vector of regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\intercal$. Each vector $\boldsymbol{x}_{ij}$ encodes student-specific categorical variables, namely: the gender of the student, the education level of her/his father and mother (primary school, secondary school, etc.), the employment typology of her/his father and mother, the regularity of the student (i.e. regular, in late, etc.), and the citizenship (italian, first generation immigrant, etc.). Moreover, the coefficients $\mu_1, \dots, \mu_{100}$ represents the school effects or, in the terminology used so far, the *added-value* of the school given a set of covariates, thus being the main quantity of interest in our analysis. Note that since the intercept term is included in $\boldsymbol{x}_{ij}$, then the coefficients $\mu_1, \dots, \mu_{100}$ are not identified. In practice, this is not a concern if inference is based on the "centred" set of parameters $\beta_0 + \mu_j$, for $j = 1, \dots, 100$, rather than the original random effects.

We aim at introducing a flexible "nonparametric" prior, for the school effects $\mu_1, \dots, \mu_{100}$, that allows for: i) borrowing information across schools; ii) arbitrary

deviations from Gaussianity, and iii) robustness under model misspecifications. Gaussian mixture models are able to capture all these three aspects and we specifically let

$$(\mu_1, \dots, \mu_{100} \mid \mathcal{P}) \overset{\text{iid}}{\sim} \mathcal{P}, \qquad \mathcal{P} = \sum_{h=1}^{H} \pi_h \mathcal{N}(\bar{\mu}_h, \bar{\sigma}_h^2), \tag{3.19}$$

for $j = 1, \dots, 100$. Selecting the appropriate number of mixture components H is typically a difficult task, and the BIC index – customarily employed in this framework – is not theoretically well justified here. Hence, overfitted mixture models can be exploited to circumvent this issue. In particular, the Dirichlet multinomial process has found great applicability as mixing measure, see e.g. Durante et al. (2017) and Rigon et al. (2019) for some recent contributions. Here we propose the usage of general NIDM processes, which amounts to have the following prior specification for the parameters in (3.19)

$$(\pi_1, \dots, \pi_{H-1}) \sim \text{NID}\left(\frac{c}{H}, \dots, \frac{c}{H}; \rho\right), \qquad (\bar{\mu}_h, \bar{\sigma}_h^2) \overset{\text{iid}}{\sim} P \qquad h = 1, \dots, H,$$

where P is a diffuse probability measure on $\mathbb{R} \times \mathbb{R}_+$. By enlarging the class of priors from the Dirichlet multinomial to general NIDM multinomial processes we are essentially acting on the robustness requirement, that is, we are ensuring that the clustering mechanism is less affected by specific choices of the total mass parameter c, whose specification is often critical (Malsiner-Walli et al., 2016). Although one might alternative mitigate this issue by placing a prior distribution on c, this is arguably a more convoluted solution. Indeed, the resulting prior would be much harder to study analytically: for instance the posterior law would not be available in closed form. Hence, albeit simple, such a specification has been missing a solid theoretical background and has mostly turned out to be a sort of "black-box" prior. Additionally, from a more practical perspective the employment of NGG priors enables exact sampling of the posterior distribution. In contrast, a Dirichlet multinomial process with a prior on c usually requires a Metropolis step in the posterior computation (Ishwaran & Zarepour, 2000).

For this specific application, we employed in (3.19) a NGG multinomial process having jump measure (3.2). We set the hyperparameters equal to $c = 0.1, \kappa = 0.1$ and $\sigma = 0.7$, and we selected a very conservative upper bound for the number of mixture components $H = 50$. The a priori effect of these values on the number of cluster $K_{n,H}$ is depicted in Figure 3.4, where it is compared with the distribution induced by a Dirichlet multinomial process ($c = 11$) having roughly the same expected value $\mathbb{E}(K_{n,H}) \approx 20$. As evidenced by Figure 3.4, the parameter σ plays a crucial role in controlling the variability of $K_{n,H}$. Hence, one can obtain flat specifications for $K_{n,H}$, which leads to more robust inference on the cluster configurations. Such an effect persists a posteriori, as we shall discuss later

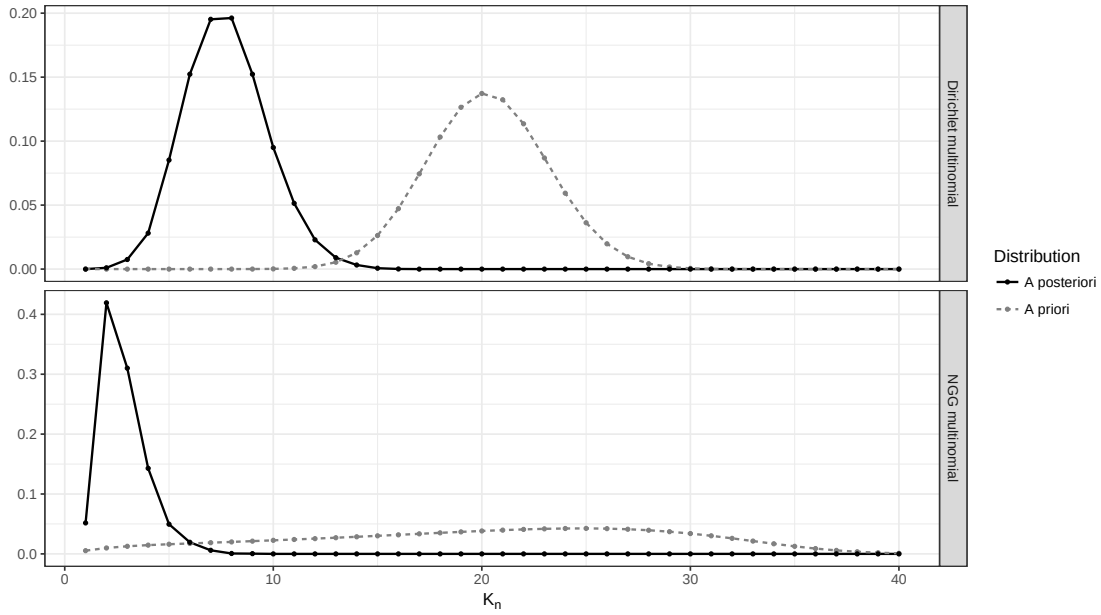


Figure 3.4: A priori and a posteriori distribution of the number of cluster $K_{n,H}$ in the INVALSI application, under a Dirichlet multinomial prior ($c = 11$) and a NGG multinomial process prior ($c = 0.1, \kappa = 0.1, \sigma = 0.7$), with $H = 50$. The comparison is limited to the interval $[1, 40]$ for graphical reasons.

on. As for the choice of P, we assume the conditionally conjugate prior

$$\tilde{\mu}_h \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\tilde{\mu}}), \qquad \tilde{\sigma}_h^{-2} \overset{\text{iid}}{\sim} \text{GA}(a_{\tilde{\sigma}}, b_{\tilde{\sigma}}), \qquad h = 1, \dots, H.$$

where we set $\sigma^2_{\tilde{\mu}} = 4$ and $a_{\tilde{\sigma}} = 1.5$, $b_{\tilde{\sigma}} = 0.05$. To conclude our Bayesian formulation we consider a multivariate Gaussian prior for the regression coefficients $\boldsymbol{\beta}$, and an inverse gamma prior for the residual variance σ^2, and we let

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \qquad \sigma^{-2} \sim \text{GAMMA}(a_\sigma, b_\sigma),$$

where we set $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \text{diag}(100, \dots, 100)$, to incorporate the neutral hypothesis of no relevant effects and $a_\sigma = b_\sigma = 1$, which induces a fairly non-informative prior.

Posterior inference was conducted through a Gibbs sampling, whose details can be found in Appendix 3.6. For comparison, we also estimated the same model under a Dirichlet multinomial prior as for Figure 3.4. We run the algorithm for 10′000 iterations after a burn-in period of 1′000 simulations; the traceplots show no evidence against convergence and an excellent mixing. Computations were performed on a standard laptop (MacBook Air with 1,3 GHz Intel Core i5 processor), and they took approximately 5 minutes for both models.

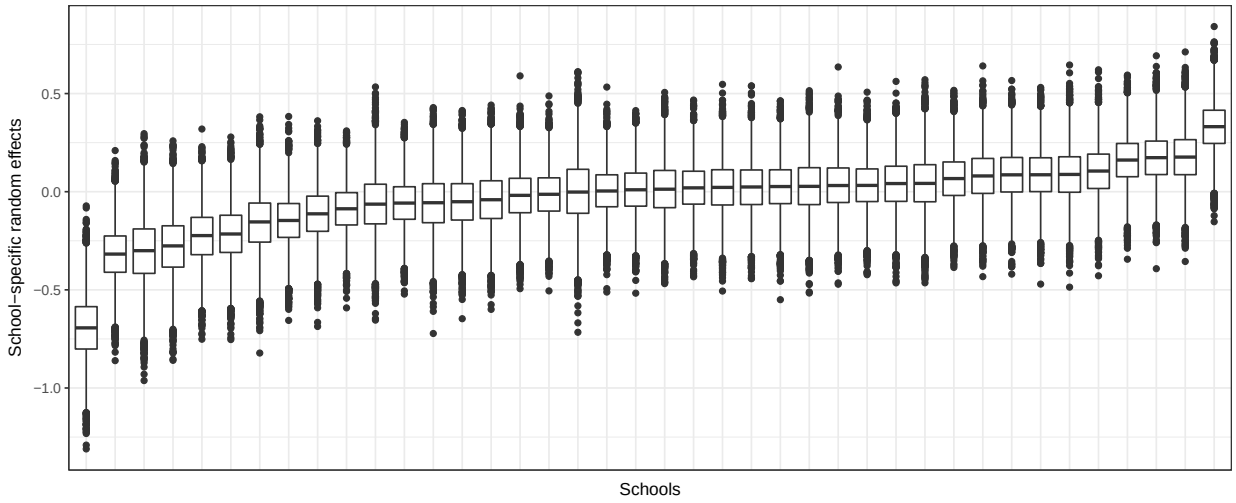


Figure 3.5: Posterior distribution of the random effects $\beta_0 + \mu_j$ for 40 randomly selected schools, ordered according to the posterior median.

From Figure 3.4, it is apparent that the a posteriori distribution of $K_{n,H}$ is heavily influenced by the prior choice. In the Dirichlet multinomial case the number of mixture components peaks at around 7 and 8, as a consequence of the strongly informative prior distribution. Conversely, when a more robust NGG prior is used, the data adaptively select a much smaller number of components, peaking at around 2 mixture components. Importantly, this implies that if a simple Gaussian random effect model were employed, the random effects $\beta_0 + \mu_j$ would be overshrunk towards the global mean, potentially affecting the quality of the analysis. We summarize our findings in Figure 3.5, where we show the posterior distribution of the $\beta_0 + \mu_j$ random effects for 40 randomly selected schools. It is evident that a certain degree of variability among schools is present and a posterior summary of each $\beta_0 + \mu_j$ might be employed to identify virtuous schools.

3.6 Appendix

Laplace functional of a IDM random measure

In first place, recall that the Laplace functional of a CRM is available in closed form and it is given in equation (3.1). The Laplace functional of a IDM random measure is then readily obtained after noting that $(\tilde{\mu}^{(H)} \mid \tilde{\theta}_1, \dots, \tilde{\theta}_H)$ is a completely random measure with a purely atomic baseline distribution, placing mass on $\tilde{\theta}_1, \dots, \tilde{\theta}_H$. Thus given the atoms $\tilde{\theta}_1, \dots, \tilde{\theta}_H$ the functional $\int_\Theta f(\theta)\tilde{\mu}^{(H)}(d\theta) = \sum_{h=1}^{H} J_h f(\tilde{\theta}_h) < \infty$ almost surely, so

that the following chains of equations hold

$$\begin{aligned}
\mathbb{E}\left(e^{-\int_\Theta f(\theta)\tilde{\mu}^{(H)}(d\theta)}\right) &= \mathbb{E}\left\{\mathbb{E}\left(e^{-\int_\Theta f(\theta)\tilde{\mu}^{(H)}(d\theta)} \mid \tilde{\theta}_1,\ldots,\tilde{\theta}_H\right)\right\} \\
&= \mathbb{E}\left\{\exp\left(-\frac{c}{H}\sum_{h=1}^{H}\int_{\mathbb{R}_+}(1-e^{-sf(\tilde{\theta}_h)})\rho(s)\,ds\right)\right\} \\
&= \mathbb{E}\left(\exp\left[-\frac{c}{H}\sum_{h=1}^{H}\psi\{f(\tilde{\theta}_h)\}\right]\right) \\
&= \prod_{h=1}^{H}\mathbb{E}\left(\exp\left[\frac{c}{H}\psi\{f(\tilde{\theta}_h)\}\right]\right) = \left\{\int_\Theta \exp\left[-\frac{c}{H}\psi\{f(\theta)\}\right]P(d\theta)\right\}^H.
\end{aligned}$$

The last two equalities follows because the locations $\tilde{\theta}_1,\ldots,\tilde{\theta}_H$ are iid from P. The Laplace transform of $\tilde{\mu}^{(H)}(A)$ readily follows having set $f = \lambda\mathbb{1}_A(x)$, for $\lambda > 0$ and $A \in \Theta$. Indeed, simple calculus lead to

$$\begin{aligned}
\int_\Theta \exp\left[-\frac{c}{H}\psi\{\lambda\mathbb{1}_A(\theta)\}\right]P(d\theta) &= \int_A \exp\left[-\frac{c}{H}\psi\{\lambda\mathbb{1}_A(\theta)\}\right]P(d\theta) \\
&\quad + \int_{\Theta\setminus A}\exp\left[-\frac{c}{H}\psi\{\lambda\mathbb{1}_A(\theta)\}\right]P(d\theta) \\
&= P(A)\exp\left\{-\frac{c}{H}\psi(\lambda)\right\} + 1 - P(A),
\end{aligned}$$

from which follows the Laplace transform in (3.9).

Proof of Proposition 3.1

First, notice that the expected value, as in any species sampling model, is simply equal to $\mathbb{E}\{\tilde{p}^{(H)}(A)\} = \sum_{h=1}^{H}\mathbb{E}(\pi_h)\mathbb{E}\{\delta_{\tilde{\theta}_h}(A)\} = P(A)\sum_{h=1}^{H}\mathbb{E}(\pi_h) = P(A)$. As an application of the well-known variance decomposition

$$\mathrm{Var}\{\tilde{p}^{(H)}(A)\} = \mathbb{E}\{\mathrm{Var}(\tilde{p}^{(H)}(A) \mid \tilde{p}_0^{(H)})\} + \mathrm{Var}\{\mathbb{E}(\tilde{p}^{(H)}(A) \mid \tilde{p}_0^{(H)})\}.$$

Let us focus on the second summand on the right-hand side of the above equation, which is equal to

$$\mathrm{Var}\{\mathbb{E}(\tilde{p}^{(H)}(A) \mid \tilde{p}_0^{(H)})\} = \mathrm{Var}\{\tilde{p}_0^{(H)}(A)\} = \frac{P(A)\{1-P(A)\}}{H}.$$

As for $\mathbb{E}\{\mathrm{Var}(\tilde{p}^{(H)}(A) \mid \tilde{p}_0^{(H)})\}$, because of Proposition 1 in James et al. (2006) we obtain

$$\begin{aligned}
\mathbb{E}\{\mathrm{Var}(\tilde{p}^{(H)}(A) \mid \tilde{p}_0^{(H)})\} &= \mathbb{E}(\tilde{p}_0^{(H)}(A)\{1-\tilde{p}_0^{(H)}(A)\}\mathcal{J}(c,\rho) \\
&= P(A)\{1-P(A)\}\left(\mathcal{J}(c,\rho) - \frac{\mathcal{J}(c,\rho)}{H}\right),
\end{aligned}$$

from which the result follows. As for the covariance, note that $\text{Var}\{\tilde{p}^{(H)}(A_1), \tilde{p}^{(H)}(A_2)\} = P(A)\{1 - P(A)\}\mathbb{E}\left(\sum_{h=1}^{H} \pi_h^2\right)$ whereas

$$\text{Cov}\{\tilde{p}^{(H)}(A_1), \tilde{p}^{(H)}(A_2))\} = \{P(C) - P(A_1)P(A_2)\}\mathbb{E}\left(\sum_{h=1}^{H} \pi_h^2\right),$$

meaning that

$$\mathbb{E}\left(\sum_{h=1}^{H} \pi_h^2\right) = \mathcal{I}(c, \rho) + \frac{1 - \mathcal{I}(c, \rho)}{H},$$

from which the result follows.

Proof of Theorem 3.1

Recall that the Laplace functional can be written as in Appendix 3.6, so that

$$\mathbb{E}\left(e^{-\int_\Theta f(\theta)\tilde{\mu}^{(H)}(d\theta)}\right) = \mathbb{E}\left(\exp\left[-\frac{c}{H}\sum_{h=1}^{H} \psi\{f(\tilde{\theta}_h)\}\right]\right).$$

Now note that the expectations of each $\psi(\tilde{\theta}_h)$ equals

$$\mathbb{E}\{\psi(f(\tilde{\theta}_h))\} = \int_\Theta \psi\{f(\theta)\}P(d\theta) < \infty,$$

which is finite by assumption. Hence, as an application of the strong law of large numbers, we get

$$\frac{1}{H}\sum_{h=1}^{H} \psi\{f(\tilde{\theta}_h)\} \xrightarrow{\text{a.s.}} \int_\Theta \psi\{f(\theta)\}P(d\theta), \qquad H \to \infty,$$

which implies that $\mathbb{E}\left(e^{-\int_\Theta f(\theta)\tilde{\mu}^{(H)}(d\theta)}\right) \to \mathbb{E}\left(e^{-\int_\Theta f(\theta)\tilde{\mu}^{(\infty)}(d\theta)}\right)$ because of bounded convergence theorem. Given the convergence of the functionals, vague convergence is implied by Kallenberg (Theorem 4.11, 2017) since the condition $\int_\Theta \psi\{f(\theta)\}P(d\theta) < \infty$ is satisfied if f is a positive, continuous and bounded function.

Proof of Theorem 3.2

The symmetry among the weights implies that

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!}\mathbb{E}\left(\prod_{j=1}^{k} \pi_j^{n_j}\right).$$

Recalling that $\tilde{\mu}(\Theta) = \sum_{h=1}^{H} J_h$, then we have

$$\begin{aligned}
\mathbb{E}\left(\prod_{j=1}^{k}\pi_j^{n_j}\right) &= \frac{1}{\Gamma(n)}\int_{\mathbb{R}_+} u^{n-1}\mathbb{E}\left(e^{-u\tilde{\mu}(\Theta)}\prod_{j=1}^{k}J_j^{n_j}\right)du \\
&= \frac{1}{\Gamma(n)}\int_{\mathbb{R}_+} u^{n-1}\prod_{j'=k+1}^{H}\mathbb{E}\left(e^{-uJ_{j'}}\right)\prod_{j=1}^{k}\mathbb{E}\left(e^{-uJ_j}J_j^{n_j}\right)du \\
&= \frac{1}{\Gamma(n)}\int_{\mathbb{R}_+} u^{n-1}e^{-c\frac{H-k}{H}\psi(u)}\prod_{j=1}^{k}(-1)^{n_j}\frac{\partial^{n_j}}{\partial u^{n_j}}e^{-\frac{c}{H}\psi(u)}du \\
&= \frac{1}{\Gamma(n)}\int_{\mathbb{R}_+} u^{n-1}e^{-c\psi(u)}\prod_{j=1}^{k}\mathcal{V}_{n_j,H}(u)du,
\end{aligned}$$

which concludes the proof, since $\mathcal{V}_{n_j,H}(u) = \frac{c}{H}\Delta_{n_j,H}(u)$. The predictive distributions of Corollary 3.2 can be obtained exploiting their relationship with the EPPF and after some algebraic manipulation. To obtain the alternative representation (3.10), recall the following equality, whose proof can be found in Camerlenghi et al. (2019), which holds for $m \geqslant 1$

$$\mathcal{V}_{m,H}(u) = \frac{c}{H}\sum_{\ell=1}^{m}\varsigma_{m,\ell,H}(u), \qquad \varsigma_{m,\ell,H}(u) = \left(\frac{c}{H}\right)^{\ell-1}\frac{1}{\ell!}\sum_{\boldsymbol{e}}\binom{m}{e_1,\ldots,e_\ell}\prod_{r=1}^{\ell}\tau_{e_r}(u), \quad (3.20)$$

for $\ell = 1,\ldots,m$, where the sum runs over all the vectors of positive integers $\boldsymbol{e} = (e_1,\ldots,e_\ell)$ such that $|\boldsymbol{e}| = m$. Thus, on the light of (3.20) we can write the EPPF as

$$\begin{aligned}
\Pi_H(n_1,\ldots,n_k) &= \frac{H!}{(H-k)!}\frac{1}{\Gamma(n)}\int_{\mathbb{R}_+} u^{n-1}e^{-c\psi(u)}\prod_{j=1}^{k}\mathcal{V}_{n_j,H}(u)du \\
&= \frac{H!}{(H-k)!}\sum_{\boldsymbol{\ell}}\frac{1}{H^{|\boldsymbol{\ell}|}}\prod_{j=1}^{k}\frac{1}{\ell_j!}\sum_{\boldsymbol{e}_j}\binom{n_j}{e_{j1},\ldots,e_{j\ell_j}}\Pi_\infty(e_{11},\ldots,e_{1\ell_1},\ldots,e_{k1},\ldots,e_{k\ell_k}),
\end{aligned}$$

where the first sum runs over all the vectors $\boldsymbol{\ell} = (\ell_1,\ldots,\ell_k)$ such that $\ell_j \in \{1,\ldots,n_j\}$, and the jth of the k sums runs over all the vectors $\boldsymbol{e}_j = (e_{j1},\ldots,e_{j\ell_j})$ such that $e_{jr} \geqslant 1$ and $|\boldsymbol{e}_j| = n_j$.

Proof of Theorem 3.3

Let us consider the ratio among the two EPPFs, which is equal for any $k \leqslant H$ to

$$\frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_\infty(n_1,\ldots,n_k)} = \frac{H!}{H^k(H-k)!}\frac{\int_{\mathbb{R}_+} u^{n-1}e^{-c\psi(u)}\prod_{j=1}^{k}\Delta_{n_j,H}(u)du}{\int_{\mathbb{R}_+} u^{n-1}e^{-c\psi(u)}\prod_{j=1}^{k}\tau_{n_j}(u)du}.$$

The result follows after noting that the ratio $\frac{H!}{H^k(H-k)!} \leqslant 1$, and also

$$\frac{\int_{\mathbb{R}_+} u^{n-1} e^{-c\psi(u)} \prod_{j=1}^{k} \Delta_{n_j,H}(u) du}{\int_{\mathbb{R}_+} u^{n-1} e^{-c\psi(u)} \prod_{j=1}^{k} \tau_{n_j}(u) du} \geqslant 1.$$

The latter inequality can be easily obtained from (3.20), from which is clear that $\Delta_{m,H}(u) = \tau_m(u) + f_m(u)$, where $f_m(u)$ is a positive function, implying that $\Delta_{m,H}(u) \geqslant \tau_m(u)$ for any $m \geqslant 1$ and $u > 0$.

Proof of Theorem 3.4

The starting point of this proof is based on Corollary 2 in Camerlenghi et al. (2018), from which one can show that

$$\mathbb{P}(K_{n,H} = k) = \sum_{s=k}^{n} \mathbb{P}(K_{n,\infty} = s)\mathbb{P}(K_{s,0} = k), \tag{3.21}$$

where $K_{n,\infty}$ and $K_{n,H}$ are defined as before, while $K_{n,0}$ for any $n \geqslant 1$ is the number of distinct values from a sample of n exchangeable observations having prior $\tilde{p}_0^{(H)}$. The distribution $\mathbb{P}(K_{n,0} = k)$ can be deduced from the associated EPPF, which is

$$\Pi_0(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} H^{-n}, \qquad k \leqslant H,$$

implying that the distribution of $K_{n,0}$ is given for any $k \leqslant \min\{H, n\}$

$$\mathbb{P}(K_{n,0} = k) = \frac{1}{k!} \sum_{(n_1,\ldots,n_k)} \binom{n}{n_1, \ldots, n_k} \Pi_0(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} H^{-n} \mathscr{S}(n, k),$$

where the sum runs over all the k-dimensional vectors of positive integers $\boldsymbol{n} = (n_1, \ldots, n_k)$ such that $|\boldsymbol{n}| = n$. The first part of the theorem then follows after the change of variable $\ell = s - k$ in (3.21). As for the second part, note that the expected value of $K_{n,H}$ can be written as

$$\begin{aligned}
\mathbb{E}(K_{n,H}) &= \mathbb{E}\{\mathbb{E}(K_{n,H} \mid \tilde{\theta}_1, \ldots, \tilde{\theta}_H)\} = \mathbb{E}\left\{\mathbb{E}\Big(\sum_{h=1}^{H} \mathbb{1}(\tilde{\theta}_h \in \{\theta_1, \ldots, \theta_n\} \mid \tilde{\theta}_1, \ldots, \tilde{\theta}_H\Big)\right\} \\
&= \sum_{h=1}^{H} \mathbb{E}\{1 - \mathbb{P}(\theta_1 \neq \tilde{\theta}_h, \ldots, \theta_n \neq \tilde{\theta}_h \mid \tilde{\theta}_1, \ldots, \tilde{\theta}_H)\} \\
&= \sum_{h=1}^{H} [1 - \mathbb{E}\{(1 - \pi_h)^n\}] = H - H\mathbb{E}\{(1 - \pi_1)^n\}.
\end{aligned}$$

The randomness in these equations is given both by $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$ and $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$, whereas in the last step we have used the symmetricity of the weights of a NIDM process. Moreover, with the same steps as for the proof of Theorem 3.2, one can easily show that

$$\mathbb{E}\{(1-\pi_1)^n\} = \sum_{\ell=1}^{n} \left(1 - \frac{1}{H}\right)^{\ell} \frac{1}{\ell!} \sum_{e} \binom{n}{e_1, \ldots, e_\ell} \Pi_\infty(e_1, \ldots, e_\ell),$$

where the sums runs over $e = (e_1, \ldots, e_\ell)$ such that $e_j \geqslant 1$ and $|e| = n$, from which the second part of the Theorem follows.

Proof of Theorem 3.5

Recall that the ratio of interest is given by

$$\frac{\mathbb{P}(K_{n,H} = k)}{\mathbb{P}(K_{n,\infty} = k)} = \frac{H!}{H^k (H-k)!} \sum_{\ell=0}^{n-k} \frac{1}{H^\ell} \mathscr{S}(\ell + k, k) \frac{\mathbb{P}(K_{n,\infty} = \ell + k)}{\mathbb{P}(K_{n,\infty} = k)},$$

and therefore the lower bound follows naturally. We will write

$$\frac{H!}{H^k (H-k)!} \leqslant \frac{\mathbb{P}(K_{n,H} = k)}{\mathbb{P}(K_{n,\infty} = k)} = \frac{H!}{H^k (H-k)!} \left(1 + \sum_{\ell=1}^{n-k} \frac{1}{H^\ell} \mathscr{S}(\ell + k, k) \frac{\mathbb{P}(K_{n,\infty} = \ell + k)}{\mathbb{P}(K_{n,\infty} = k)}\right).$$

Now recall the well-known inequality due to Rennie & Dobson (1969), for which for any $n \geqslant 2$ and $1 \leqslant k \leqslant n-1$ a Stirling number of the second kind can be bounded above by

$$\mathscr{S}(n, k) \leqslant \frac{1}{2} \binom{n}{k} k^{n-k},$$

implying that we can further bound the summation of the above equation for $1 \leqslant k \leqslant \min\{H, n-1\}$ in the following way

$$\sum_{\ell=1}^{n-k} \frac{1}{H^\ell} \mathscr{S}(\ell + k, k) \frac{\mathbb{P}(K_{n,\infty} = \ell + k)}{\mathbb{P}(K_{n,\infty} = k)} \leqslant \frac{1}{2} \sum_{\ell=1}^{n-k} \left(\frac{k}{H}\right)^{\ell} \binom{\ell + k}{k} \frac{\mathbb{P}(K_{n,\infty} = \ell + k)}{\mathbb{P}(K_{n,\infty} = k)}.$$

Hence, the result follows.

Proof of Theorem 3.6

We first derive the posterior distribution of $\tilde{p}_0^{(H)} = \frac{1}{H} \sum_{h=1}^{H} \delta_{\tilde{\theta}_h}$ given the $\boldsymbol{\theta}$, when P is assumed to be diffuse. This fact is summarized in the following proposition.

Lemma 3.1. *Let $\theta_1, \ldots, \theta_n$ be a draw from an exchangeable sequence directed by a* NIDM *process and let P be diffuse. Then, the posterior distribution of $\tilde{p}_0^{(H)}$, defined as in (3.10), for any $A \in$*

$\mathscr{B}(\Theta)$ is

$$(\tilde{p}_0^{(H)} \mid \boldsymbol{\theta}) \stackrel{d}{=} \frac{1}{H}\left[\sum_{j=k+1}^{H} \delta_{\tilde{\theta}_j} + \sum_{j=1}^{k} \delta_{\theta_j^*}\right],$$

where the atoms $\tilde{\theta}_{k+1}, \ldots, \tilde{\theta}_H$ are iid draws from P.

Proof. Since the weights of $\tilde{p}_0^{(H)}$ are fixed and equal, we only need to determine the posterior law of the atoms $(\tilde{\theta}_1, \ldots, \tilde{\theta}_H \mid \boldsymbol{\theta})$. Recall that a NIDM process, when P is diffuse, is a species sampling model, meaning that k out of H atoms are necessarily equal almost surely to one of the previously observed values $\theta_1^*, \ldots, \theta_k^*$, while the remaining $H-k$ are iid draws from the baseline measure P. Notice that the actual order of the weights is irrelevant, because of the symmetry of the weights of $\tilde{p}_0^{(H)}$. Hence, the result in Proposition 3.1 follows.

Because of symmetry of the weights, we can assume without loss of generality that the distinct values $\theta_1^*, \ldots, \theta_k^*$ are associated to the first k random weights $\pi_1, \ldots, \pi_k$ of the process $\tilde{p}^{(H)}$. The Laplace functional of $\tilde{\mu}^{(H)}$, given $\boldsymbol{\theta}$ and $\tilde{p}_0^{(H)}$ is given by

$$\mathbb{E}\left(e^{-\tilde{\mu}^{(H)}(f)} \mid \boldsymbol{\theta}, \tilde{p}_0^{(H)}\right) = \frac{\mathbb{E}\left(e^{-\tilde{\mu}^{(H)}(f)} \prod_{j=1}^{k} \pi_j^{n_j} \mid \tilde{p}_0^{(H)}\right)}{\mathbb{E}\left(\prod_{j=1}^{k} \pi_j^{n_j} \mid \tilde{p}_0^{(H)}\right)},$$

and hence, with similar steps as for Theorem 3.2, both at the numerator and the denominator, we obtain

$$\begin{aligned}
&\mathbb{E}\left(e^{-\tilde{\mu}^{(H)}(f)} \mid \boldsymbol{\theta}, \tilde{p}_0^{(H)}\right) = \\
&= \frac{\int_{\mathbb{R}_+} u^{n-1} e^{-\frac{c}{H}\sum_{j=1}^{k}\psi(f(\theta_j^*)+u)}\, e^{-\frac{c}{H}\sum_{h=k+1}^{H}\psi(f(\tilde{\theta}_h)+u)} \prod_{j=1}^{k} \Delta_{n_j,H}(f(\theta_j^*)+u)du}{\int_{\mathbb{R}_+} u^{n-1} e^{-c\psi(u)} \prod_{j=1}^{k} \Delta_{n_j,H}(u)du} \\
&= \int_{\mathbb{R}_+} e^{-\frac{c}{H}\sum_{j=1}^{k}\psi^{(u)}(f(\theta_j^*))}\, e^{-\frac{c}{H}\sum_{h=k+1}^{H}\psi^{(u)}(f(\tilde{\theta}_h))} \prod_{j=1}^{k} \frac{\Delta_{n_j,H}(f(\theta_j^*)+u)}{\Delta_{n_j,H}(u)}\, p^{(H)}(u)\, du \\
&= \int_{\mathbb{R}_+} \prod_{h=k+1}^{H} \mathbb{E}\left(e^{-f(\tilde{\theta}_h)J_h^*}\right) \prod_{j=1}^{k} \mathbb{E}\left(e^{-f(\theta_j^*)(J_j^*+I_j)}\right) p^{(H)}(u)\, du
\end{aligned}$$

where we used the fact that $\psi(f(\theta)+u) = \psi^{(u)}(f(\theta)) + \psi(u)$, with $\psi^{(u)}(\lambda)$ denoting the Laplace exponent associated to the tilted jump measure $\rho^*(s) = e^{-us}\rho(s)$. It remains to show that any ratio $\Delta_{m,H}(\lambda+u)/\Delta_{m,H}(u)$ is indeed the Laplace transform associated to some nonnegative random variable, for any $m \geqslant 1$ and $\lambda > 0$. This is immediately evident from equation (3.20), because each $\Delta_{m,H}(u)$ can be expressed as a linear combination of Laplace exponents of the form $\tau_m(\lambda+u)/\tau_m(u)$, meaning that each random variable

I_j can be interpreted as a mixture of convolutions of random variables. By combining Proposition 3.1 with the above Laplace functional the result follows.

Proof of Corollary 3.4

By exploiting equation (3.20), one can easily notice that $\mathbb{E}\left(e^{-\tilde{\mu}^{(H)}(f)} \mid \boldsymbol{\theta}, \tilde{p}_0^{(H)}\right)$ obtained in the proof of Theorem 3.6 can be interpreted as a mixture over the table configurations. Thus, by augmenting and subsequently conditioning on the table frequencies, one can easily obtain

$$\begin{aligned}
&\mathbb{E}\left(e^{-\tilde{\mu}^{(H)}(f)} \mid \boldsymbol{\theta}, \mathfrak{T}, \tilde{p}_0^{(H)}\right) = \\
&= \int_{\mathbb{R}_+} e^{-\frac{c}{H}\sum_{h=k+1}^{H}\psi^{(u)}(f(\tilde{\theta}_h))}\, e^{-\frac{c}{H}\sum_{j=1}^{k}\psi^{(u)}(f(\theta_j^*))} \times \prod_{j=1}^{k}\prod_{r=1}^{\ell_j}\frac{\tau_{e_{jr}}(f(\theta_j^*)+u)}{\tau_{e_{jr}}(u)} p^{(\infty)}(u)du \\
&= \int_{\mathbb{R}_+} \prod_{h=k+1}^{H}\mathbb{E}\left(e^{-f(\tilde{\theta}_h)J_h^*}\right)\prod_{j=1}^{k}\prod_{r=1}^{\ell_j}\mathbb{E}\left(e^{-f(\theta_j^*)(J_j^*+I_{jr})}\right) p^{(\infty)}(u)du,
\end{aligned}$$

from which the result follows, by combining the above equation with Proposition 3.1.

Dirichlet multinomial process

In order to derive the EPPF od the Dirichlet multinomial from Theorem 3.2 one just need to notice that when $\rho(s)ds = s^{-1}e^{-s}ds$, then for any $m \geqslant 1$ and $u > 0$ it holds

$$\mathcal{V}_{m,H}(u) = \frac{c}{H}\Delta_{m,H}(u) = \frac{\Gamma(m+c/H)}{\Gamma(m)\Gamma(c/H)}\tau_m(u),$$

which can be verified directly from the definition of $\mathcal{V}_{m,H}(u)$ and $\tau_m(u)$. Substituting the above quantity in general formula of Theorem 3.2, one has simply that for $k \leqslant H$

$$\Pi_H(n_1,\ldots,n_k) = \frac{H!}{(H-k)!}\frac{1}{c^k\Gamma(c/H)^k}\prod_{j=1}^{k}\left(\frac{\Gamma(n_j+c/H)}{\Gamma(n_j)}\right) \times \Pi_\infty(n_1,\ldots,n_k),$$

where $\Pi_\infty(n_1,\ldots,n_k) = c^k/(c)_n\prod_{j=1}^{k}\Gamma(n_j)$ is the EPPF of a Dirichlet process. Hence the desired EPPF can be obtained with some simple algebra. Notice that one could also obtain this result specializing the general EPPF of the NGG multinomial process, by letting $\sigma \to 0$. Indeed, recall that in the Dirichlet process case $\mathscr{V}_{n,k} = c^k/(c)_n$, and that as $\sigma \to 0$ one has $\sigma^{-k}\mathscr{C}(n,k;\sigma) \to |s(n,k)|$, the sign-less Stirling number of the first kind. The distribution of $K_{n,H}$ is also obtained by exploiting properties of Stirling numbers. Indeed, specializing

Theorem 3.4 and after a change of variable

$$\begin{aligned}\mathbb{P}(K_{n,H}=k) &= \frac{H!}{(H-k)!}\frac{1}{(c)_n}\sum_{t=k}^{n}\left(\frac{c}{H}\right)^t \mathscr{S}(t,k)|s(n,t)| \\ &= \frac{H!}{(H-k)!}\frac{(-1)^n}{(c)_n}\sum_{t=k}^{n}\left(-\frac{c}{H}\right)^t \mathscr{S}(t,k)s(n,t) \\ &= \frac{H!}{(H-k)!}\frac{(-1)^k}{(c)_n}\mathscr{C}(n,k;-c/H).\end{aligned}$$

NGG multinomial process

Substituting the EPPF of a generalized gamma NRMI in (3.11), and focusing on the summation one has

$$\begin{aligned}&\frac{1}{\ell_j!}\sum_{e_j}\binom{n_j}{e_{j1},\ldots,e_{j\ell_j}}\Pi_\infty(e_{11},\ldots,e_{1\ell_1},\ldots,e_{k1},\ldots,e_{k\ell_k}) = \\ &= \mathscr{V}_{n,|\ell|}\frac{1}{\ell_j!}\sum_{e_j}\binom{n_j}{e_{j1},\ldots,e_{j\ell_j}}\prod_{r=1}^{\ell_j}(1-\sigma)_{e_{jr}-1} = \mathscr{V}_{n,|\ell|}\frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\sigma^{\ell_j}},\end{aligned}$$

from which the EPPF of a NGG multinomial process follows. With the same reasoning, one also obtain the explicit relation for $\Delta_{m,H}(u)$ after recalling (3.20).

Simulation of U in the NGG multinomial case

We devise here a simple acceptance-rejection method for sampling the latent variable U in the NGG multinomial case, whose density was denoted with $p^{(H)}(u)$. Let us focus on the limiting case $H \to \infty$, and suppose we want to simulate a random variable having density proportional to

$$p^{(\infty)}(u) \propto u^{n-1}(\kappa+u)^{-n+k\sigma}\exp\left\{-\frac{c}{\sigma}\left[(\kappa+u)^\sigma-\kappa^\sigma\right]\right\}.$$

As also discussed in Favaro & Teh (2013), rather than handling U directly it is convenient to draw samples from $Z := \log U$, whose density function is readily available after a change of variable:

$$p^{(\infty)}(z) \propto e^{vn}(\kappa+e^z)^{-n+k\sigma}\exp\left\{-\frac{c}{\sigma}\left[(\kappa+e^z)^\sigma-\kappa^\sigma\right]\right\}.$$

The distribution of Z is log-concave, that is, the logarithm of its density is concave, as one can readily verify through direct calculation of the second derivative. This is a major computational advantage and it implies, for instance, that the distribution of Z

is unimodal. Moreover, we note that everal black-box techniques were developed for sampling log-concave distributions.

We propose a simple sampling algorithm which has the advantage of being straightforward to implement, and it can be easily extended to the finite-dimensional setting. As a matter of fact, it is just an application of the well-known ratio-of-uniform method, which we recall here for convenience. Set

$$b = \sqrt{\sup\{p^{(\infty)}(z) : z \in \mathbb{R}\}},$$

and

$$b_- = -\sqrt{\sup\{z^2 p^{(\infty)}(z) : z \leqslant 0\}}, \quad b_+ = \sqrt{\sup\{z^2 p^{(\infty)}(z) : z \geqslant 0\}}.$$

Log-concavity of Z ensures that the above constants are finite. Unfortunately, there are no closed form expressions for b, b_- and b_+, but they can be readily computed via univariate numerical maximization, which is a particularly simple problem in this log-concave setting. Then, a draw from U can be obtained as follows:

Step 1. Sample independently Z_1, Z_2 uniformly on $(0, b)$ and (b_-, b_+), respectively.
Step 2. Set the candidate value $Z^* = Z_2/Z_1$.
Step 3. If $Z_1^2 \leqslant p^{(\infty)}(Z^*)$ then accept Z^* and set $Z = Z^*$, otherwise repeat the whole procedure.
Step 4. Set $U = \exp Z$.

The simulation from $p^{(H)}(u)$ proceeds in a similar manner, with the obvious modifications. A good degree of tractability is preserved because $p^{(H)}(u)$, and equivalently $p^{(H)}(z)$, is a finite mixture of densities having the kernel of $p^{(\infty)}(u)$, namely

$$\begin{aligned} p^{(H)}(u) &\propto \sum_{\ell} \left[\prod_{j=1}^{k} \left(\frac{c}{H} \right)^{\ell_j - 1} \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{\sigma^{\ell_j}} \right] u^{n-1} (\kappa + u)^{-n+|\ell|\sigma} \exp\left\{ -\frac{c}{\sigma} \left[(\kappa + u)^{\sigma} - \kappa^{\sigma} \right] \right\}, \\ &\propto u^{n-1} \exp\left\{ -\frac{c}{\sigma} \left[(\kappa + u)^{\sigma} - \kappa^{\sigma} \right] \right\} \prod_{j=1}^{k} \sum_{\ell_j=1}^{n_j} \varsigma_{n_j, \ell_j, H}(u), \end{aligned}$$

which implies that the constants b, b_- and b_+ involved in the simulation of $p^{(H)}(z)$ are finite also in this case. Moreover, as $H \to \infty$ the density $p^{(H)}(z)$ converges to $p^{(\infty)}(z)$, implying that log-concavity is recovered at the limit.

Gibbs sampling algorithm for the INVALSI application

We describe here a Gibbs sampling algorithm for posterior computation of the model described in Section 3.5. Let $G_j \in \{1, \ldots, H\}$ be an indicator function denoting to which

mixture component each school is allocated, for $j = 1, \dots, 100$. The Gibbs sampling algorithm alternates between the following full conditional steps:

Step 1. Exploiting standard results of Gaussian linear models, the full conditional for the coefficients $\boldsymbol{\beta}$ is multivariate Gaussian with

$$(\boldsymbol{\beta} \mid -) \sim \mathcal{N}\left\{(\boldsymbol{X}^\intercal \boldsymbol{X}/\sigma^2 + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}(\boldsymbol{X}^\intercal \boldsymbol{\eta}_{\boldsymbol{\beta}}/\sigma^2 + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}), (\boldsymbol{X}^\intercal \boldsymbol{X}/\sigma^2 + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}\right\},$$

where $\boldsymbol{\eta}_{\boldsymbol{\beta}}$ is a vector with entries $\eta_{ij\boldsymbol{\beta}} = S_{ij} - \mu_j$, for $i = 1, \dots, N_j$ and $j = 1, \dots, 100$, whereas $\boldsymbol{X}$ is the corresponding design matrix having row entries $\boldsymbol{x}_{ij}^\intercal$.
Step 2. The full conditional for the residual variance is

$$(\sigma^{-2} \mid -) \sim \text{GA}\left(a_\sigma + N/2, b_\sigma + \frac{1}{2}\sum_{j=1}^{100}\sum_{i=1}^{N_j}(S_{ij} - \mu_j - \boldsymbol{x}_{ij}^\intercal \boldsymbol{\beta})^2\right),$$

which can be obtained through standard calculations involved in Gaussian linear models.
Step 3. We update the cluster indicators $G_j \in \{1, \dots, H\}$ from their full conditional categorical random variables

$$\mathbb{P}(G_j = h \mid -) = \frac{\pi_h \mathcal{N}(\mu_j; \bar{\mu}_h, \bar{\sigma}_h^2)}{\sum_{h'=1}^{H} \pi_{h'} \mathcal{N}(\mu_j; \bar{\mu}_{h'}, \bar{\sigma}_{h'}^2)}, \qquad h = 1, \dots, H,$$

for any $j = 1, \dots, 100$.
Step 4. The full conditional for the school-specific parameters, given the above cluster assignments, is easily available as

$$(\mu_j \mid -) \overset{\text{ind}}{\sim} \mathcal{N}\left(\frac{\sum_{i=1}^{N_j}(S_{ij} - \boldsymbol{x}_{ij}^\intercal \boldsymbol{\beta})/\sigma^2 + \bar{\mu}_{G_j}/\bar{\sigma}_{G_j}^2}{1/\bar{\sigma}_{G_j}^2 + N_j/\sigma^2}, \frac{1}{1/\bar{\sigma}_{G_j}^2 + N_j/\sigma^2}\right),$$

independently for every $j = 1, \dots, 100$.
Step 5. The full conditional for $\bar{\mu}_h$ and $\bar{\sigma}_h^2$ are given by

$$(\bar{\mu}_h \mid -) \overset{\text{ind}}{\sim} \mathcal{N}\left(\frac{\sum_{j:G_j=h} \mu_j/\bar{\sigma}_h^2}{1/\sigma_{\bar{\mu}}^2 + 1/\bar{\sigma}_h^2 \sum_{j=1}^{100} \mathbb{1}(G_j = h)}, \frac{1}{1/\sigma_{\bar{\mu}}^2 + 1/\bar{\sigma}_h^2 \sum_{j=1}^{100} \mathbb{1}(G_j = h)}\right),$$

independently for $h = 1, \dots, H$ and

$$(\bar{\sigma}_h^{-2} \mid -) \overset{\text{ind}}{\sim} \text{GA}\left(a_{\bar{\sigma}} + \frac{1}{2}\sum_{j=1}^{100} I(G_j = h), b_{\bar{\sigma}} + \frac{1}{2}\sum_{j:G_j=h}(\mu_j - \bar{\mu}_{G_j})^2\right),$$

again independently for $h = 1, \dots, H$.

Step 6. Update the weights $(\pi_1, \dots, \pi_H)$ from their full conditional distribution by exploiting the posterior characterization of Theorem 3.6, with the necessary modifications. More precisely, the frequencies of the distinct values are given by the vector

$$\left(\sum_{j=1}^{100} \mathbb{1}(G_j = 1), \dots, \sum_{j=1}^{100} \mathbb{1}(G_j = H) \right).$$

Chapter 4

Functional clustering via finite-dimensional enriched priors

4.1 Summary

The chapter is organized as follows. In Section 4.2 we introduced the enriched Dirichlet multinomial mixture model for functional data, while in Section 4.3 we discuss some theoretical properties. Specifically, we investigate the underlying clustering mechanism, we present a novel enriched Pólya-urn scheme and we prove the convergence of our proposal to some well-defined infinite-dimensional process. In Section 4.4 a variational Bayes algorithm for posterior inference is developed and it is tested on a simulation study in Section 4.5. In Section 4.6 we apply the proposed method to a real dataset from e-commerce, as outlined in Section 1.3.3, and we discuss the empirical findings.

4.2 A Bayesian functional mixture model

In the additive representation (1.12) we consider standardized functional observations. That is, the empirical mean of $Y_i(t)$ evaluated on the time grid $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{iT_i})^\intercal$ for $i = 1, \ldots, n$, equals zero, whereas the empirical variance equals one. In fact, in this specific application we are interested in grouping functions with similar shapes and not in capturing their average levels. Then, for each standardized route and time value $t \in \mathbb{R}_+$, we let

$$Y_i(t) = f_i(t) + \epsilon_i(t), \qquad i = 1, \ldots, n,$$

where each $f_i : \mathbb{R}_+ \to \mathbb{R}$ is an unknown function to be estimated, and where $\epsilon_i(t)$ is a Gaussian local error measurement with zero mean and variance σ^2, in turn having a conditionally conjugate gamma prior distribution $\sigma^{-2} \sim \text{GA}(a_\sigma, b_\sigma)$. Consistent with the discussion of Section 1.3.3, we employ a discrete prior law $\tilde{p}$ to borrow information

across the latent trajectories f_i and to induce functional clustering, namely we assume

$$(f_i \mid \tilde{p}^{(H)}) \overset{\text{iid}}{\sim} \tilde{p}^{(H)}, \qquad \tilde{p}^{(H)} = \sum_{h=1}^{H} \xi_h \delta_{\tilde{\phi}_h}, \tag{4.1}$$

independently for $i = 1, \dots, n$, with δ_x denoting the point mass function at x. The collections of weights $\xi_1, \dots, \xi_H$ are random probabilities such that $\sum_{h=1}^{H} \xi_h = 1$ almost surely, whereas each atom $\tilde{\phi}_h$ is the realization of a random function. Hence, each f_i can be formally regarded as a random function defined on a suitable complete and separable metric space $\mathbb{F}$ endowed with its Borel σ-algebra $\mathscr{F}$. From representation (4.1) it is apparent that a *discrete* prior induces ties among the functions f_i. We will say that two different functional observations $Y_i(t)$ and $Y_j(t)$ belong to same group whenever they possess the same functional atom $\tilde{\phi}_h$, i.e. when they share the same latent trajectory $f_i = f_j$. Clearly, the choice of the prior law for $\tilde{p}^{(H)}$ has a strong impact on the clustering procedure. A popular class of models, arising in the infinite case $H \to \infty$, is given by stick-breaking priors (Ishwaran & James, 2001), of which the functional Dirichlet process (FDP) is a special case. However, as discussed in the Introduction, such a choice might be unsuitable for our goals, and we rather want to upper-bound the model complexity by selecting a finite value for H. Furthermore, we aim at adapting (4.1) to incorporate prior information about functional shapes.

Suppose it is known that each f_i possesses specific shapes or features. For example, we may know in advance that a subset of the functional observations f_i is monotone, cyclical or it is bounded by some constant. In our application, for instance, we know that a subset of routes presents a strong cyclical pattern. More formally, we assume that each function f_i belongs to a functional class among a finite collection $\{\mathbb{F}_1, \dots, \mathbb{F}_L\}$ of L specifications, with each $\mathbb{F}_l \in \mathscr{F}$ being a measurable subset of $\mathbb{F}$. These functional classes have to be specified in consultation with subject matter experts or as a consequence of exploratory analyses. Splines are particularly convenient in accommodating a variety of constraints such as monotonicity (Ramsay, 1988), but there are endless modeling possibilities. For example, Gaussian processes are a flexible and widely used prior for functional modeling (e.g. Petrone et al., 2009), and one may select for each class a different covariance function. A computationally convenient class of functions which includes the aforementioned examples is discussed in Section 4.2.1.

Let P_l for $l = 1, \dots, L$ be a collection of *diffuse* and fixed probability measures defined over the space $(\mathbb{F}, \mathscr{F})$ and placing mass only on the corresponding class space $\mathbb{F}_l$, so that $P_l(\mathbb{F}_l) = 1$. The diffuseness assumption amount to have $P_l(\{f\}) = 0$ for any $f \in \mathbb{F}$. Then,

our enriched formulation specializes the general model (4.1) as follow

$$\begin{aligned} \tilde{p}^{(H)} &= \sum_{l=1}^{L} \Upsilon_l \sum_{h=1}^{H_l} \pi_{lh} \delta_{\tilde{\theta}_{lh}}, \\ \tilde{\theta}_{lh} &\overset{\text{ind}}{\sim} P_l, \qquad h = 1, \dots, H_l, \qquad l = 1, \dots, L. \end{aligned} \tag{4.2}$$

Such a construction can be readily interpreted as a *mixture of mixtures*. Differently from common mixture models, the atoms $\tilde{\theta}_{lh}$ are independent and identically distributed (iid) within the feature class, but only independent across them. Exploiting standard hierarchical representation for mixture models, let us introduce a set of latent cluster indicators $\boldsymbol{G} = (G_1, \dots, G_n)$ whose values are the pairs (l, h) for any $h = 1, \dots, H_l$ and $l = 1, \dots, L$, so that each function f_i is associated to the corresponding atom $\tilde{\theta}_{G_i}$. Therefore, two functional observations f_i and f_j belong to the same cluster if and only if $G_i = G_j$. Moreover, let us define an additional set of latent indicators $F_i \in \{1, \dots, L\}$, for $i = 1, \dots, n$, representing the membership of each f_i to the corresponding functional class. Then, the mixing probabilities in (4.2) have a simple and useful interpretation, which is outlined in the following scheme:

$$\begin{aligned} \textit{Functional class allocation:} \quad & \mathbb{P}(F_i = l) = \Upsilon_l, \\ \textit{Within-class allocation:} \quad & \mathbb{P}(G_i = (l, h) \mid F_i = l) = \pi_{lh}, \quad h = 1, \dots, H_l, \\ \textit{Cluster allocation:} \quad & \mathbb{P}(G_i = (l, h)) = \Upsilon_l \pi_{lh}, \quad h = 1, \dots, H_l, \end{aligned}$$

for any $l = 1, \dots, L$ and unit $i = 1, \dots, n$. To summarize, each membership indicator G_i might be obtained as the result of a two-step procedure. In the first step, the functional class indicator F_i associated to the ith unit is sampled according to the probabilities $\boldsymbol{\Upsilon} = (\Upsilon_1, \dots, \Upsilon_L)$. Then, conditionally on $F_i = l$, each cluster membership G_i is drawn according to the within-class probabilities $\boldsymbol{\pi}_l = (\pi_{l1}, \dots, \pi_{lH_l})$. To allow uncertainty in such probabilities, we let

$$(\Upsilon_1, \dots, \Upsilon_{L-1}) \sim \textsc{dirichlet}(\alpha_1, \dots, \alpha_L), \tag{4.3}$$

whereas for the within-class step, independently on (4.3), we let

$$(\pi_{l1}, \dots, \pi_{lH_l - 1}) \overset{\text{ind}}{\sim} \textsc{dirichlet}\left(\frac{c_l}{H_l}, \dots, \frac{c_l}{H_l}\right), \qquad l = 1, \dots, L. \tag{4.4}$$

The Dirichlet distribution in equation (4.4) is symmetric because the atoms $\tilde{\theta}_{\ell h}$ are iid within the functional class. Altogether, equations (4.2)-(4.4) describe what we will term an enriched functional Dirichlet multinomial process (E-FDMP).

Such a nested clustering mechanism characterizes general enriched priors, like the E-FDP and other enriched stick-breaking priors (Scarpa & Dunson, 2014). As we will show in Section 4.3, there is a sharp connection between the E-FDP and our E-FDMP, since the former can be recovered as limiting case of the latter. Beside constituting a more flexible class compared to classical mixtures, enriched processes allow the estimation of "groups of clusters", which are identified by the functional class indicators F_i. Indeed, we might want to group the routes characterized by cyclical patterns or increasing trends, irrespectively of their within-class allocation. Moreover, even when the G_i indicators are of interests, it might be useful to split the clustering solution into homogeneous classes, e.g. to facilitate their presentation to the stakeholders. These are major interpretative advantages of enriched priors which do not have a direct equivalent in classical mixture models.

4.2.1 Baseline measures specification

The specification of the baseline measures P_l has clearly a crucial impact on inference. A priori, each P_l can be interpreted as a "functional prior guess", because the expected value of $\tilde{p}^{(H)}$ is a mixture of the baseline measures $P_1, \dots, P_L$. Indeed, for any $A \in \mathscr{F}$

$$\mathbb{E}\{\tilde{p}^{(H)}(A)\} = \sum_{l=1}^{L} \mathbb{E}(\Upsilon_l) P_l(A) = \frac{1}{\alpha} \sum_{l=1}^{L} \alpha_l P_l(A), \qquad \alpha = \sum_{l=1}^{L} \alpha_l.$$

The role of the hyperparameters $\alpha_1/\alpha, \dots, \alpha_L/\alpha$ is hence clear, being the prior proportions of each mixture component. For the remaining of the chapter, we will focus on a broad subclass of baseline probability measures which are characterized by a significantly improved computational and analytical tractability. More precisely, we assume that $\tilde{\theta}_{lh}(t)$ is linear in the parameters, with a Gaussian prior on the regression coefficients, namely

$$\tilde{\theta}_{lh}(t) = \sum_{m=1}^{M_l} \mathcal{B}_{ml}(t) \tilde{\beta}_{mlh}, \qquad \tilde{\boldsymbol{\beta}}_{lh} = (\tilde{\beta}_{1lh}, \dots, \tilde{\beta}_{M_l lh})^\intercal \overset{\text{ind}}{\sim} \mathcal{N}_{M_l}(\boldsymbol{\mu}_{\beta_l}, \boldsymbol{\Sigma}_{\beta_l}), \tag{4.5}$$

where each $\mathcal{B}_{1l}(t), \dots, \mathcal{B}_{M_l l}(t)$ for $l = 1, \dots, L$ is a set of pre-specified basis functions and where $\tilde{\boldsymbol{\beta}}_{lh} \in \mathbb{R}^{M_l}$ is an unknown vector of regression coefficients having multivariate Gaussian prior with mean $\boldsymbol{\mu}_{\beta_l} = (\mu_{1l}, \dots, \mu_{M_l l})^\intercal$ and covariance matrix $\boldsymbol{\Sigma}_{\beta_l}$. Under such a choice, the a priori expected value of each function $f_i(t)$ for $i = 1, \dots, n$ and $t \in \mathbb{R}_+$ simplifies

$$\mathbb{E}\{f_i(t)\} = \sum_{l=1}^{L} \frac{\alpha_l}{\alpha} \sum_{m=1}^{M_l} \mathcal{B}_{ml}(t) \mu_{ml},$$

thus being a weighted average of the expected values of the regression coefficients. Note that Bayesian penalized splines (Lang & Brezger, 2004) fall within specification (4.5).

We shall remark that if inference on the functional classes $F_1, \dots, F_n$ is of interest, the measures $P_1, \dots, P_L$ must be distinguishable a priori, in the sense that they should characterize to quite different functional shapes. Otherwise, it might be difficult to infer the functional classes from the data. Indeed, while very flexible specifications might be employed for each P_l, these choices would lead to identifiability issues across functional classes. However, this is not a concern if one is interested in the cluster memberships.

4.3 Random partitions and clustering

In this section we investigate the a priori random partition mechanism of the E-FDMP model. Our proposal can be viewed as a middle ground between finite and infinite mixture models. Indeed, it is closely related to proper nonparametric priors while being finite dimensional. These features have several important implications for clustering.

A key property of the E-FDMP model is that the number of clusters is bounded by $H = \sum_{l=1}^{L} H_l$. However, this does not imply that the actual number of clusters is equal to H, because some partitions might be empty. Indeed, to circumvent the issue of selecting the number of mixture components, one might consider a mixture model with a large H and employ a sparse prior, thus effectively deleting the redundant mixture weights. Such an approach has been advocated by Malsiner-Walli et al. (2016), on the ground of the asymptotic results of Rousseau & Mengersen (2011). The amount of shrinkage towards the upper bound H or towards the single cluster solution is regulated by the sparse prior (4.4). Hence, the E-FDMP should not be regarded as a classical finite mixture model, because the number of clusters is inferred from the data and it should not be specified in advance.

We begin our discussion by first pointing out relevant connections of our proposal with both the E-FDP and the FDP processes, and by providing some first intuitions about the role of each H_l. Consider the probability that two functions are assigned to the same cluster. More precisely, let f_i and f_j be two draws from a E-FDMP with $i \neq j$, then it is easy to check that a priori

$$\mathbb{P}(f_i = f_j) = \sum_{l=1}^{L} \frac{\alpha_l(\alpha_l + 1)}{\alpha(\alpha + 1)} \frac{c_l + H_l}{c_l H_l + H_l}. \tag{4.6}$$

The a priori probability of co-clustering of equation (4.6) is decreasing over H_l, i.e. the within-class upper bounds, and increasing over c_l, the within-class total mass parameter. Importantly, as each $H_l \to \infty$ for $l = 1, \dots, L$, the probability of co-clustering converges

to a strictly positive constant

$$\lim_{H_l \to \infty} \mathbb{P}(f_i = f_j) = \sum_{l=1}^{L} \frac{\alpha_l(\alpha_l + 1)}{\alpha(\alpha+1)} \frac{1}{1+c_l},$$

which unsurprisingly coincides with the co-clustering probability of the E-FDP, given in Scarpa & Dunson (2014). Indeed, one can show that a E-FDMP (weakly) converges to a E-FDP as each $H_l \to \infty$. This convergence result has relevant practical implications: broadly speaking, it means that if we augment the model complexity indefinitely by increasing H_l, we nonetheless obtain a well-defined model, whose probability of co-clustering does not goes to zero. However, this is not to say that we should choose H_l as large as possible, because this might lead to uninterpretable clustering solutions. Rather, the bounds H_l should be selected as the largest value maintaining the model sufficiently tractable.

We now provide a formal statement of the aforementioned convergence result, which rely on the notion of weak convergence for random measures; we refer to Kallenberg (Chap. 4, 2017) for a rigorous treatment. Let $\tilde{q}^{(\infty)} \sim \text{DP}(cP)$ denote a Dirichlet process having total mass parameter c and baseline probability distribution P (Ferguson, 1973).

Theorem 4.1. *Let $\tilde{p}^{(H)}$ be a* E-FDMP *defined by equations (4.2)-(4.4) and let $\tilde{p}^{(\infty)}$ be a* E-FDP *(Scarpa & Dunson, 2014), which is defined as*

$$\tilde{p}^{(\infty)} = \sum_{l=1}^{L} \Upsilon_l \tilde{q}_l^{(\infty)}, \qquad \tilde{q}_l^{(\infty)} \overset{\text{ind}}{\sim} \text{DP}(c_l P_l),$$

where the probabilities $(\Upsilon_1, \dots, \Upsilon_L)$ are distributed as in (4.3). Then,

$$\tilde{p}^{(H)} \xrightarrow{\text{wd}} \tilde{p}^{(\infty)}, \quad \textit{as} \quad H_l \to \infty, \quad l = 1, \dots, L,$$

where $\xrightarrow{\text{wd}}$ denotes weak convergence.

Proof. Note that we can write $\tilde{p}^{(H)} = \sum_{l=1}^{L} \Upsilon_l \tilde{p}_l^{(H_l)}$, where each $\tilde{p}_l^{(H_l)}$ follows a Dirichlet multinomial process. It is well known that $\tilde{p}_l^{(H_l)}$ weakly converges to a Dirichlet process $q_l^{(\infty)}$ (e.g. Ishwaran & Zarepour, 2000) as $H_l \to \infty$, implying that for any finite collection of sets $A_1, \dots, A_d \in \mathscr{F}$

$$\{\tilde{p}^{(H)}(A_1), \dots, \tilde{p}^{(H)}(A_d)\} \xrightarrow{\text{d}} \{\tilde{p}^{(\infty)}(A_1), \dots, \tilde{p}^{(\infty)}(A_d)\}.$$

Weak convergence of the process follows from Theorem 4.11 in Kallenberg (2017). □

Theorem 4.1 is important also on the light of the following connection between the E-FDP and the FDP which, to the best of our knowledge, was not made explicit elsewhere.

If $L = 1$, then the E-FDP trivially reduces to a FDP. However, this occurs also under specific hyperparameter settings. Indeed, the next corollary implies that if $\alpha_l = c_l$ for $l = 1, \dots, L$, then the limiting process $\tilde{p}^{(\infty)}$ will be distributed according to a Dirichlet process whose baseline probability measure is a mixture of the class-specific measures $P_1, \dots, P_L$. Such a result is stated as a corollary of Theorem 4.1 for the sake of the exposition, but it is actually a property of the E-FDP; see the proof for details.

Corollary 4.1. *Suppose additionally to Theorem 4.1 that* $\alpha_l = c_l$ *for any* $l = 1, \dots, L$. *Then* $\tilde{p}^{(H)} \xrightarrow{wd} \tilde{p}^{(\infty)}$ *as each* $H_l \to \infty$ *and moreover*

$$\tilde{p}^{(\infty)} \sim \text{DP}\left(\sum_{l=1}^{L} \alpha_l P_l\right).$$

Proof. The proof rely on the finite-dimensional characterization of the Dirichlet process (Ferguson, 1973). Specifically, for any finite partition $B_1, \dots, B_d \in \mathscr{F}$ we have

$$\{\tilde{q}_l^{(\infty)}(B_1), \dots, \tilde{q}_l^{(\infty)}(B_d)\} \overset{\text{ind}}{\sim} \text{DIRICHLET}\{\alpha_l P_l(B_1), \dots, \alpha_l P_l(B_d)\}, \qquad l = 1, \dots, L.$$

Note in particular that $\{\tilde{p}^{(\infty)}(B_1), \dots, \tilde{p}^{(\infty)}(B_d)\} = \sum_{l=1}^{L} \Upsilon_l \{\tilde{q}_l^{(\infty)}(B_1), \dots, \tilde{q}_l^{(\infty)}(B_d)\}$, and

$$\{\tilde{p}^{(\infty)}(B_1), \dots, \tilde{p}^{(\infty)}(B_d)\} \sim \text{DIRICHLET}\left\{\sum_{l=1}^{L} \alpha_l P_l(B_1), \dots, \sum_{l=1}^{L} \alpha_l P_l(B_d)\right\},$$

thanks to well-know properties of the Dirichlet distribution. □

4.3.1 Enriched Pólya urn scheme

Similar to Blackwell & MacQueen (1973) in the Dirichlet process case, our E-FDMP is characterized by a Pólya urn scheme, whose description greatly facilitates the understanding of the underlying clustering mechanism. Conditionally on the latent class indicators $F_1, \dots, F_n$, our enriched formulation reduces to a collection of Dirichlet multinomial processes. Recalling equation (4.2), we can rewrite the E-FDMP as follows

$$\tilde{p}^{(H)} = \sum_{l=1}^{L} \Upsilon_l \tilde{p}_l^{(H_l)}, \qquad \tilde{p}_l^{(H_l)} = \sum_{h=1}^{H_l} \pi_{lh} \delta_{\tilde{\theta}_{lh}}.$$

Then, we can augment the above specification by including the set of latent class indicators $\boldsymbol{F} = (F_1, \dots, F_n)$. In this hierarchical representation, the functions belonging the same class $f_i : i \in \mathbb{I}_l$ with $\mathbb{I}_l = \{i = 1, \dots, n : F_i = l\}$ are iid draws from $\tilde{p}_l^{(H_l)}$, a Dirichlet multinomial process. More precisely, we can equivalently represent our E-FDMP

hierarchically as

$$
\begin{aligned}
(F_i \mid \boldsymbol{\gamma}) &\overset{\text{iid}}{\sim} \textsc{multinom}(\gamma_1, \dots, \gamma_L), && i = 1, \dots, n, \\
(f_i \mid F_i = l, \tilde{p}_l^{(H_l)}) &\overset{\text{iid}}{\sim} \tilde{p}_l^{(H_l)}, && i \in \mathbb{I}_l
\end{aligned}
$$

with prior distributions as in equations (4.3)-(4.4). Such a hierarchical representation naturally leads to the definition of a sequential mechanism for generating both $f_1, \dots, f_n$ and $F_1, \dots, F_n$. Let $n_l = \sum_{i=1}^{n} \mathbb{1}(F_i = l)$ be the number of elements belonging to the lth functional class and let $k_l \leqslant n_l$ be the number of distinct values observed among the functions of the lth class. Moreover, let $f^*_{11}, \dots, f^*_{1n_1}, \dots, f^*_{L1}, \dots, f^*_{Ln_L}$ represent the distinct values observed in the whole sample $\boldsymbol{f} = (f_1, \dots, f_n)$, having frequencies n_{jl} for $j = 1, \dots, k_l$ and $l = 1, \dots, L$, so that $n_l = \sum_{j=1}^{k_l} n_{jl}$ and $n = \sum_{l=1}^{L} n_l$. Then, the enriched Pólya urn scheme is characterized by the following two steps, so that for any $n \geqslant 1$ and any $A \in \mathscr{F}$ we have

$$
\begin{aligned}
\mathbb{P}(F_{n+1} = l \mid \boldsymbol{F}) &= \frac{\alpha_l + n_l}{\alpha + n}, \qquad l = 1, \dots, L, \\
\mathbb{P}(f_{n+1} \in A \mid \boldsymbol{f}, \boldsymbol{F}, F_{n+1} = l) &= \left(1 - \frac{k_l}{H_l}\right) \frac{c_l}{c_l + n_l} P_l(A) + \sum_{j=1}^{k_l} \frac{n_{jl} + c_l/H_l}{c_l + n_l} \delta_{f^*_{jl}}(A).
\end{aligned}
$$

At the first step, one draws the F_{n+1} functional class indicator with a probability depending on the observed frequencies $n_1, \dots, n_L$ and the $\alpha_1, \dots, \alpha_L$ coefficients, which can be naturally interpreted as a priori frequencies. Then, at the second step and given $F_{n+1} = l$, one either draw a novel functional observation from P_l or she samples one of the previously observed functions with probability proportional to $n_{jl} + c_l/H_l$.

On the light of Theorem 4.1, it is not surprising that the second step converges to the classical scheme of Blackwell & MacQueen (1973) as $H_l \to \infty$, conditionally on the lth functional class. Moreover, if $\alpha_l = c_l$ the classical Pólya urn scheme is recovered also marginally, a consequence of Corollary 4.1. Furthermore, such an enriched Pólya urn scheme is reminiscent of the one presented in Wade et al. (2011), and indeed it can be essentially regarded as its finite-dimensional counterpart.

Let us focus on the conditional probability of obtaining a new cluster, given the functions $\boldsymbol{f}$ and the class indicators $\boldsymbol{F}$. From the enriched Pólya urn scheme one can easily get

$$
\mathbb{P}(f_{n+1} = \text{“new"} \mid \boldsymbol{f}, \boldsymbol{F}) = \sum_{l=1}^{L} \frac{\alpha_l + n_l}{\alpha + n} \left(1 - \frac{k_l}{H_l}\right) \frac{c_l}{c_l + n_l}. \tag{4.7}
$$

The above predictive probability provides a clear guidance about the role of the hyperparameters. In first place, note that the probability of drawing a new function

decreases the more clusters k_l we observe, and it equals zero whenever $k_l = H_l$. Hence, the E-FDMP penalizes partitions with a large number of clusters, effectively bounding the model complexity, one of the overarching goals of our analysis. Note that as $H_l \to \infty$ the aforementioned penalization disappears. Moreover, the parameters c_l control the creation of a new cluster—the larger each c_l the more cluster we should expect.

4.4 Posterior computations

Bayesian mixture models are routinely estimated using Markov chain Monte Carlo (MCMC). While this approach is supported by strong theoretical guarantees, it has some drawbacks when performing clustering. The first concern is scalability: MCMC sampling might face computational bottlenecks when the sample size grows. This is a severe limitation because in practice one would like to conduct the clustering algorithm on a weekly basis, and perhaps on several different datasets. In addition, a further difficulty arises when performing clustering with MCMC. As discussed in Lau & Green (2007), at each step of the chain one samples a different partition of the observations; however, it is hard to provide a point estimate, essentially because of the label switching phenomenon. Existing solutions rely either on ad-hoc procedures (Medvedovic & Sivaganesan, 2002), or on post-process optimizations problems (Lau & Green, 2007; Fritsch & Ickstadt, 2009; Wade & Ghahramani, 2018). In both cases, this implies an additional layer of difficulty that one might want to avoid.

To address these issues we employ a mean-field variational approximation of the posterior distribution, which is nowadays a well-established inferential tool (Blei et al., 2017). The involved computations are much faster than MCMC, and the variational Bayes (VB) approach is particularly well suited for clustering purposes, since it is not affected by label switching, thus ruling out the aforementioned additional steps. In addition, variational inference for the E-FDMP is straightforward to implement because such a model belongs to the conditionally conjugate exponential family, for which efficient optimization algorithms are available (Blei et al., 2017).

Unfortunately, these advantages do not come without some drawbacks: indeed, the variational posterior is only an approximation of the proper posterior law, and it is well known that VB generally leads to accurate point estimates but also it typically underestimate the variability. If uncertainty quantification were of interest, a Gibbs sampling algorithm for the E-FDMP could be easily devised, since the full conditional distributions are be available in closed form. However, in our motivating application we are only interested in a single cluster solution and therefore VB represents an appealing choice.

Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_L)$ be the collection of the within-class probabilities of equation (4.4) and let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_{11}, \dots, \tilde{\boldsymbol{\beta}}_{1H_1}, \dots, \tilde{\boldsymbol{\beta}}_{L1}, \dots, \tilde{\boldsymbol{\beta}}_{LH_L})$ be the set of regression coefficients appearing in equation (4.5). We seek an optimal variational distribution $q^{(*)}(\boldsymbol{G}, \boldsymbol{\Upsilon}, \boldsymbol{\pi}, \tilde{\boldsymbol{\beta}}, \sigma^2)$ that best approximates the joint posterior, while maintaining simple computations. This can be obtained by minimizing the Kullback-Leibler divergence between the variational distribution and the full posterior, or equivalently by maximizing the so-called evidence lower bound (ELBO), so that $q^{(*)}(\boldsymbol{G}, \boldsymbol{\Upsilon}, \boldsymbol{\pi}, \tilde{\boldsymbol{\beta}}, \sigma^2) = \arg\max_{q \in \mathcal{Q}} \text{ELBO}\{q(\boldsymbol{G}, \boldsymbol{\Upsilon}, \boldsymbol{\pi}, \tilde{\boldsymbol{\beta}}, \sigma^2)\}$; see Blei et al. (2017) and the discussion in Chapter 7.

Without further restrictions, the Kullback-Leibler divergence is minimized when the variational distribution is equal to the true posterior distribution, which is analytically intractable. Hence, a common strategy is to assume that the variational distribution belongs to a mean-field family $\mathcal{Q}$. Such a class of distributions incorporate a posteriori independence among distinct groups of parameters, meaning that the variational distribution factorizes as

$$q^{(*)}(\boldsymbol{G}, \boldsymbol{\Upsilon}, \boldsymbol{\pi}, \tilde{\boldsymbol{\beta}}, \sigma^2) = q^{(*)}(\sigma^2) \prod_{i=1}^{n} q^{(*)}(G_i) q^{(*)}(\boldsymbol{\Upsilon}) \prod_{l=1}^{L} q^{(*)}(\boldsymbol{\pi}_l) \prod_{l=1}^{L} \prod_{h=1}^{H_l} q^{(*)}(\tilde{\boldsymbol{\beta}}_{lh}).$$

Under such an assumption, the optimal variational distributions can be found exploiting an iterative algorithm called coordinate ascent variational inference (CAVI). Its full derivation entails standard calculations which are omitted for the sake of the exposition; we report in Algorithm 1 only the resulting CAVI algorithm. One may refer to Bishop (Chap. 10, 2006) for detailed illustrations on similar models.

We define here some additional notation necessary for the description of the CAVI Algorithm 1. As mentioned in Section 4.2, recall that each functional observation $Y_i(t)$ is only available on a finite grid of points $\boldsymbol{t}_i = (t_{i1}, \dots, t_{iT_i})^\intercal$. The observed values associated to these time grids are stacked into a single $\sum_{i=1}^{n} T_i$-dimensional vector

$$\boldsymbol{Y} = (Y_1(t_{11}), \dots, Y_1(t_{1T_1}), \dots, Y_n(t_{n1}), \dots, Y_n(t_{nT_n}))^\intercal.$$

Similarly, we define the $\sum_{i=1}^{n} T_i \times M_l$ matrices $\boldsymbol{B}_l$ for $l = 1, \dots, L$, which are paired to the data $\boldsymbol{Y}$ and whose entries are the values of the basis functions $\mathcal{B}_m(t_{is})$ of equation (4.5), for $m = 1, \dots, M_l$ over the columns and for $s = 1, \dots, T_i$ and $i = 1, \dots, n$ over the rows. Moreover, note that in Algorithm 1 the density functions are identified by the same symbols that are used to characterize distributions. Finally, the expected values appearing in Algorithm 1 are taken with respect to the variational distributions $q^{(r)}(\cdot)$ at the rth step of the cycle, motivating the notation $\mathbb{E}_{q^{(r)}}$. The CAVI algorithm, at convergence, returns the optimal variational distribution $q^{(*)}(\cdot)$.

From the output of the CAVI algorithm, it is straightforward to derive a posteriori variational estimates for the cluster memberships $G_1, \ldots, G_n$, for the class-specific membership $F_1, \ldots, F_n$, and for the cluster-specific trajectories $\tilde{\theta}_{lh}$. If $\varrho^{(*)}_{i\ell h}$ denote the variational probabilities computed at ***Step 1*** of Algorithm 1, at convergence, then a natural variational Bayes estimate $G^{(*)}_1, \ldots, G^{(*)}_n$ for the cluster memberships is given by

$$G^{(*)}_i = \arg\max_{l,h} \varrho^{(*)}_{ilh} = \arg\max_{l,h} q^{(*)}\{G_i = (l,h)\}, \qquad i = 1, \ldots, n,$$

and similarly a variational estimate $F^{(*)}_1, \ldots, F^{(*)}_n$ for the functional classes is

$$F^{(*)}_i = \arg\max_{l} \sum_{h=1}^{H_l} \varrho^{(*)}_{ilh} = \arg\max_{l} q^{(*)}(F_i = l), \qquad i = 1, \ldots, n.$$

These natural estimators can not be easily computed when performing MCMC because of the label-switching phenomenon. Finally, an estimate $\tilde{\theta}^{(*)}_{lh}(t)$ for the cluster-specific functions is given by its variational expectation, which equals

$$\tilde{\theta}^{(*)}_{lh}(t) = \mathbb{E}_{q^{(*)}}\{\tilde{\theta}_{lh}(t)\} = \sum_{m=1}^{M_l} \mathcal{B}_{ml}(t)\mathbb{E}_{q^{(*)}}(\tilde{\beta}_{mlh}).$$

The estimate $\tilde{\theta}^{(*)}_{lh}(t)$ will be useful for the interpretation of the clusters.

4.5 Simulated illustration

In this section we assess the empirical performance of the E-FDMP—and the associated CAVI algorithm—by conducting a simple simulation study. Such a simulation is far from being extensive and it serves mainly as an illustration of the concepts presented in Section 4.3. Specifically, we aim at showing the ability of our model to effectively recover the true number of groups, as well as the cluster memberships, thereby empirically validating the role of each parameter H_l as the upper bound for the total number of clusters.

For this illustrative example, we consider identical and equally spaced time grids $\boldsymbol{t}_i = (1/T_i, \ldots, T_i/T_i)^\intercal$ for $i = 1, \ldots, n$, ranging over the unit interval $[0, 1]$, and we let the number of observations $n = 100$ and each grid length $T_1 = \cdots = T_n = 50$. Among the functions $f_1, \ldots, f_n$ there are only four distinct values $f^*_1, \ldots, f^*_4$, defined as

$$f^*_1(t) = 1 - 2t, \qquad f^*_2(t) = \frac{1}{2}\{\cos(2\pi t) + \sin(2\pi t)\},$$
$$f^*_3(t) = 2t^4 - 1, \qquad f^*_4(t) = \frac{1}{2}\{\cos(4\pi t) + \sin(4\pi t)\}.$$

Algorithm 1: CAVI algorithm for the E-FDMP with baseline measures (4.5)

begin

Let $q^{(r)}(\cdot)$ denote the generic variational distribution at iteration r and let $\mathbb{E}_{q^{(r)}}$ denote the expected value taken with respect to it. At every step of the algorithm, update each block of $q^{(r)}(\cdot)$ according to the following steps:

Step 1. Update $q^{(r)}(G_i)$ for each $i = 1, \dots, n$;

for i *from* 1 *to* n **do**

Update the variational probabilities $q^{(r)}\{G_i = (l,h)\} = \varrho^{(r)}_{ilh}$ according to

$$\begin{aligned}\varrho^{(r)}_{ilh} &\propto \exp\left[\mathbb{E}_{q^{(r-1)}}\{\log(\Upsilon_l \pi_{lh})\} + \sum_{s=1}^{T_i} \mathbb{E}_{q^{(r-1)}}\{\log \mathcal{N}(Y_i(t_{is}); \tilde{\theta}_{lh}(t_{is}), \sigma^2)\}\right], \\ &\propto \exp\left(\mathbb{E}_{q^{(r-1)}}\{\log(\Upsilon_l \pi_{lh})\} - \frac{1}{2}\mathbb{E}_{q^{(r-1)}}(\sigma^{-2}) \sum_{s=1}^{T_i} \mathbb{E}_{q^{(r-1)}}\left[\{Y_i(t_{is}) - \tilde{\theta}_{lh}(t_{is})\}^2\right]\right),\end{aligned}$$

for any $h = 1, \dots, H_l$ and $l = 1, \dots, L$.

Step 2. Update the variational distribution $q(\boldsymbol{\Upsilon})$ according to

$$q^{(r)}(\boldsymbol{\Upsilon}) = \text{DIRICHLET}\left(\boldsymbol{\Upsilon}; \alpha_1 + \sum_{i=1}^{n}\sum_{h=1}^{H_1} \varrho^{(r)}_{i1h}, \dots, \alpha_L + \sum_{i=1}^{n}\sum_{h=1}^{H_L} \varrho^{(r)}_{iLh}\right).$$

Step 3. Update $q^{(r)}(\boldsymbol{\pi}_l)$ for each $l = 1, \dots, L$;

for l *from* 1 *to* L **do**

Update the variational distribution of each $q^{(r)}(\boldsymbol{\pi}_l)$ according to

$$q^{(r)}(\boldsymbol{\pi}_l) = \text{DIRICHLET}\left(\boldsymbol{\pi}_l; \frac{c_l}{H_l} + \sum_{i=1}^{n} \varrho^{(r)}_{il1}, \dots, \frac{c_l}{H_l} + \sum_{i=1}^{n} \varrho^{(r)}_{ilH_l}\right).$$

Step 4. Update $q^{(r)}(\tilde{\boldsymbol{\beta}}_{lh})$ for each $h = 1, \dots, H_l$ and $l = 1, \dots, L$;

for l *from* 1 *to* L **do**

for h *from* 1 *to* H_l **do**

Update the variational distribution of each $q^{(r)}(\tilde{\boldsymbol{\beta}}_{lh})$ according to

$$q^{(r)}(\tilde{\boldsymbol{\beta}}_{lh}) = \mathcal{N}_{M_l}\left(\tilde{\boldsymbol{\beta}}_{lh}; \boldsymbol{\mu}^{(r)}_{lh}, \boldsymbol{\Sigma}^{(r)}_{lh}\right),$$

where $\boldsymbol{\Sigma}^{(r)}_{lh} = (\boldsymbol{B}_l^\intercal \boldsymbol{\Gamma}^{(r)}_{lh} \boldsymbol{B}_l + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}_l})^{-1}$ and $\boldsymbol{\mu}^{(r)}_{lh} = \boldsymbol{\Sigma}^{(r)}_{lh}(\boldsymbol{B}_l \boldsymbol{\Gamma}^{(r)}_{lh} \boldsymbol{Y} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_l} \boldsymbol{\mu}_{\boldsymbol{\beta}_l})$, and with $\boldsymbol{\Gamma}^{(r)}_{lh} = \mathbb{E}_{q^{(r-1)}}(\sigma^{-2}) \text{diag}(\varrho^{(r)}_{1lh}, \dots, \varrho^{(r)}_{1lh}, \dots, \varrho^{(r)}_{nlh}, \dots, \varrho^{(r)}_{nlh})$.

Step 5. Let $\bar{T} = 1/n \sum_{i=1}^{n} T_i$. Update the variational distribution $q^{(r)}(\sigma^{-2})$ according to

$$q^{(r)}(\sigma^{-2}) = \text{GA}\left(\sigma^{-2}; a_\sigma + \frac{n\bar{T}}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}\sum_{s=1}^{T_i}\sum_{l=1}^{L}\sum_{h=1}^{H_l} \varrho^{(r)}_{ilh} \mathbb{E}_{q^{(r)}}[\{Y_i(t_{is}) - \tilde{\theta}_{lh}(t_{is})\}^2]\right).$$

The first $f_1, \dots, f_{25}$ functions are set equal to f_1^*, while each element of the second block $f_{26}, \dots, f_{50}$ is set equal f_2^*, and similarly for the third and fourth blocks of functions $f_{51}, \dots, f_{75}$ and $f_{76}, \dots, f_{100}$, whose elements are equal to f_3^* and f_4^*, respectively. Summarizing, we let the number of cluster be equal to 4 and we assume that each partition has 25 elements, for a total of $n = 100$ functional observations. Recall that we observe error prone realizations $Y_i(t)$ of these functions, for $i = 1, \dots, n$, as for equation (1.12). Clearly, the clustering performance is affected by the amount of noise in the observed data. To emphasize this aspect we consider two different scenarios. In the first simulated setting, the variance of the error is relatively small ($\sigma^2 = 0.1^2$), while in the second scenario the functions are perturbed by a much higher amount ($\sigma = 1.5^2$). The simulated trajectories are depicted in Figure 4.1: in the first scenario the four functions $f_1^*, \dots, f_4^*$ are clearly distinguishable, whereas in the latter the underlying signal is less evident. Consequently, the clustering algorithm is expected to perform better in the small variance setting than in the high variance one.

Although the true number of clusters is 4, we set the total number of mixture components $H = 20$, to empirically demonstrate the ability of the E-FDMP to recover the correct number of distinct functions. Moreover, we let the number of class functions $L = 4$ and each within-class upper bound $H_l = 5$ for $l = 1, \dots, 4$. The functional atom specifications, as for equation (4.5), are the following

$$\begin{aligned} \tilde{\theta}_{1h}(t) &= \tilde{\beta}_{11h} + \tilde{\beta}_{21h} t, \quad \tilde{\theta}_{2h}(t) = \tilde{\beta}_{12h} + \tilde{\beta}_{22h} \cos(2\pi t) + \tilde{\beta}_{32h} \sin(2\pi t), \\ \tilde{\theta}_{3h}(t) &= \tilde{\beta}_{13h} + \tilde{\beta}_{23h} t^4, \quad \tilde{\theta}_{4h}(t) = \tilde{\beta}_{14h} + \tilde{\beta}_{24h} \cos(4\pi t) + \tilde{\beta}_{34h} \sin(4\pi t), \end{aligned}$$

with iid prior distributions $\tilde{\beta}_{mlh} \overset{\text{iid}}{\sim} \mathcal{N}(0, 10)$. The prior specification is concluded by setting $\alpha_1 = \cdots = \alpha_L = 1$, $c_1 = \cdots = c_L = 1$ and $a_\sigma = b_\sigma = 1$.

The optimization of the ELBO might be troublesome due to the presence of local maxima. To mitigate this issue, the CAVI algorithm was initialized at several different starting points; the solution achieving the highest value of the ELBO was retained (Blei et al., 2017). Remarkably, each run of the CAVI required only few seconds for the computations on a standard laptop and with a naïve implementation in the R statistical software. The results are depicted in Figure 4.1 for both the scenarios.

In the small variance setting (top graph of Figure 4.1), the CAVI algorithm applied to the E-FDMP model performs remarkably well. Indeed, it correctly identifies 4 clusters—meaning that among the estimated memberships $G_1^{(*)}, \dots, G_n^{(*)}$ there are only 4 distinct values—even though a conservative upper bound $H = 20$ was selected. Moreover, the observed curves are always allocated to the correct cluster, as summarized in Table 4.1a, up to a label permutation. Finally, the estimated curves $\hat{\theta}_{lh}$ depicted in Figure 4.1 closely

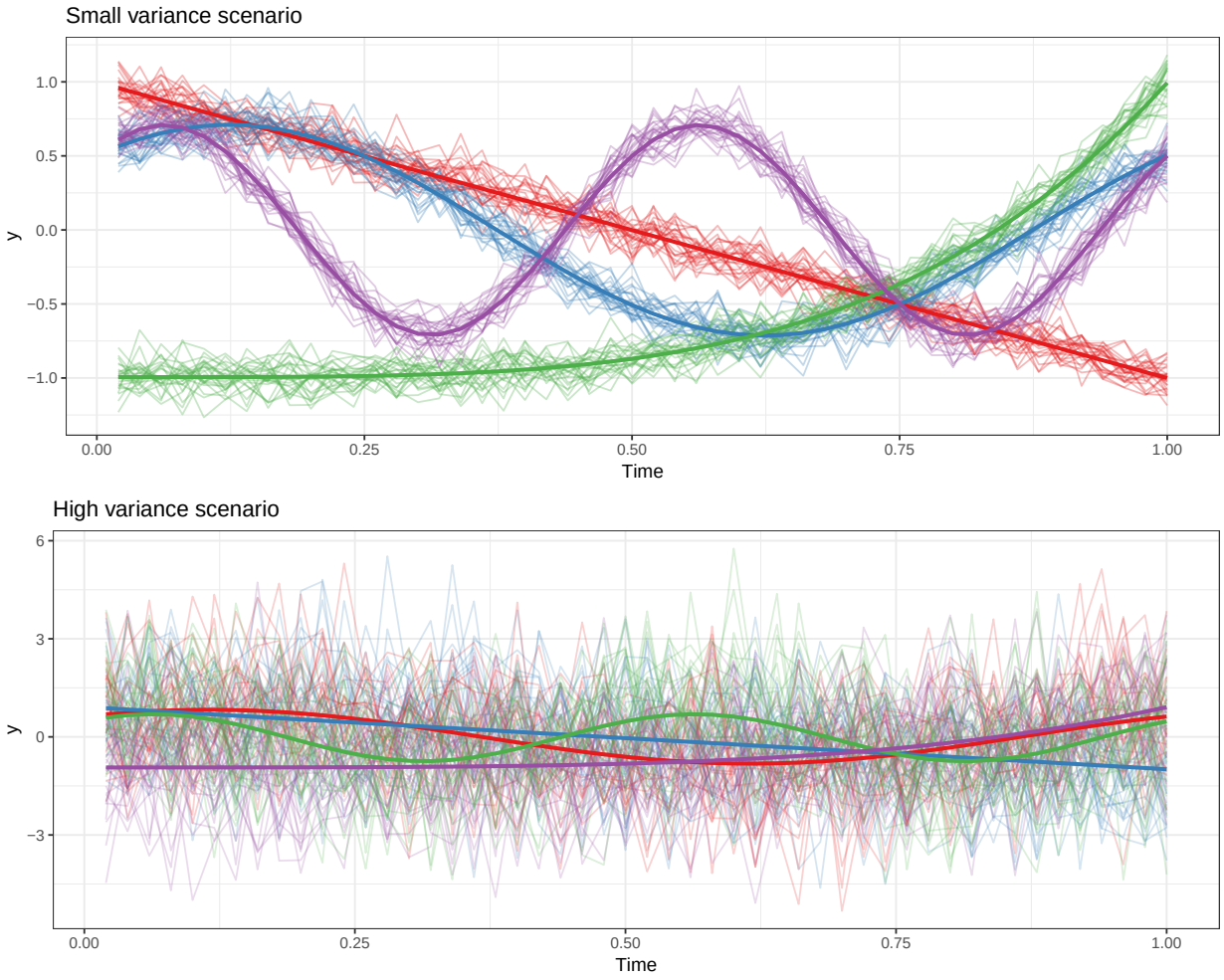


Figure 4.1: Simulated trajectories $Y_1(t), \ldots, Y_n(t)$ in the small variance scenario (top graph, $\sigma^2 = 0.1^2$), and high variance scenario (bottom graph, $\sigma^2 = 1.5^2$). Different colors refer to the estimated cluster memberships $G_1^{(*)}, \ldots, G_n^{(*)}$ whereas the corresponding solid lines are the estimated cluster-specific functions $\tilde{\theta}_{lh}^{(*)}(t)$.

Class label	1	2	3	4
Within-class label	1	2	3	3
f_1^*	25	0	0	0
f_2^*	0	25	0	0
f_3^*	0	0	25	0
f_4^*	0	0	0	25

(a) Small variance scenario.

Class label	1	2	3	4
Within-class label	4	2	2	4
f_1^*	22	1	0	2
f_2^*	3	19	1	2
f_3^*	0	2	23	0
f_4^*	1	0	0	24

(b) High variance scenario.

Table 4.1: Contingency tables showing the true cluster memberships $G_1, \dots, G_n$ against the estimated memberships $G_1^{(*)}, \dots, G_n^{(*)}$ in the small variance (a) and in the high variance (b) scenarios. The functional class and the within-class labels are reported. The cluster labels having zero frequencies are omitted.

resemble the true functions $f_1^*, \dots, f_4^*$. Similar remarks can be made also in the high variance scenario (bottom graph of Figure 4.1), although the performance are less striking, as one would expect. In particular, according to Table 4.1b the estimated memberships $G_1^{(*)}, \dots, G_n^{(*)}$ are correct in the 88% of the cases. However, it should be emphasized that in both cases the correct number of cluster is automatically identified, without the need of a post-processing step. This corroborates the usage of each H_l as an upper bound, implying that one should not be worried to overfit the data when selecting large H, as long as the $c_1, \dots, c_L$ parameters are well calibrated.

4.6 E-commerce application

4.6.1 Prior specifications

Recall that in our motivating application we aim at grouping flight routes according to the searches on the website of the company. From the original dataset at our disposal—concerning only Italian airports—we retained the flight routes having the highest number of searches within the period under consideration. As a result, the final dataset comprises $n = 214$ different flight routes accounting for the 94% of the total counts. Each $Y_i(t)$ is observed over a weekly time grid ranging from the 1st March 2017 ($t = 1$) to the 14th March 2018 ($t = 55$), so that each time grid equals $\boldsymbol{t}_i = (1, \dots, 55)^\intercal$, for $i = 1, \dots, n$. Hence, the dataset can be represented as a 214×55 matrix having 11770 entries.

We set the number of functional classes $L = 2$ and we select P_1 and P_2 so that they have interpretable but yet sufficiently flexible forms. The number of basis functions for both the functional classes is $M_1 = M_2 = 6$. The first functional class ($l = 1$) captures yearly cyclical patterns and characterizes the routes having e.g. a peak of web-searches

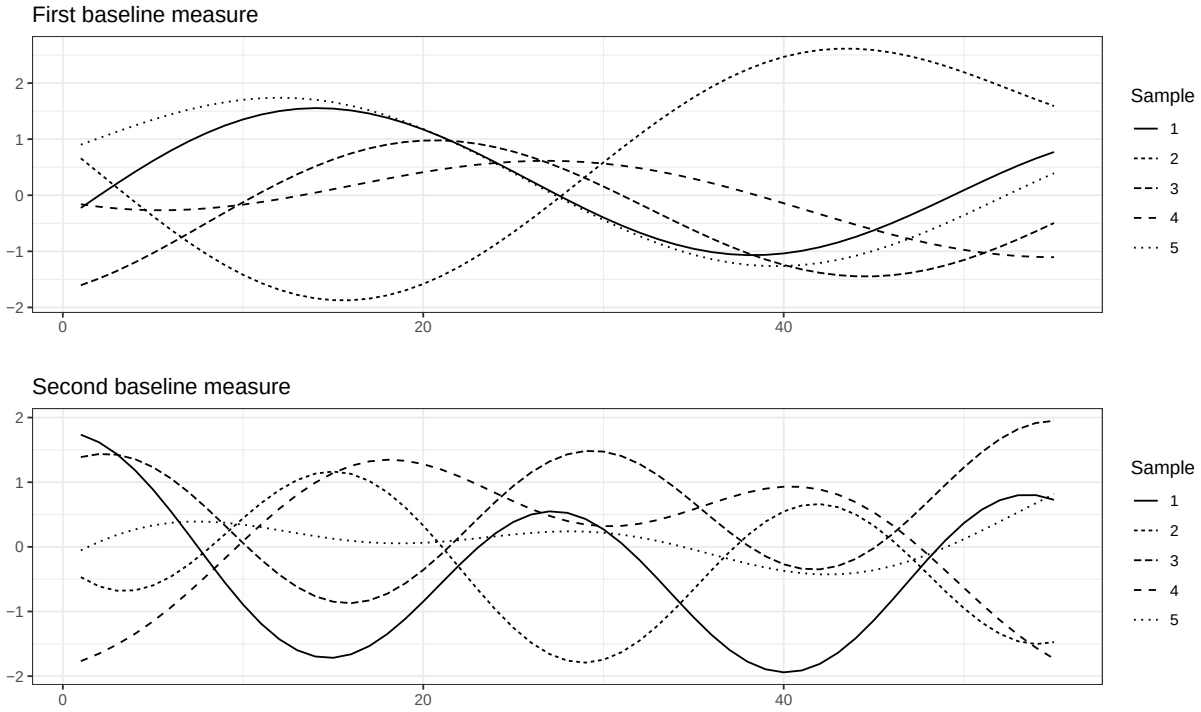


Figure 4.2: Prior samples for the $L = 2$ baseline probability measures P_1 (top graph) and P_2 (bottom graph) according to equations (4.8)-(4.9).

during either the summer or the winter. This is the case for example of the `MIL-AHO` route—from Milan to Alghero, a small city in Sardinia—as apparent from Figure 1.1. We increase the flexibility of this functional class by including also a semi-parametric component, thus allowing moderate deviations from this cyclical behavior. Specifically, we let

$$\tilde{\theta}_{1h}(t) = \sum_{m=1}^{4} \tilde{\beta}_{m1h} \mathcal{S}_m(t) + \tilde{\beta}_{51h} \cos\left(2\pi \frac{7}{365} t\right) + \tilde{\beta}_{61h} \sin\left(2\pi \frac{7}{365} t\right), \tag{4.8}$$

where $\mathcal{S}_1(t), \ldots, \mathcal{S}_4(t)$ are deterministic cubic spline basis functions. The second functional class ($l = 2$) has a mathematical formulation similar to (4.8), but with an important practical distinction. In particular, it characterizes functions having two peaks per year, which amounts to let

$$\tilde{\theta}_{2h}(t) = \sum_{m=1}^{4} \tilde{\beta}_{m2h} \mathcal{S}_m(t) + \tilde{\beta}_{52h} \cos\left(2\pi \frac{14}{365} t\right) + \tilde{\beta}_{62h} \sin\left(2\pi \frac{14}{365} t\right), \tag{4.9}$$

The `MIL-NAP` route—from Milan to Naples, depicted in Figure 1.1—is presumably a member of this functional class. As for the prior distributions $\tilde{\boldsymbol{\beta}}_{lh} \sim \mathcal{N}_{M_l}(\boldsymbol{\mu}_{\beta_l}, \boldsymbol{\Sigma}_{\beta_l})$, we set the prior means $\boldsymbol{\mu}_{\beta_1} = \boldsymbol{\mu}_{\beta_2} = \mathbf{0}$ and the covariance matrices $\boldsymbol{\Sigma}_{\beta_1} = \boldsymbol{\Sigma}_{\beta_2}$ to be equal and diagonal, having entries $\text{diag}(\boldsymbol{\Sigma}_{\beta_1}) = \text{diag}(\boldsymbol{\Sigma}_{\beta_2}) = (1, \ldots, 1)$, which were chosen to

induce a fairly uninformative prior, considered that the data were standardized. Few simulated draws from the prior baselines P_1 and P_2 are shown in Figure 4.2, which confirms that these two functional classes are both sufficiently flexible but distinct.

To induce a priori a moderate amount of clusters we select $c_1 = c_2 = 1$, whereas we specify a uniform prior for functional class probabilities $\boldsymbol{\Upsilon} = (\Upsilon_1, \Upsilon_2)$ by letting $\alpha_1 = \alpha_2 = 1$. The latter choice corresponds to the a priori indifference between the two functional classes. Moreover, by virtue of Corollary 4.1, it also implies that for H_l large enough the E-FDMP is approximately a FDP with baseline measure $\frac{1}{2}(P_1 + P_2)$. Finally, we let $a_\sigma = b_\sigma = 1$ for the residual precision σ^{-2}, a fairly uninformative setting.

4.6.2 Selection of the upper bounds

The theoretical findings of Section 4.3 as well as the simulation study of Section 4.5 seem to suggest that each H_l should be taken as large possible, being limited only by computational constraints. Indeed, the redundant clusters would be automatically deleted by the shrinkage prior in equation (4.4). Taken to the extreme (i.e. as each $H_l \to \infty$), this argument would lead to a proper Bayesian nonparametric prior; see Section 4.3. Although such an approach is theoretically sounding, its direct application might be troublesome on certain statistical problems. Indeed, real data are far more heterogeneous than those typically considered in simulations, meaning that the "true" number of clusters could be large with respect to the sample size. This effect is particularly marked within the context of functional clustering, because even small local oscillations lead to mathematically distinct functions. Hence, flexible priors with very large upper bounds—as well as infinite dimensional nonparametric priors—might constitute a better fit for the data, at the price of more complex cluster solutions. The strength of the E-FDMP formulation—especially in comparison with nonparametric priors—is in that one can balance the flexibility and the complexity of the model by tuning the bounds H_l.

On the basis of the above discussion, we let $H = \sum_{l=1}^{L} H_l$ be the largest value for which the resulting clustering solution is still useful in practice. Such a value is evidently quite subjective and it depends on the specific statistical problem. In our e-commerce application—in consultation with the stakeholders of the company—we let the upper bounds $H_1 = 20$ and $H_2 = 5$. Indeed, the second baseline measure is more prone to capture specificities of the functional observations compared to the first one, and this might lead to highly similar clusters. As discussed in the next section, such an effect is present even under the tight choice $H_2 = 5$. Note that the values H_l still preserve their interpretation of upper bounds for the within-class number of clusters: if less than H_l clusters are needed, then the redundant mixture components will be neglected.

Within-class label	2	3	5	6	10	14	16	17	20
Frequency	8	7	1	2	40	1	4	13	41
Volume ($\times 10^5$)	4.49	2.54	0.51	0.78	51.45	0.44	26.61	15.46	33.43

(a) First functional class ($l = 1$).

Within-class label	1	2	3	4	5
Frequency	27	9	28	21	12
Volume ($\times 10^5$)	35.24	8.27	23.93	26.96	16.16

(b) Second functional class ($l = 2$).

Table 4.2: For both the functional classes $l = 1$ and $l = 2$ the frequencies of the estimated clusters, as well as the traffic volumes associated to these groups, are reported. The traffic volumes represent the summation of the within-cluster number of web-searches over the period of consideration. The cluster labels having zero frequencies are omitted.

4.6.3 Flight routes segmentation

We run the CAVI Algorithm 1 multiple times, starting from different initialization points to mitigate the issue of local maxima. Such a procedure required only few minutes of computations on a standard laptop. From the ouput of the CAVI algorithm, we estimate the group memberships $G_1^{(*)}, \dots, G_n^{(*)}$ as discussed in Section 4.4. In Table 4.2 the frequencies of the resulting clusters are reported. Note that only 14 clusters are obtained out of $H = 25$ and furthermore some of them are composed only by few functional observations. Moreover, all the $H_2 = 5$ groups of the second functional class are occupied, which suggests that by selecting a larger upper bound one would probably get more clusters. However, this would be of little practical interest because—as evidenced in Figure 4.3—these 5 groups are already highly similar. This is an important practical advantage of the E-FDMP with respect to nonparametric priors, namely the ability of bounding the model complexity by avoiding the exploration of complex and less relevant partition structures.

Together with the cluster frequencies, we report in Table 4.2 also the traffic volumes associated to these groups, namely the within-cluster summation of the number of web-searches. Such a metric is far more important than the cluster frequencies: for example, cluster 16 of class 1—which has only 4 observations and a sensible traffic volume—is much more relevant from a business perspective than cluster 3 of class 1. Unsurprisingly, cluster 16 of class 1 identifies flights from the cities Milan and Bologna to Palermo and Catania, whose airports are among the biggest in Italy.

In Figure 4.3 we depict the raw standardized observations $Y_i(t)$ of the 10 most relevant clusters—i.e. those having the highest traffic volumes—overlaid with the corresponding estimated curves $\tilde{\theta}_{ln}^{(*)}(t)$. A direct graphical inspection confirms that the

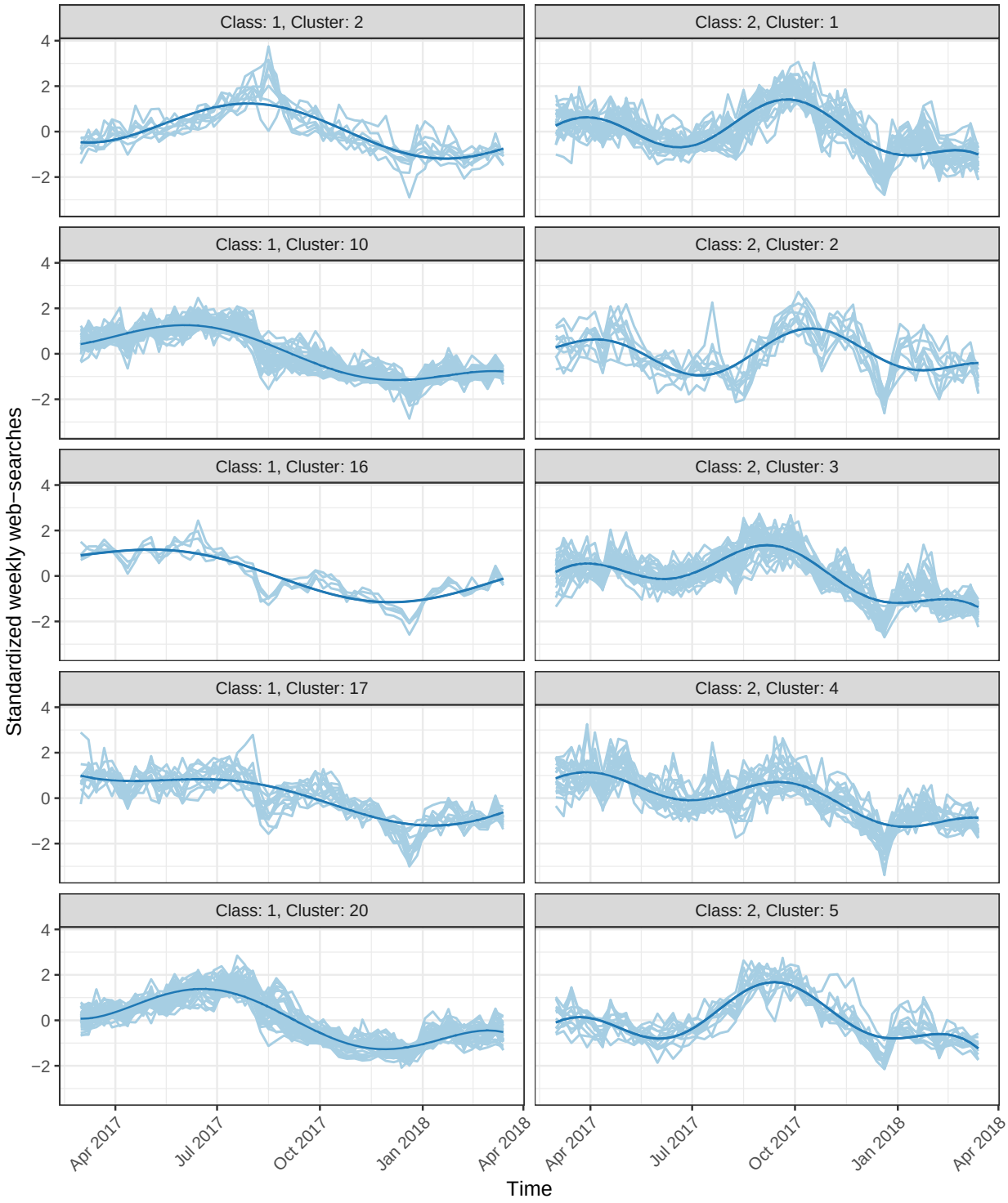


Figure 4.3: The standardized functional observations $Y_i(t)$ of the 10 most relevant clusters (according to the volumes of Table 4.2) are depicted. The solid dark lines represent the associated cluster-specific estimated trajectories $\tilde{\theta}^{(*)}_{lh}(t)$.

		Arrival		
		North	Center	South & Islands
Departure	North	0	2	49
	Center	0	0	24
	South & Islands	6	3	12

(a) Macro cluster A. Labels $\{10, 20\}$ of the first functional class $(l = 1)$.

		Arrival		
		North	Center	South & Islands
Departure	North	0	7	6
	Center	10	0	0
	South & Islands	47	21	7

(b) Macro cluster B. Labels $\{1, \ldots, 5\}$ of the second functional class $(l = 2)$.

Table 4.3: Contingency tables for the regions associated to the departure and arrival airports, for the flight routes belonging to macro clusters A and B.

baseline specifications of equations (4.8)-(4.9) are indeed flexible enough to capture the main tendencies of the data. Moreover, the differences between the two functional classes are evident also a posteriori: indeed, the clusters of the first column in Figure 4.3 are characterized by single peaked functions, while the other groups display two-peaked functions.

As previously mentioned, the clusters of the second functional class are mathematically different but quite similar, since all the corresponding functions have a first peak around April and a second one between September and October. Between functional classes, and within the first functional class, however, there is much more heterogeneity. For instance, the functions belonging to cluster 2 of class 1 have a single peak in August, while those belonging to clusters 10 and 20 of class 1 have a single peak between June and July. Moreover, functions of cluster 17, class 1, are quite stationary at the beginning and then they drop around August.

We now investigate in more detail the features of clusters 10 and 20 of the first functional class, termed henceforth macro cluster A, as well as those of the second functional class, which we will call macro cluster B. Indeed, these macro clusters are fairly homogeneous and they are also characterized by the highest traffic volumes. Recall that the airports of our dataset are located in Italy, which can be conveniently divided in three areas (North, Center and South & Islands), following standard administrative divisions. Arrival and departure airports of the flight routes belong to one of these areas. Remarkably, both the macro clusters A and B can be well described in terms of these administrative borders, as it is apparent from Table 4.3. In particular, the vast majority

of flight routes belonging to macro cluster A arrive to an airport located in the South & Island region. Conversely, in the macro cluster B most of the flight routes depart from the South & Islands area and are directed to the North and to the Center regions. These findings further corroborate the quality of the obtained cluster solution and they provide useful intuitions about the role of each cluster. Indeed, these qualitative descriptions might help marketing specialists in designing effective cluster-specific policies.

Chapter 5

Computational advances for hierarchical processes

5.1 Summary

The chapter is organized as follows. In Section 5.2 we review some background material on the Pitman-Yor process and on homogeneous normalized random measures with independent increments (NRMIs) that has not been covered in Chapters 2 and 3. In Section 5.3, we propose a particular instance of hierarchical process and we discuss a finite dimensional approximation based on a deterministic truncation of $\tilde{p}_0$. In Section 5.4 the truncated process is employed to define an infinite mixture model for partially exchangeable data. The novel conditional Gibbs sampler to conduct posterior inference is derived and described in detail. To assess the practical performance of both the novel algorithm and the aforementioned infinite mixture model, we conduct a simulation study in Section 5.5. Finally, as an illustration, we apply our algorithm on real data in Section 5.6.

5.2 Preliminaries and background

Throughout the chapter we will make extensive use of the notion of homogeneous normalized completely random measures (NRMIs), and of the Pitman-Yor process (PY). The definition of the Pitman–Yor process was given in Chapter 2 whereas a preliminary and concise background on completely random measures can be found in Chapter 3.

As discussed in Chapter 2, if $\tilde{p}^{(\infty)} \sim \text{PY}(c, \sigma; P)$ with $\tilde{p}^{(\infty)} = \sum_{h=1}^{\infty} \xi_h \delta_{\tilde{\phi}_h}$, then collection of the weights $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots)$ follows a stick-breaking construction. In the sequel we will only consider a subset of the collection of parameters (σ, c) for which $\sigma \in [0, 1)$ and $c \geqslant 0$, excluding the degenerate case $(\sigma, c) = (0, 0)$. Clearly, setting $\sigma = 0$ one obtain the stick-breaking construction of Sethuraman (1994) for the Dirichlet process,

whereas for $c = 0$ one is able to recover the stick-breaking construction of the σ-stable process given in Perman (1990). See also Perman et al. (1992). The distribution of the weights $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots)$ will be denoted with $\boldsymbol{\xi} \sim \text{GEM}(\sigma, c)$ after Griffiths, Engen, and McCloskey, and is also referred to as the *two-parameter Poisson–Dirichlet process*.

5.2.1 NRMI with finitely supported base measure

The hierarchical specification of discrete random probability measures given in (1.4) entails that each $\tilde{p}_l$ has, conditionally on $\tilde{p}_0$, an atomic base measure. In our case the $\tilde{p}_l$'s are homogeneous NRMIs and this motivates our interest in discussing specific features of NRMIs whose base measure is purely atomic. Accordingly, in this Section we will suppose that $\tilde{p}^{(H)} \sim \text{NRMI}(c, \rho; P)$ and that for some $H \geqslant 1$, there exists $\{\tilde{\theta}_1, \dots, \tilde{\theta}_H\} \subset \Theta$ such that $P(\{\tilde{\theta}_h\}) > 0$ for any $h \in \{1, \dots, H\}$ and $\sum_{h=1}^{H} P(\{\tilde{\theta}_h\}) = 1$. This corresponds to normalizing a CRM with fixed points of discontinuity. Because of this fact, the following discussion will partially overlap with some notions already discussed in Chapter 3, which are recalled here, with the appropriate modifications, for the ease of the exposition.

Let us first consider a finite collection $\{J_1, \dots, J_d\}$ of independent and infinitely divisible positive random variables such that for any $\lambda > 0$, one has $\mathbb{E}\{\exp(-\lambda J_i)\} = \exp\{-c_i \psi(\lambda)\}$, where ψ is the Laplace exponent corresponding to the jump measure ρ—as for equation (3.6) in Chapter 3—and $c_i > 0$ for any $i = 1, \dots, d$.

Definition 5.1. If $\bar{J} = \sum_{i=1}^{d} J_i$ and we let $\pi_i = J_i/\bar{J}$, then we say that $(\pi_1, \dots, \pi_{d-1})$ identifies a *normalized infinitely divisible* distribution and will use the notation

$$(\pi_1, \dots, \pi_{d-1}) \sim \text{NID}(c_1, \dots, c_d; \rho).$$

These distributions have been discussed at length in Favaro et al. (2011) and they are the building block of NIDM processes introduced in Chapter 3. If $\tilde{p}^{(H)} \sim \text{NRMI}(c, \rho; P)$ and P is purely atomic with H atoms, for any finite and measurable partition $\{B_1, \dots, B_d\}$ of Θ the vector $\{\tilde{p}^{(H)}(B_1), \dots, \tilde{p}^{(H)}(B_{d-1})\}$ clearly identifies a probability distribution on the simplex $S_{d-1} = \{(w_1, \dots, w_{d-1}) : w_i \geqslant 0; \ \sum_{i=1}^{d-1} w_i \leqslant 1\}$. Moreover, by virtue of Definition 5.1 one has

$$\{\tilde{p}^{(H)}(B_1), \dots, \tilde{p}^{(H)}(B_{d-1})\} \sim \text{NID}(cP(B_1), \dots, cP(B_d); \rho),$$

with the proviso that $\tilde{p}(B_i) = 0$, almost surely, if $P(B_i) = 0$. If we set $c_h = cP(\{\tilde{\theta}_h\})$ for each $h = 1, \dots, H$, and note that $P(\Theta \setminus \{\tilde{\theta}_1, \dots, \tilde{\theta}_H\}) = 0$, the random probability measure $\tilde{p}^{(H)} \sim \text{NRMI}(c, \rho; P)$ is fully characterized by the random vector

$$\left\{\tilde{p}^{(H)}(\{\tilde{\theta}_1\}), \dots, \tilde{p}^{(H)}(\{\tilde{\theta}_{H-1}\})\right\} \sim \text{NID}(c_1, \dots, c_H; \rho),$$

and the support of $\tilde{p}^{(H)}$ is the finite set of points $\{\tilde{\theta}_1, \ldots, \tilde{\theta}_H\}$, almost surely. This motivates the shorter notation $\tilde{p}^{(H)} \sim \text{NRMI}(c_1, \ldots, c_H; \rho)$ we use in this setting. We move on presenting some examples of homogeneous NRMIs, the associated NID distributions and their densities, that will play a relevant role in the sequel.

Example 5.1 (Dirichlet process). If $\rho(s) = s^{-1}e^{-s}$ then $\tilde{p}^{(H)} \sim \text{NRMI}(c, \rho; P)$ is a Dirichlet process and for any measurable partition $\{B_1 \ldots, B_d\}$ of Θ

$$\{\tilde{p}^{(H)}(B_1), \ldots, \tilde{p}^{(H)}(B_{d-1})\} \sim \text{DIRICHLET}\,(c_1, \ldots, c_d)\,, \qquad c_i = cP(B_i), \qquad i = 1, \ldots, d.$$

If $c_i > 0$ for each $i = 1, \ldots, d$, its density function is

$$p(\boldsymbol{w}) = \frac{\Gamma\,(c_1 + \cdots + c_d)}{\Gamma(c_1) \times \cdots \times \Gamma(c_d)}\, w_1^{c_1 - 1} \cdots w_{d-1}^{c_{d-1}-1}(1 - |\boldsymbol{w}|)^{c_d - 1} I_{S_{d-1}}(\boldsymbol{w}), \quad |\boldsymbol{w}| = \sum_{i=1}^{d-1} w_i.$$

Example 5.2 (Normalized inverse Gaussian process). If $\rho(s) = (\sqrt{2\pi})^{-1}\, s^{-3/2}\, e^{-s/2}$ then for any measurable partition $\{B_1 \ldots, B_d\}$ of Θ

$$\{\tilde{p}^{(H)}(B_1), \ldots, \tilde{p}^{(H)}(B_{d-1})\} \sim \text{N-IG}\,(c_1, \ldots, c_d)\,, \qquad c_i = cP(B_i), \qquad i = 1, \ldots, d,$$

and if $c_i > 0$ for any $i = 1, \ldots, d$, its density function can be obtained in closed form (Lijoi et al., 2005) and coincides with

$$p(\boldsymbol{w}) = \frac{e^{\sum_{i=1}^{d} c_i} \prod_{i=1}^{d} c_i}{2^{d/2-1}\Gamma(1/2)^d} \frac{\mathscr{K}_{-d/2}\left(\sqrt{\mathcal{A}_d(\boldsymbol{w})}\right)}{\mathcal{A}_d(\boldsymbol{w})^{d/4}} \left\{w_1 \cdots w_{d-1}(1 - |\boldsymbol{w}|)\right\}^{-3/2} I_{S_{d-1}}(\boldsymbol{w}),$$

where $\mathcal{A}_d(\boldsymbol{w}) = \sum_{i=1}^{d-1}(c_i^2/w_i) + c_d^2/(1 - |\boldsymbol{w}|)$ and $\mathscr{K}_d(\cdot)$ denotes the modified Bessel function of the third type.

Example 5.3 (1/2 stable process). If $\rho(s) = (\sqrt{2\pi})^{-1}\, s^{-3/2}$ then for any measurable for any measurable partition $\{B_1 \ldots, B_d\}$ of Θ

$$\{\tilde{p}^{(H)}(B_1), \ldots, \tilde{p}^{(H)}(B_{d-1})\} \sim \text{N-STABLE}\,(c_1, \ldots, c_d)\,, \qquad c_i = cP(B_i), \qquad i = 1, \ldots, d,$$

and if $c_i > 0$ for any $i = 1, \ldots, d$, its density function is

$$p(\boldsymbol{w}) = \frac{\Gamma(d/2) \prod_{i=1}^{d} c_i}{\Gamma(1/2)^d\, \mathcal{A}_d(\boldsymbol{w})^{d/2}} \left\{w_1 \cdots w_{d-1}(1 - |\boldsymbol{w}|)\right\}^{-3/2} I_{S_{d-1}}(\boldsymbol{w}),$$

where, as before $\mathcal{A}_d(\boldsymbol{w}) = \sum_{i=1}^{d-1}(c_i^2/w_i) + c_d^2/(1 - |\boldsymbol{w}|)$. See Carlton (2002). A well-known property of the normalized stable process is that it does not depend on the total mass c and this is clearly reflected by the expression of the density function above.

5.2.2 NID processes

While NIDs have been defined on a finite-dimensional simplex, they can be easily extended to an infinite dimensional setting. This is illustrated in the following.

Definition 5.2. Let $\boldsymbol{c} = (c_1, c_2, \dots)$ be an infinite collection of non-negative numbers such that $\sum_{h=1}^{\infty} c_h < \infty$. An infinite random vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$ such that $\sum_{h=1}^{\infty} \pi_h = 1$, almost surely, is a *normalized infinitely divisible process* (NIDP) with parameters ρ and $\boldsymbol{c}$ if, for any $d \geqslant 2$ and finite partition $\mathcal{H}_1, \dots, \mathcal{H}_d$ of $\mathbb{N}$, one has

$$\Big(\sum_{j \in \mathcal{H}_1} \pi_j, \dots, \sum_{j \in \mathcal{H}_{d-1}} \pi_j \Big) \sim \text{NID}\Big(\sum_{j \in \mathcal{H}_1} c_j, \dots, \sum_{j \in \mathcal{H}_d} c_j; \rho \Big),$$

and it will be denoted $\boldsymbol{\pi} \sim \text{NIDP}(\boldsymbol{c}, \rho)$.

If we take $\tilde{p}^{(\infty)} \sim \text{NRMI}(c, \rho; P)$ with $cP = \sum_{h=1}^{\infty} c_h \delta_{\tilde{\theta}_h}$, $\tilde{p}^{(\infty)} = \sum_{h=1}^{\infty} (\mathcal{J}_h / \bar{\mathcal{J}}) \delta_{\tilde{\phi}_h}$ and let, for any $h \geqslant 1$

$$\pi_h = \tilde{p}(\{\tilde{\theta}_h\}) = \sum_{\{j \colon \tilde{\phi}_j = \tilde{\theta}_h\}} \mathcal{J}_j / \bar{\mathcal{J}},$$

then $\boldsymbol{\pi} \sim \text{NIDP}(\boldsymbol{c}, \rho)$ with $c_h = cP(\{\tilde{\theta}_h\})$ for each $h \geqslant 1$. Because of their connection with NRMIs with countable baseline measure, NIDP processes will play a central role also in the description of general hierarchical processes.

5.3 Hierarchical processes

5.3.1 The hierarchical NRMI-PY process

In order to define the prior Q_L that governs a L-dimensional partially exchangeable array $\{(\theta_{li})_{i \geqslant 1} : l = 1, \dots, L\}$, according to (1.3), we rely on (1.4) and resort to a special instance of hierarchical discrete random probabilities. More specifically, we will deal with the following setting

$$\begin{aligned} (\theta_{li} \mid \tilde{p}_l^{(\infty)}) &\overset{\text{iid}}{\sim} \tilde{p}_l^{(\infty)}, && i = 1, \dots, n^{(l)}, \quad l = 1, \dots, L, \\ (\tilde{p}_l^{(\infty)} \mid \tilde{p}_0^{(\infty)}) &\overset{\text{iid}}{\sim} \text{NRMI}(c, \rho, \tilde{p}_0^{(\infty)}), && l = 1, \dots, L, \\ \tilde{p}_0^{(\infty)} &\sim \text{PY}(\sigma_0, c_0, P), \end{aligned} \tag{5.1}$$

where P is a *diffuse* probability measure defined on Θ. We will identify this model as a hierarchical NRMI-PY process. Notice that both the HDP (Teh et al., 2006) and the hierarchical stable process (Camerlenghi et al., 2019) can be recovered as particular cases.

A key feature of hierarchical species sampling models (1.4), and consequently also of the NRMI-PY process (5.1), is that with positive probability they induce ties among

the θ_{li}'s, because of the almost sure discreteness of both $(\tilde{p}_l^{(\infty)} \mid \tilde{p}_0^{(\infty)})$ and $\tilde{p}_0$. Ties might occur both within and across groups, because the $(\tilde{p}_l^{(\infty)} \mid \tilde{p}_0^{(\infty)})$ share the same discrete baseline measure, for $l = 1, \dots, L$. Thus, investigating the a priori clustering mechanism is of greater importance to highlight possible limitations induced by specific choices of $(\tilde{p}_l^{(\infty)} \mid \tilde{p}_0^{(\infty)})$ and $\tilde{p}_0^{(\infty)}$. Indeed, compared to the HDP, specification (5.1) allows for a more flexible modeling of the clustering mechanism while still preserving analytical tractability: one can resort to the general theory set forth in Camerlenghi et al. (2019) in order to derive the partially exchangeable partition function, the full posterior characterization, and a closed form expression for the distribution of the number of clusters. See also Bassetti et al. (2018) for further developments in this direction. In addition, formulation (5.1) is also a suitable choice for computational reasons, as we will discuss in Section 5.4. Indeed, the stick-breaking construction of the PY process $\tilde{p}_0$ leads to a simple simulation strategy, both a priori and a posteriori, whereas NRMIS are a good candidate for each $(\tilde{p}_l^{(\infty)} \mid \tilde{p}_0^{(\infty)})$ whenever it is relatively simple to study their finite-dimensional distribution, as discussed in Section 5.2.

An alternative representation of the model in (5.1) highlights a direct connection with a hierarchical collection of random weights following NIDP and GEM distributions, respectively. This approach provides a deeper understanding of the model and, in addition, has relevant computational advantages. Let us first recall that, in view of the definition of homogeneous NRMI given in Chapter 3, one has

$$\tilde{p}_l^{(\infty)} = \sum_{h=1}^{\infty} (J_{lh}/\bar{J}_l)\delta_{\tilde{\phi}_{lh}}, \qquad l = 1, \dots, L, \tag{5.2}$$

where $(\tilde{\phi}_{lh} \mid \tilde{p}_0^{(\infty)}) \overset{\text{iid}}{\sim} \tilde{p}_0^{(\infty)}$, for $h \geqslant 1$ and $l = 1, \dots, L$. Moreover, the sequences of random jumps $\{(J_{lh})_{h\geqslant 1} : l = 1, \dots, L\}$ are independent from the locations $\{(\tilde{\phi}_{lh})_{h\geqslant 1} : l = 1, \dots, L\}$ and conditionally independent across groups, given $\tilde{p}_0^{(\infty)}$. As for the random baseline distribution, we let $\tilde{p}_0^{(\infty)} \sim \text{PY}(\sigma_0, c_0; P)$ implying that

$$\tilde{p}_0^{(\infty)} = \sum_{h=1}^{\infty} \xi_{0h}\delta_{\tilde{\phi}_{0h}}, \qquad \tilde{\phi}_{0h} \overset{\text{iid}}{\sim} P,$$

with $\boldsymbol{\xi}_0 = (\xi_1, \xi_2, \dots)$ following the stick-breaking construction. From the above construction, each random probability measures $\tilde{p}_l^{(\infty)}$ charges locations that are sampled from $\tilde{p}_0^{(\infty)}$. Because of the almost sure discreteness of $\tilde{p}_0^{(\infty)}$, one can equivalently rewrite (5.2) as follows

$$\tilde{p}_l^{(\infty)} = \sum_{h=1}^{\infty} \pi_{lh}\delta_{\tilde{\phi}_{0h}}, \qquad l = 1, \dots, L, \tag{5.3}$$

in which, conditionally on $\tilde{p}_0^{(\infty)}$, the locations $\tilde{\phi}_{0h}$ are fixed whereas the "modified weights" $\boldsymbol{\pi}_l = (\pi_{l1}, \pi_{l2}, \dots)$ are

$$\pi_{lh} = \sum_{\{j:\, \tilde{\phi}_{lj} = \tilde{\phi}_{0h}\}} \jmath_{lj} / \bar{\jmath}_l, \qquad h \geqslant 1,$$

for any $l = 1, \dots, L$. Remarkably, the conditional law of the perturbed weights $\boldsymbol{\pi}_l$, given $\tilde{p}_0^{(\infty)}$, can be derived and it follows a NIDP process. This can be easily seen from additivity of NRMIS, since for any finite and measurable partition $\{B_1, \dots, B_d\}$ of Θ,

$$\Big(\tilde{p}_l^{(\infty)}(B_1), \dots, \tilde{p}_l^{(\infty)}(B_{d-1}) \mid \tilde{p}_0^{(\infty)}\Big) = \Big(\sum_{j \in \mathcal{H}_1} \pi_{lj}, \dots, \sum_{j \in \mathcal{H}_{d-1}} \pi_{lj} \mid \tilde{p}_0^{(\infty)}\Big),$$

where $\mathcal{H}_i = \{h \geqslant 1 : \ \tilde{\phi}_{0h} \in B_i\}$, for $i = 1, \dots, d$, form a partition of $\mathbb{N}$. Then, we have that

$$\Big(\sum_{j \in \mathcal{H}_1} \pi_{lj}, \dots, \sum_{j \in \mathcal{H}_{d-1}} \pi_{lj} \mid \tilde{p}_0^{(\infty)}\Big) \sim \text{NID}\Big(c \sum_{j \in \mathcal{H}_1} \xi_{0j}, \dots, c \sum_{j \in \mathcal{H}_d} \xi_{0j}; \rho\Big),$$

since $\tilde{p}_0^{(\infty)}(B_i) = \sum_{j \in \mathcal{H}_i} \xi_{0j}$, for any $i = 1, \dots, d$. This implies, by definition of a NIDP, that $(\boldsymbol{\pi}_l \mid \boldsymbol{\xi}_0) \overset{\text{iid}}{\sim} \text{NIDP}(c\, \boldsymbol{\xi}_0, \rho)$, for any $l = 1, \dots, L$. Now let us introduce a collection of assignment variables $G_{li} \in \{1, 2, \dots\}$, denoting the cluster membership of each observation, namely $\theta_{li} = \tilde{\phi}_{0G_{li}}$. Then, we express model (5.1) in the following equivalent form

$$\begin{aligned} \boldsymbol{\xi}_0 &\sim \text{GEM}(\sigma_0, c_0), & \tilde{\phi}_{0h} &\overset{\text{iid}}{\sim} P, \qquad h \geqslant 1, \\ (\boldsymbol{\pi}_l \mid \boldsymbol{\xi}_0) &\overset{\text{iid}}{\sim} \text{NIDP}(c\, \boldsymbol{\xi}_0, \rho), & (G_{li} \mid \boldsymbol{\pi}_l) &\overset{\text{iid}}{\sim} \text{MULTINOMIAL}(\boldsymbol{\pi}_l), \end{aligned} \tag{5.4}$$

for $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, L$. Specification (5.4) in the particular case of the HDP is already available from Teh et al. (2006) and it is extended here to the NRMI-PY process. Moreover, such a construction does not use peculiar properties of the PY process, and it would hold for any other discrete random probability measure $\tilde{p}_0^{(\infty)}$. The major advantage of the PY process relies on the fact that the GEM distribution appearing in (5.4) is analytically and computationally tractable.

5.3.2 Deterministic truncation of the infinite process

Posterior inference for the NRMI-PY hierarchical processes of equation (5.1) is complicated by the infinite amount of parameters involved in the prior specification. A possible strategy for circumventing the problem is the marginalization with respect to the random probability measures $\tilde{p}_1^{(\infty)}, \dots, \tilde{p}_d^{(\infty)}, \tilde{p}_0^{(\infty)}$ to obtain generalized Pólya urn schemes that

are building blocks of Gibbs samplers of the type proposed in Camerlenghi et al. (2019). This approach is very effective when one wants to approximate Bayesian point estimators under squared error loss or, more generally, evaluate linear functionals of the underlying posterior distribution. On the contrary, it is not ideal if one is interested in non-linear functionals such as those needed for determining credible intervals that are relevant for uncertainty quantification.

In order to address the issue, we first introduce a deterministic truncation of the stick-breaking construction of the PY process. This obviously has a cascade effect also on the conditional distributions of the $\tilde{p}_l^{(\infty)}$'s, given such a truncated version of $\tilde{p}_0^{(\infty)}$, since they boil down to finite-dimensional random elements, without the need of further approximations. More precisely, we approximate model (5.1) with the following truncated specification

$$\begin{aligned} (\theta_{li} \mid \tilde{p}_l^{(H)}) &\overset{\text{ind}}{\sim} \tilde{p}_l^{(H)}, && i = 1, \dots, n^{(l)}, \quad l = 1, \dots, L, \\ (\tilde{p}_l^{(H)} \mid \tilde{p}_{0,tr}^{(H)}) &\overset{\text{iid}}{\sim} \text{NRMI}\left(c\tilde{p}_{0,tr}^{(H)}; \rho\right), && l = 1, \dots, L, \\ \tilde{p}_{0,tr}^{(H)} &\sim \text{PY}_H(\sigma_0, c_0, P), \end{aligned} \tag{5.5}$$

where $\tilde{p}_{0,tr}^{(H)} \sim \text{PY}_H(\sigma_0, c_0, P)$ denotes a truncated PY process with H components, as in Chapter 1 and Chapter 2. Clearly, the truncated measure $\text{PY}_H(\sigma_0, c_0, P)$ converges weakly, almost surely, to a proper Pitman-Yor process as $H \to \infty$, and hence implying also the weak convergence, almost surely, of the bottom level NRMIs.

An assessment of the effect of such a deterministic truncation can be obtained by determining an upper bound of the total variation distance between $\tilde{p}_l^{(\infty)}$ of the hierarchical process (5.1) and its finite-dimensional approximation $\tilde{p}_l^{(H)}$ in (5.5), for each $l = 1, \dots, L$. This can provide some guidance on the value at which H can be fixed. It is apparent that such an upper bound turns out to be random and we will rely on its expected value in order to gain some intuitive insight on the accuracy of the proposed truncation. To this end, we need to recall $\tau_2(u) = \int_{\mathbb{R}_+} s^2 e^{-us} \rho(s) ds$ and recall that $(a)_n = a(a+1)\cdots(a+n-1)$ denotes the Pochammer symbol. Moreover, we recall that ψ is the Laplace exponent associated to the jump intensity ρ, i.e. $\psi(u) = \int_{\mathbb{R}_+} (1 - e^{-us})\, \rho(s) ds$ for any $u > 0$.

Theorem 5.1. *Let* $(\tilde{p}_1^{(\infty)}, \dots, \tilde{p}_d^{(\infty)})$ *be a hierarchical* NRMI-PY *process as in* (5.1) *and let* $(\tilde{p}_1^{(H)}, \dots, \tilde{p}_d^{(H)})$ *be the truncated version defined in* (5.5)*. Then, for any* $l = 1, \dots, L$,

$$d_{TV}\left(\tilde{p}_l^{(\infty)}, \tilde{p}_l^{(H)}\right) = \sup_{A \in \mathscr{B}(\Theta)} \left|\tilde{p}_l(A) - \tilde{p}_l^{(H)}(A)\right| \leqslant \mathscr{R}_{lH} = \sum_{h > H} \pi_{lh} \qquad a.s.,$$

implying that

$$\mathbb{E}\left\{d_{TV}\left(\tilde{p}_l^{(\infty)}, \tilde{p}_l^{(H)}\right)\right\} \leqslant \mathbb{E}(\mathscr{R}_{lH}) = \prod_{h=1}^{H} \frac{c_0 + \sigma_0 h}{c_0 + \sigma_0(h-1)+1}.$$

In addition, set $\mathcal{R}_1(H) = \prod_{h=1}^{H} \frac{c_0+\sigma_0 h}{c_0+\sigma_0(h-1)+1}$ *and* $\mathcal{R}_2(H) = \prod_{h=1}^{H} \frac{(c_0+\sigma_0 h)_2}{(c_0+\sigma_0(h-1)+1)_2}$, *then*

$$\mathrm{Var}\,(\mathscr{R}_{lH}) = \mathscr{I}(c,\rho)\mathcal{R}_1(H) + (1-\mathscr{I}(c,\rho))\mathcal{R}_2(H) - \mathcal{R}_1(H)^2,$$

where $\mathscr{I}(c,\rho) = c\int_{\mathbb{R}_+} u e^{-c\psi(u)}\tau_2(u)du$.

The upper bound $\mathscr{R}_{lH}$ has a simple interpretation: broadly speaking, it consists on the part of π_l neglected by the truncation, and hence it is sometimes called *truncation error* in the exchangeable setting (Arbel et al., 2018). As a natural and intuitive consequence of Theorem 5.1, we get that $d_{TV}\left(\tilde{p}_l^{(\infty)}, \tilde{p}_l^{(H)}\right) \xrightarrow{a.s.} 0$ as $H \to \infty$. More importantly, the first two moments of $\mathscr{R}_{lH}$ can be used to determine a suitable truncation level H; for example, one might select the value of H such that the expected value of $\mathscr{R}_{lH}$ is below a certain threshold. Some further insight on $\mathscr{R}_{lH}$ may be gained by using the fact that

$$(\mathscr{R}_{lH} \mid \boldsymbol{\xi}_0) \sim \textsc{nid}\Big(c\Big(1-\sum_{h=1}^{H}\xi_{0h}\Big), c\sum_{h=1}^{H}\xi_{0h}; \rho\Big),$$

so that one can simulate its realizations, conditionally on $\boldsymbol{\xi}_0$. When $\tilde{p}_0^{(\infty)}$ is a Dirichlet process the expected value of the random variable $\mathscr{R}_{lH}$ goes to zero exponentially fast, meaning that H has not to be very large in practice. This is illustrated in the following example.

Example 5.4 (Truncated HDP). If $\rho(s) = s^{-1}e^{-s}$ and $\sigma_0 = 0$, then $\tilde{p}_{0,tr}^{(H)}$ in (5.5) is a truncated Dirichlet process and the $\tilde{p}_l^{(H)}$ are, conditionally on $\tilde{p}_{0,tr}^{(H)}$, iid draws from a Dirichlet distribution. Specializing Theorem 5.1 we get

$$\mathbb{E}\left\{d_{TV}\left(\tilde{p}_l^{(\infty)}, \tilde{p}_l^{(H)}\right)\right\} \leqslant \left(\frac{c_0}{c_0+1}\right)^H.$$

Therefore, on average, the total variation distance $d_{TV}\left(\tilde{p}_l^{(\infty)}, \tilde{p}_l^{(H)}\right)$ goes to zero exponentially fast as a function of H. Moreover,

$$\mathrm{Var}\,(\mathscr{R}_{lH}) = \frac{1}{c+1}\left\{c\left(\frac{c_0}{c_0+2}\right)^H - (c+1)\left(\frac{c_0}{c_0+1}\right)^{2H} + \left(\frac{c_0}{c_0+1}\right)^H\right\},$$

which is, again, exponentially decreasing as a function of H, implying that the upper bound $\mathscr{R}_{lH}$ is quite concentrated on its expected value for reasonably large values of H.

As apparent from Theorem 5.1, the parameters (c_0, σ_0) of the (truncated) PY process $\tilde{p}_0^{(H)}$ directly impact the quality of the approximation. Indeed, the expectation $\mathbb{E}(\mathscr{R}_{lH})$ increases as a function of both c_0 and σ_0. However, if $\sigma_0 > 0$ the decay is not anymore exponential, implying that to achieve reasonable approximations we need a larger H, especially for values of σ_0 close to one. This is consistent with the discussions in Ishwaran & James (2001) and Arbel et al. (2018) in the exchangeable case.

Another natural aspect that is worth pointing out is the dependence between $\tilde{p}_l^{(H)}$ and $\tilde{p}_{l'}^{(H)}$, for any $l \neq l'$, and how this differ from the one associated to the original hierarchical process specification in (1.4). To this end one can, for instance, evaluate the correlation between $\tilde{p}_l^{(H)}(A)$ and $\tilde{p}_{l'}^{H}(A)$ for any $A \in \mathscr{X}$ and truncation level H.

Theorem 5.2. *Let* $(\tilde{p}_1^{(H)}, \ldots, \tilde{p}_d^{(H)})$ *be a hierarchical approximate* NRMI-PY *process as in* (5.5). *Then, for any* $A \in \mathscr{B}(\Theta)$ *such that* $0 < P(A) < 1$ *and any* $l \neq l'$

$$\operatorname{Corr}\left\{\tilde{p}_l^{(H)}(A), \tilde{p}_{l'}^{(H)}(A)\right\} = \frac{\mathscr{I}_0(\sigma_0, c_0, H)}{\mathscr{I}(c, \rho) + \mathscr{I}_0(\sigma_0, c_0, H)(1 - \mathscr{I}(c, \rho))}, \tag{5.6}$$

where $\mathscr{I}(c, \rho)$ *is as in Theorem 5.1 and*

$$\mathscr{I}_0(\sigma_0, c_0, H) = \sum_{h=1}^{H-1} \frac{(1-\sigma_0)_2}{(1+c_0+(h-1)\sigma_0)_2} \prod_{l=1}^{h-1} \frac{(c_0+l\sigma_0)_2}{(1+c_0+(l-1)\sigma_0)_2} + \prod_{h=1}^{H-1} \frac{(c_0+h\sigma_0)_2}{(1+c_0+(h-1)\sigma_0)_2}.$$

Moreover, taking the limit we get $\lim_{H\to\infty} \mathscr{I}_0(\sigma_0, c_0, H) = (1-\sigma_0)/(1+c_0)$, which entails that $\operatorname{Corr}(\tilde{p}_l^{(H)}(A), \tilde{p}_{l'}^{(H)}(A))$ converges to the actual $\operatorname{Corr}(\tilde{p}_l^{(\infty)}(A), \tilde{p}_{l'}^{(\infty)}(A))$, implied by the model (5.1), as $H \to \infty$. It is apparent that the correlation coefficient is always positive and, unsurprisingly, does not depend on the specific set A as a consequence of homogeneity of the underlying random probability measures at the different levels of the hierarchy. As such, it is generally interpreted as an overall measure of dependence between the random probability measures.

Remark 5.1. Note that the parameters c_0 and σ_0 do not play the same role as in the infinite-dimensional case. Indeed, one can show that $\lim_{c_0\to\infty} \mathscr{I}_0(\sigma_0, c_0, H) = 1$, which clearly entails that $\lim_{c_0\to\infty} \operatorname{Corr}(\tilde{p}_l^{(H)}(A), \tilde{p}_{l'}^{(H)}(A)) = 1$. On the other hand, it is clear that when $H = \infty$ one has the opposite limiting behavior, namely $\lim_{c_0\to\infty} \operatorname{Corr}(\tilde{p}_l^{(\infty)}(A), \tilde{p}_{l'}^{(\infty)}(A)) = 0$, for any $l \neq l'$. Similar can be determined when considering $\sigma_0 \to 1$. The truncation effect that explains this different limiting dependence

structure is quite intuitive: when either σ_0 or c_0 increase more mass is placed on the Hth atom of the stick-breaking construction, so that $\tilde{p}_{0,tr}^{(H)}$ eventually converges to a point mass at $\tilde{\phi}_{0H}$. To sum up, if we let σ_0 (or c_0) be fixed and consider the correlation as a function of c_0 (or of σ_0) it first decreases until it reaches a minimum and, then, increases.

Example 5.5 (Truncated HDP, cont'd.). In the HDP case, the above correlation can be significantly simplified. Indeed, a straightforward application of Theorem 5.2 yields

$$\mathrm{Corr}\{\tilde{p}_l^{(H)}(A), \tilde{p}_{l'}^{(H)}(A)\} = \frac{(1+c)\left(1+c_0\left(\frac{c_0}{c_0+2}\right)^{H-1}\right)}{1+c_0+c\left(1+c_0\left(\frac{c_0}{c_0+2}\right)^{H-1}\right)}.$$

In the infinite case $H \to \infty$ the correlation reduces to $(1+c)/(1+c_0+c)$, as already obtained in Camerlenghi et al. (2019). Thus, the truncation of $\tilde{p}_0$ induces a perturbation of the correlation of the HDP through a factor which is exponentially decreasing in H.

5.4 Hierarchical NRMI-PY mixture model

5.4.1 Infinite mixture model

In several applied contexts the discreteness of the hierarchical NRMI-PY prior is not a realistic assumption. Nonetheless, we can adapt formulation (5.1) by adding a further level in the hierarchy, giving rise to a mixture model for partially exchangeable observations. Within the exchangeable framework, this idea was firstly suggested by Lo (1984), and discussed in practice for instance in Escobar & West (1995) in the Dirichlet process case, and by Barrios et al. (2013) for general homogeneous NRMIs.

Let Y_{li} for $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, L$ be a sample of observations taking values in a complete and separable metric space $\mathbb{Y}$ and let $\mathcal{K} : \mathbb{Y} \times \Theta \to \mathbb{R}_+$ a transition kernel such that $y \mapsto \mathcal{K}(y;\theta)$ is a density function on $\mathbb{Y}$, for any $\theta \in \Theta$, with respect to some dominating σ-finite measure. Exploiting representation (5.4), for any truncation level H the approximate hierarchical NRMI-PY mixture model is

$$\begin{aligned}
\boldsymbol{\xi}_0^{(H)} &\sim \mathrm{GEM}_H(\sigma_0, c_0), & \tilde{\phi}_{0h} &\overset{iid}{\sim} P, \qquad h = 1, \dots, H,\\
(\boldsymbol{\pi}_l \mid \boldsymbol{\xi}_0^{(H)}) &\overset{iid}{\sim} \mathrm{NID}(c\xi_{01}, \dots, c\xi_{0H}; \rho), & (G_{li} \mid \boldsymbol{\pi}_l) &\overset{iid}{\sim} \mathrm{MULTINOMIAL}(\pi_{l1}, \dots, \pi_{lH}),\\
(Y_{li} \mid G_{li}, \tilde{\boldsymbol{\phi}}_0) &\overset{ind}{\sim} \mathcal{K}(y; \tilde{\phi}_{0G_{li}}),
\end{aligned} \tag{5.7}$$

for $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, L$, with $\tilde{\boldsymbol{\phi}}_0 = (\tilde{\phi}_{01}, \dots, \tilde{\phi}_{0H})$ and $\boldsymbol{\pi}_l = (\pi_{l1}, \dots, \pi_{lH-1})$, and where $\mathrm{GEM}_H(\sigma_0, c_0)$ denotes the truncated sequence $\boldsymbol{\xi}_0^{(H)} = (\xi_{01}, \dots, \xi_{0H-1})$, associated to the aforementioned truncated PY process. Also, we set $\pi_{lH} = 1 - |\boldsymbol{\pi}_l|$ for

$l = 1, \dots, L$. Marginalizing over the cluster indicators G_{li}, we obtain a finite mixture representation

$$(Y_{li} \mid \boldsymbol{\pi}_l, \tilde{\phi}_0) \overset{\text{ind}}{\sim} \sum_{h=1}^{H} \pi_{lh} \mathcal{K}(y; \tilde{\phi}_{0h}), \tag{5.8}$$

for $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, L$. As apparent from equations (5.7)-(5.8), under this hierarchical constructions the distributions $\sum_{h=1}^{H} \pi_{lh} \mathcal{K}(y; \tilde{\phi}_{0h})$ for $l = 1, \dots, L$ share the same mixture components $\mathcal{K}(y; \tilde{\phi}_{0h})$. However, they have different mixing weights $\boldsymbol{\pi}_l$, accounting for heterogeneity across groups. We remark that the conditional density $\sum_{h=1}^{H} \pi_{lh} \mathcal{K}(y; \tilde{\phi}_{0h})$ is often of direct inferential interest and one may want to obtain its posterior distribution rather than just confining herself to a point estimate. In this case, one cannot rely on marginal algorithms that integrate out the random weights π_{lh} and a different (conditional) sampler must be adopted.

5.4.2 Blocked Gibbs sampler

In this Section we propose a simple Markov Chain Monte Carlo (MCMC) scheme that makes use of the approximate specification in equation (5.7) and enables posterior inference. The algorithms originally proposed for the HDP in Teh et al. (2006) are of marginal type, thus being characterized by their pros and cons: very effective for point estimation, but unreliable when it comes to uncertainty quantification. In the supplementary material of Fox et al. (2011) a conditional algorithm for the HDP is discussed, and it is based on a finite-dimensional approximation of $\tilde{p}_0^{(\infty)}$; however, its applicability is limited to the HDP case. A general marginal algorithm for hierarchical NRMI processes and hierarchical PY processes was proposed by Camerlenghi et al. (2019). In this very same paper, the authors discuss also a conditional algorithm based on a representation of CRMs that can be traced back to Ferguson & Klass (1972). Its actual implementation must still rely on some truncation of the underlying infinite-dimensional process that can be achieved through a specific approach as the one suggested, e.g., in Arbel & Prünster (2017). Since the representation in Ferguson & Klass (1972) displays jumps arranged in decreasing order, any truncation rule will retain the most relevant jumps. On the other hand, any computational procedure based on this construction will require the inversion of an underlying Lévy measure attainable and this may cause some computational issues.

The blocked Gibbs sampler we propose does not rely on the augmented scheme proposed in Camerlenghi et al. (2019) nor it makes use of the (suitably truncated) Ferguson & Klass representation, while still being a conditional algorithm. Furthermore, the effect of the approximations can be explicitly assessed a priori thanks to Theorem 5.1. The main relevant constraint implied by our proposal is the availability of the density

function $p(\pi_l \mid \xi_0^{(H)})$ in closed form since it needs to be evaluated. Nonetheless there are some noteworthy examples of NRMIs that comply with this requirement, namely the Dirichlet process, the normalized inverse-Gaussian process, and the 1/2-stable process. See Section 5.2.

We now review the steps of the blocked Gibbs sampler, outlined in Algorithm 2, highlighting practical difficulties and suggesting possible solutions. Each step represents a full conditional distribution for a block of random variables, and we will denote with a "$-$" the conditioning to all the other variables.

Step 1. Observations are randomly and independently allocated to different clusters. Since we have truncated the sequence of weights $\xi_0^{(H)}$ up to the Hth term, the number of mixture component is finite. In turns, this implies that the normalizing constant can be obtained as a simple summation of the involved quantities.

Step 2. The mixing probabilities π_l are sampled independently for $l = 1, \dots, L$. Unfortunately, the full conditional $p(\pi_l \mid -)$ is typically not available in closed form. The only exception occurs when the prior $p(\pi_l \mid \xi_0^{(H)})$ is the conditionally conjugate Dirichlet distribution, that is, when we assume that $(\tilde{p}_l^{(\infty)} \mid \tilde{p}_0^{(\infty)})$ is distributed according to a Dirichlet process. Beside the latter particular case, in general we must resort to a Metropolis-Hastings step. Having tried several different proposal distributions, we obtained very good performance by working in the unconstrained space $\log(\pi_{lh}/\pi_{lH})$, for any $h = 1, \dots, H-1$—and then by applying a componentwise Gaussian random walk. The variances on the Gaussian proposal were adaptively and automatically selected as in Roberts & Rosenthal (2009).

Step 3. The baseline mixing weights $\xi_0^{(H)}$ are sampled. Notice that the vector $\xi_0^{(H)}$ is a particular instance of a generalized Dirichlet distribution (Connor & Mosimman, 1969), and its density is

$$p(\boldsymbol{w}) = \frac{(1-|\boldsymbol{w}|)^{c_0+\sigma_0(H-1)-1}}{\prod_{h=1}^{H-1} B(1-\sigma_0, c_0+h\sigma_0)} \prod_{h=1}^{H-1} \left[w_h^{-\sigma_0} \left(\sum_{j=h}^{H} w_j \right)^{-1} \right] I_{S_{H-1}}(\boldsymbol{w}).$$

where $B(p, q)$ is the beta function evaluated at $p, q > 0$. While the full conditional $p(\pi_l \mid -)$ has no closed form—even in the Dirichlet case—we can follow the same sampling strategy of the previous step, which has been proven to be effective even in this case.

Step 4. The atoms $\tilde{\phi}_{0h}$ are sampled independently for $h = 1, \dots, H$, proceeding as in the exchangeable setting and considering only within-cluster observations. The complexity of this sampling step depends both on the chosen kernel K and on the prior distribution P. However, if the kernel belongs to an exponential family, then one might adopt a conjugate prior distribution (Diaconis & Ylvisaker, 1979), and hence simplify the computations.

Algorithm 2: Steps of the Gibbs sampler

begin

Step 1. Assign each unit $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, L$, to a mixture component;

for l *from* 1 *to* d **do**

for i *from* 1 *to* $n^{(l)}$ **do**

Sample $G_i \in (1, \dots, H)$ independently from the categorical variable with probabilities

$$\mathbb{P}(G_{li} = h \mid -) = \frac{\pi_{lh} \mathcal{K}(Y_{li}; \tilde{\phi}_{0h})}{\sum_{h'=1}^{H} \pi_{lh'} \mathcal{K}(Y_{li}; \tilde{\phi}_{0h'})},$$

for every $h = 1, \dots, H$.

Step 2. Update the mixing parameters $\boldsymbol{\pi}_l$, for any $l = 1, \dots, d$;

for l *from* 1 *to* d **do**

Sample $\boldsymbol{\pi}_l$ independently from the full conditional having density proportional to

$$p(\boldsymbol{\pi}_l \mid -) \propto p(\boldsymbol{\pi}_l \mid \boldsymbol{\xi}_0^{(H)}) \prod_{h=1}^{H} \pi_{lh}^{n_{lh}},$$

where $n_{lh} = \sum_{i=1}^{n^{(l)}} \mathbb{1}(G_{li} = h)$, and where $\mathbb{1}(\cdot)$ denotes the indicator function.

Step 3. Sample the baseline mixing parameter $\boldsymbol{\xi}_0^{(H)}$ from the full conditional having density proportional to

$$p(\boldsymbol{\xi}_0^{(H)} \mid -) \propto p(\boldsymbol{\xi}_0^{(H)}) \prod_{l=1}^{d} p(\boldsymbol{\pi}_l \mid \boldsymbol{\xi}_0^{(H)}).$$

Step 4. Update the kernel parameters $\tilde{\phi}_{0h}$, for any $h = 1, \dots, H$;

for h *from* 1 *to* H **do**

Sample the kernel parameters $\tilde{\phi}_{0h}$ independently from the full conditional having density proportional to

$$p(\tilde{\phi}_{0h} \mid -) \propto p(\tilde{\phi}_{0h}) \prod_{(l,i) \in \mathcal{G}_h} \mathcal{K}(Y_{li}; \tilde{\phi}_{0h}),$$

where $\mathcal{G}_h = \{i = 1, \dots, n^{(l)}, l = 1, \dots, L : G_{li} = h\}$.

As a final remark, we notice that the deterministic truncation allows for the implementation of other well-established MCMC techniques, essentially because it shifts the original nonparametric formulation to a finite-dimensional problem, whose likelihood and prior distribution can be readily evaluated. As such, automatic tools like STAN (Carpenter et al., 2017) might be used for posterior inference.

5.5 Simulation study

To assess the empirical performance of model (5.7) and the associated Gibbs sampling algorithm, we conduct a simple simulation study. The target of this analysis is the comparison between the HDP and more general hierarchical processes in terms of inference on the clustering structure of the data.

We consider a total of $n = 2500$ observations divided in $L = 5$ different groups, each having a different sample size, precisely $(n^{(1)}, \dots, n^{(5)}) = (750, 50, 750, 200, 750)$. Within group, the simulated data are iid draws from a group-specific finite mixture of Gaussian distributions, whereas across groups they are independently sampled. The Gaussian mixtures densities were chosen so that different groups share some mixture components. In particular, there are a total of 7 latent Gaussian mixture components having mean parameters $(-2.5, -1.5, -1, 0, 1, 1.5, 2.5)$ and standard deviations $(1.2, 0.7, 0.25, 0.25, 0.25, 0.7, 1.2)$, which are split over the $L = 5$ groups, as reported in Table 5.1. For instance, the mixture component with 0 mean and standard deviation 0.25 is shared by all the groups. The mixing proportions are not uniform within groups nor equal across groups: this means, for example, that some mixture components are specific of two groups but they are not shared by the other three.

		Mixture component						
		1	2	3	4	5	6	7
Group	1	0.0	0.1	0.0	0.6	0.3	0.0	0.0
	2	0.1	0.0	0.0	0.5	0.4	0.0	0.0
	3	0.1	0.0	0.3	0.3	0.0	0.3	0.0
	4	0.0	0.2	0.2	0.5	0.0	0.1	0.0
	5	0.0	0.0	0.0	0.4	0.4	0.0	0.2

Table 5.1: True mixing proportions of the simulated data for each group $l = 1, \dots, 5$, and for each of the 7 mixture components.

In the hierarchical mixture model (5.7), we employ a Gaussian kernel $\mathcal{K}(y; \theta) = \mathcal{N}(y; \mu, \tau^{-1})$, and we choose a conditionally conjugate prior distribution for the parame-

ters (μ, τ), so that their baseline measure is

$$P(d\mu, d\tau) = P_1(d\mu)P_2(d\tau),$$

where P_1 is a Gaussian distribution with mean 0 and standard deviation 10, whereas P_2 is a Gamma distribution with parameters $(1, 1)$. To simplify our treatment, we decided not to place any hyperprior distribution on the parameters in P, although this further hierarchical layer could be easily handled with a straightforward modification of the blocked Gibbs sampler in Algorithm 2.

We fitted four different hierarchical mixture models to the same simulated dataset, for different choices of the jump intensity $\rho(s)$ and of the hyperparameters c, c_0 and σ_0, whose value are presented in Table 5.2. These models include: i) a hierarchical Dirichlet Process (HDP); ii) a hierarchical Dirichlet and Pitman-Yor process (HDP-PY); iii) a hierarchical 1/2-stable and Pitman-Yor process (HST-PY); iv) a hierarchical normalized inverse Gaussian and Pitman-Yor process (HIG-PY). Notice that in the 1/2-stable case the total mass parameter is irrelevant and therefore it was omitted. We fixed a common truncation level $H = 250$, which we found to be sufficiently large to guarantee a good approximation of the infinite hierarchical mixture model. Indeed, in Table 5.2 we also report the expected value of upper bound $\mathscr{R}_{lH}$, defined as in Theorem 5.1, which in the worst case scenario is approximately equal to 0.06.

Model	c	c_0	σ_0	Correlation	Expected # of clusters	$\mathbb{E}(\mathscr{R}_{lH})$	H
HDP	18	13	0	0.59	≈ 41	$< 10^{-6}$	250
HDP-PY	7	5	0.5	0.43	≈ 40	0.042	250
HST-PY	-	7	0.5	0.12	≈ 39	0.057	250
HIG-PY	2.5	2	0.5	0.50	≈ 40	0.020	250

Table 5.2: Hyperparameter settings for each hierarchical mixture model. The correlation coefficient is evaluated using Theorem 5.2. The expected number of cluster is obtained via Monte Carlo simulations, averaging over 100'000 values from the truncated prior. The expected value of the upper bound $\mathscr{R}_{lH}$, defined as in Theorem 5.1, is also reported.

The hyperparameters c, c_0 and σ_0 were selected so that peculiar characteristics of each model can be appreciated—especially compared to the HDP. In particular, the a priori expected number of cluster—obtained via Monte Carlo after averaging over 100'000 draws from the truncated prior in (5.5)—is centered approximately around 40, as reported in Table 5.2 and depicted in Figure 5.1. That is, we set on purpose the a priori expected number of cluster to be much higher than the true number of mixture components. An extensive description of the underlying clustering behaviors is beyond the aim of this thesis, and one can refer e.g. to Lijoi et al. (2007); De Blasi et al. (2015)

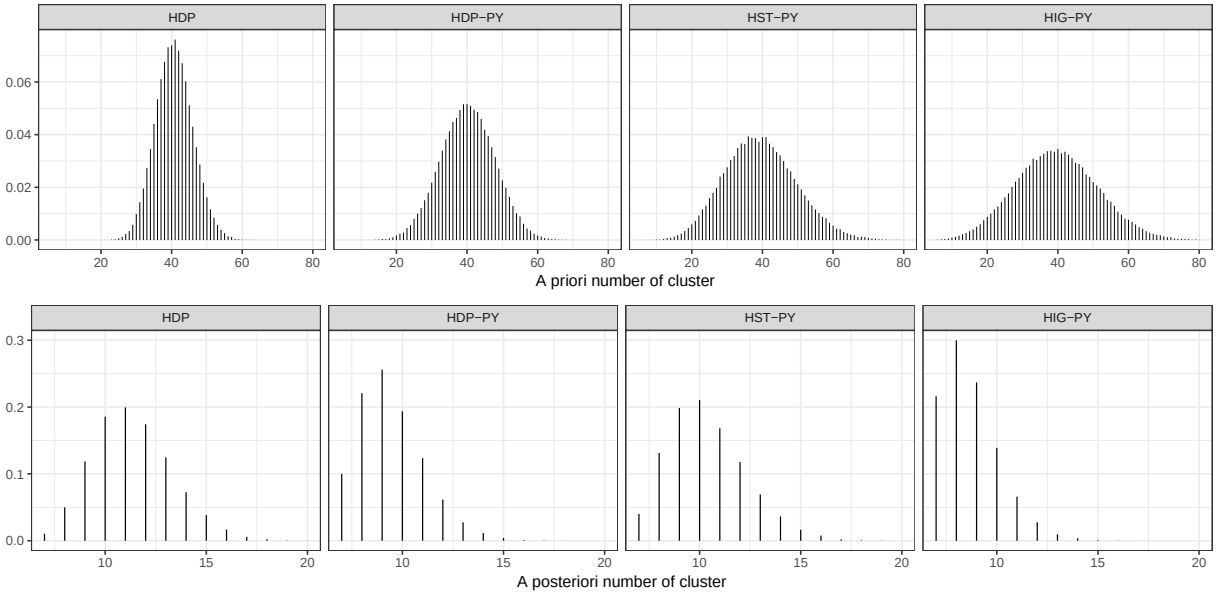


Figure 5.1: Top figures: a priori distribution of the number of cluster, based on 100′000 simulations from the truncated prior. Bottom figures: a posteriori distribution of the number of clusters, based on 20′000 MCMC draws. Both top and bottom figures refers to the models in Table 5.2.

in the exchangeable case and to Camerlenghi et al. (2019) in the partially exchangeable setting with hierarchical processes. To our purposes, it suffices to notice that the a priori distribution of the number of cluster is much "flatter"—i.e. less informative—in general hierarchical mixture models compared to the one of the HDP, as empirically evidenced in Figure 5.1. This is due to the stable parameter σ_0 in the Pitman-Yor specification, but also to the specific choice of jump measure ρ. For example, as mentioned in Section 5.2, the normalized inverse Gaussian distribution might be regarded as less informative compared to the Dirichlet, essentially leading to a flatter cluster configuration. Thus, we aim at showing that hierarchical models beyond the HDP might be more robust in identifying a suitable number of cluster components, especially in severely misspecified prior settings. This behavior was already noticed in Lijoi et al. (2007) for exchangeable data, and extend to the case of truncated hierarchical processes.

We run the chain for 200′000 iterations—after a burn-in period of 100′000 draws—and we thin the chain every 10 iterations, thus comprising a total of 20′000 posterior samples. The traceplots show good mixing and no evidence against convergence. As expected, the posterior distribution of the number of clusters—depicted in the bottom row of Figure 5.1—differs across models: in the HDP the values having highest probabilities are located between 10 and 12, whereas in all the other cases the posterior distribution is shifted towards 7, the correct number of mixture components. This is particularly evident in the HIG-PY case, whose a priori distribution was indeed the less informative.

5.6 Illustration

To further corroborate the practical relevancy of the proposed conditional algorithm, in this section we discuss an application of the NRMI-PY process to latent class analysis, in presence of qualitative covariates (Lazarsfeld & Henry, 1968; Goodman, 1974; Hagenaars & McCutcheon, 2002). As an illustration, we analyze the dataset presented in Stouffer & Toby (1951) and reported in Appendix 5.7. This has been the object of several investigations (e.g. Goodman, 1974, 1975; Clogg & Goodman, 1986; Hagenaars & McCutcheon, 2002) through latent class analysis, and from a frequentist perspective. The data are based on a short questionnaire completed by $n = 648$ undergraduate students at Harvard and Radcliffe, in 1950. Four ethical dilemmas, denoted as A,B,C and D, were posed to these students: a response coded as 1 represents a preference towards particularistic values, and viceversa 0 indicates a preference towards universalistic values. The questions were presented in slightly different forms to $L = 3$ independent and equally sized groups of students, meaning $n^{(1)} = n^{(2)} = n^{(3)} = 216$. The first group received each dilemma so that it refers to themselves (`EGO`), the second group so that it refers to a stranger (`SMITH`), and the third group so that it refers to a friend (`FRIEND`).

Clearly, some degree of agreement of the responses among different groups is expected, since the ethical dilemmas are the same. Nonetheless, the three groups should not be treated as identical, because the way in which each dilemma is posed might influence the response. Hence, within a Bayesian framework, the partial exchangeability assumption seem fairly natural in this setting, and it provides practical advantages. In particular, it allows to borrow information across groups and therefore to take stronger inferential conclusion compared to single-group analyses. Relying on the notation of Section 5.4, we assume that our observations are drawn from a collection of partially exchangeable binary random vectors $Y_{li} = (Y_{li1}, \ldots, Y_{li4}) \in \{0,1\}^4$, for $i = 1, \ldots, 216$ and $l = 1, 2, 3$, where the components of each Y_{li} refer to items A,B,C and D, respectively.

Latent class models are essentially mixture models in which, given a latent class (cluster) indicator G_{li}, the qualitative random variables $(Y_{li1}, \ldots, Y_{li4})$ are mutually independent. However, as noted in Dunson & Xing (2009), in this setting it is not straightforward to obtain a well-justified estimate for the number of mixture components. In addition, they proved that any probability mass function $\mathbb{P}(Y_{li} = y_{li})$ can be represented in terms of a latent class mixture model, when the number of mixture components is large enough. This leads us to assuming a mixture model with infinitely many components that we truncate up to the Hth term, thus obtaining a flexible and theoretically justified model for contingency tables. Hence, we can extend the approach of Dunson & Xing (2009) to the partially exchangeable setting, whereby d groups of

contingency tables are observed and a product of multinomial kernel in the NRMI-PY mixture model of equations (5.7)-(5.8) is specified. More precisely

$$Y_{li} \overset{\text{ind}}{\sim} \mathbb{P}(Y_{l1} = y_{l1}, \dots, Y_{l4} = y_{l4} \mid \boldsymbol{\pi}_l, \tilde{\boldsymbol{\phi}}_0) = \sum_{h=1}^{H} \pi_{lh} \left(\prod_{j=1}^{4} \tilde{\phi}_{0hj}^{y_{lj}} (1 - \tilde{\phi}_{0hj})^{1-y_{lj}} \right), \qquad (5.9)$$

independently for $i = 1, \dots, 216$, and $l = 1, 2, 3$, where $\boldsymbol{\pi}_l = (\pi_{l1}, \dots, \pi_{lH})$ has the same hierarchical prior distribution as in equation (5.7) and where $\tilde{\boldsymbol{\phi}}_0 = (\tilde{\phi}_{01}, \dots, \tilde{\phi}_{0H})$ is such that $\tilde{\phi}_{0h} = (\tilde{\phi}_{0h1}, \dots, \tilde{\phi}_{0h4})$ for any $h = 1, \dots, H$. As for the baseline measure P, we selected a uniform prior over the space $(0, 1)^4$, which is conditionally conjugate and hence facilitates posterior computations. A possible alternative specification for P consists of independent beta distributions, for $j = 1, \dots, 4$, which would still preserve conjugacy while allowing for the inclusion of more specific prior information in the model.

As for the prior setting of $\boldsymbol{\pi}_l$, we specified a hierarchical normalized inverse-Gaussian and stable process (NIG-ST), with hyperparameter settings $c = 1/2$, $c_0 = 0$ and $\sigma_0 = 3/10$ and with a truncation level $H = 150$. We achieve a good approximation of the infinite dimensional process, since $\mathbb{E}(\mathscr{R}_{lH}) < 10^{-4}$. Moreover, this specification induces high correlation a priori (to be meant in terms of the statement of Theorem 5.2) among the random probability measures $\tilde{p}_l^{(H)}$ (≈ 0.86): this is consistent with our prior belief that the same ethical dilemma should lead to very similar responses, regardless the way it was presented. The a priori expected number of cluster, evaluated via Monte Carlo, is approximately 3.9; however, the a priori distribution of the number of cluster is quite dispersed, consistently with the findings of previous analyses, which indeed do not provide a univocal recommendation about the number of latent components (Stouffer & Toby, 1951; Goodman, 1974, 1975; Clogg & Goodman, 1986).

Posterior inference was conducted via MCMC, using the blocked Gibbs sampler described in Section 5.4. We run the chain for 200′000 iterations—after a burn-in period of 50′000 draws— and we thin the chain every 10 iterations, thus comprising a total of 20′000 posterior samples. The traceplots show good mixing and no evidence against convergence.

In Clogg & Goodman (1986) it is suggested that these dilemmas can be ordered (D $\rightarrow$ C $\rightarrow$ B $\rightarrow$ A), according to a Guttman scale. This means, for instance, that a negative answer to C should imply, on average, also a negative response to dilemmas B and A. While such an assumption greatly simplifies the analysis, it seem clear from the subsequent results that it can only provide a reasonable approximation of the phenomenon. Indeed, we aim at studying for instance the conditional probability of $B = 1$ given that $C = 0, D = 1$ for each group of respondents, which should be close to zero under the Guttman scale assumption. As it will turn out, these probabilities not

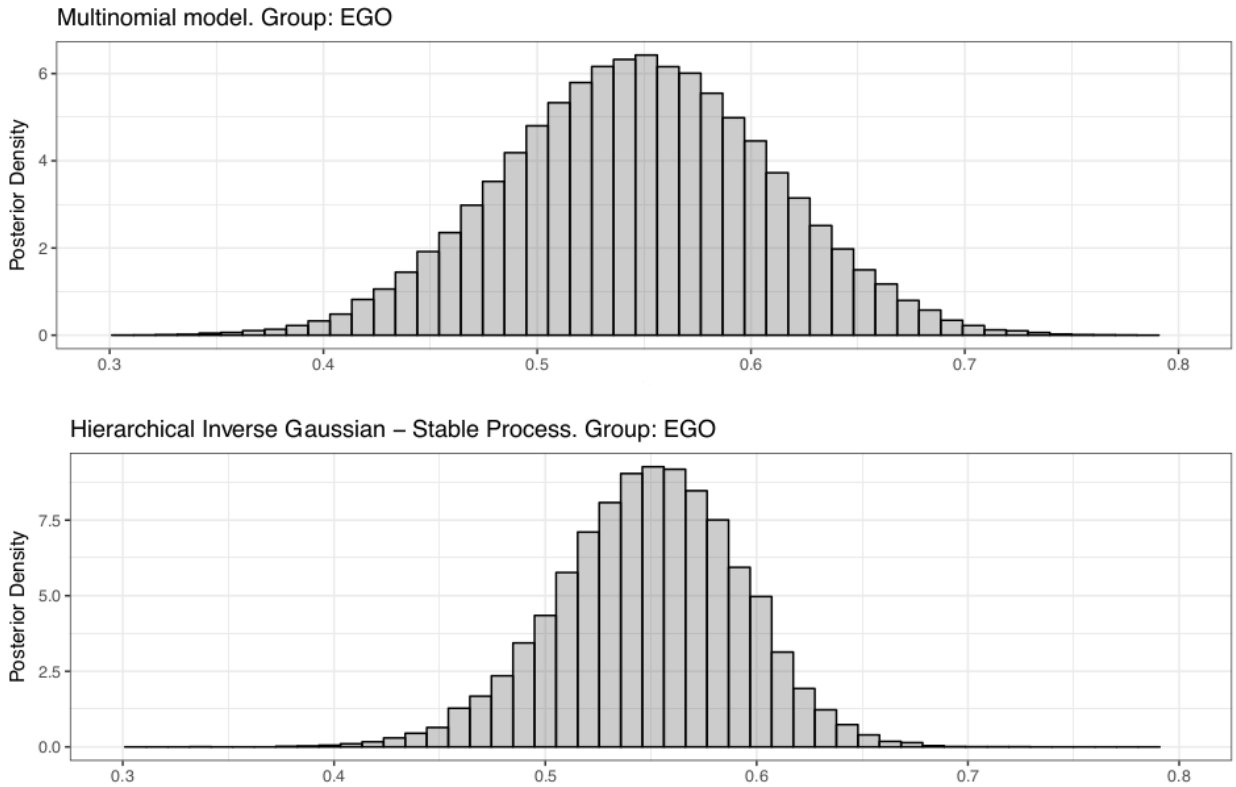


Figure 5.2: Posterior distribution of $\neg_1$, group: EGO. Top: posterior distribution of $\neg_1$ under the alternative multinomial formulation. Bottom: posterior distribution of $\neg_1$ under the HIG-ST mixture model of equation (5.9).

only are away from zero, but they are also significantly greater than 1/2. More formally, we are interested in the posterior distribution of

$$\neg_l = \mathbb{P}(Y_{l2} = 1 \mid Y_{l3} = 0, Y_{l4} = 1, \pi_l, \tilde{\phi}_0),$$

for any group $l = 1, 2, 3$, and given the data. Once more, we remark that the posterior distribution of each $\neg_l$ can be obtained only through conditional algorithms, which therefore represent the only possible choice to conduct inference in this specific application.

In Figure 5.2 we compare the posterior distribution of $\neg_1$ (EGO group) obtained using the aforementioned HIG-ST model in equation (5.9), with the posterior distribution of $\neg_1$ obtained under a much simpler multinomial model. More precisely, under the alternative model we treat the $2^4 = 16$ possible combination of responses as mutually exclusive categories. Among groups, we assume full heterogeneity—i.e. independence—whereas within group observations are conditionally iid draws from a multinomial distribution having 16 possible outcomes, and with a uniform prior. In both cases, the posterior distribution of $\neg_1$ is far from zero, suggesting that the Guttman scaling adopted in Clogg & Goodman (1986) should be interpreted with care. However, as apparent from

Figure 5.2, our HIG-ST model is able to substantially reduce the posterior uncertainty compared to the benchmark multinomial model. Essentially, this is due to two reasons: i) the latent class representation of equation (5.9), albeit flexible, allows for a parsimonious characterization of the distribution function $\mathbb{P}(Y_l = y_l)$ compared to the alternative multinomial formulation (Dunson & Xing, 2009); ii) in our hierarchical formulation we flexibly borrow information across the three groups, and this translates in a lower variability of the the posterior distribution. The posterior distributions of γ_2, γ_3 for other groups (SMITH, FRIEND), lead to similar conclusions.

5.7 Appendix

Proof of Theorem 5.1

Recall that $(\tilde{p}_1^{(\infty)}, \ldots, \tilde{p}_d^{(\infty)})$ comes from a hierarchical NRMI-PY process as in (5.1). Moreover, let $(\tilde{p}_1^{(H)}, \ldots, \tilde{p}_d^{(H)})$ be the hierarchical approximate process NRMI-PY defined in (5.5), with truncation level H. Then for any $A \in \mathscr{X}$, and exploiting representation (5.3), we have that almost surely

$$\begin{aligned}
\left|\tilde{p}_l^{(\infty)}(A) - \tilde{p}_l^{(H)}(A)\right| &= \left|\sum_{h=1}^{\infty} \pi_{lh}\delta_{\tilde{\phi}_{0h}}(A) - \left(\sum_{h=1}^{H-1} \pi_{lh}\delta_{\tilde{\phi}_{0h}}(A) + \left(1 - \sum_{h=1}^{H-1}\pi_{lh}\right)\delta_{\tilde{\phi}_{0H}}(A)\right)\right| \\
&= \left|\pi_{lH}\delta_{\tilde{\phi}_{0H}}(A) + \sum_{h>H}\pi_{lh}\delta_{\tilde{\phi}_{0h}}(A) - \left(1 - \sum_{h=1}^{H-1}\pi_{lh}\right)\delta_{\tilde{\phi}_{0H}}(A)\right| \\
&= \left|\delta_{\tilde{\phi}_{0H}}(A)\sum_{h>H}\pi_{lh} - \sum_{h>H}\pi_{lh}\delta_{\tilde{\phi}_{0h}}(A)\right| \\
&\leqslant \sum_{h>H}\pi_{lh} = \mathscr{R}_{lH}.
\end{aligned}$$

Note that $\sum_{h>H}\pi_{lh} \geqslant \sum_{h>H}\pi_{lh}\delta_{\tilde{\phi}_{0h}}(A)$ almost surely. Hence, if $\delta_{\tilde{\phi}_{0H}}(A) = 0$ a.s., then the last inequality easily follows, and the same holds true if $\delta_{\tilde{\phi}_{0H}}(A) = 1$ almost surely. Hence,

$$d_{TV}\left(\tilde{p}_l^{(\infty)}, \tilde{p}_l^{(H)}\right) = \sup_{A\in\mathscr{X}}\left|\tilde{p}_l^{(\infty)}(A) - \tilde{p}_l^{(H)}(A)\right| \leqslant \mathscr{R}_{lH} = \sum_{h>H}\pi_{lh},$$

almost surely. Moreover, notice that

$$\left(\sum_{h>H}\pi_{lh} \mid \xi_0\right) \sim \text{NID}\left(c\left(1 - \sum_{h=1}^{H}\xi_{0h}\right), c\sum_{h=1}^{H}\xi_{0h}; \rho\right),$$

from which it follows that the expected value is equal to

$$\mathbb{E}\left(\sum_{h>H}\pi_{lh}\right)=\mathbb{E}\left\{\mathbb{E}\left(\sum_{h>H}\pi_{lh}\mid\boldsymbol{\xi}_0\right)\right\}=\mathbb{E}\left(\sum_{h>H}\xi_{0h}\right)=\prod_{h=1}^{H}\frac{c_0+\sigma_0 h}{c_0+\sigma_0(h-1)+1}.$$

Now recall that $\mathscr{I}(c,\rho)=c\int_{\mathbb{R}_+}ue^{-c\psi(u)}\tau_2(u)du$ with $\tau_2(u)=\int_{\mathbb{R}_+}s^2e^{-us}\rho(s)ds$ and let $\mathscr{R}_{0H}=\sum_{h>H}\xi_{0h}$, then

$$\begin{aligned}\mathrm{Var}\left(\mathscr{R}_{lH}\right)&=\mathbb{E}\left\{\mathrm{Var}\left(\sum_{h>H}\pi_{lh}\mid\boldsymbol{\xi}_0\right)\right\}+\mathrm{Var}\left\{\mathbb{E}\left(\sum_{h>H}\pi_{lh}\mid\boldsymbol{\xi}_0\right)\right\}\\&=\mathscr{I}(c,\rho)\mathbb{E}(\mathscr{R}_{0H})+\{1-\mathscr{I}(c,\rho)\}\mathbb{E}\left(\mathscr{R}_{0H}^2\right)-\mathbb{E}(\mathscr{R}_{0H})^2,\end{aligned}$$

where $\mathbb{E}(\mathscr{R}_{0H})$ can be computed as before and

$$\mathbb{E}\left(\mathscr{R}_{0H}^2\right)=\prod_{h=1}^{H}\mathbb{E}\left\{(1-v_{0h})^2\right\}=\prod_{h=1}^{H}\frac{(c_0+\sigma_0 h)_2}{(c_0+\sigma_0(h-1)+1)_2}.$$

Proof of Theorem 5.2

First of all, notice that the expected value of the truncated Pitman-Yor process $\tilde{p}_{0,tr}^{(H)}\sim \mathrm{PY}_H(\sigma_0,c_0,P)$, for any $A\in\mathscr{B}(\Theta)$ and any $H\geqslant 1$, is equal to the baseline measure

$$\mathbb{E}(\tilde{p}_{0,tr}^{(H)}(A))=\sum_{h=1}^{H}\mathbb{E}(\xi_{0h})\mathbb{E}\{\delta_{\tilde{\phi}_{0h}}(A)\}=P(A)\sum_{h=1}^{H}\mathbb{E}(\xi_{0h})=P(A).$$

Moreover, one can show that

$$\mathrm{Var}(\tilde{p}_{0,tr}^{(H)}(A))=P(A)\{1-P(A)\}\sum_{h=1}^{H}\mathbb{E}(\xi_{0h}^2),$$

for any $H=1,2,\ldots$, and $A\in\mathscr{X}$. Define $\mathscr{I}_0(\sigma_0,c_0,H)=\sum_{h=1}^{H}\mathbb{E}(\xi_{0h}^2)$ and recall that $\mathscr{I}(c,\rho)=c\int_{\mathbb{R}_+}ue^{-c\psi(u)}\tau_2(u)du$ with $\tau_2(u)=\int_{\mathbb{R}_+}s^2e^{-us}\rho(s)ds$. From Proposition 1 of James et al. (2006), one has that $\mathrm{Var}(\tilde{p}_l^{(H)}(A)\mid\tilde{p}_{0,tr}^{(H)})=P(A)\{1-P(A)\}\mathscr{I}(c,\rho)$ for any $A\in\mathscr{B}(\Theta)$. Hence, for any $l=1,\ldots,L$,

$$\begin{aligned}\mathrm{Var}(\tilde{p}_l^{(H)}(A))&=\mathbb{E}(\mathrm{Var}(\tilde{p}_l^{(H)}(A)\mid\tilde{p}_{0,tr}^{(H)}))+\mathrm{Var}(\tilde{p}_{0,tr}^{(H)}(A))\\&=\mathscr{I}(c,\rho)\mathbb{E}[\tilde{p}_{0,tr}^{(H)}(A)\{1-\tilde{p}_{0,tr}^{(H)}(A)\}]+P(A)\{1-P(A)\}\mathscr{I}_0(\sigma_0,c_0,H)\\&=P(A)\{1-P(A)\}\{\mathscr{I}(c,\rho)-\mathscr{I}(c,\rho)\mathscr{I}_0(\sigma_0,c_0,H)+\mathscr{I}_0(\sigma_0,c_0,H)\}.\end{aligned}$$

Moreover, following Camerlenghi et al. (2019, Appendix A.1), for any $l \neq l'$

$$\text{Cov}\{\tilde{p}_l^{(H)}(A), \tilde{p}_{l'}^{H}(A)\} = \text{Var}\{\tilde{p}_{0,tr}^{(H)}(A)\} = P(A)\{1 - P(A)\}\mathscr{I}_0(\sigma_0, c_0, H),$$

from which it follows that

$$\text{Corr}\{\tilde{p}_l^{(H)}(A), \tilde{p}_{l'}^{H}(A)\} = \frac{\mathscr{I}_0(\sigma_0, c_0, H)}{\mathscr{I}(c, \rho) + \mathscr{I}_0(\sigma_0, c_0, H)(1 - \mathscr{I}(c, \rho))}.$$

It remains to find the explicit formulation of $\mathscr{I}_0(\sigma_0, c_0, H)$, being equal to

$$\begin{aligned}
\mathscr{I}_0(\sigma_0, c_0, H) &= \sum_{h=1}^{H} \mathbb{E}(\xi_{0h}^2) = \sum_{h=1}^{H} \mathbb{E}\left\{\nu_{0h}^2 \prod_{l=1}^{h-1}(1-\nu_{0l})^2\right\} \\
&= \sum_{h=1}^{H} \mathbb{E}(\nu_{0h}^2) \prod_{l=1}^{h-1} \mathbb{E}\left\{(1-\nu_{0l})^2\right\} \\
&= \sum_{h=1}^{H-1} \left\{ \frac{(1-\sigma_0)_2}{(1+c_0+(h-1)\sigma_0)_2} \left(\prod_{l=1}^{h-1} \frac{(c_0+l\sigma_0)_2}{(1+c_0+(l-1)\sigma_0)_2} \right) \right\} \\
&\qquad + \left(\prod_{l=1}^{H-1} \frac{(c_0+l\sigma_0)_2}{(1+c_0+(l-1)\sigma_0)_2} \right).
\end{aligned}$$

Notice that all the above results hold also for the infinite case, having replaced $\mathscr{I}_0(\sigma_0, c_0, H)$ with its limit $\mathscr{I}_0(\sigma_0, c_0)$, so that

$$\lim_{H \to +\infty} \mathscr{I}_0(\sigma_0, c_0, H) = \mathscr{I}_0(\sigma_0, c_0) = \mathbb{E}\left(\sum_{h=1}^{\infty} \xi_{0h}^2\right) = \sum_{h=1}^{\infty} \mathbb{E}\left(\xi_{0h}^2\right) = \frac{1-\sigma_0}{1+c_0},$$

where the last equality follows for instance from Ishwaran & James (2001, Appendix A.2).

Dataset

We report in Table 5.3 the dataset used in the illustrative analysis of Section 5.6 and originally presented in Stouffer & Toby (1951).

A	B	C	D	EGO	SMITH	FRIEND
0	0	0	0	42	37	35
0	0	0	1	23	31	17
0	0	1	0	6	6	9
0	0	1	1	25	15	26
0	1	0	0	6	5	3
0	1	0	1	24	29	27
0	1	1	0	7	6	3
0	1	1	1	38	25	32
1	0	0	0	1	2	3
1	0	0	1	4	4	5
1	0	1	0	1	3	2
1	0	1	1	6	4	5
1	1	0	0	2	3	0
1	1	0	1	9	23	20
1	1	1	0	2	3	3
1	1	1	1	20	20	26
Total				216	216	216

Table 5.3: The Stouffer & Toby (1951) dataset. We report the frequencies for each possible combination of the $2^4 = 16$ responses, divided over the the three groups `EGO`, `SMITH` and `FRIEND`.

Chapter 6

Computational advances for logit stick-breaking priors

6.1 Summary

The chapter is organized as follows. In Section 6.2 we introduce the logit stick-breaking prior process (LSBP) and we formalize its sequential characterization. In Section 6.3 three computational routines for the LSBP are derived, namely a Gibbs sampling, an EM algorithm and a variational Bayes algorithm . All these methods are based on the sequential representation and on the Pólya-gamma data-augmentation. Theoretical developments about the Pólya-gamma data-augmentation will be presented in Chapter 7. In Section 6.4 these methodologies are illustrated in a toxicological application.

6.2 Logit stick-breaking prior

This section presents a formal construction of the LSBP via *continuation-ratio* logistic regressions. As a natural extension of model (1.7), we consider the general class of predictor-dependent infinite mixture models

$$\int_{\Theta} \mathcal{K}_{\boldsymbol{x}}(y;\theta)\tilde{p}_{\boldsymbol{x}}(d\theta) = \sum_{h=1}^{\infty} \xi_h(\boldsymbol{x})\mathcal{K}_{\boldsymbol{x}}(y;\tilde{\phi}_h), \tag{6.1}$$

where $\xi_h(\boldsymbol{x}) = \nu_h(\boldsymbol{x})\prod_{l=1}^{h-1}\{1-\nu_l(\boldsymbol{x})\}$ are predictor-dependent mixing probabilities having a *stick-breaking representation*, whereas $\mathcal{K}_{\boldsymbol{x}}(y;\theta)$ denotes a predictor-dependent kernel, indexed by the parameters θ and the covariates.

Let us first consider an equivalent formulation of the predictor-dependent mixture model in (6.1). In particular, following standard hierarchical representations of mixture models, independent samples $Y_1,\ldots,Y_n$ of the variable with density function displayed

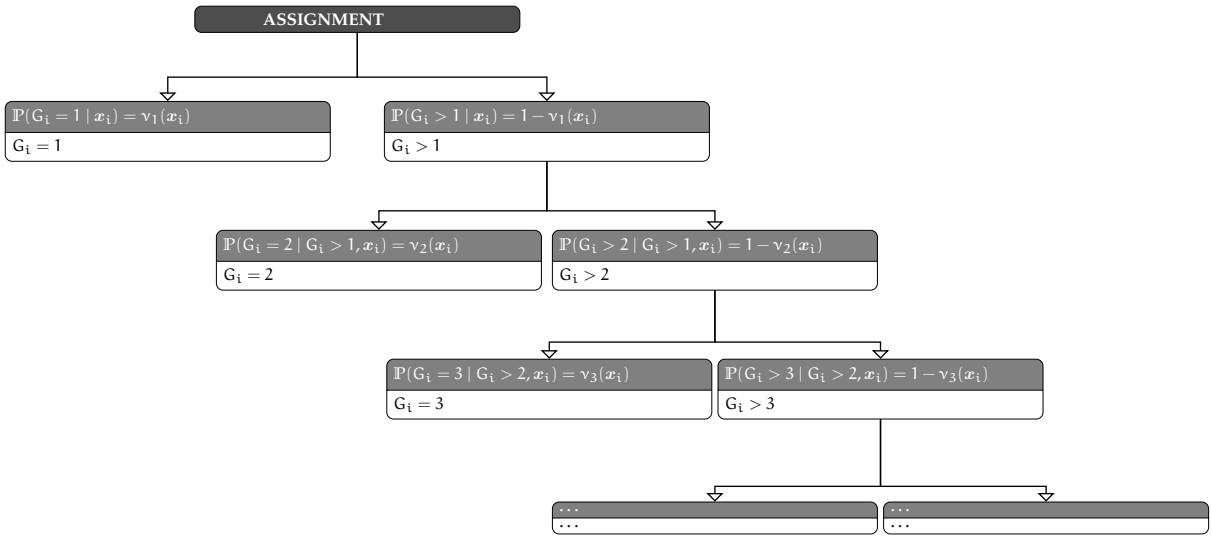


Figure 6.1: Representation of the sequential mechanism to sample G_i.

in (6.1), can be obtained from

$$(Y_i \mid G_i = h, \boldsymbol{x}_i) \overset{\text{ind}}{\sim} \mathcal{K}_{\boldsymbol{x}_i}(y; \tilde{\phi}_h), \quad \mathbb{P}(G_i = h \mid \boldsymbol{x}_i) = \nu_h(\boldsymbol{x}_i) \prod_{l=1}^{h-1} \{1 - \nu_l(\boldsymbol{x}_i)\}, \tag{6.2}$$

for each unit $i = 1, \ldots, n$, where $\tilde{\phi}_h \overset{\text{iid}}{\sim} P$, whereas $G_i \in \mathbb{N}$ is the categorical variable denoting the mixture component associated with the ith unit. According to (6.2), every G_i has probability mass function $p(G_i \mid \boldsymbol{x}_i) = \prod_{h=1}^{\infty} \pi_h(\boldsymbol{x}_i)^{\mathbb{1}(G_i = h)}$, where $\mathbb{1}(\cdot)$ denotes the indicator function. Hence, re-writing $\{\nu_h(\boldsymbol{x}_i)\}_{h \geqslant 1}$ as a function of the mixing probabilities $\{\xi_h(\boldsymbol{x}_i)\}_{h \geqslant 1}$ via

$$\nu_h(\boldsymbol{x}_i) = \frac{\xi_h(\boldsymbol{x}_i)}{1 - \sum_{l=1}^{h-1} \pi_l(\boldsymbol{x}_i)} = \frac{\mathbb{P}(G_i = h \mid \boldsymbol{x}_i)}{\mathbb{P}(G_i > h - 1 \mid \boldsymbol{x}_i)}, \qquad h \geqslant 1, \tag{6.3}$$

allows to interpret each $\nu_h(\boldsymbol{x}_i)$ as the probability of being allocated to component h, conditionally on the event of surviving to the previous $1, \ldots, h-1$ components, namely $\nu_h(\boldsymbol{x}_i) = \mathbb{P}(G_i = h \mid G_i > h - 1, \boldsymbol{x}_i)$. This result provides a formal characterization of the stick-breaking construction in (6.2) as the continuation-ratio parameterization (Tutz, 1991) of the probability mass function for each component membership variable G_i. This connection with the literature on sequential inference for categorical data is common to all the stick-breaking priors—as mentioned also by Rodriguez & Dunson (2011) in the probit case.

As we will describe in Section 6.3, the above result facilitates the implementation of different routine-use algorithms in Bayesian inference, and provides a simple generative process for each G_i. In particular, as illustrated in Figure 6.1, in the first step of this

continuation-ratio generative mechanism, unit i is either assigned to the first component with probability $\nu_1(\boldsymbol{x}_i)$ or to one of the others with complement probability. If $G_i = 1$ the process stops, otherwise it continues considering the reduced set $\{h : h > 1\}$. A generic step h is reached if i has not been assigned to $1, \ldots, h-1$, and the decision at this step will be to either allocate i to component h with probability $\nu_h(\boldsymbol{x}_i)$, or to one of the subsequent components with probability $1 - \nu_h(\boldsymbol{x}_i)$, conditioned on $G_i > h-1$. Based on this representation, the assignment indicator $\zeta_{ih} = \mathbb{1}(G_i = h)$ can be expressed, for every unit $i = 1, \ldots, n$, as

$$\zeta_{ih} = z_{ih} \prod_{l=1}^{h-1} (1 - z_{il}), \qquad h \geqslant 1, \tag{6.4}$$

where the generic z_{ih}, $h \geqslant 1$, is a Bernoulli variable $(z_{ih} \mid \boldsymbol{x}_i) \sim \text{Bern}\{\nu_h(\boldsymbol{x}_i)\}$ denoting the decision at the hth step to either allocate i to component h or to one of the subsequents. Hence, according to (6.4), the sampling of each G_i, under the predictor-dependent stick-breaking representation for each $\xi_h(\boldsymbol{x}_i)$ in (6.2), can be reformulated as a set of sequential Bernoulli choices with natural parameters $\eta_h(\boldsymbol{x}_i) = \text{logit}\{\nu_h(\boldsymbol{x}_i)\} = \log[\nu_h(\boldsymbol{x}_i)/\{1 - \nu_h(\boldsymbol{x}_i)\}]$ under an exponential family representation. Hence, we can write

$$\xi_h(\boldsymbol{x}_i) = \frac{\exp\{\eta_h(\boldsymbol{x}_i)\}}{1 + \exp\{\eta_h(\boldsymbol{x}_i)\}} \prod_{l=1}^{h-1} \left[\frac{1}{1 + \exp\{\eta_l(\boldsymbol{x}_i)\}} \right], \qquad h \geqslant 1, \tag{6.5}$$

allowing each $\eta_h(\boldsymbol{x}_i)$ to be explicitly interpreted as the log-odds of the probability of being allocated to component h, conditionally on the event of surviving to the first $1, \ldots, h-1$ components. This result might be helpful in driving prior specification for the stick-breaking weights, while allowing recent computational advances in Bayesian logistic regression (Polson et al., 2013) to be inherited in our density regression problem.

To conclude our Bayesian representation, we require priors for the log-odds $\eta_h(\boldsymbol{x}_i)$, $h \geqslant 1$ in the continuation-ratio logistic regressions. A natural choice, which is consistent with classical generalized linear models (e.g. Nelder & Wedderburn, 1972), is to define $\eta_h(\boldsymbol{x}_i)$ as a linear combination of selected functions of the covariates $\boldsymbol{\mathcal{B}}_2(\boldsymbol{x}_i) = \{\mathcal{B}_{21}(\boldsymbol{x}_i), \ldots, \mathcal{B}_{2M_2}(\boldsymbol{x}_i)\}^\intercal$ and consider Gaussian priors for the coefficients, thus obtaining

$$\eta_h(\boldsymbol{x}_i) = \boldsymbol{\mathcal{B}}_2(\boldsymbol{x}_i)^\intercal \boldsymbol{\gamma}_h, \quad \text{with } \boldsymbol{\gamma}_h \sim \mathcal{N}_{M_2}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma), \qquad h \geqslant 1. \tag{6.6}$$

Although the linearity assumption in (6.6) may seem restrictive, note that flexible formulations for $\eta_h(\boldsymbol{x}_i)$, including regression via splines and Gaussian processes, induce

linear relations in the coefficients. Moreover, as we will outline in Section 6.3, the linearity assumption simplifies computations, while inducing a logistic-normal prior for each $v_h(\boldsymbol{x}_i)$. Although such a prior can closely approximate Dirichlet distributions (Aitchison & Shen, 1980), the logit stick-breaking does not induce beta distributed stick-breaking weights, and therefore it cannot be included in the class discussed by Ishwaran & James (2001). However, one can easily adapt the theoretical results in Rodriguez & Dunson (2011) to our logit link. For example, the infinite summation of the mixing weights is such that $\sum_{h=1}^{\infty} \xi_h(\boldsymbol{x}_i) = 1$ almost surely for any $\boldsymbol{x} \in \mathbb{X}$; see the Appendix for details. Moreover, the LSBP is highly similar in its probabilistic nature and properties to other popular predictor-dependent stick-breaking constructions. In particular, PSBP can be approximated by LSBP, and viceversa, up to a simple transformation of the prior for each γ_h. This is a natural consequence of the well known relationship between the probit and the logit function (Amemiya, 1981), since the mapping $\{1 + \exp(-\mathcal{B}_2(\boldsymbol{x})^\intercal \gamma_h)\}^{-1}$ can be roughly approximated by $\Phi\{\mathcal{B}_2(\boldsymbol{x})^\intercal \gamma_h \sqrt{\pi/8}\}$. This is summarized in Remark 6.1.

Remark 6.1. The logit stick-breaking prior in (6.6), can be approximated by a probit stick-breaking process $v_h(\boldsymbol{x}) \approx \Phi\{\mathcal{B}_2(\boldsymbol{x})^\intercal \bar{\gamma}_h\}$, with $\bar{\gamma}_h = \gamma_h \sqrt{\pi/8} \sim \mathcal{N}_{M_2}\{\sqrt{\pi/8}\boldsymbol{\mu}_\gamma, (\pi/8)\boldsymbol{\Sigma}_\gamma\}$, for every $\boldsymbol{x} \in \mathbb{X}$ and $h \geqslant 1$.

Hence, a researcher considering a PSBP could perform approximate inference leveraging our algorithms, after rescaling the prior for each γ_h by $\sqrt{8/\pi}$. Moreover, this link suggests that the $\mathcal{O}(\log n)$ growth of the number of clusters found in empirical studies on the PSBP, should hold also for LSBP.

6.3 Bayesian computational methods

Although the LSBP and the associated computational procedures apply to a wider set of dependent mixture models and kernels, we focus, for the sake of clarity, on the general class of predictor-dependent infinite mixtures of Gaussians

$$\int \mathcal{N}(y; \mathcal{B}_1(\boldsymbol{x})^\intercal \boldsymbol{\beta}, \tau) \tilde{p}_{\boldsymbol{x}}(d\boldsymbol{\beta}, d\tau) = \sum_{h=1}^{\infty} \xi_h(\boldsymbol{x}) \mathcal{N}(y; \mathcal{B}_1(\boldsymbol{x})^\intercal \tilde{\boldsymbol{\beta}}_h, \tilde{\tau}_h^{-1}), \tag{6.7}$$

where $\tilde{\tau}_h = \tilde{\sigma}_h^2$ is the precision parameter, whereas $\tilde{\boldsymbol{\beta}}_h = (\tilde{\beta}_{1h}, \ldots, \tilde{\beta}_{Mh})^\intercal$ denotes a vector of coefficients linearly related to selected functions of the observed predictors $\mathcal{B}_1(\boldsymbol{x}) = \{\mathcal{B}_{11}(\boldsymbol{x}), \ldots, \mathcal{B}_{1M_1}(\boldsymbol{x})\}^\intercal$. Formulation (6.7) provides a flexible construction (Barrientos et al., 2012; Pati et al., 2013), and is arguably the most widely used in Bayesian density regression. As mentioned in Section 6.1, we provide here a detailed derivation of three computational methods for Bayesian density regression under model (6.7), with logit stick-breaking prior (6.6) for the mixing weights. In particular, we consider

a Gibbs sampler converging to the exact posterior, an expectation-maximization (EM) algorithm for point estimation, and a mean-field variational Bayes (VB) approximation for scalable posterior inference. The algorithms associated with these methods are available at https://github.com/tommasorigon/LSBP, along with the code to reproduce the application in Section 6.4.

In the classical predictor-independent mixture of Gaussians framework, these computational methods are closely related, and relevant connections can be drawn also with k-means and Bayesian k-means algorithms (Bishop, 2006; Kurihara & Welling, 2009). A summary of these relations is depicted in Figure 1 of Kurihara & Welling (2009). Broadly speaking, these strategies differ in how they handle unknown parameters and the involved latent quantities, either through maximization or by taking expectations. These connections are paralleled in the LSBP model, although our focus is mainly on Gibbs sampling, EM and VB.

Before providing a detailed derivation of these different algorithms, we first study a truncated version of the random probability measure $\tilde{p}_{\boldsymbol{x}}$, which will be employed as an approximation of the infinite process. Indeed, although Gibbs samplers for infinite representations are available (Kalli et al., 2011), developing EM and VB algorithms is not straightforward. In line with Rodriguez & Dunson (2011) and Ren et al. (2011), we develop detailed routines based on a finite representation. In particular, we model the first $H-1$ weights $\nu_1(\boldsymbol{x}), \ldots, \nu_{H-1}(\boldsymbol{x})$ and let $\nu_H(\boldsymbol{x}) = 1$ for any $\boldsymbol{x} \in \mathbb{X}$, so that $\sum_{h=1}^{H} \xi_h(\boldsymbol{x}) = 1$. Based on Theorem 6.1 below, this choice provides an accurate approximation of the infinite representation for sufficiently large truncations H.

Theorem 6.1. *For a sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\intercal$ with covariates $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\intercal$, let*

$$m_{\boldsymbol{X}}^{(H)}(\boldsymbol{Y}) = \mathbb{E}\left\{ \prod_{i=1}^{n} \sum_{h=1}^{H} \xi_h(\boldsymbol{x}) \mathcal{N}(Y_i; \mathcal{B}_1(\boldsymbol{x})^\intercal \tilde{\boldsymbol{\beta}}_h, \tilde{\tau}_h^{-1}) \right\},$$

be the marginal joint density arising from a truncated LSBP prior with H components, and define with $m_{\boldsymbol{X}}^{(\infty)}(\boldsymbol{Y})$ the same quantity in the infinite case. Note that in the above formula the expectation is taken with respect to the LSBP prior law. Then

$$\| m_{\boldsymbol{X}}^{(H)}(\boldsymbol{Y}) - m_{\boldsymbol{X}}^{(\infty)}(\boldsymbol{Y}) \|_1 \leqslant 4 \sum_{i=1}^{n} [1 - \mathbb{E}\{\nu_1(\boldsymbol{x}_i)\}]^{H-1},$$

where $\| \cdot \|_1$ denotes the L^1–norm.

According to Theorem 6.1, for fixed sample size n and covariates $\boldsymbol{X}$, the L^1 distance between $m_{\boldsymbol{X}}^{(H)}(\boldsymbol{Y})$ and $m_{\boldsymbol{X}}^{(\infty)}(\boldsymbol{Y})$ vanishes as $H \to \infty$, implying that the marginal density $m_{\boldsymbol{X}}^{(H)}(\boldsymbol{Y})$ converges to $m_{\boldsymbol{X}}^{(\infty)}(\boldsymbol{Y})$. This rate of decay is exponential in H, and therefore

the number of components does not have to be very large in practice to accurately approximate the infinite representation, thus motivating computational methods based on truncated versions.

6.3.1 MCMC via Gibbs sampling

In deriving a Gibbs sampler for model (6.7) we focus on a dependent mixture of Gaussians with fixed H, and exploit the hierarchical representation (6.2) along with the continuation-ratio characterization of the logit stick-breaking prior, given in Section 6.2. Under these constructions, the joint law for the augmented model (6.2) and its parameters becomes

$$p(\gamma)p(\tilde{\beta})p(\tilde{\tau})\prod_{i=1}^{n}\prod_{h=1}^{H}\mathcal{N}(Y_i;\mathcal{B}_1(\boldsymbol{x}_i)^\intercal\tilde{\beta}_h,\tilde{\tau}_h^{-1})^{\mathbb{1}(G_i=h)}\prod_{h=1}^{H-1}v_h(\boldsymbol{x}_i)^{\mathbb{1}(G_i=h)}\{1-v_h(\boldsymbol{x}_i)\}^{\mathbb{1}(G_i>h)}, \tag{6.8}$$

with $p(\gamma)p(\tilde{\beta})p(\tilde{\tau})=\prod_{h=1}^{H-1}p(\gamma_h)\prod_{h=1}^{H}p(\tilde{\beta}_h)p(\tilde{\tau}_h)$ denote the prior laws of the parameters comprising γ, β and $\tilde{\tau}$. As is clear from (6.8), given $\boldsymbol{G}=(G_1,\ldots,G_n)$, sampling of $\tilde{\beta}_h$ and $\tilde{\tau}_h$, for $h=1,\ldots,H$, requires standard methods for Gaussian linear regression within each mixture component, as long as conditionally conjugate priors $\tilde{\beta}_h\sim\mathcal{N}_{M_1}(\boldsymbol{\mu}_\beta,\boldsymbol{\Sigma}_\beta)$ and $\tilde{\tau}_h\sim\text{GA}(a_\sigma,b_\sigma)$, or normal-gammas for the pair $(\tilde{\beta}_h,\tilde{\tau}_h)$, are employed. Here we focus on the first choice to keep notation more compact.

The updating of the γ_h parameters, for $h=1,\ldots,H-1$, relies instead on a set of separate Bayesian logistic regressions with responses $z_{ih}=1$ when $G_i=h$ and $z_{ih}=0$ if $G_i>h$, for those units i having $G_i>h-1$, thus allowing parallel sampling from the full-conditional of each γ_h. Adapting results from the recent Pólya-gamma data augmentation scheme (Polson et al., 2013) to our statistical model, these updatings can be easily accomplished by noticing that $v_h(\boldsymbol{x}_i)^{z_{ih}}\{1-v_h(\boldsymbol{x}_i)\}^{1-z_{ih}}=\int_{\mathbb{R}_+}p_{\boldsymbol{x}_i}(z_{ih})p_{\boldsymbol{x}_i}(\omega_{ih})d\omega_{ih}$, with laws $p_{\boldsymbol{x}_i}(z_{ih})$ and $p_{\boldsymbol{x}_i}(\omega_{ih})$ defined as

$$p_{\boldsymbol{x}_i}(z_{ih})=\frac{0.5\exp\{(z_{ih}-0.5)\mathcal{B}_2(\boldsymbol{x}_i)^\intercal\gamma_h\}}{\cosh\{0.5\mathcal{B}_2(\boldsymbol{x}_i)^\intercal\gamma_h\}},\quad p_{\boldsymbol{x}_i}(\omega_{ih})=\frac{\exp[-0.5\{\mathcal{B}_2(\boldsymbol{x}_i)^\intercal\gamma_h\}^2\omega_{ih}]p(\omega_{ih})}{[\cosh\{0.5\mathcal{B}_2(\boldsymbol{x}_i)^\intercal\gamma_h\}]^{-1}}, \tag{6.9}$$

for every $i:G_i>h-1$ and $h=1,\ldots,H-1$. In (6.9), $p_{\boldsymbol{x}_i}(\omega_{ih})$ and $p(\omega_{ih})$ are the density functions of the Pólya-gamma random variables $\text{PG}\{1,\mathcal{B}_2(\boldsymbol{x}_i)^\intercal\gamma_h\}$, and $\text{PG}(1,0)$, respectively. Hence, based on (6.9), the contribution to the augmented likelihood for each pair (z_{ih},ω_{ih}) is proportional to a Gaussian kernel for transformed data $(z_{ih}-0.5)/\omega_{ih}$, provided that $p_{\boldsymbol{x}_i}(z_{ih})p_{\boldsymbol{x}_i}(\omega_{ih})\propto\exp[(z_{ih}-0.5)\mathcal{B}_2(\boldsymbol{x}_i)^\intercal\gamma_h-0.5\{\mathcal{B}_2(\boldsymbol{x}_i)^\intercal\gamma_h\}^2\omega_{ih}]$. This allows conditionally conjugate updating steps for each γ_h under a classical Bayesian linear regression framework. Refer to Choi & Hobert (2013); Wang & Roy (2018a,b) for

further theoretical properties of the Pólya-gamma scheme. Finally, note that in (6.9), the latent indicators z_{ih} and the Pólya-gamma random variables ω_{ih} are conditionally independent given the coefficients γ_h for $i : G_i > h-1$. This is in contrast with the data augmentation underlying the PSBP, which would lead to more complex calculations, especially in the EM and VB algorithms discussed in Sections 6.3.2 and 6.3.3.

The detailed steps of the Gibbs sampler for the truncated representation of model (6.7) are outlined in Algorithm 3. In this routine, $\boldsymbol{B}_{1h}$ and $\boldsymbol{B}_{2h}$ denote the $n_h \times M_1$ and the $\bar{n}_h \times M_2$ predictor matrices in (6.7) and (6.6) having row entries $\mathcal{B}_1(\boldsymbol{x}_i)^\intercal$ and $\mathcal{B}_2(\boldsymbol{x}_i)^\intercal$, for only those statistical units i such that $G_i = h$ and $G_i > h-1$, respectively. We shall also emphasize that ***Step 1*** can be run in parallel across units $i = 1, \ldots, n$, whereas parallel computing for the different mixture components can be easily implemented in ***Step 2***, ***Step 3*** and ***Step 4***.

6.3.2 EM algorithm

In high-dimensional studies, the Gibbs sampler described in Section 6.3.1 could face computational bottlenecks. If a point estimate of model (6.7) is the main quantity of interest, for example for prediction purposes, one possibility is to rely on a more efficient procedure specifically designed for this goal, such as the EM (Dempster et al., 1977). The implementation of a simple EM for a finite representation of model (6.7) under the LSBP prior benefits from the Pólya-gamma data augmentation, which has analytical expectation and allows direct maximization within a Gaussian linear regression framework. Note that, although the EM algorithm is commonly implemented for maximum likelihood estimation, it can be modified to estimate posterior modes (e.g. Dempster et al., 1977).

The proposed EM in Algorithm 4 alternates between a maximization step for the parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\boldsymbol{\tau}})$ and an expectation step for the augmented data $(\boldsymbol{\zeta}_i, \bar{\boldsymbol{\omega}}_i)$, $i = 1, \ldots, n$, with $\boldsymbol{\zeta}_i = \{\zeta_{i1} = \mathbb{1}(G_i = 1), \ldots, \zeta_{iH} = \mathbb{1}(G_i = H)\}^\intercal$ the vector of binary indicators denoting the membership to a mixture component, and $\bar{\boldsymbol{\omega}}_i = (\bar{\omega}_{i1}, \ldots, \bar{\omega}_{iH-1})^\intercal$ the corresponding Pólya-gamma augmented data. Although this data augmentation parallels the one described for the Gibbs sampler, we adopt a slightly different notation for the Pólya-gamma random variables $\bar{\omega}_{ih}$, to emphasize that we are considering n units, and not only those for which the cluster indicators $G_i > h-1$. Indeed, in line with the EM rationale, we do not condition on the membership indicators and on the Pólya-gamma latent random variables, but we rather take expectations with respect to their conditional distributions. For the same reason, in this case we work directly with the component indicator variables $\boldsymbol{\zeta}_i$ instead of the binary vectors $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH-1})^\intercal$ in (6.4).

Algorithm 3: Steps of the Gibbs sampler for the LSBP

begin

Step 1. Assign each unit $i = 1, \dots, n$ to a mixture component $h = 1, \dots, H$;

for i from 1 *to* n **do**

Sample $G_i \in \{1, \dots, H\}$ from the categorical variable with probabilities

$$\mathbb{P}(G_i = h \mid -) \propto \left[\nu_h(\boldsymbol{x}_i) \prod_{l=1}^{h-1} \{1 - \nu_l(\boldsymbol{x}_i)\} \right] \mathcal{N}(Y_i; \boldsymbol{\mathcal{B}}_1(\boldsymbol{x}_i)^\intercal \tilde{\boldsymbol{\beta}}_h, \tilde{\tau}_h^{-1}]],$$

for every $h = 1, \dots, H$.

Step 2. Update the parameters $\boldsymbol{\gamma}_h$ for $h = 1, \dots, H-1$ exploiting the continuation-ratio representation and the results from the Pólya-gamma data augmentation in (6.9);

for h from 1 *to* $H-1$ **do**

for *every* i such that $G_i > h-1$ **do**

Sample the Pólya-gamma data ω_{ih} from $(\omega_{ih} \mid -) \sim \text{PG}\{1, \boldsymbol{\mathcal{B}}_2(\boldsymbol{x}_i)^\intercal \boldsymbol{\gamma}_h\}$.

Given the Pólya-gamma data, update $\boldsymbol{\gamma}_h$ from the full conditional

$$(\boldsymbol{\gamma}_h \mid -) \sim \mathcal{N}_{M_2}(\boldsymbol{\mu}_{\boldsymbol{\gamma}_h}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_h}),$$

$\boldsymbol{\mu}_{\boldsymbol{\gamma}_h} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_h}\{\boldsymbol{B}_{2h}^\intercal \boldsymbol{\aleph}_h + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\gamma}}\}$, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_h} = \{\boldsymbol{B}_{2h}^\intercal \text{diag}(\omega_{i1}, \dots, \omega_{i\bar{n}_h}) \boldsymbol{B}_{2h} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\}^{-1}$, where $\boldsymbol{\aleph}_h = (z_{i1} - 0.5, \dots, z_{i\bar{n}_h} - 0.5)^\intercal$, with $z_{ih} = 1$ if $G_i = h$ and $z_{ih} = 0$ if $G_i > h$.

Step 3. Update the kernel parameters $\boldsymbol{\beta}_h$, $h = 1, \dots, H$, in (6.7), leveraging standard Bayesian linear regression;

for h from 1 *to* H **do**

Sample the coefficients comprising $\boldsymbol{\beta}_h$ from the full conditional

$$(\boldsymbol{\beta}_h \mid -) \sim \mathcal{N}_{M_1}(\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}_h}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_h}),$$

with $\boldsymbol{\mu}_{\tilde{\boldsymbol{\beta}}_h} = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_h}\{\tilde{\tau}_h \boldsymbol{B}_{1h}^\intercal \boldsymbol{Y}_h + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}}\}$, $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_h} = \{\tilde{\tau}_h \boldsymbol{B}_{1h}^\intercal \boldsymbol{B}_{1h} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\}^{-1}$, and $\boldsymbol{Y}_h$ the $n_h \times 1$ vector containing the responses for all the units with $G_i = h$.

Step 4. Update the precision parameters $\tilde{\tau}_h$, $h = 1, \dots, H$ of each kernel in (6.7);

for h from 1 *to* H **do**

Sample $\tilde{\tau}_h$ from

$$(\tilde{\tau}_h \mid -) \sim \text{GA}[a_\sigma + 0.5 \sum_{i=1}^{n} \mathbb{1}(G_i = h), b_\sigma + 0.5 \sum_{i:G_i=h} \{Y_i - \boldsymbol{\mathcal{B}}_1(\boldsymbol{x}_i)^\intercal \tilde{\boldsymbol{\beta}}_h\}^2].$$

Algorithm 4: Steps of the EM algorithm for the LSBP

begin

Let $(\boldsymbol{\gamma}^{(r)}, \tilde{\boldsymbol{\beta}}^{(r)}, \tilde{\boldsymbol{\tau}}^{(r)})$ denote the values of the parameters at iteration r.

***Step 1*: Expectation.** Exploiting results in (6.11), the expectation of (6.10) with respect to the augmented data $(\zeta_i, \bar{\omega}_i)$, for each $i = 1, \ldots, n$, can be obtained by plugging in $\zeta_i^{(r)} = \mathbb{E}(\zeta_i \mid y_i, \boldsymbol{x}_i, \tilde{\boldsymbol{\beta}}^{(r-1)}, \tilde{\boldsymbol{\tau}}^{(r-1)})$ and $\bar{\omega}_i^{(r)} = \mathbb{E}(\bar{\omega}_i \mid \boldsymbol{x}_i, \zeta_i^{(r)}, \boldsymbol{\gamma}^{(r-1)})$ in (6.11).

for i from 1 *to* n **do**

for h from 1 *to* H **do**

Compute $\zeta_{ih}^{(r)}$ by applying the following expression

$$\zeta_{ih}^{(r)} \propto \left[\nu_h^{(r-1)}(\boldsymbol{x}_i) \prod_{l=1}^{h-1} \{1 - \nu_l^{(r-1)}(\boldsymbol{x}_i)\} \right] \mathcal{N}(Y_i; \mathcal{B}_1(\boldsymbol{x}_i)^\intercal \tilde{\boldsymbol{\beta}}_h^{(r-1)}, 1/\tilde{\tau}_h^{(r-1)}),$$

and compute
$\bar{\omega}_{ih}^{(r)} = \{2\mathcal{B}_2(\boldsymbol{x}_i)^\intercal \boldsymbol{\gamma}_h^{(r-1)}\}^{-1} \tanh\{0.5\mathcal{B}_2(\boldsymbol{x}_i)^\intercal \boldsymbol{\gamma}_h^{(r-1)}\} \sum_{l=h}^{H} \zeta_{il}^{(r)}$.

***Step 2*: Maximization.** According to (6.10)–(6.11), modes $\boldsymbol{\gamma}^{(r)}$ and $(\boldsymbol{\beta}^{(r)}, \tilde{\boldsymbol{\tau}}^{(r)})$ can be obtained separately as follow:

for h from 1 *to* $H - 1$ **do**

To compute $\boldsymbol{\gamma}_h^{(r)}$, note that since $\boldsymbol{\gamma}_h$ has Gaussian prior, and provided that the second term in (6.11) is based on Gaussian kernels, the estimated $\boldsymbol{\gamma}_h$ at step $t+1$ coincides with the mean of a full conditional Gaussian, similar to the one in ***Step 2*** of Algorithm 3.

$$\boldsymbol{\gamma}_h^{(r)} = \{\boldsymbol{B}_2^\intercal \mathrm{diag}(\bar{\omega}_{1h}^{(r)}, \ldots, \bar{\omega}_{nh}^{(r)}) \boldsymbol{B}_2 + \boldsymbol{\Sigma}_\gamma^{-1}\}^{-1} \{\boldsymbol{B}_2^\intercal (\bar{\aleph}_{1h}^{(r)}, \ldots, \bar{\aleph}_{nh}^{(r)})^\intercal + \boldsymbol{\Sigma}_\gamma^{-1} \boldsymbol{\mu}_\gamma\},$$

where each $\bar{\aleph}_{ih}^{(r)} = \zeta_{ih}^{(r)} - 0.5 \sum_{l=h}^{H} \zeta_{ih}^{(r)}$ and $\boldsymbol{B}_2$ is the design matrix of the logistic regression based on all units.

for h from 1 *to* H **do**

A similar approach can be considered to compute $\tilde{\boldsymbol{\beta}}_h^{(r)}$ and $\tilde{\tau}_h^{(r)}$ under the Gaussian and inverse-gamma priors for these parameters and the Gaussian kernel characterizing the first term in (6.11). Hence, adapting ***Step 3*** and ***Step 4*** in Algorithm 3 to the EM setting, provides:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_h^{(r)} &= \{\tilde{\tau}_h^{(r-1)} \boldsymbol{B}_1^\intercal \mathrm{diag}(\zeta_{1h}^{(r)}, \ldots, \zeta_{nh}^{(r)}) \boldsymbol{B}_1 + \boldsymbol{\Sigma}_\beta^{-1}\}^{-1} \\
&\qquad \times \{\tilde{\tau}_h^{(r-1)} \boldsymbol{B}_1^\intercal \mathrm{diag}(\zeta_{1h}^{(r)}, \ldots, \zeta_{nh}^{(r)}) \boldsymbol{Y} + \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta\}, \\
\tilde{\tau}_h^{(r)} &= \max\{0, [a_\sigma + 0.5 \sum_{i=1}^{n} \zeta_{ih}^{(r)} - 1][b_\sigma + 0.5 \sum_{i=1}^{n} \zeta_{ih}^{(r)} \{Y_i - \mathcal{B}_1(\boldsymbol{x}_i)^\intercal \tilde{\boldsymbol{\beta}}_h^{(r)}\}^2]^{-1}\},
\end{aligned}$$

where $\boldsymbol{B}_1$ is the design matrix of the Gaussian regression within each kernel based on all units.

Based on the data augmentations outlined in (6.2) and (6.9), the complete log-posterior $\log p_{\boldsymbol{x}}(\gamma, \beta, \tilde{\tau} \mid \boldsymbol{Y}, \zeta, \bar{\omega})$ underlying the proposed EM routine, can be written as

$$\sum_{i=1}^{n} \ell_{\boldsymbol{x}_i}(\gamma, \beta, \tilde{\tau}; y_i, \zeta_i, \bar{\omega}_i) + \sum_{h=1}^{H-1} \log p(\gamma_h) + \sum_{h=1}^{H} \log p(\tilde{\beta}_h) + \sum_{h=1}^{H} \log p(\tau_h) + \text{const}, \quad (6.10)$$

where $\ell_{\boldsymbol{x}_i}(\gamma, \beta, \tilde{\tau}; y_i, \zeta_i, \bar{\omega}_i)$ is the contribution of unit i to the complete log-likelihood. Working on the complete log-likelihood has relevant benefits. Indeed, exploiting equations (6.2) and (6.4), and the results in Polson et al. (2013) summarized in (6.9), the term $\ell_{\boldsymbol{x}_i}(\gamma, \beta, \tilde{\tau}; y_i, \zeta_i, \bar{\omega}_i) = \ell_{\boldsymbol{x}_i}(\beta, \tilde{\tau}; y_i, \zeta_i) + \ell_{\boldsymbol{x}_i}(\gamma; \zeta_i, \bar{\omega}_i)$, can be factorized as

$$\begin{aligned} \sum_{h=1}^{H} \zeta_{ih} \left[-\frac{\tilde{\tau}_h \{Y_i - \mathcal{B}_1(\boldsymbol{x}_i)^\intercal \tilde{\beta}_h\}^2}{2} + \frac{1}{2} \log(\tilde{\tau}_h) \right] \\ + \sum_{h=1}^{H-1} \left[\bar{\kappa}_{ih} \mathcal{B}_2(\boldsymbol{x}_i)^\intercal \gamma_h - \bar{\omega}_{ih} \frac{\{\mathcal{B}_2(\boldsymbol{x}_i)^\intercal \gamma_h\}^2}{2} \right] + \text{const}, \end{aligned} \quad (6.11)$$

where $\bar{\kappa}_{ih} = \zeta_{ih} - 0.5 \sum_{l=h}^{H} \zeta_{ih}$. Hence, both terms in equation (6.11) are linear in the augmented data $(\zeta_i, \bar{\omega}_i)$, and represent the sum of Gaussian kernels. This linearity property simplifies computations in the expectation step for the complete log-posterior in equation (6.10), whereas the Gaussian structure allows simple maximizations. Since the joint maximization of the expected complete log-posterior with respect to $(\beta, \tilde{\tau})$ is intractable, we rely on a conditional maximization procedure (Meng & Rubin, 1993) in the last step of Algorithm 4, which provides analytical solutions.

6.3.3 Mean-field variational Bayes

Section 6.3.2 provides a scalable procedure for estimation of posterior modes in large-scale problems. However, an appealing aspect of the Bayesian approach is in allowing uncertainty quantification via inference on the entire posterior. The Gibbs sampler in Section 6.3.1 represents an appealing procedure which converges to the exact posterior, but faces computational bottlenecks. This motivates scalable variational methods for approximate Bayesian inference (Bishop, 2006; Blei et al., 2017). Clearly, these computational gains do not come without some drawbacks. For example, variational approximations typically underestimate posterior variability. This issue might be mitigated via a post-processing operation as in Giordano et al. (2015), at the cost of an additional computational step.

Due to the Pólya-gamma data augmentation, our variational strategy is framed within the well-established exponential family setting, for which there exists a closed-form coordinate ascent variational inference algorithm (CAVI). Compared to more accurate

black-box variational strategies (e.g. Ranganath et al., 2014), the CAVI algorithm is appealing because it requires no tuning. Moreover, recent theoretical properties for this class of computational methods (Blei et al., 2017) are inherited by our variational algorithm. This seems in contrast with the variational strategy discussed by Ren et al. (2011), which considers a local approximation based on the lower bound of Jaakkola & Jordan (2000). However, the recent contribution of Durante & Rigon (2019), illustrated in Chapter 7, allows to draw a sharp connection between the Pólya-gamma data augmentation and the Jaakkola & Jordan (2000) lower bound. As a consequence, the VB approach we propose relies on the same optimization problem considered by Ren et al. (2011).

Compared to the Gibbs sampler in Section 6.3.1, here we augment the entire model (6.7) with respect to the binary vectors $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH-1})^\intercal$, $i = 1, \ldots, n$ comprising $\boldsymbol{z}$, rather than using the membership indicators $\mathbf{G}$. Hence, the joint law $p_{\boldsymbol{x}}(\boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\boldsymbol{\tau}}, \boldsymbol{z}, \boldsymbol{\omega}) = p_{\boldsymbol{x}}(\boldsymbol{Y} \mid \boldsymbol{z}, \boldsymbol{\beta}, \tilde{\boldsymbol{\tau}}) p_{\boldsymbol{x}}(\boldsymbol{z} \mid \boldsymbol{\gamma}) p_{\boldsymbol{x}}(\boldsymbol{\omega} \mid \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) p(\tilde{\boldsymbol{\beta}}) p(\tilde{\boldsymbol{\tau}})$ is equal to

$$
\begin{aligned}
p(\boldsymbol{\gamma}) p(\tilde{\boldsymbol{\beta}}) p(\tilde{\boldsymbol{\tau}}) \prod_{i=1}^{n} \prod_{h=1}^{H} \mathcal{N}(Y_i; \mathcal{B}_1(\boldsymbol{x}_i)^\intercal \tilde{\boldsymbol{\beta}}_h, \tilde{\tau}_h^{-1})^{z_{ih} \prod_{l=1}^{h-1}(1-z_{il})} \\
\times \prod_{i=1}^{n} \prod_{h=1}^{H-1} \frac{p(\omega_{ih})}{2} \frac{\exp\{(z_{ih} - 0.5)\mathcal{B}_2(\boldsymbol{x}_i)^\intercal \boldsymbol{\gamma}_h\}}{\exp\{0.5 \omega_{ih} (\mathcal{B}_2(\boldsymbol{x}_i)^\intercal \boldsymbol{\gamma}_h)^2\}},
\end{aligned}
\tag{6.12}
$$

where $z_{iH} = 1$. Our goal is to find an optimal variational distribution $q_{\boldsymbol{x}}^{(*)}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}, \boldsymbol{z}, \boldsymbol{\omega})$ that best approximates the joint posterior $p_{\boldsymbol{x}}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}, \boldsymbol{z}, \boldsymbol{\omega} \mid \mathbf{y})$, while maintaining simple computations. This can be obtained by minimizing the Kullback-Leibler divergence between the variational distribution and the full posterior, or, alternatively, by maximizing the evidence lower bound (ELBO) of the log-marginal density $\log m_{\boldsymbol{X}}^{(H)}(\boldsymbol{Y})$, provided that $\log m_{\boldsymbol{X}}^{(H)}(\boldsymbol{Y})$ can be analytically expressed as the sum of the ELBO and the positive KL divergence. Refer to Chapter 7 for details about this decomposition and the formal definition of the ELBO. The optimal variational distribution will be obtained so that

$$
q_{\boldsymbol{x}}^{(*)}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}, \boldsymbol{z}, \boldsymbol{\omega}) = \arg\max_{q \in \mathcal{Q}} \text{ELBO}\{q_{\boldsymbol{x}}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}, \boldsymbol{z}, \boldsymbol{\omega})\}.
$$

Without further restrictions, the Kullback-Leibler divergence is minimized when the variational distribution is equal to the true posterior, which is intractable. To address this issue, a common strategy is to assume that the variational distribution $q_{\boldsymbol{x}}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}, \boldsymbol{z}, \boldsymbol{\omega})$ belongs to a mean-field family $\mathcal{Q}$ (see e.g. Blei et al., 2017). This incorporates a posteriori independence among distinct groups of parameters, implying that the variational distribution can be expressed as the product of marginal laws. Specifically, we consider

Algorithm 5: Steps of the CAVI algorithm for the LSBP

begin

Let $q^{(r)}(\cdot)$ denote the generic variational distribution at iteration r and let $\mathbb{E}_{q^{(r)}}$ denote the expected value taken with respect to it.

Step 1. Update the variational probabilities $q^{(r)}_{\boldsymbol{x}_i}(z_{ih})$;

for i *from* 1 *to* n **do**

for h *from* 1 *to* $H-1$ **do**

Update the variational probabilities $q^{(r)}_{\boldsymbol{x}_i}(z_{ih}=1)=\varrho^{(r)}_{ih}$, for each $i=1,\ldots,n$ and $h=1,\ldots,H-1$. First set each $\varrho^{(r)}_{ih}=\varrho^{(r-1)}_{ih}$, then update

$$\text{logit}(\varrho^{(r)}_{ih})=\boldsymbol{\mathcal{B}}_2(\boldsymbol{x}_i)^\intercal\mathbb{E}_{q^{(r-1)}}(\boldsymbol{\gamma}_h)+ + \sum_{l=h}^{H}\zeta^{(r,h)}_{il}\left[0.5\cdot\mathbb{E}_{q^{(r-1)}}(\log\tilde{\tau}_l)-0.5\cdot\mathbb{E}_{q^{(r-1)}}(\tilde{\tau}_l)\mathbb{E}_{q^{(r-1)}}\{(Y_i-\boldsymbol{\mathcal{B}}_1(\boldsymbol{x}_i)^\intercal\tilde{\boldsymbol{\beta}}_l)^2\}\right],$$

where $\zeta^{(r,h)}_{il}=\prod_{r=1}^{l-1}(1-\varrho^{(r)}_{ir})$ if $l=h$, and $\zeta^{(r,h)}_{il}=-\varrho^{(r)}_{il}\prod_{r=1,r\neq h}^{l-1}(1-\varrho^{(r)}_{ir})$ otherwise. Note also that $\varrho^{(r)}_{iH}=1$.

Step 2. Update the variational distributions $q^{(r)}_{\boldsymbol{x}}(\boldsymbol{\gamma}_h)$, for each $h=1,\ldots,H-1$;

for h from 1 *to* $H-1$ **do**

Update the variational distribution of each $\boldsymbol{\gamma}_h$, being the density of the Gaussian random variable

$$q^{(r)}_{\boldsymbol{x}}(\boldsymbol{\gamma}_h)=\mathcal{N}_{M_2}\{\boldsymbol{\gamma}_h;(\boldsymbol{B}_2^\intercal\boldsymbol{\Omega}_h^{(r-1)}\boldsymbol{B}_2+\boldsymbol{\Sigma}_\gamma^{-1})^{-1}(\boldsymbol{B}_2^\intercal\varrho_h^{(r)}+\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma),(\boldsymbol{B}_2^\intercal\boldsymbol{\Omega}_h^{(r-1)}\boldsymbol{B}_2+\boldsymbol{\Sigma}_\gamma^{-1})^{-1}\}$$

$$\boldsymbol{\Omega}_h^{(r-1)}=\text{diag}\{\mathbb{E}_{q^{(r-1)}}(\omega_{1h}),\ldots,\mathbb{E}_{q^{(r-1)}}(\omega_{nh})\},\ \varrho_h^{(r)}=(\varrho^{(r)}_{1h}-0.5,\ldots,\varrho^{(r)}_{nh}-0.5)^\intercal.$$

Step 3. Update the variational distribution $q^{(r)}_{\boldsymbol{x}_i}(\omega_{ih})$;

for i *from* 1 *to* n **do**

for h *from* 1 *to* $H-1$ **do**

Update the variational distribution $q^{(r)}_{\boldsymbol{x}_i}(\omega_{ih})$ for each $i=1,\ldots,n$ and $h=1,\ldots,H-1$ according to

$$q^{(r)}_{\boldsymbol{x}_i}(\omega_{ih})=\text{PG}\left(\omega_{ih};1,\varphi^{(r)}_{ih}\right),\quad \varphi^{(r)}_{ih}=\{\boldsymbol{\mathcal{B}}_2(\boldsymbol{x}_i)^\intercal\mathbb{E}_{q^{(r)}}(\boldsymbol{\gamma}_h\boldsymbol{\gamma}_h^\intercal)\boldsymbol{\mathcal{B}}_2(\boldsymbol{x}_i)\}^{1/2}.$$

Recall that $\mathbb{E}_{q^{(r)}}(\omega_{ih})=0.5/\varphi^{(r)}_{ih}\tanh(0.5\varphi^{(r)}_{ih})$.

Step 4. Update the variational distributions $q^{(r)}_{\boldsymbol{x}}(\tilde{\boldsymbol{\beta}}_h)$ and $q^{(r)}_{\boldsymbol{x}}(\tilde{\tau}_h)$, for each $h=1,\ldots,H$;

for h from 1 *to* H **do**

Update the variational distributions $q^{(r)}_{\boldsymbol{x}}(\tilde{\boldsymbol{\beta}}_h)$ and $q^{(r)}_{\boldsymbol{x}}(\tilde{\tau}_h)$, for each $h=1,\ldots,H$ according to

$$q^{(r)}_{\boldsymbol{x}}(\tilde{\boldsymbol{\beta}}_h)=\mathcal{N}_{M_1}\{\tilde{\boldsymbol{\beta}}_h;(\boldsymbol{B}_1^\intercal\boldsymbol{\Gamma}_h^{(r)}\boldsymbol{B}_1+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}(\boldsymbol{B}_1^\intercal\boldsymbol{\Gamma}_h^{(r)}\boldsymbol{Y}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}),(\boldsymbol{B}_1^\intercal\boldsymbol{\Gamma}_h^{(r)}\boldsymbol{B}_1+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}\}$$

$$q^{(r)}_{\boldsymbol{x}}(\tilde{\tau}_h)=\text{GA}\{\tilde{\tau}_h;a_\sigma+0.5\sum_{i=1}^{n}\mathbb{E}_{q^{(r)}}(\zeta_{ih}),b_\sigma+0.5\sum_{i=1}^{n}\mathbb{E}_{q^{(r)}}(\zeta_{ih})\mathbb{E}(Y_i-\boldsymbol{\mathcal{B}}_1(\boldsymbol{x}_i)^\intercal\tilde{\boldsymbol{\beta}}_h)^2\}$$

with $\boldsymbol{\Gamma}_h^{(r)}=\mathbb{E}_{q^{(r)}}(\tilde{\tau}_h)\text{diag}\{\mathbb{E}_{q^{(r)}}(\zeta_{1h}),\ldots,\mathbb{E}_{q^{(r)}}(\zeta_{nh})\}$ and $\zeta_{ih}=z_{ih}\prod_{l=1}^{h-1}(1-z_{il})$, $i=1,\ldots,n$.

the following factorization for the variational distribution

$$q_{\boldsymbol{x}}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}, \boldsymbol{z}, \boldsymbol{\omega}) = \prod_{h=1}^{H-1} q_{\boldsymbol{x}}(\boldsymbol{\gamma}_h) \prod_{h=1}^{H} q_{\boldsymbol{x}}(\tilde{\boldsymbol{\beta}}_h) \prod_{h=1}^{H} q_{\boldsymbol{x}}(\tilde{\tau}_h) \prod_{h=1}^{H-1} \prod_{i=1}^{n} q_{\boldsymbol{x}_i}(z_{ih}) \prod_{h=1}^{H-1} \prod_{i=1}^{n} q_{\boldsymbol{x}_i}(\omega_{ih}). \tag{6.13}$$

Note that we are not making specific assumptions about the functional form of the variational distributions. Combining (6.12) with (6.13), we obtain a tractable expression for the ELBO, which can be easily maximized as in Bishop (2006, Ch. 10). In particular, the optimal solutions are provided by the following system of equations

$$\begin{aligned}
&\log q_{\boldsymbol{x}}^{(*)}(\tilde{\boldsymbol{\beta}}_h) = \mathbb{E}_{q^{(*)}(\tilde{\boldsymbol{\tau}}, \boldsymbol{z})}[\log\{p_{\boldsymbol{x}}(\boldsymbol{Y} \mid \boldsymbol{z}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}) p(\tilde{\boldsymbol{\beta}}_h)\}] + \text{const}, && h = 1, \dots, H, \\
&\log q_{\boldsymbol{x}}^{(*)}(\tau_h) = \mathbb{E}_{q^{(*)}(\tilde{\boldsymbol{\beta}}, \boldsymbol{z})}[\log\{p_{\boldsymbol{x}}(\boldsymbol{Y} \mid \boldsymbol{z}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}) p(\tilde{\tau}_h)\}] + \text{const}, && h = 1, \dots, H, \\
&\log q_{\boldsymbol{x}}^{(*)}(\boldsymbol{\gamma}_h) = \mathbb{E}_{q^{(*)}(\boldsymbol{z}, \boldsymbol{\omega})}[\log\{p_{\boldsymbol{x}}(\boldsymbol{z}, \boldsymbol{\omega} \mid \boldsymbol{\gamma}) p(\boldsymbol{\gamma}_h)\}] + \text{const}, && h = 1, \dots, H-1, \\
&\log q_{\boldsymbol{x}_i}^{(*)}(z_{ih}) = \mathbb{E}_{q^{(*)}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}, \boldsymbol{z}_{i,-h})}[\log p_{\boldsymbol{x}}(\boldsymbol{Y}, \boldsymbol{z} \mid \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}}, \boldsymbol{\gamma})] + \text{const}, && h = 1, \dots, H-1, \\
&\log q_{\boldsymbol{x}_i}^{(*)}(\omega_{ih}) = \mathbb{E}_{q^{(*)}(\boldsymbol{\gamma})}[\log p_{\boldsymbol{x}}(\omega_{ih} \mid \boldsymbol{\gamma})] + \text{const}, && h = 1, \dots, H-1,
\end{aligned}$$

for $i = 1, \dots, n$, where $\boldsymbol{z}_{i,-h}$ denotes the vector of binary indicators $\boldsymbol{z}_i$ without considering the hth one, whereas the const terms are additive constants with respect to the argument in the corresponding variational distribution. Each expectation in the above equations is evaluated with respect to the variational distribution of the other parameters, and therefore we need to rely on iterative methods to find the optimal solution. We consider the coordinate ascent variational inference (CAVI) iterative procedure—described in Algorithm 5—which maximizes the variational distribution of each parameter based on the current estimate for the remaining ones (e.g. Bishop, 2006, Ch. 10). This procedure generates a monotone sequence for the ELBO, which ensures convergence to a local joint maximum. As shown in Algorithm 5, the normalizing constants in the above equations have not to be computed numerically, since kernels of well known distributions can be recognized.

6.4 Epidemiology application

We compare the performance of the three computational methods developed in Section 6.3, in a toxicology study. Consistent with recent interests in Bayesian density regression (e.g. Dunson & Park, 2008; Hwang & Pennell, 2014; Canale et al., 2018), we focus on a dataset aimed at studying the relationship between the DDE concentration in maternal serum, and the gestational days at delivery (Longnecker et al., 2001). Such a dataset was considered also in Chapter 2.

The DDE is a metabolite of DDT, which is still used against malaria-transmitting mosquitoes in certain developing countries—according to the Malaria Report 2015 from the World Health Organization—thus raising concerns about its adverse effects on premature delivery. Popular studies in reproductive epidemiology address this goal by dichotomizing the gestational age at delivery (GAD) with a clinical threshold, so that births occurred before the 37th week are considered preterm. Although this approach allows for a simpler modeling strategy, it leads to a clear loss of information. In particular, a greater risk of mortality and morbidity is associated with preterm birth, which increases rapidly as the GAD decreases. This has motivated an increasing interest in modeling how the entire distribution of GAD changes with DDE exposure (e.g. Dunson & Park, 2008; Hwang & Pennell, 2014; Canale et al., 2018).

Data are composed by $n = 2312$ measurements (x_i, Y_i), $i = 1, \ldots, n$, where x_i denotes the DDE concentration, and Y_i is the gestational age at delivery for woman i. Our goal is to reproduce the analyses in Dunson & Park (2008) on this dataset, and compare the inference and computational performance of the MCMC via Gibbs sampling, the EM algorithm, and the VB routine proposed in Section 6.3. Note that, consistent with the main novelty of this contribution, we do not attempt to improve the flexibility and the efficiency of the available statistical models for Bayesian density regression—such as the kernel stick-breaking (Dunson & Park, 2008), and the PSBP (Rodriguez & Dunson, 2011). Indeed, as discussed in Sections 6.1 and 6.2, these representations are expected to provide a comparable performance to our LSBP in terms of inference. However, unlike current models for Bayesian density regression, inference under the LSBP is available under a broader variety of simple computational methods, thus facilitating implementation of the same model in a wider range of applications—including large M_1, M_2 and n settings. Due to this, the main focus is on providing an empirical comparison of the algorithms in Section 6.3, while using results in Dunson & Park (2008) as a benchmark to provide reassurance that inference under the LSBP is comparable to alternative representations.

We apply the predictor-dependent mixture of Gaussians (6.7) with LSBP (6.5)–(6.6), to a normalized version of the DDE and GAD $(\bar{x}_i, \bar{y}_i)$, $i = 1, \ldots, n$, and then show results for $p_x(y)$ on the original scale of the data. Consistent with previous works (Dunson & Park, 2008; Canale et al., 2018), we let $M_1 = 2$, with $\mathcal{B}_{11}(\bar{x}_i) = 1$ and $\mathcal{B}_{12}(\bar{x}_i) = \bar{x}_i$, for every $i = 1, \ldots, n$, and rely instead on a flexible representation for $\eta_h(\bar{x}_i)$ to characterize changes in the stick-breaking weights with DDE. In particular, each $\eta_h(\bar{x}_i)$ is defined via a natural cubic spline basis $\mathcal{B}_2(\bar{x}_i) = \{1, \mathcal{B}_{21}(\bar{x}_i), \ldots, \mathcal{B}_{25}(\bar{x}_i)\}^\intercal$, for every $h = 1, \ldots, H-1$. Bayesian posterior inference—under the three computational methods developed in Section 6.3—is instead performed with default hyperparameters $\boldsymbol{\mu}_\beta = (0, 0)^\intercal$, $\boldsymbol{\Sigma}_\beta = I_{2\times 2}$, $\boldsymbol{\mu}_\gamma = (0, \ldots, 0)^\intercal$, $\boldsymbol{\Sigma}_\gamma = I_{6\times 6}$ and $a_\sigma = b_\sigma = 1$. For the total number of mixture

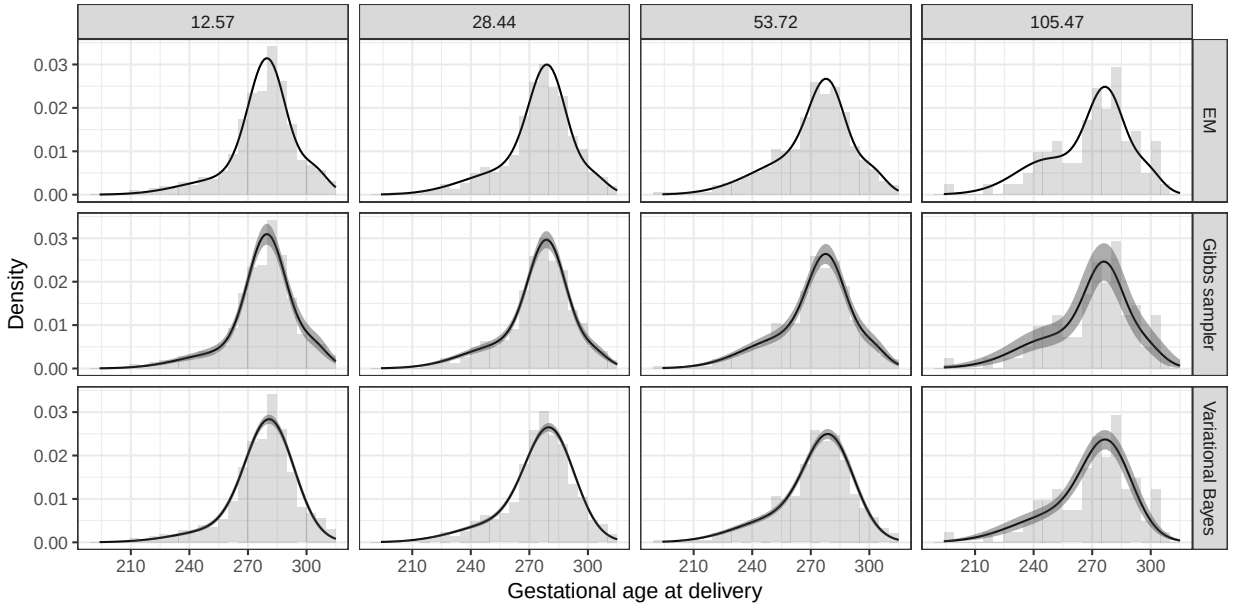


Figure 6.2: For selected quantiles of DDE $\in (12.57, 28.44, 53.72, 105.47)$, graphical representation of the posterior mean of the conditional density for GAD given DDE, obtained from the Gibbs sampler and the VB, together with 0.95 pointwise credibility intervals (shaded area). Since the EM provides only a mode for the conditional density, we consider a graphical representation of the plug-in estimate for the density in (6.7). The histograms represent the observations of GAD, having DDE in the intervals $(-\infty, 20.505)$, $[20.505, 41.08)$, $[41.08, 79.6)$, $[79.6, \infty)$, respectively.

components we consider $H = 20$, and allow the shrinkage induced by the stick-breaking prior to adaptively delete redundant components not required to characterize the data. As shown in Figure 6.2, these choices allows accurate inference on the density (6.7).

In providing posterior inference under the Gibbs sampling algorithm described in Section 6.3.1, we rely on 30,000 iterations, after discarding the first 5,000 as a burn-in, and initialize the routine from random starting values sampled from the prior. Analysis of the traceplots for the quantities discussed in Figures 6.2 and 6.3 showed that this choice is sufficient for good convergence. The EM algorithm and the VB procedures discussed in Sections 6.3.2 and 6.3.3, respectively, are instead run until convergence to a modal solution. Since such modes could be local, we run both algorithms for different initial values, and consider the solutions having the highest log-posterior and ELBO, respectively. We also controlled the monotonicity of the sequences for these quantities, in order to further validate the correctness of our derivations. In this study, the EM and the VB reach convergence in about 2 and 6 seconds, respectively, whereas the Gibbs sampler requires 5 minutes, using a MacBook Air with a Intel Core i5.

Similarly to Figure 3 in Dunson & Park (2008), Figure 6.2 provides posterior inference for the conditional density (6.7) evaluated at the 0.1, 0.6, 0.9, 0.99 quantiles of DDE, for the three algorithms. Histograms for the GAD, are instead obtained by grouping the response

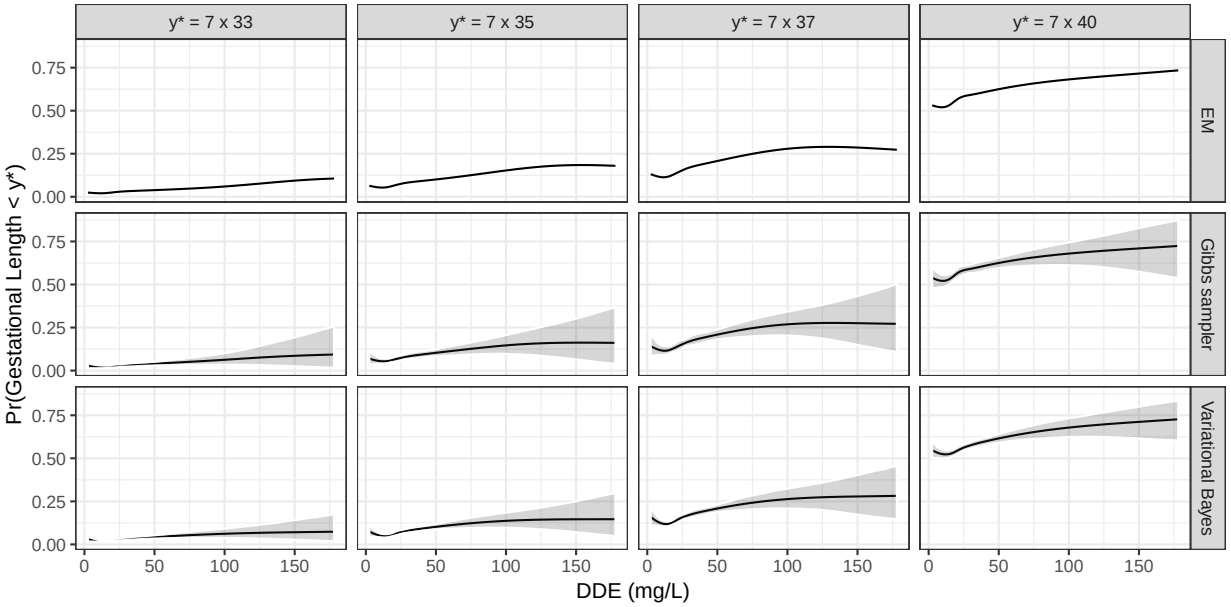


Figure 6.3: For the Gibbs sampler and the VB, posterior means of four different conditional probabilities $\mathbb{P}(y < y^* \mid x)$—based on thresholds $y^* \in (7 \times 33, 7 \times 35, 7 \times 37, 7 \times 40)$—along with 0.95 pointwise credibility intervals (shaded area). These quantities are not available from the EM algorithm, for which a plug-in estimate of $\mathbb{P}(y < y^* \mid x)$ is displayed.

data according to a binning of the DDE with cut-offs at the central values of subsequent quantiles, so that the conditional density can be plotted alongside the corresponding histogram. Results in Figure 6.2 confirm accurate fit to the data and suggest that the left tail of the GAD distribution—associated with preterm deliveries—increasingly inflates as DDE grows. Moreover, as seen in Figure 6.2, the three algorithms have similar results, thus providing empirical reassurance for the goodness of the proposed routines. As expected, the point estimate from the EM matches the posterior mean of the Gibbs sampler, whereas the VB tends to over-smooth some modes of the conditional distribution. This is likely due to the fact that the VB outputs a mean-field approximation of the posterior distribution, instead of the exact one. However, differently from the EM, this routine allows uncertainty quantification, and provides a much scalable methodology compared to the Gibbs sampler, thus representing a valid candidate in high-dimensional inference when the focus is on specific functionals of the density (6.7). Indeed, as shown in Figure 6.3, when the aim is infer conditional preterm probabilities $\tilde{F}_x(y^*) = \mathbb{P}(y < y^* \mid x)$ with $y^* \in (7 \times 33, 7 \times 35, 7 \times 37, 7 \times 40)$ denoting a clinical threshold, the VB provides very similar conclusions.

Prior to conclude our analysis, note that the results in Figures 6.2 and 6.3 are similar to those obtained under the kernel stick-breaking prior in Dunson & Park (2008). This provides empirical guarantee that the flexibility characterizing popular Bayesian nonparametric models for density regression is maintained also under LSBP,

which has the additional relevant benefit of facilitating computational implementation of these methodologies. Minor differences are found at extreme DDE exposures, but this is mainly due to the sparsity of the data in this subset of the predictor space.

6.5 Appendix

Proposition 6.1. *For any fixed* $\boldsymbol{x} \in \mathbb{X}$, $\sum_{h=1}^{\infty} \xi_h(\boldsymbol{x}) = 1$ *almost surely, with* $\xi_h(\boldsymbol{x})$ *factorized as in* (6.5) *and* $\boldsymbol{\gamma}_h \sim \mathcal{N}_{M_2}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$ *independently for every* $h \geqslant 1$. *Hence, the* LSBP *provides a well defined predictor-dependent random probability measure* $\tilde{p}_x$ *at every* $x \in \mathbb{X}$.

Proof. Recalling results in Ishwaran & James (2001), we have that $\sum_{h=1}^{\infty} \xi_h(\boldsymbol{x}) = 1$ almost surely if and only if the equality $\sum_{h=1}^{\infty} \mathbb{E}[\log\{1 - v_h(\boldsymbol{x})\}] = -\infty$ holds. Since $\log\{1 - v_h(\boldsymbol{x})\}$ is concave in $v_h(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{X}$ and $h \geqslant 1$, by the Jensen inequality $\mathbb{E}[\log\{1 - v_h(\boldsymbol{x})\}] \leqslant \log[1 - \mathbb{E}\{v_h(\boldsymbol{x})\}]$. Therefore, since $v_h(\boldsymbol{x}) \in (0, 1)$, we have that $0 < \mathbb{E}\{v_h(\boldsymbol{x})\} = \mu_v(\boldsymbol{x}) < 1$, thereby providing $\log\{1 - \mu_v(\boldsymbol{x})\} < 0$. Leveraging these results, the proof of Proposition 6.1 follows after noticing that $\sum_{h=1}^{\infty} \mathbb{E}[\log\{1 - v_h(\boldsymbol{x})\}] \leqslant \sum_{h=1}^{\infty} \log[1 - \mathbb{E}\{v_h(\boldsymbol{x})\}] = -\infty$.

Proof of Theorem 6.1

Adapting the proof of Theorem 1 in Ishwaran & James (2002) to our representation we have

$$\|m_{\boldsymbol{X}}^{(H)}(\boldsymbol{Y}) - m_{\boldsymbol{X}}^{(\infty)}(\boldsymbol{Y})\|_1 \leqslant 4\left[1 - \mathbb{E}\left\{\prod_{i=1}^{n}\sum_{h=1}^{H-1}\xi_h(\boldsymbol{x}_i)\right\}\right] = 4\mathbb{E}\left[1 - \prod_{i=1}^{n}\sum_{h=1}^{H-1}\xi_h(\boldsymbol{x}_i)\right].$$

Since $\sum_{h=1}^{H-1}\xi_h(\boldsymbol{x}_i) \leqslant 1$, and $1 = \prod_{i=1}^{n} 1$, we can write $1 - \prod_{i=1}^{n}\sum_{h=1}^{H-1}\xi_h(\boldsymbol{x}_i) = \prod_{i=1}^{n} 1 - \prod_{i=1}^{n}\sum_{h=1}^{H-1}\xi_h(\boldsymbol{x}_i) \leqslant \sum_{i=1}^{n}\{1 - \sum_{h=1}^{H-1}\xi_h(\boldsymbol{x}_i)\}$ (Billingsley, 1995, pp. 358). Hence $\|m_{\boldsymbol{X}}^{(H)}(\boldsymbol{Y}) - m_{\boldsymbol{X}}^{(\infty)}(\boldsymbol{Y})\|_1 \leqslant 4[n - \sum_{i=1}^{n}\sum_{h=1}^{H-1}\mathbb{E}\{\xi_h(\boldsymbol{x}_i)\}]$, with $\sum_{h=1}^{H-1}\mathbb{E}\{\xi_h(\boldsymbol{x}_i)\} = \sum_{h=1}^{H-1}\mathbb{E}\{v_1(\boldsymbol{x})\{1 - \mathbb{E}\{v_1(\boldsymbol{x})\}^{h-1} = 1 - \{1 - \mathbb{E}\{v_1(\boldsymbol{x})\}^{H-1}$. Substituting this quantity in $4[n - \sum_{i=1}^{n}\sum_{h=1}^{H-1}\mathbb{E}\{\xi_h(\boldsymbol{x}_i)\}]$, we obtain the final bound $4\sum_{i=1}^{n}[1 - \mathbb{E}\{v_h(\boldsymbol{x})\}]^{H-1}$.

Chapter 7

Conditionally conjugate variational Bayes for logistic models

7.1 Summary

The chapter is organized as follows. In Section 7.2 we introduce some basic concepts about mean-field variational inference, with particular emphasis on variational methods for Bayesian logistic regression. In Section 7.3 we provide a strong theoretical connection between the Jaakkola & Jordan (2000) approach and the Pólya-gamma data-augmentation. In Section 7.4 we discuss a conditionally conjugate CAVI algorithm based on our theoretical findings. Concluding remarks are given in Section 7.5. Codes and additional empirical assessments are available at `https://github.com/tommasorigon/logisticVB`.

7.2 Variational inference for logistic models

The increasing availability of massive and high-dimensional datasets has motivated a wide interest in strategies for Bayesian learning of posterior distributions, beyond classical MCMC methods (e.g. Gelfand & Smith, 1990). Indeed, sampling algorithms can face severe computational bottlenecks in complex statistical models, thus motivating alternative solutions based on scalable and efficient optimization of approximate posterior distributions. Notable methods within this class are the Laplace approximation (e.g. Bishop, 2006, Ch. 4.4), variational Bayes (e.g. Bishop, 2006, Ch. 10.1) and expectation propagation (e.g. Bishop, 2006, Ch. 10.7), with variational inference providing a standard choice in several fields, as discussed in recent reviews by Blei et al. (2017) and Ormerod & Wand (2010). Refer also to Jordan et al. (1999) for a seminal introduction of variational inference from a statistical perspective.

Variational Bayes aims at obtaining a tractable approximation $q^{(*)}(\boldsymbol{\theta})$ for the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$ of the random coefficients $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\intercal$, in the model having

joint density $p(\boldsymbol{Y}, \boldsymbol{\theta}) = p(\boldsymbol{Y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ and the observed data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\intercal$, with $p(\boldsymbol{\theta})$ denoting the prior distribution for $\boldsymbol{\theta}$. This optimization problem is formally addressed by minimizing the Kullback–Leibler (KL) divergence

$$\text{KL}\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta} \mid \boldsymbol{Y})\} = \int_\Theta q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{Y})} d\boldsymbol{\theta} = \int_\Theta q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta}) m(\boldsymbol{Y})}{p(\boldsymbol{Y}, \boldsymbol{\theta})} d\boldsymbol{\theta}, \tag{7.1}$$

with respect to $q(\boldsymbol{\theta}) \in \mathcal{Q}$, where $\mathcal{Q}$ denotes a tractable, yet sufficiently flexible, class of approximating distributions. As is clear from (7.1), the calculation of the KL divergence between $q(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$ requires the evaluation of the marginal density $m(\boldsymbol{Y})$, whose intractability is actually the main reason motivating approximate Bayesian methods. Due to this, the above minimization problem is commonly translated into the maximization of the evidence lower bound (ELBO) function

$$\text{ELBO}\{q(\boldsymbol{\theta})\} = \int_\Theta q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{Y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} = -\text{KL}\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta} \mid \boldsymbol{Y})\} + \log m(\boldsymbol{Y}), \tag{7.2}$$

which does not require the evaluation of $m(\boldsymbol{Y})$. In fact, since $\log m(\boldsymbol{Y})$ does not depend on $\boldsymbol{\theta}$, maximizing (7.2) is equivalent to minimizing (7.1). Re-writing (7.2) as $\log m(\boldsymbol{Y}) = \text{ELBO}\{q(\boldsymbol{\theta})\} + \text{KL}\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta} \mid \boldsymbol{Y})\}$ it can be additionally noticed that the ELBO provides a lower bound of $\log m(\boldsymbol{Y})$ for any $q(\boldsymbol{\theta})$, since the Kullback–Leibler divergence is always non-negative.

The above set-up defines the general rationale underlying VB but, as is clear from (7.2), the practical feasibility of the variational optimization requires a tractable form for the joint density $p(\boldsymbol{Y}, \boldsymbol{\theta})$ along with a simple, yet flexible, variational family $\mathcal{Q}$. This is the case of mean-field VB for conditionally conjugate exponential family models with global and local variables (Wang & Titterington, 2004; Bishop, 2006; Hoffman et al., 2013; Blei et al., 2017). Recalling Hoffman et al. (2013), these methods focus on obtaining a mean-field approximation

$$\begin{aligned} q^{(*)}(\boldsymbol{\theta}) = q^{(*)}(\boldsymbol{\beta}, \boldsymbol{\omega}) = q^{(*)}(\boldsymbol{\beta}) \prod_{i=1}^n q^{(*)}(\omega_i) &= \arg\min_{q \in \mathcal{Q}} \text{KL}\{q(\boldsymbol{\beta}) \prod_{i=1}^n q(\omega_i) \,\|\, p(\boldsymbol{\beta}, \boldsymbol{\omega} \mid \boldsymbol{Y})\}, \\ &= \arg\max_{q \in \mathcal{Q}} \text{ELBO}\{q(\boldsymbol{\beta}) \prod_{i=1}^n q(\omega_i)\} \end{aligned} \tag{7.3}$$

for the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\omega} \mid \boldsymbol{Y})$ of the global coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ and the local variables $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\intercal$ in the statistical model having joint density

$$p(\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\omega}) = p(\boldsymbol{\beta}) \prod_{i=1}^n p(\omega_i \mid \boldsymbol{\beta}) p(Y_i \mid \omega_i, \boldsymbol{\beta}) = p(\boldsymbol{\beta}) \prod_{i=1}^n p(Y_i, \omega_i \mid \boldsymbol{\beta}), \tag{7.4}$$

with $p(Y_i, \omega_i \mid \boldsymbol{\beta})$ from an exponential family and $p(\boldsymbol{\beta})$ being a conjugate prior for this density. The latent quantities $\boldsymbol{\omega}$, when present, typically denote random effects or unit-specific augmented data within some hierarchical formulation, such as in mixture models.

Although the above assumptions appear restrictive, the factorization of $q(\boldsymbol{\beta}, \boldsymbol{\omega})$, characterizing the mean-field variational family, provides a flexible class in several applications and allows direct implementation of simple coordinate ascent variational inference (CAVI) routines (Bishop, 2006, Ch. 10.1.1) which sequentially maximize the ELBO in (7.3) with respect to each factor in $q(\boldsymbol{\beta}, \boldsymbol{\omega}) = q(\boldsymbol{\beta}) \prod_{i=1}^{n} q(\omega_i)$—fixing the others at their most recent update. Instead, the exponential family and conjugacy assumptions further simplify calculations by providing approximating densities $q^{(*)}(\boldsymbol{\beta})$ and $q^{(*)}(\omega_i)$, $i = 1, \ldots, n$ from tractable classes of random variables. These advantages have also motivated recent computational improvements (Hoffman et al., 2013) and theoretical studies (Wang & Titterington, 2004). We refer to Hoffman et al. (2013) and Blei et al. (2017) for details on the methods related to the general formulation in (7.3)-(7.4), and focus here on models having logistic likelihoods as building-blocks. Indeed, although the conjugacy and exponential family assumptions are common to a variety of machine learning representations (e.g. Blei et al., 2003; Airoldi et al., 2008; Hoffman et al., 2013), classical Bayesian logistic regression models of the form

$$p(Y_i \mid \boldsymbol{\beta}) = \frac{\{\exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}^{Y_i}}{1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})}, \quad Y_i \in \{0, 1\}, \; i = 1, \ldots, n, \quad \text{with } \boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{7.5}$$

do not enjoy direct conjugacy between the likelihood for the binary response data and the Gaussian prior for the coefficients in the linear predictor (e.g. Wang & Blei, 2013). This apparently notable exception to conditionally conjugate exponential family models also holds, as a direct consequence, for a wide set of formulations which incorporate Bayesian logistic regressions at some layer of the hierarchical specification. Some relevant examples are classification via Gaussian processes (Rasmussen & Williams, 2006), supervised nonparametric clustering (Ren et al., 2011) and hierarchical mixture of experts (Bishop & Svensén, 2003).

To allow tractable VB for non-conjugate models, several alternatives beyond conjugate mean-field VB have been proposed (see e.g. Jaakkola & Jordan, 2000; Braun & McAuliffe, 2010; Wand et al., 2011; Wang & Blei, 2013). Within the context of logistic regression, Jaakkola & Jordan (2000) developed a seminal VB algorithm which relies on the quadratic

Algorithm 6: EM algorithm for approximate Bayesian inference by Jaakkola & Jordan (2000).

Initialize $\varphi_1^{(0)}, \ldots, \varphi_n^{(0)}$.

for $r = 1$ until convergence of (7.7) **do**

***Step 1.* Expectation.** Update $q^{(r)}(\boldsymbol{\beta}) = \bar{p}^{(r-1)}(\boldsymbol{\beta} \mid \boldsymbol{Y}) \propto p(\boldsymbol{\beta}) \prod_{i=1}^{n} \bar{p}^{(r-1)}(Y_i \mid \boldsymbol{\beta})$ to obtain a Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}^{(r)}, \boldsymbol{\Sigma}^{(r)})$ with

$$\boldsymbol{\Sigma}^{(r)} = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \boldsymbol{X}^\intercal \boldsymbol{\Omega}^{(r-1)} \boldsymbol{X})^{-1}, \qquad \boldsymbol{\mu}^{(r)} = \boldsymbol{\Sigma}^{(r)} \{\boldsymbol{X}^\intercal (\boldsymbol{Y} - 0.5 \cdot \boldsymbol{1}_n) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}}\},$$

where $\boldsymbol{1}_n = (1, \ldots, 1)^\intercal$ and $\boldsymbol{\Omega}^{(r-1)}$ is a diagonal matrix with entries $\{0.5/\varphi_i^{(r-1)} \tanh(0.5\varphi_i^{(r-1)})$ for $i = 1, \ldots, n$. Note that the quadratic form of (7.6), restores conjugacy between the Gaussian prior for $\boldsymbol{\beta}$ and the approximated likelihood. To clarify this result, note that, for every φ_i, $\bar{p}(Y_i \mid \boldsymbol{\beta})$ is proportional to the kernel of a Gaussian variable having mean $\boldsymbol{x}_i^\intercal \boldsymbol{\beta}$ and variance $2\varphi_i \tanh(0.5\varphi_i)^{-1}$ for the "data" $2\varphi_i \tanh(0.5\varphi_i)^{-1}(Y_i - 0.5)$.

***Step 2.* Maximization.** Compute $\boldsymbol{\varphi}^{(r)} = \operatorname{argmax}_{\boldsymbol{\varphi}} \int_{\mathbb{R}^p} q^{(r)}(\boldsymbol{\beta}) \log \bar{p}(\boldsymbol{Y}, \boldsymbol{\beta}) d\boldsymbol{\beta}$ to obtain the solutions

$$\varphi_i^{(r)} = \{\mathbb{E}_{q^{(r)}(\boldsymbol{\beta})}[(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2]\}^{\frac{1}{2}} = \{\boldsymbol{x}_i^\intercal \boldsymbol{\Sigma}^{(r)} \boldsymbol{x}_i + (\boldsymbol{x}_i^\intercal \boldsymbol{\mu}^{(r)})^2\}^{\frac{1}{2}}, \quad \text{for every } i = 1, \ldots, n.$$

Note that $\int_{\mathbb{R}^p} q^{(r)}(\boldsymbol{\beta}) \log \bar{p}(\boldsymbol{Y}, \boldsymbol{\beta}) d\boldsymbol{\beta} = \text{const} + \sum_{i=1}^{n} \int_{\mathbb{R}^p} q^{(r)}(\boldsymbol{\beta}) \log \bar{p}(Y_i \mid \boldsymbol{\beta}) d\boldsymbol{\beta}$. Hence, it is possible to maximize the expected log-likelihood associated with every Y_i separately, as a function of each φ_i, for $i = 1, \ldots, n$. This result leads to the above solution.

Output at the end of the algorithm: $\boldsymbol{\varphi}^{(*)}$ and, as a byproduct, the approximate posterior $q^{(*)}(\boldsymbol{\beta}) = \bar{p}^{(*)}(\boldsymbol{\beta} \mid \boldsymbol{Y})$.

lower bound

$$\begin{aligned}\log \bar{p}(Y_i \mid \boldsymbol{\beta}) =& (Y_i - 0.5)\boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \\ & - 0.5\varphi_i - 0.25\varphi_i^{-1}\tanh(0.5\varphi_i)\{(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2 - \varphi_i^2\} - \log\{1 + \exp(-\varphi_i)\},\end{aligned} \tag{7.6}$$

for the log-likelihood $\log p(Y_i \mid \boldsymbol{\beta}) = Y_i(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}) - \log[1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})] \geqslant \log \bar{p}(Y_i \mid \boldsymbol{\beta})$ of every Y_i from a logistic regression. In (7.6), the vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ comprises the covariates measured for unit i, whereas $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ are the associated coefficients. The vector $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)^\intercal$ denotes instead unit-specific variational parameters defining the location where $\log \bar{p}(Y_i \mid \boldsymbol{\beta})$ is tangent to $\log p(Y_i \mid \boldsymbol{\beta})$. In fact, $\log \bar{p}(Y_i \mid \boldsymbol{\beta}) = \log p(Y_i \mid \boldsymbol{\beta})$ when $\varphi_i^2 = (\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2$. Leveraging equation (7.6), Jaakkola & Jordan (2000) proposed an expectation-maximization (EM) algorithm (Dempster et al., 1977) to approximate $p(\boldsymbol{\beta} \mid \boldsymbol{Y})$. At the generic iteration r, this routine alternates between an E-step in which the conditional distribution of the random coefficients $\boldsymbol{\beta}$ given the current $\boldsymbol{\varphi}^{(r-1)}$ is updated to obtain $q^{(r)}(\boldsymbol{\beta})$, and an M-step which calculates the expectation of the augmented approximate log-likelihood $\log \bar{p}(\boldsymbol{Y}, \boldsymbol{\beta}) = \log p(\boldsymbol{\beta}) + \sum_{i=1}^n \log \bar{p}(Y_i \mid \boldsymbol{\beta})$ with respect to $q^{(r)}(\boldsymbol{\beta})$ and maximizes it as a function of $\boldsymbol{\varphi}$. Recalling the general presentation of EM by Bishop (2006, Ch. 9.4) and Appendices A-B in Jaakkola & Jordan (2000), this strategy ultimately maximizes $\log \bar{m}(\boldsymbol{Y}) = \log \int_{\mathbb{R}^p} p(\boldsymbol{\beta}) \prod_{i=1}^n \bar{p}(Y_i \mid \boldsymbol{\beta}) d\boldsymbol{\beta}$ with respect to $\boldsymbol{\varphi}$, by sequentially optimizing the lower bound

$$\int_{\mathbb{R}^p} q(\boldsymbol{\beta}) \log \frac{p(\boldsymbol{\beta}) \prod_{i=1}^n \bar{p}(Y_i \mid \boldsymbol{\beta})}{q(\boldsymbol{\beta})} d\boldsymbol{\beta}, \tag{7.7}$$

as a function of the unknown distribution $q(\boldsymbol{\beta})$ and the fixed parameters $\boldsymbol{\varphi}$, where $p(\boldsymbol{\beta})$ is the density of the Gaussian prior for $\boldsymbol{\beta}$. Hence, as is clear from Algorithm 6, this EM produces an optimal estimate $\boldsymbol{\varphi}^{(*)}$ of $\boldsymbol{\varphi}$ and, as a byproduct, also a distribution $q^{(*)}(\boldsymbol{\beta})$, which is regarded as an approximate posterior in Jaakkola & Jordan (2000). Indeed, recalling the EM structure, $q^{(*)}(\boldsymbol{\beta})$ coincides with the conditional distribution $\bar{p}^{(*)}(\boldsymbol{\beta} \mid \boldsymbol{Y})$ obtained by updating the prior $p(\boldsymbol{\beta})$ with the approximate likelihood $\prod_{i=1}^n \bar{p}^{(*)}(Y_i \mid \boldsymbol{\beta})$ induced by (7.6) and evaluated at the optimal variational parameters $\varphi_1^{(*)}, \ldots, \varphi_n^{(*)}$. However, although being successfully implemented in the machine learning and statistical literature (e.g. Bishop & Svensén, 2003; Rasmussen & Williams, 2006; Lee et al., 2010; Ren et al., 2011; Carbonetto & Stephens, 2012; Tang et al., 2015; Wand, 2017), it is not clear how the solution $q^{(*)}(\boldsymbol{\beta})$ relates to the formal VB set-up in (7.1)-(7.2). Indeed, $\bar{p}^{(*)}(\boldsymbol{\beta} \mid \boldsymbol{Y})$ is not the posterior induced by a Bayesian logistic regression. This is due to the fact that each $p(Y_i \mid \boldsymbol{\beta})$ in the kernel of $p(\boldsymbol{\beta} \mid \boldsymbol{Y})$ is replaced with the approximate likelihood $\bar{p}^{(*)}(Y_i \mid \boldsymbol{\beta})$ evaluated at the optimal variational parameters maximizing $\log \bar{m}(\boldsymbol{Y})$. This last result, which is inherent to the EM (Dempster et al., 1977), suggests a heuristic

intuition for why $q^{(*)}(\boldsymbol{\beta})$ may still provide a reasonable approximation. Indeed, since $\log \bar{p}(Y_i \mid \boldsymbol{\beta}) \leqslant \log p(Y_i \mid \boldsymbol{\beta})$ for every φ_i and $i = 1, \ldots, n$, the same holds for $\log \bar{m}(\boldsymbol{Y})$ and $\log m(\boldsymbol{Y})$. Thus, since $\log m(\boldsymbol{Y})$ does not vary with $\boldsymbol{\varphi}$, maximizing $\log \bar{m}(\boldsymbol{Y})$ with respect to $\boldsymbol{\varphi}$ is expected to provide the tightest approximation of each $\log p(Y_i \mid \boldsymbol{\beta})$ via the lower bound in (7.6) evaluated at the optimum $\varphi_i^{(*)}$, for $i = 1, \ldots, n$, thereby guaranteeing similar predictive densities $m(\boldsymbol{Y})$ and $\bar{m}^{(*)}(\boldsymbol{Y})$. Hence, in correspondence to the optimum $\boldsymbol{\varphi}^{(*)}$, the minimization of $\text{KL}\{\bar{q}(\boldsymbol{\beta}) \parallel \bar{p}^{(*)}(\boldsymbol{\beta} \mid \boldsymbol{Y})\}$ in the E-step, would hopefully provide a solution $q^{(*)}(\boldsymbol{\beta}) = \bar{p}^{(*)}(\boldsymbol{\beta} \mid \boldsymbol{Y})$ close to the true posterior $p(\boldsymbol{\beta} \mid y)$.

Although the above discussion provides an intuition for the excellent performance of the methods proposed by Jaakkola & Jordan (2000), it shall be noticed that finding the tightest bound within a class of functions might not be sufficient if this class is not flexible enough. Indeed, the quadratic form of (7.6) might be restrictive for logistic log-likelihoods, and hence even the optimal approximation may fail to mimic $\log p(Y_i \mid \boldsymbol{\beta})$. Moreover, according to (7.1), a formal VB set-up requires the minimization of a well-defined KL divergence between an exact posterior and an approximating density from a given variational family. Instead, Jaakkola & Jordan (2000) seem to minimize the divergence between an approximate posterior and a pre-specified density. If this were the case, then their methods could be only regarded as approximate solutions to formal VB. Indeed, although (7.6) has been recently studied (De Leeuw & Lange, 2009; Browne & McNicholas, 2015), this is currently the main view of the EM in Algorithm 6 (e.g. Blei et al., 2017; Wang & Blei, 2013; Bishop, 2006).

In Section 7.3 we prove that this is not true and that (7.6), although apparently supported by purely mathematical arguments, has indeed a clear probabilistic interpretation related to a recent Pólya-gamma data augmentation for logistic regression (Polson et al., 2013). In particular, let $q(\omega_i)$ be the density of a Pólya-gamma $\text{PG}(1, \varphi_i)$, then (7.6) is a proper evidence lower bound associated with a VB approximation of the posterior for ω_i in the conditional model $p(Y_i, \omega_i \mid \boldsymbol{\beta})$ for data Y_i from (7.5) and the Pólya-gamma variable $(\omega_i \mid \boldsymbol{\beta}) \sim \text{PG}(1, \boldsymbol{x}_i^\intercal \boldsymbol{\beta})$, with $\boldsymbol{\beta}$ kept fixed. Combining this result with the objective function in equation (7.7), allows us to formalize Algorithm 6 as a pure CAVI which approximates the joint posterior of $\boldsymbol{\beta}$ and the augmented Pólya-gamma data $\omega_1, \ldots, \omega_n$, under a mean-field variational approximation within a conditionally conjugate exponential family framework.

7.3 Conditionally conjugate variational representation

This section discusses the theoretical connection between equation (7.6) and a recent Pólya-gamma data augmentation for conditionally conjugate inference in Bayesian logistic

regression (Polson et al., 2013), thus allowing us to recast the methods proposed by Jaakkola & Jordan (2000) within the wider framework of mean-field variational inference for conditionally conjugate exponential family models. We shall emphasize that, in a recent manuscript, Scott & Sun (2013) proposed an EM for maximum a posteriori estimation of $\boldsymbol{\beta}$ in (7.5), discussing connections with the variational methods in Jaakkola & Jordan (2000). Their findings are however limited to computational differences and similarities among the two methods and the associated algorithms. We instead provide a fully probabilistic connection between the contribution by Jaakkola & Jordan (2000) and the one of Polson et al. (2013), thus opening new avenues for advances in VB for logistic models.

To anticipate Lemma 7.1, note that the core contribution of Polson et al. (2013) is in showing that $p(Y_i \mid \boldsymbol{\beta})$ in model (7.5) can be expressed as a scale-mixture of Gaussians with respect to a Pólya-gamma density. This result facilitates the implementation of MCMC methods which update $\boldsymbol{\beta}$ and the Pólya-gamma augmented data $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\intercal$ from conjugate full conditionals. In fact, the joint density $p(\boldsymbol{Y}, \boldsymbol{\omega} \mid \boldsymbol{\beta})$ has a Gaussian kernel in $\boldsymbol{\beta}$, thus restoring Gaussian-Gaussian conjugacy in the full conditional. As discussed in Lemma 7.1, this data augmentation, although developed a decade later, was implicitly hidden in the bound of Jaakkola & Jordan (2000).

Lemma 7.1. *Let* $\log \tilde{p}(Y_i \mid \boldsymbol{\beta})$ *be the quadratic lower bound in* (7.6) *proposed by Jaakkola & Jordan (2000) for the logistic log-likelihood* $\log p(Y_i \mid \boldsymbol{\beta})$ *in* (7.5). *Then, for every unit* $i = 1, \ldots, n$, *we have*

$$\begin{aligned} \log \tilde{p}(Y_i \mid \boldsymbol{\beta}) &= \int_{\mathbb{R}_+} q(\omega_i) \log \frac{p(Y_i, \omega_i \mid \boldsymbol{\beta})}{q(\omega_i)} d\omega_i \\ &= \mathbb{E}_{q(\omega_i)}\{\log p(Y_i, \omega_i \mid \boldsymbol{\beta})\} - \mathbb{E}_{q(\omega_i)}\{\log q(\omega_i)\}, \end{aligned} \tag{7.8}$$

with $p(Y_i, \omega_i \mid \boldsymbol{\beta}) = p(Y_i \mid \boldsymbol{\beta}) p(\omega_i \mid \boldsymbol{\beta})$ *and* $p(Y_i \mid \boldsymbol{\beta}) = \exp(Y_i \boldsymbol{x}_i^\intercal \boldsymbol{\beta}) \{1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}^{-1}$, *whereas* $q(\omega_i)$ *and* $p(\omega_i \mid \boldsymbol{\beta})$ *are the densities of the Pólya-gamma variables* $\text{PG}(1, \varphi_i)$ *and* $\text{PG}(1, \boldsymbol{x}_i^\intercal \boldsymbol{\beta})$, *respectively.*

Proof. To prove Lemma 7.1, first notice that $0.5\varphi_i + \log\{1 + \exp(-\varphi_i)\} = \log\{2\cosh(0.5\varphi_i)\}$ and $0.5(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}) = \log\{1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\} - \log[2\cosh\{0.5(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}]$. Replacing such quantities in (7.6), we obtain

$$\begin{aligned} Y_i \boldsymbol{x}_i^\intercal \boldsymbol{\beta} - \log\{1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\} - 0.25 \varphi_i^{-1} \tanh(0.5\varphi_i)\{(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2 - \varphi_i^2\} \\ + \log[\cosh(0.5\varphi_i)^{-1} \cosh\{0.5(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}]. \end{aligned}$$

To highlight equation (7.8) in the above function, note that, recalling Polson et al. (2013), the quantity $-0.25\varphi_i^{-1}\tanh(0.5\varphi_i)\{(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2 - \varphi_i^2\}$ is equal to $\mathbb{E}\{-0.5\omega_i(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2\} -$

$\mathbb{E}(-0.5\omega_i\varphi_i^2)$, where the expectation is taken with respect to $\omega_i \sim \text{PG}(1, \varphi_i)$. Hence, $\log \bar{p}(Y_i \mid \boldsymbol{\beta})$ can be expressed as

$$\int_{\mathbb{R}_+} \frac{\exp(-0.5\omega_i\varphi_i^2)p(\omega_i)}{\cosh(0.5\varphi_i)^{-1}} \times \log \frac{\exp(Y_i\boldsymbol{x}_i^\intercal\boldsymbol{\beta})\{1+\exp(\boldsymbol{x}_i^\intercal\boldsymbol{\beta})\}^{-1}\exp\{-0.5\omega_i(\boldsymbol{x}_i^\intercal\boldsymbol{\beta})^2\}\cosh\{0.5(\boldsymbol{x}_i^\intercal\boldsymbol{\beta})\}p(\omega_i)}{\exp(-0.5\omega_i\varphi_i^2)\cosh(0.5\varphi_i)p(\omega_i)} d\omega_i.$$

Based on the above expression, the proof is concluded after noticing that $\exp(Y_i\boldsymbol{x}_i^\intercal\boldsymbol{\beta})\{1+\exp(\boldsymbol{x}_i^\intercal\boldsymbol{\beta})\}^{-1} = p(Y_i \mid \boldsymbol{\beta})$, whereas the term $\exp\{-0.5\omega_i(\boldsymbol{x}_i^\intercal\boldsymbol{\beta})^2\}\cosh\{0.5(\boldsymbol{x}_i^\intercal\boldsymbol{\beta})\}p(\omega_i)$ and $\exp(-0.5\omega_i\varphi_i^2)\cosh(0.5\varphi_i)p(\omega_i)$ are the densities $p(\omega_i \mid \boldsymbol{\beta})$ and $q(\omega_i)$ of the Pólya-gamma random variables $\text{PG}(1, \boldsymbol{x}_i^\intercal\boldsymbol{\beta})$ and $\text{PG}(1, \varphi_i)$, respectively, with $p(\omega_i)$ the density of a $\text{PG}(1, 0)$. □

According to Lemma 7.1, the expansion in equation (7.6) is a proper ELBO related to a VB approximation of the posterior for ω_i in the conditional model $p(Y_i, \omega_i \mid \boldsymbol{\beta})$ for response data Y_i from (7.5) and the local variable $(\omega_i \mid \boldsymbol{\beta}) \sim \text{PG}(1, \boldsymbol{x}_i^\intercal\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ kept fixed. Note that, although some intuition on the relation between $\log \bar{p}(Y_i \mid \boldsymbol{\beta})$ and $\mathbb{E}_{q(\omega_i)}\{\log p(Y_i, \omega_i \mid \boldsymbol{\beta})\}$ can be deduced from Scott & Sun (2013), the authors leave out additive constants not depending on $\boldsymbol{\beta}$ in $\log \bar{p}(Y_i \mid \boldsymbol{\beta})$ when discussing this connection. Indeed, according to Lemma 7.1, these quantities are crucial to formally interpret $\log \bar{p}(Y_i \mid \boldsymbol{\beta})$ as a genuine ELBO, since they coincide with $-\mathbb{E}_{q(\omega_i)}\{\log q(\omega_i)\}$. Besides this result, Lemma 7.1 provides a formal characterization for the approximation error $\log p(Y_i \mid \boldsymbol{\beta}) - \log \bar{p}(Y_i \mid \boldsymbol{\beta})$. Indeed, adapting (7.2) to this setting, such a quantity is the KL divergence between a generic Pólya-gamma variable and the one obtained by conditioning on $\boldsymbol{\beta}$. This allows to complete $\log p(Y_i \mid \boldsymbol{\beta}) \geqslant \log \bar{p}(Y_i \mid \boldsymbol{\beta})$, as

$$\begin{aligned} \log p(Y_i \mid \boldsymbol{\beta}) &= \log \bar{p}(Y_i \mid \boldsymbol{\beta}) + \text{KL}\{q(\omega_i) \,\|\, p(\omega_i \mid Y_i, \boldsymbol{\beta})\} \\ &= \log \bar{p}(Y_i \mid \boldsymbol{\beta}) + \text{KL}\{q(\omega_i) \,\|\, p(\omega_i \mid \boldsymbol{\beta})\}, \end{aligned} \tag{7.9}$$

where the last equality follows from the fact that $p(Y_i, \omega_i \mid \boldsymbol{\beta}) = p(Y_i \mid \boldsymbol{\beta})p(\omega_i \mid \boldsymbol{\beta})$, and hence $p(\omega_i \mid Y_i, \boldsymbol{\beta}) = p(\omega_i \mid \boldsymbol{\beta})$. This result sheds light on the heuristic interpretation of $q^{(*)}(\boldsymbol{\beta})$ in Section 7.2. Indeed, as is clear from (7.9), if $q(\omega_i)$ evaluated at the optimal $\varphi_i^{(*)}$ is globally close to $p(\omega_i \mid \boldsymbol{\beta})$ for every $\boldsymbol{\beta}$ and $i = 1, \ldots, n$, then (7.6) ensures accurate approximation of $\log p(Y_i \mid \boldsymbol{\beta})$, thus providing approximate posteriors $q^{(*)}(\boldsymbol{\beta})$ close to the target $p(\boldsymbol{\beta} \mid \boldsymbol{Y})$. Exploiting Lemma 7.1, Theorem 7.1 formalizes this discussion by proving that the EM in Algorithm 6 maximizes the ELBO of a well-defined model under a mean-field VB.

Theorem 7.1. *The lower bound in* (7.7) *maximized by Jaakkola & Jordan (2000) in their* EM *for approximate Bayesian inference in model* (7.5) *coincides with a genuine evidence lower bound*

$$\begin{aligned}\text{ELBO}\{q(\boldsymbol{\beta},\boldsymbol{\omega})\} &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^n_+} q(\boldsymbol{\beta},\boldsymbol{\omega}) \log \frac{p(\boldsymbol{Y},\boldsymbol{\beta},\boldsymbol{\omega})}{q(\boldsymbol{\beta},\boldsymbol{\omega})} d\boldsymbol{\omega} d\boldsymbol{\beta}, \\ &= \mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\omega})}\{\log p(\boldsymbol{Y},\boldsymbol{\beta},\boldsymbol{\omega})\} - \mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\omega})}\{\log q(\boldsymbol{\beta},\boldsymbol{\omega})\},\end{aligned} \tag{7.10}$$

where $p(\boldsymbol{Y},\boldsymbol{\beta},\boldsymbol{\omega}) = p(\boldsymbol{\beta})\prod_{i=1}^n p(Y_i \mid \boldsymbol{\beta})p(\omega_i \mid \boldsymbol{\beta})$ *and* $q(\boldsymbol{\beta},\boldsymbol{\omega}) = q(\boldsymbol{\beta})\prod_{i=1}^n q(\omega_i)$*, with* $q(\omega_i)$ *and* $p(\omega_i \mid \boldsymbol{\beta})$ *denoting the densities of the Pólya-gamma variables* $\text{PG}(1,\varphi_i)$ *and* $\text{PG}(1,\boldsymbol{x}_i^\intercal\boldsymbol{\beta})$*, respectively.*

Proof. The proof is a direct consequence of Lemma 7.1. In particular, let

$$\int_{\mathbb{R}^p} q(\boldsymbol{\beta}) \log\{p(\boldsymbol{\beta})q(\boldsymbol{\beta})^{-1}\} d\boldsymbol{\beta} + \int_{\mathbb{R}^p} q(\boldsymbol{\beta}) \sum_{i=1}^n \log \bar{p}(Y_i \mid \boldsymbol{\beta}) d\boldsymbol{\beta}$$

denote an expanded representation of (7.7). Then, replacing $\log \bar{p}(Y_i \mid \boldsymbol{\beta})$ with its probabilistic definition in (7.8) and performing simple mathematical calculations, we obtain

$$\int_{\mathbb{R}^p} q(\boldsymbol{\beta}) \log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})} d\boldsymbol{\beta} + \sum_{i=1}^n \int_{\mathbb{R}^p}\int_{\mathbb{R}_+} q(\boldsymbol{\beta})q(\omega_i) \log \frac{p(Y_i \mid \boldsymbol{\beta})p(\omega_i \mid \boldsymbol{\beta})}{q(\omega_i)} d\omega_i d\boldsymbol{\beta}.$$

Note now that the first summand does not depend on $\boldsymbol{\omega}$, thus allowing us to replace this integral with $\int_{\mathbb{R}^p}\int_{\mathbb{R}^n_+} \log\{p(\boldsymbol{\beta})q(\boldsymbol{\beta})^{-1}\} q(\boldsymbol{\beta})\prod_{i=1}^n q(\omega_i) d\boldsymbol{\omega} d\boldsymbol{\beta}$. Similar arguments can be made to include $\prod_{i=1}^n q(\omega_i)$ in the second integral. Making these substitutions in the above equation we obtain

$$\begin{aligned}&\int_{\mathbb{R}^p}\int_{\mathbb{R}^n_+} \left[\log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})} + \log \frac{\prod_{i=1}^n p(Y_i \mid \boldsymbol{\beta})p(\omega_i \mid \boldsymbol{\beta})}{\prod_{i=1}^n q(\omega_i)}\right] q(\boldsymbol{\beta}) \prod_{i=1}^n q(\omega_i) d\boldsymbol{\omega} d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^n_+} q(\boldsymbol{\beta},\boldsymbol{\omega}) \log \frac{p(\boldsymbol{\beta})\prod_{i=1}^n p(Y_i \mid \boldsymbol{\beta})p(\omega_i \mid \boldsymbol{\beta})}{q(\boldsymbol{\beta},\boldsymbol{\omega})} d\boldsymbol{\omega} d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^n_+} q(\boldsymbol{\beta},\boldsymbol{\omega}) \log \frac{p(\boldsymbol{Y},\boldsymbol{\beta},\boldsymbol{\omega})}{q(\boldsymbol{\beta},\boldsymbol{\omega})} d\boldsymbol{\omega} d\boldsymbol{\beta},\end{aligned}$$

thus proving Theorem 7.1. Note that $q(\boldsymbol{\beta},\boldsymbol{\omega}) = q(\boldsymbol{\beta})\prod_{i=1}^n q(\omega_i)$ and $\int_{\mathbb{R}_+} q(\omega_i) d\omega_i = 1$. □

As is clear from Theorem 7.1, the variational strategy proposed by Jaakkola & Jordan (2000) is a pure VB minimizing $\text{KL}\{q(\boldsymbol{\beta},\boldsymbol{\omega}) \parallel p(\boldsymbol{\beta},\boldsymbol{\omega} \mid \boldsymbol{Y})\}$ under a mean-field variational family $\mathcal{Q} = \{q(\boldsymbol{\beta},\boldsymbol{\omega}) : q(\boldsymbol{\beta},\boldsymbol{\omega}) = q(\boldsymbol{\beta})\prod_{i=1}^n q(\omega_i)\}$ in the conditionally

conjugate exponential family model with

$$\begin{aligned}
&\text{Global variables} && \boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \\
&\text{Local variables} && (\omega_i \mid \boldsymbol{\beta}) \sim \text{PG}(1, \boldsymbol{x}_i^\intercal \boldsymbol{\beta}), \quad i = 1, \dots, n, \\
&\text{Data} && (Y_i \mid \boldsymbol{\beta}) \sim \text{BERN}[\exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\{1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}^{-1}], \quad i = 1, \dots, n.
\end{aligned} \tag{7.11}$$

We refer to Choi & Hobert (2013, Sect. 2) for this specific formulation of the Pólya-gamma data augmentation scheme which highlights how, unlike the general specification in (7.4), the conditional distribution of Y_i does not depend on ω_i. As discussed in Section 7.2, this is not a necessary requirement. Indeed, what is important is that the joint likelihood $p(Y_i, \omega_i \mid \boldsymbol{\beta})$ is within an exponential family and the prior $p(\boldsymbol{\beta})$ is conjugate to it. Recalling Choi & Hobert (2013, Sect. 2) and noticing that $\cosh\{0.5(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\} = 0.5[1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})] \exp\{-0.5(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}$, this is the case of (7.11). In fact

$$\begin{aligned}
p(Y_i, \omega_i \mid \boldsymbol{\beta}) &= p(Y_i \mid \boldsymbol{\beta}) p(\omega_i \mid \boldsymbol{\beta}) \\
&= \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{Y_i} \{1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}^{-1} \exp\{-0.5\omega_i (\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2\} \cosh\{0.5(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\} p(\omega_i), \\
&= 0.5 \exp\{(Y_i - 0.5)\boldsymbol{x}_i^\intercal \boldsymbol{\beta} - 0.5\omega_i (\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2\} p(\omega_i),
\end{aligned} \tag{7.12}$$

is proportional to the Gaussian kernel $\exp[(Y_i - 0.5)\boldsymbol{x}_i^\intercal \boldsymbol{\beta} - 0.5\omega_i (\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2]$, which is conjugate to $p(\boldsymbol{\beta})$.

7.4 Coordinate ascent variational inference (CAVI)

As discussed in Section 7.2, the mean-field assumption allows the implementation of a simple CAVI (Blei et al., 2017; Bishop, 2006, Ch. 10.1.1) which sequentially maximizes the evidence lower bound in (7.10) with respect to each factor in $q(\boldsymbol{\beta}) \prod_{i=1}^n q(\omega_i)$, via the following updates

$$\begin{aligned}
q^{(r)}(\boldsymbol{\beta}) &= \exp\left[\mathbb{E}_{q^{(r-1)}(\boldsymbol{\omega})} \log\{p(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{\omega})\}\right] c_{\boldsymbol{\beta}}(\boldsymbol{Y})^{-1}, \\
q^{(r)}(\omega_i) &= \exp\left[\mathbb{E}_{q^{(r)}(\boldsymbol{\beta})} \log\{p(\omega_i \mid \boldsymbol{Y}, \boldsymbol{\omega}_{-i}, \boldsymbol{\beta})\}\right] c_{\omega_i}(\boldsymbol{Y})^{-1}, \quad i = 1, \dots, n,
\end{aligned} \tag{7.13}$$

at iteration r, until convergence of the ELBO. In the above expressions, $c_{\boldsymbol{\beta}}(\boldsymbol{Y})$ and $c_{\omega_i}(\boldsymbol{Y})$, $i = 1, \dots, n$, denote constants leading to proper densities. Note that in our case $p(\omega_i \mid \boldsymbol{Y}, \boldsymbol{\omega}_{-i}, \boldsymbol{\beta}) = p(\omega_i \mid \boldsymbol{Y}, \boldsymbol{\beta})$.

To clarify why (7.13) provides a routine which iteratively improves the ELBO, and ultimately maximizes it, note that, keeping fixed $q^{(r-1)}(\omega_1), \dots, q^{(r-1)}(\omega_n)$, equation

(7.10) can be re-written as

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{\beta})}\left\{\mathbb{E}_{q^{(r-1)}(\boldsymbol{\omega})}\log\frac{p(\boldsymbol{\beta})\prod_{i=1}^{n}p(Y_i\mid\boldsymbol{\beta})p(\omega_i\mid\boldsymbol{\beta})p(\boldsymbol{Y},\boldsymbol{\omega})}{q(\boldsymbol{\beta})p(\boldsymbol{Y},\boldsymbol{\omega})}\right\}+\text{const}\\
&=\mathbb{E}_{q(\boldsymbol{\beta})}\left\{\mathbb{E}_{q^{(r-1)}(\boldsymbol{\omega})}\log\frac{p(\boldsymbol{\beta}\mid\boldsymbol{Y},\boldsymbol{\omega})}{q(\boldsymbol{\beta})}\right\}+\mathbb{E}_{q^{(r-1)}(\boldsymbol{\omega})}\log p(\boldsymbol{Y},\boldsymbol{\omega})+\text{const},\\
&=\mathbb{E}_{q(\boldsymbol{\beta})}\left(\log\frac{\exp\left[\mathbb{E}_{q^{(r-1)}(\boldsymbol{\omega})}\log\{p(\boldsymbol{\beta}\mid\boldsymbol{Y},\boldsymbol{\omega})\}\right]}{q(\boldsymbol{\beta})c_{\boldsymbol{\beta}}(\boldsymbol{Y})}\right)+\mathbb{E}_{q^{(r-1)}(\boldsymbol{\omega})}\log c_{\boldsymbol{\beta}}(\boldsymbol{Y})p(\boldsymbol{Y},\boldsymbol{\omega})+\text{const},
\end{aligned}\tag{7.14}$$

where the first term in the last equation is the only quantity which depends on $\boldsymbol{\beta}$ and is equal to the negative KL divergence between $q(\boldsymbol{\beta})$ and the distribution $\exp\left[\mathbb{E}_{q^{(r-1)}(\boldsymbol{\omega})}\log\{p(\boldsymbol{\beta}\mid\boldsymbol{Y},\boldsymbol{\omega})\}\right]c_{\boldsymbol{\beta}}(\boldsymbol{Y})^{-1}$, thus motivating the CAVI update for $q(\boldsymbol{\beta})$. Similar derivations can be done to obtain the solutions for $q(\omega_1),\dots,q(\omega_n)$ in (7.13). As is clear from (7.13), the CAVI solution identifies both the form of the approximating densities—without pre-specifying them as part of the mean-field assumption—and the optimal parameters of such densities. As discussed in Section 7.2, these solutions are particularly straightforward in conditionally conjugate exponential family representations (Hoffman et al., 2013), including model (7.11). In fact, recalling Polson et al. (2013), the full conditionals for the local and global variables in model (7.11) can be obtained via conditional conjugacy properties, which lead to

$$\begin{aligned}
&(\boldsymbol{\beta}\mid\boldsymbol{Y},\boldsymbol{\omega})\sim\mathcal{N}_p\{(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}+\boldsymbol{X}^\intercal\boldsymbol{\Omega}\boldsymbol{X})^{-1}(\boldsymbol{X}^\intercal(\boldsymbol{Y}-0.5\cdot\mathbf{1}_n)+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}),(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}+\boldsymbol{X}^\intercal\boldsymbol{\Omega}\boldsymbol{X})^{-1}\},\\
&(\omega_i\mid\boldsymbol{Y},\boldsymbol{\omega}_{-i},\boldsymbol{\beta})\sim\text{PG}(1,\boldsymbol{x}_i^\intercal\boldsymbol{\beta}),\qquad i=1,\dots,n,
\end{aligned}\tag{7.15}$$

with $\boldsymbol{\Omega}=\text{diag}(\omega_1,\dots,\omega_n)$ and $\boldsymbol{X}$ the $n\times p$ design matrix with rows $\boldsymbol{x}_i^\intercal$, $i=1,\dots,n$. Substituting these expressions in (7.13), it can be immediately noticed that the CAVI solutions have the same density of the corresponding full-conditionals.

As shown in Algorithm 7, the above expectations can be computed in closed-form since $q(\boldsymbol{\beta})$ and $q(\omega_1),\dots,q(\omega_n)$ are already known to be Gaussian and Pólya-gammas, thus requiring only sequential optimizations of natural parameters. This form of CAVI, which is discussed in Hoffman et al. (2013) and is known in the literature as variational Bayesian EM (Beal & Ghahramani, 2003), clarifies the link between CAVI and the EM in Jaakkola & Jordan (2000). Indeed, recalling Section 7.3, both methods optimize the same objective function and rely, implicitly, on the same steps. In particular, due to Lemma 7.1, the E-step in Algorithm 6 is in fact maximizing the conditional $\text{ELBO}[q(\boldsymbol{\beta})\prod_{i=1}^{n}q^{(r-1)}(\omega_i)]$ with respect to $q(\boldsymbol{\beta})$ as in the first maximization of Algorithm 7. Similarly, the M-step solution for $\boldsymbol{\varphi}$ in Algorithm 6 is actually the one maximizing the conditional

Algorithm 7: CAVI for logistic regression.

Initialize $\varphi_1^{(0)}, \ldots, \varphi_n^{(0)}$.

for $r = 1$ until convergence of the evidence lower bound $\text{ELBO}\{q(\boldsymbol{\beta}, \boldsymbol{\omega})\}$ **do**

Step 1. Maximize $\text{ELBO}\{q(\boldsymbol{\beta}) \prod_{i=1}^n q^{(r-1)}(\omega_i)\}$ with respect to $q(\boldsymbol{\beta})$. As discussed in Section 7.4, this maximization provides $q^{(r)}(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\beta}; \boldsymbol{\mu}^{(r)}, \boldsymbol{\Sigma}^{(r)})$ with

$$\boldsymbol{\Sigma}^{(r)} = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \boldsymbol{X}^\intercal \boldsymbol{\Omega}^{(r-1)} \boldsymbol{X})^{-1}, \qquad \boldsymbol{\mu}^{(r)} = \boldsymbol{\Sigma}^{(r)} \{\boldsymbol{X}^\intercal (\boldsymbol{Y} - 0.5 \cdot \mathbf{1}_n) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}}\},$$

with $\boldsymbol{\Omega}^{(r-1)} = \text{diag}\{\text{E}_{q^{(r-1)}(\omega_1)}(\omega_1), \ldots, \text{E}_{q^{(r-1)}(\omega_n)}(\omega_n)\}$.

Step 2. Maximize $\text{ELBO}\{q^{(r)}(\boldsymbol{\beta}) \prod_{i=1}^n q(\omega_i)\}$ with respect to $\prod_{i=1}^n q(\omega_i)$. As discussed in Section 7.4, this maximization provides $q^{(r)}(\omega_i) = \text{PG}(\omega_i; \varphi_i^{(r)})$ for $i = 1, \ldots, n$, with

$$\varphi_i^{(r)} = \{\boldsymbol{x}_i^\intercal \boldsymbol{\Sigma}^{(r)} \boldsymbol{x}_i + (\boldsymbol{x}_i^\intercal \boldsymbol{\mu}^{(r)})^2\}^{1/2}, \quad i = 1, \ldots, n.$$

Note that φ_i and $-\varphi_i$ induce the same Pólya-gamma density. Hence, there is no ambiguity in the above square root. A similar remark, from a different perspective, is found in footnote 3 of Jaakkola & Jordan (2000).

Output at the end of the algorithm: $q^{(*)}(\boldsymbol{\beta}, \boldsymbol{\omega}) = q^{(*)}(\boldsymbol{\beta}) \prod_{i=1}^n q^{(*)}(\omega_i)$.

$\text{ELBO}[q^{(r)}(\boldsymbol{\beta}) \prod_{i=1}^n q(\omega_i)]$ with respect to $\prod_{i=1}^n q(\omega_i)$ in the second optimization of the CAVI in Algorithm 7.

7.5 Discussion

Motivated by the success of the lower bound developed by Jaakkola & Jordan (2000) for logistic log-likelihoods, and by the lack of formal justifications for its excellent performance, we introduced a novel connection between their construction and a Pólya-gamma data augmentation developed in recent years for logistic regression (Polson et al., 2013). Besides providing a probabilistic interpretation of the bound derived by Jaakkola & Jordan (2000), this connection crucially places the variational methods associated with the proposed lower bound in a more general framework having desirable properties. More specifically, the EM for variational inference proposed by Jaakkola & Jordan (2000) maximizes a well-defined ELBO associated with a conditionally conjugate exponential family model and, hence, provides the same approximation of the CAVI for VB in this model.

The above result motivates further generalizations to novel computational methods, including the stochastic variational inference algorithm in Hoffman et al. (2013). On a similar line of research, an interesting direction is to incorporate the method of Giordano et al. (2015) to correct the variance-covariance matrix in $q^{(*)}(\boldsymbol{\beta})$ from Algorithm 6, which is known to underestimate variability. Finally, we shall also emphasize that although

our focus is on classical Bayesian logistic regression, the results in Section 7.3 can be easily generalized to more complex learning procedures incorporating logistic models as a building-block, as for the LSBP prior in Chapter 6, as long as such formulations admit conditionally conjugate exponential family representations.

Bibliography

Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9**, 1981–2014.

Aitchison, J. (1985). A general class of distributions on the simplex. *J. R. Statist. Soc. B* **47**, 136–146.

Aitchison, J. & Shen, S. M. (1980). Logistic-normal distributions: some properties and uses. *Biometrika* **67**, 262–272.

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.* **88**, 669–79.

Amemiya, T. (1981). Qualitative response models: a survey. *J. Ecom. Liter.* **19**, 1483–536.

Antoniano-Villalobos, I., Wade, S. & Walker, S. (2014). A Bayesian nonparametric regression model with normalized weights: a study of hippocampal atrophy in Alzheimer's disease. *J. Am. Statist. Assoc.* **109**, 477–90.

Arbel, J., De Blasi, P. & Prünster, I. (2018). Stochastic approximations to the Pitman-Yor process. *Bayesian Anal.* **In press**, 1–19.

Arbel, J. & Prünster, I. (2017). A moment-matching Ferguson & Klass algorithm. *Statist. Comp.* **27**, 3–17.

Barrientos, A. F., Jara, A. & Quintana, F. A. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Anal.* **7**, 277–310.

Barrios, E., Lijoi, A., Nieto-Barajas, L. E. & Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statist. Sc.* **28**, 313–34.

Bassetti, F., Casarin, R. & Rossini, L. (2018). Hierarchical species sampling models. *ArXiv:1803.05793* , 1–51.

Beal, M. J. & Ghahramani, Z. (2003). The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statist.* **7**, 453–64.

Bigelow, J. L. & Dunson, D. B. (2009). Bayesian semiparametric joint models for functional predictors. *J. Am. Statist. Assoc.* **104**, 26–36.

Billingsley, P. (1995). *Probability and Measure*. New York: John Wiley & Sons, 3rd ed.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bishop, C. M. & Svensén, M. (2003). Bayesian hierarchical mixtures of experts. In *Conference on Uncertainty in Artificial Intelligence*.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–5.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Statist. Assoc.* **112**, 859–77.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022.

Braun, M. & McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *J. Am. Statist. Assoc.* **105**, 324–35.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. in Appl. Prob.* **31**, 929–53.

Browne, R. P. & McNicholas, P. D. (2015). Multivariate sharp quadratic bounds via σ-strong convexity and the fenchel connection. *Electron. J. Stat.* **9**, 1913–38.

Camerlenghi, F., Lijoi, A., Orbanz, P. & Prünster, I. (2019). Distribution theory for hierarchical processes. *Ann. Statist.* **47**, 67–92.

Camerlenghi, F., Lijoi, A. & Prünster, I. (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scand. J. Statist.* **45**, 1062–91.

Canale, A., Durante, D. & Dunson, D. (2018). Convex mixture regression for quantitative risk assessment. *Biometrics* **74**, 1331–40.

Canale, A., Lijoi, A., Nipoti, B. & Prünster, I. (2017). On the Pitman-Yor process with spike and slab base measure. *Biometrika* **104**, 681–97.

Canale, A. & Prünster, I. (2017). Robustifying Bayesian nonparametric mixtures for count data. *Biometrics* **73**, 174–84.

Caponera, A., Denti, F., Rigon, T., Sottosanti, A. & Gelfand, A. E. (2018). Hierarchical spatio-temporal modeling of resting state fMRI data. In *Studies in Neural Data Science*, A. Canale, D. Durante, P. Paci & B. Scarpa, eds. Springer, pp. 111–30.

Carbonetto, P. & Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7**, 73–108.

Carlton, M. A. (2002). A family of densities derived from the three-parameter Dirichlet process. *J. Appl. Prob.* **39**, 764–74.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddel, A. (2017). Stan: a probabilistic programming language. *J. Statist. Soft.* **76**, 1–32.

Charalambides, C. A. (2002). *Enumerative Combinatorics*. New York: Chapman and Hall / CRC.

Choi, H. M. & Hobert, J. P. (2013). The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electron. J. Stat.* **7**, 2054–64.

Cifarelli, D. & Regazzini, E. (1978). Problemi statistici non parametrici in condizioni di scambiabilità parziale. Technical report, Quaderni Istituto Matematica Finanziaria, Università di Torino Serie III, 12.

Clogg, C. C. & Goodman, L. A. (1986). On scaling models applied to data from several groups. *Psychometrika* **51**, 123–35.

Connor, R. J. & Mosimman, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Statist. Assoc.* **64**, 194–206.

Daley, D. J. & Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes*, vol. II: General Theory and Structure. New York: Springer, 2nd ed.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. & Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pat. Anal. Mach. Intel.* **37**, 212–29.

De Iorio, M., Müller, P., Rosner, G. L. & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Am. Statist. Assoc.* **99**, 205–15.

De la Cruz-Mesía, R., Quintana, F. A. & Müller, P. (2007). Semiparametric Bayesian classification with longitudinal markers. *J. R. Statist. Soc. C* **56**, 119–137.

De Leeuw, J. & Lange, K. (2009). Sharp quadratic majorization in one dimension. *Comp. Statist. Data Anal.* **53**, 2471–84.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1–38.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer.

Diaconis, P. & Ylvisaker, D. (1979). Conjugate prior for exponential families. *Ann. Statist.* **7**, 269–92.

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics*, N. L. Hjort, C. C. Holmes, P. Muller & S. G. Walker, eds. Cambridge University Press, pp. 223–73.

Dunson, D. B. & Park, J. H. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–23.

Dunson, D. B. & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Am. Statist. Assoc.* **104**, 1042–51.

Durante, D., Canale, A. & Rigon, T. (2019). A nested expectation-maximization algorithm for latent class models. *Statist. Prob. Lett.* **146**, 97–103.

Durante, D., Dunson, D. B. & Vogelstein, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *J. Am. Statist. Assoc.* **112**, 1516–30.

Durante, D. & Rigon, T. (2019). Conditionally conjugate mean-field variational bayes for logistic models. *Statistical Science* **34**, 472–485.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90**, 577–88.

Favaro, S., Hadjicharalambous, G. & Prünster, I. (2011). On a class of distributions on the simplex. *J. Statist. Plann. Infer.* **141**, 2987–3004.

Favaro, S. & Teh, Y. W. (2013). MCMC for normalized random measure mixture models. *Statist. Sc.* **28**, 335–59.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–30.

Ferguson, T. S. & Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Statist.* **43**, 1634–43.

Fox, E. B., Sudderth, E. B., Jordan, M. I. & Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *Ann. Appl. Statist.* **5**, 1020–56.

Fritsch, A. & Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.* **4**, 367–92.

Gelfand, A. E., Kottas, A. & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Am. Statist. Assoc.* **100**, 1021–35.

Gelfand, A. E. & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

Ghosal, S., Ghosh, J. K. & Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143–58.

Ghosal, S. & Van Der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35**, 697–723.

Giordano, R. J., Broderick, T. & Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational bayes. *Adv. in Neur. Inf. Proc. Sys.* , 1441–9.

Gnedin, A. & Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov, POMI* **325**, 83–102.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unindentifiable models. *Biometrika* **61**, 215–31.

Goodman, L. A. (1975). A new model for scaling response patterns: an application of quasi independence concept. *J. Am. Statist. Assoc.* **70**, 755–68.

Green, P. J. & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Statist.* **28**, 355–75.

Griffin, J. E. & Leisen, F. (2017). Compound random measures and their use in Bayesian non-parametrics. *J. R. Statist. Soc. B* **79**, 525–45.

Griffin, J. E. & Steel, M. (2006). Order-based dependent Dirichlet processes. *J. Am. Statist. Assoc.* **10**, 179–94.

Griffin, J. E. & Steel, M. F. (2011). Stick-breaking autoregressive processes. *J. Econom.* **162**, 383–96.

Hagenaars, J. A. & McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press.

Heard, N. A., Holmes, C. C. & Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *J. Am. Statist. Assoc.* **101**, 18–29.

Hjort, N. L., Holmes, C. C., Müller, P. & Walker, S. G. (2010). *Bayesian Nonparametrics*. New York: Cambridge University Press.

Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14**, 1303–1347.

Hwang, B. S. & Pennell, M. L. (2014). Semiparametricbayesian joint modeling of a binary and continuous outcome with applications in toxicological risk assessment. *Statist. Med.* **33**, 1162–75.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc.* **96**, 161–73.

Ishwaran, H. & James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures. *J. Comp. Graph. Statist.* **11**, 508–32.

Ishwaran, H. & Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371–90.

Ishwaran, H. & Zarepour, M. (2002). Exact and approximate sum representation for the Dirichlet process. *Canad. J. Statist.* **30**, 269–83.

Jaakkola, T. S. & Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statist. Comp.* **10**, 25–37.

James, L. F., Lijoi, A. & Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.* **33**, 105–20.

James, L. F., Lijoi, A. & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.* **36**, 76–97.

Johndrow, J. E., Smith, A., Pillai, N. & Dunson, D. B. (2018). MCMC for imbalanced categorical data. *J. Am. Statist. Assoc.* **In press**.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37**, 183–233.

Kallenberg, O. (2017). *Random Measures, Theory and Applications*. Cham: Springer.

Kalli, M., Griffin, J. E. & Walker, S. G. (2011). Slice sampling mixture models. *Statist. Comp.* **21**, 93–105.

Kingman, J. F. C. (1967). Completely random measures. *Pacific J. Math.* **21**, 59–78.

Kingman, J. F. C. (1975). Random discrete distributions. *J. R. Statist. Soc. B* **37**, 1–22.

Kurihara, K. & Welling, M. (2009). Bayesian k-means as a "Maximization-Expectation" algorithm. *Neur. Comp.* **21**, 1145–72.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *J. Comp. Graph. Statist.* **13**, 183–212.

Lau, J. W. & Green, P. J. (2007). Bayesian model-based clustering procedures. *J. Comp. Graph. Statist.* **16**, 526–58.

Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lee, S., Huang, J. Z. & Hu, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Statist.* **4**, 1579–601.

Lijoi, A., Mena, R. H. & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Am. Statist. Assoc.* **100**, 1278–91.

Lijoi, A., Mena, R. H. & Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Statist. Soc. B* **69**, 715–40.

Lijoi, A. & Nipoti, B. (2014). A class of hazard rate mixtures for combining survival data from different experiments. *J. Am. Statist. Assoc.* **109**, 802–14.

Lijoi, A., Nipoti, B. & Prünster, I. (2014a). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20**, 1260–91.

Lijoi, A., Nipoti, B. & Prünster, I. (2014b). Dependent mixture models: clustering and borrowing information. *Comp. Statist. Data Anal.* **71**, 417–33.

Lijoi, A. & Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, N. L. Hjort, C. C. Holmes, P. Muller & S. G. Walker, eds. Cambridge: Cambridge University Press, pp. 80–136.

Lijoi, A., Prünster, I. & Walker, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Prob.* **18**, 1519–47.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *Ann. Statist.* **12**, 351–7.

Longnecker, M. P., Klebanoff, M. A., Zhou, H. & Brock, J. W. (2001). Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *Lancet* **358**, 110–4.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, Alexandria, VA: American Statistical Association.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Tech. rep., Department of Statistics, Ohio State University.

Malsiner-Walli, G., Frühwirth-Schnatter, S. & Grün, B. (2016). Model based clustering based on sparse finite Gaussian mixtures. *Statist. Comp.* **26**, 303–24.

Medvedovic, M. & Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–206.

Meng, X.-L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–78.

Miller, J. W. & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *J. Am. Statist. Assoc.* **113**, 340–56.

Muliere, P. & Secchi, P. (1996). Bayesian nonparametric predictive inference and bootstrap techniques. *Ann. I. Statist. Math.* **48**, 663–73.

Muliere, P. & Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canad. J. Statist.* **26**, 283–97.

Müller, P. & Mitra, R. (2013). Bayesian nonparametric inference - why and how. *Bayesian Anal.* **8**, 269–302.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Statist.* **9**, 249–65.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc. A* **135**, 370–84.

Ormerod, J. T. & Wand, M. P. (2010). Explaining variational approximations. *Am. Stat.* **64**, 140–53.

Pati, D., Dunson, D. B. & Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Mult. Anal.* **116**, 456–72.

Perman, M. (1990). *Random discrete distributions derived from subordinators*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–University of California, Berkeley.

Perman, M., Pitman, J. & Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Prob. Theory Rel. Fields* **92**, 21–39.

Petrone, S., Guindani, M. & Gelfand, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *J. R. Statist. Soc. B* **71**, 755–82.

Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. In *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, T. S. Ferguson, L. S. Shapley & J. B. MacQueen, eds., vol. 30 of *IMS Lecture notes, Monograph Series*. Hayward: Institute of Mathematical Statistics, pp. 245–67.

Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Prob.* **25**, 855–900.

Polson, N. G., Scott, J. G. & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Am. Statist. Assoc.* **108**, 1339–49.

Ramsay, J. & Silverman, B. W. (2005). *Functional data analysis*. Springer.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statist. Sc.* **3**, 425–61.

Ranganath, R., Gerrish, S. & Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*.

Rasmussen, C. E. & Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ray, S. & Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *J. R. Statist. Soc. B* **68**, 305–32.

Regazzini, E., Lijoi, A. & Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31**, 560–85.

Ren, L., Du, L., Carin, L. & Dunson, D. B. (2011). Logistic stick-breaking process. *J. Mach. Learn. Res.* **12**, 203–39.

Rennie, B. & Dobson, A. J. (1969). On Stirling numbers of the second kind. *J. Combinat. Th.* **6**, 116–21.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B* **59**, 731–92.

Ridout, M. S. (2009). Generating random numbers from a distribution specified by its Laplace transform. *Statist. Comp.* **19**, 439–50.

Rigon, T., Durante, D. & Torelli, N. (2019). Bayesian semiparametric modelling of contraceptive behaviour in India via sequential logistic regressions. *J. R. Statist. Soc. A* **182**, 225–47.

Roberts, G. O. & Rosenthal, J. S. (2009). Examples of adaptive MCMC. *J. Comp. Graph. Statist.* **18**, 349–67.

Rodriguez, A. & Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6**, 145–78.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Statist. Soc. B* **73**, 689–710.

Sato, K.-I. (1999). *Lévy processes and infinitely divisible distributions*. Cambridge University Press.

Scarpa, B. & Dunson, D. B. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics* **65**, 772–80.

Scarpa, B. & Dunson, D. B. (2014). Enriched stick-breaking processes for functional data. *J. Am. Statist. Assoc.* **109**, 647–60.

Scott, J. G. & Sun, L. (2013). Expectation-maximization for logistic regression. *arXiv:1306.0040* .

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sin.* **4**, 639–50.

Stouffer, S. A. & Toby, J. (1951). Role conflict and personality. *Am. J. Soc.* **56**, 395–406.

Tang, Y., Browne, R. P. & McNicholas, P. D. (2015). Model based clustering of high-dimensional binary data. *Comp. Statist. Data Anal.* **87**, 84–101.

Teh, Y. W. & Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*, N. L. Hjort, C. C. Holmes, P. Muller & S. G. Walker, eds. Cambridge University Press, pp. 158–207.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006). Hierarchical Dirichlet processes. *J. Am. Statist. Assoc.* **101**, 1–41.

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhya: Ind. J. Statist.* , 90–110.

Tutz, G. (1991). Sequential models in categorical regression. *Comp. Statist. Data Anal.* **11**, 275–95.

Wade, S., Dunson, D. B., Petrone, S. & Trippa, L. (2014). Improving Prediction from Dirichlet Process Mixtures via Enrichment. *J. Mach. Learn. Res.* **15**, 1041–71.

Wade, S. & Ghahramani, Z. (2018). Bayesian cluster analysis: point estimation and credible balls. *Bayesian Anal.* **13**, 559–626.

Wade, S., Mongelluzzo, S. & Petrone, S. (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Anal.* **6**, 359–86.

Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *J. Am. Statist. Assoc.* **112**, 137–68.

Wand, M. P., Ormerod, J. T., Padoan, S. A. & Frühwirth, R. (2011). Mean field variational bayes for elaborate distributions. *Bayesian Anal.* **6**, 847–900.

Wang, B. & Titterington, D. M. (2004). Convergence and asymptotic normality of variational bayesian approximations for exponential family models with missing values. *Proceedings of the Conference on Uncertainty in Artificial Intelligence* , 577–84.

Wang, C. & Blei, D. M. (2013). Variational inference in nonconjugate models. *J. Mach. Learn. Res.* **14**, 1005–1031.

Wang, X. & Roy, V. (2018a). Analysis of the Pólya-gamma block Gibbs sampler for Bayesian logistic linear mixed models. *Statist. Prob. Lett.* **137**, 251–6.

Wang, X. & Roy, V. (2018b). Geometric ergodicity of Pólya-Gamma Gibbs sampler for Bayesian logistic regression with a flat prior. *Electron. J. Stat.* **12**, 3295–311.

Zhang, L., Guindani, M., Versace, F., Engelmann, J. M. & Vannucci, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. *Ann. Appl. Statist.* **10**, 638–66.