Posterior Asymptotics for Boosted Hierarchical Dirichlet Process Mixtures

Marta Catalano MARTA.CATALANO@WARWICK.AC.UK
University of Warwick
Coventry, UK

Pierpaolo De Blasi PIERPAOLO.DEBLASI@UNITO.IT
University of Torino and Collegio Carlo Alberto
Torino, Italy

Antonio Lijoi ANTONIO.LIJOI@UNIBOCCONI.IT
Bocconi University and BIDSA
Milan, Italy

Igor Prünster IGOR@UNIBOCCONI.IT
Bocconi University and BIDSA
Milan, Italy

Abstract

Bayesian hierarchical models are powerful tools for learning common latent features across multiple data sources. The Hierarchical Dirichlet Process (HDP) is invoked when the number of latent components is a priori unknown. While there is a rich literature on finite sample properties and performance of hierarchical processes, the analysis of their frequentist posterior asymptotic properties is still at an early stage. Here we establish theoretical guarantees for recovering the true data generating process when the data are modeled as mixtures over the HDP or a generalization of the HDP, which we term *boosted* because of the faster growth in the number of discovered latent features. By extending Schwartz's theory to partially exchangeable sequences we show that posterior contraction rates are crucially affected by the relationship between the sample sizes corresponding to the different groups. The effect varies according to the smoothness level of the true data distributions. In the supersmooth case, when the generating densities are Gaussian mixtures, we recover the parametric rate up to a logarithmic factor, provided that the sample sizes are related in a polynomial fashion. Under ordinary smoothness assumptions more caution is needed as a polynomial deviation in the sample sizes could drastically deteriorate the convergence to the truth.

Keywords: Bayesian asymptotics, Dirichlet process, hierarchical process, nonparametric density estimation, partial exchangeability, Pitman-Yor process, posterior contraction rates.

1. Introduction

The probabilistic modeling of data with complex forms of dependence, beyond the simplistic exchangeability assumption, has been a major focus in Bayesian Statistics and Machine Learning for the last two decades. Among the wealth of contributions to the area, the hierarchical Dirichlet process (HDP), introduced in Teh et al. (2006), stands out as a powerful tool for unsupervised learning. It assumes that the generative process encodes the original sample into meaningful latent features or categories. A typical example comes from informa-

tion retrieval (Cowans, 2004), where the content of a collection of documents is unraveled by assigning each word of a document to one out of several topics, i.e. latent features. Topics are shared across different documents and their number is learned directly from the data, providing an effective nonparametric version of probabilistic topic models such as the latent Dirichlet allocation (Blei et al., 2003). These key properties have spurred numerous applications beyond information retrieval, including statistical genetics (Xing et al., 2007; Elliott et al., 2019), computer vision (Sudderth et al., 2008; Haines and Xiang, 2011), cognitive science (Griffiths et al., 2007) and robotics (Nakamura et al., 2011; Taniguchi et al., 2018), where it is typically not possible to fix or bound *a priori* the number of features. Each element of a group or subpopulation is assigned to a latent feature with an unknown probability which is learned from the data through a posterior distribution, thus allowing for mixed membership and borrowing of information. This is achieved through a hierarchy of Dirichlet processes (DPs) by modeling the feature distribution in each subpopulation as a DP conditionally on a common parent distribution that is also a DP. The hierarchical structure impacts the rate at which new latent features are discovered among the observations, as the overall sample size n increases: in fact, for the HDP it leads to a slowdown from the classical $\log(n)$ rate of the DP to the iterated $\log(\log(n))$. By considering a natural generalization of the HDP where, in place of a DP, the parent distribution is a Pitman-Yor process (PY) (Pitman and Yor, 1997), one gains modeling flexibility and is also able to recover the $\log(n)$ growth. We term this class of models *boosted hierarchical Dirichlet process* (bHDP) given it speeds up the growth rate of the number of features, while preserving the DP at the subpopulation level. These growth rates follow from the general distribution theory for hierarchical processes in Camerlenghi et al. (2019).

It is important to remark that the HDP and general hierarchical processes (Teh et al., 2006; Teh and Jordan, 2010; Camerlenghi et al., 2017, 2019) fall within the framework of partial exchangeability, which represents the natural probabilistic dependence structure for multiple populations. In fact, a wide array of partially exchangeable models has been proposed in the literature and we refer the reader to the reviews in, e.g., Foti and Williamson (2013) and Quintana et al. (2022). However, these extensive studies focus on distributional properties, while the analysis of their frequentist asymptotic behaviour is still at its early stage. This paper provides a two-fold contribution in this direction: (1) we generalize Schwartz's theory (Schwartz, 1965), which is pivotal to frequentist asymptotics in the exchangeable case, to the partially exchangeable setup; (2) we derive posterior contraction rates for multivariate mixtures with respect to the bHDP, which includes the HDP as a special case, and establish that they crucially depend on the relation between the sample sizes corresponding to different groups.

As far as the extension of Schwartz's theory is concerned, it is worth to recall its key ideas and modern use, first pioneered in Ghosal et al. (1999) for DP mixtures, in the simple exchangeable setup. The basic form of frequentist validation of Bayesian inference is posterior consistency, that is, the convergence of the posterior to the true data generating distribution. Schwartz's theory provides a general framework for dealing with consistency when the posterior distribution is not available in closed form. It relies on two ingredients: (a) the existence of a sequence of tests that separates the true P_0 from all the probability distributions that do not lie in any neighborhood of P_0 with exponentially small errors as the number of observations grows (i.e., an *exponentially consistent test*); (b) sufficient prior

mass on neighborhoods of P_0 defined in terms of the Kullback-Leibler divergence. In a nutshell, one has to find an exponentially consistent test by appointing a high-mass-low-entropy sieve and, then, combine it with the Kullback-Leibler support condition to ensure consistency. Suitable refinements allowed to derive also contraction rates for densities on the real line (Ghosal and van der Vaart, 2001; Ghosal and van Der Vaart, 2007) and corresponding results for multivariate densities (Tokdar, 2006; Shen et al., 2013; Canale and De Blasi, 2017). As pointed out in Wu and Ghosal (2010), the extension to the multivariate setting is highly non–trivial and requires the construction of a new sieve with low entropy and high mass. In Section 4 this theory is extended to partially exchangeable models, with special focus on the case of different sample sizes across groups.

As for the derivation of posterior contraction rates, our asymptotic analysis aims at providing theoretical guarantees for multivariate mixtures with respect to hierarchical processes to recover the true data generating distributions, when data are recorded under different, though related, experimental conditions. Each experimental condition identifies a population, or a group: while the number of groups is kept fixed, asymptotics is investigated as the number of observations in each group diverges at possibly different rates. To the best of our knowledge, only Nguyen (2016) dealt with this topic: they focused on the recovery of the parent mixing distribution of the HDP rather than on the true data-generating processes, and under the different asymptotic regime of a divergent number of populations. However, in many applications the number of groups is bounded by the experiment (for example, data clustered by blood types or by logfiles) and, thus, the natural asymptotic regime corresponds to letting the number of observations within each group diverge, as pursued here. More precisely, we determine conditions that yield posterior convergence rates for the bHDP mixture model according to metrics on the product space that are built on popular distances such as the total variation and the Hellinger. We stress that, since the bHDP includes the HDP as a special case, this also provides an asymptotic validation of the HDP mixture model. Our extension of Schwartz's theory to partially exchangeable models proves crucial for these achievements, which may be summarized as follows: (i) the simultaneous reproduction of the distributions of all subpopulations is very different from reproducing them one at a time; (ii) if all groups have the same number of observations (asymptotically), Schwartz's theory for exchangeable sequences may be easily extended to partially exchangeable sequences. (iii) if the groups have a different number of observations, one has to develop a non-trivial extension of the classical theory. A crucial difference with respect to the exchangeable case is that we can not directly build a frequentist test that separates the true distributions with exponentially bounded errors with respect to the total number of observations. However, in Lemma 5 we manage to build an exponentially consistent test with respect to the minimum of the group sizes (n). We then require the prior to put sufficient mass on a *reinforced* multivariate Kullback–Leibler neighborhood, as will be made clear in Section 4. This is of great importance in many applied settings where the observations in each group grow at different rates. For instance, for patients grouped according to blood type one cannot assume the same growth rate across groups as there will be consistently more patients with type $0+$ than $AB-$. Although in a different modeling framework, different sequence lengths have been shown to play a role also in Wei and Nguyen (2021). When applied to the bHDP mixture model, the fulfillment of the reinforced Kullback–Leibler condition forces a maximum discrepancy between the smallest and the

largest cardinality of the subpopulations (n and $n_\vee$, respectively). In the supersmooth case (see Definition 2 below), when the largest increases at most polynomially with respect to the smallest, i.e. $n_\vee \lesssim n^k$ for some $k > 0$, we find a parametric rate of convergence $\sqrt{n}$ up to a logarithmic factor. If the rate of increase is faster than polynomial, the contraction rate progressively deteriorates and fails to converge whenever $n_\vee \gtrsim e^{n^\gamma}$, where γ is a constant that depends on the dimension of the sample space and on the tails of the true mixing distributions. In the ordinary smooth case, the asymptotic divergence of the ratio $n_\vee/n$ enters even more explicitly in determining the posterior convergence rate. The minimax rate is achieved when the sample sizes are asymptotically of the same order and even a polynomial deviation in the sample sizes suffices to yield drastically deteriorated rates of convergence to the truth. This suggests to use particular care when analyzing data about populations that differ greatly in size, such as the spread of a virus in countries where artificial immunization may be available or not.
The paper is organized as follows. After recalling some preliminary notions in Section 2, in Section 3 we state the main result on the contraction rates for the bHDP mixture model (Theorem 3). The following sections develop the analytical tools for achieving Theorem 3. In Section 4 we extend Schwartz's theory to partially exchangeable models, with a particular focus on the case of different cardinalities between groups. This is applied to the bHDP mixture model in Section 5, whereas future developments are discussed in Section 6. All proofs are deferred to Section 7.

2. Background and preliminaries

In this section we give a concise summary of the preliminary concepts and notation that underlie the rest of the work.

Sets and real functions. Given a set A, we indicate with A^c its complement and with $|A|$ its cardinality. Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two sequences on $\mathbb{R}$ such that $a_n \to \infty$ and $b_n \to \infty$ as $n \to +\infty$. We write $a_n \asymp b_n$ if they are of the same order as a function of n, i.e., $a_n b_n^{-1} \to K$ as $n \to +\infty$, where $K \neq 0$; if $K = 0$, we write $a_n \ll b_n$. Moreover, we use the notation $a_n \lesssim b_n$ if $a_n \leq c\, b_n$ holds for all n and some absolute constant $c > 0$; similarly for $a_n \gtrsim b_n$. The negative part of the logarithm is denoted by $\log_-(\cdot) = \max(-\log(\cdot), 0)$. The ascending factorial of $\beta \in \mathbb{R}$ is $\beta^{[n]} = \Gamma(\beta + n)/\Gamma(\beta)$ for any integer $n \geq 0$, where Γ is the gamma function. Given $(a_i)_{i=1}^m \in \mathbb{R}^m$, we write $a = \min(a_1, \ldots, a_m)$ and $a_+ = a_1 + \cdots + a_m$. For $x > 0$, let $\lceil x \rceil = \min\{n \in \mathbb{N} : n \geq x\}$ and $\lfloor x \rfloor = \lceil x - 1 \rceil$.

Measure theory. Let $\mathbb{X}$ denote a Polish space and let $\mathcal{P}_\mathbb{X}$ indicate the space of probability distributions on $\mathbb{X}$, endowed with the topology of weak convergence. In most cases we will write $\mathcal{P} = \mathcal{P}_\mathbb{X}$. We denote by $\mathcal{P}^m = \prod_{i=1}^m \mathcal{P}$ the Cartesian product space with product topology. Let $P \in \mathcal{P}$. When P is absolutely continuous with respect to a measure λ, p denotes its density function. The n-fold product probability measure is $P^n = \prod_{i=1}^n P \in \mathcal{P}^n$ and p^n the corresponding density. If $\phi : \mathbb{X} \to \mathbb{R}$ is a measurable function such that $\int |\phi| dP < \infty$, then $P(\phi)$ denotes the expected value of ϕ with respect to the probability $P \in \mathcal{P}$. A *test* is any measurable function $\phi : \mathbb{X}^k \to [0, 1]$ for some $k \in \mathbb{N}$. The Lebesgue measure on $A \subset \mathbb{R}^n$ is denoted by $\mathscr{L}_n(A)$.

Metric spaces. The Kullback–Leibler divergence between two densities p_1 and p_2 is $\mathrm{KL}(p_1;p_2) = P_1(\log(p_1/p_2))$; the Hellinger distance is $d_H(p_1,p_2) = \int(\sqrt{p_1}-\sqrt{p_2})^2\,d\lambda$; the total variation distance is $\mathrm{TV}(p_1,p_2) = 2^{-1}\int|p_1-p_2|\,d\lambda$. The L_s–norm of a density p is $\|p\|_s = (\int|p|^s\,d\lambda)^{1/s}$, so that $\mathrm{TV}(p_1,p_2) = 2^{-1}\|p_1-p_2\|_1$. Similarly, the ℓ_s–norm in $\mathbb{R}^n$ is $\|(q_i)_{i=1}^n\|_s = (\sum_{i=1}^n|q_i|^s)^{1/s}$. For any $\epsilon>0$, a subset T_ϵ of a metric space (T,d) is an ϵ-net if for every $t\in T$ there exists $t^*\in T_\epsilon$ such that $d(t^*,t)<\epsilon$. We call the ϵ-covering of T the minimal cardinality of ϵ-nets T_ϵ and denote such number by $\mathscr{N}(\epsilon,T,d)$.

Gaussian mixtures. Let ϕ_Σ be the density of the d–dimensional Gaussian distribution centered in the origin with covariance matrix Σ, where Σ is a positive-definite matrix of dimension $d\times d$. We denote its eigenvalues by $\mathrm{eig}_1(\Sigma)\le\cdots\le\mathrm{eig}_d(\Sigma)$. For any mixing distribution $F\in\mathcal{P}_{\mathbb{R}^d}$, the multivariate Gaussian mixture density is defined as $p_{F,\Sigma}(\boldsymbol{x}) = \int_{\mathbb{R}^d}\phi_\Sigma(\boldsymbol{x}-\boldsymbol{y})\,F(d\boldsymbol{y})$, where $\boldsymbol{x}=(x_1,\dots,x_d)$, $\boldsymbol{y}=(y_1,\dots,y_d)$.

Dirichlet process (mixtures). The Dirichlet process is a probability distribution on the space of distributions $\mathcal{P}$ first introduced by Ferguson (1973). Among the various possible constructions, we here present the stick–breaking representation provided by Sethuraman (1994). Here and after we will indicate independent and identically distributed random variables as iid. Let $\theta>0$ and $F^*\in\mathcal{P}$. Consider $Z_i\overset{\text{iid}}{\sim}F^*$ and $V_i\overset{\text{iid}}{\sim}\mathrm{Beta}(1,\theta)$, with $(V_i)_{i\ge1}$ independent from $(Z_i)_{i\ge1}$, and define $W_i = V_i\prod_{j=1}^{i-1}V_j$. A random probability $\tilde F$ is a Dirichlet process of base probability F^* and concentration parameter θ if $\tilde F = \sum_{i\ge1}W_i\,\delta_{Z_i}$ in distribution. We write $\tilde F\sim\mathrm{DP}(\theta,F^*)$.

In the context of density estimation, the Dirichlet process and other discrete priors are most notably used within mixture models, an avenue first pioneered by Lo (1984) and extensively reviewed in Hjort et al. (2010). Here, one models a random density function as a (Gaussian) mixture $p_{F,\Sigma}$, with $F\sim\mathrm{DP}(\theta,F^*)$ and $\Sigma\sim G$ independently, where G is a probability measure on the space of $d\times d$ positive–definite real matrices.

Pitman-Yor process. The Pitman–Yor process or two parameters Dirichlet process (Perman et al., 1992; Pitman and Yor, 1997) is a generalization of the Dirichlet process that accounts for slower decreasing weights. Let $\alpha\in[0,1)$, $\theta>-\alpha$ and $F^*\in\mathcal{P}$. Consider $Z_i\overset{\text{iid}}{\sim}F^*$ and $V_i\overset{\text{ind}}{\sim}\mathrm{Beta}(1-\alpha,\theta+i\alpha)$, with $(V_i)_{i\ge1}$ independent from $(Z_i)_{i\ge1}$ and define $W_i = V_i\prod_{j=1}^{i-1}V_j$. A random probability $\tilde F\sim\mathrm{PY}(\alpha,\theta,F^*)$ if $\tilde F = \sum_{i\ge1}W_i\,\delta_{Z_i}$ in distribution. In particular, when $\alpha=0$ we recover the Dirichlet process.

Partial exchangeability. Let $\boldsymbol{X}_i=(X_{i,1},\dots,X_{i,n_i})$ be a sample of size n_i on $\mathbb{X}$. Then $\boldsymbol{X}=(\boldsymbol{X}_1,\dots,\boldsymbol{X}_m)$ is partially exchangeable if

$$\begin{aligned}\boldsymbol{X}_i|(\tilde P_1,\dots,\tilde P_m)&\overset{\text{ind}}{\sim}\tilde P_i^{n_i}\qquad\text{for } i=1,\dots,m;\\(\tilde P_1,\dots,\tilde P_m)&\sim\Pi,\end{aligned}\tag{1}$$

where Π is a probability on the product space of probability distributions $\mathcal{P}^m$. In particular, each group of observations is exchangeable and such that $\boldsymbol{X}_i|\tilde P_i\overset{\text{iid}}{\sim}\tilde P_i$ with $\tilde P_i\sim\Pi_i$, the i-th marginal distribution of Π. This work focuses on partially exchangeable models (1) where the realizations of Π are dominated by a σ–finite measure almost surely. For example, when dealing with Gaussian mixture models the dominating measure is the Lebesgue measure on $\mathbb{R}^d$. Hereafter, we denote by $\Pi(\cdot|\boldsymbol{X})$ a version of the posterior distribution of $(\tilde P_1,\dots,\tilde P_m)$.

3. Posterior contraction rates for boosted hierarchical Dirichlet process mixtures

In this section we define the boosted HDP mixture model and we state the main result of the paper (Theorem 3), which provides its posterior contraction rates. The proof of Theorem 3 will then be built in Sections 4 and 5.

3.1 Boosted HDP mixtures

We introduce the boosted hierarchical Dirichlet process (bHDP) as a model for dependent random probabilities that naturally extends the hierarchical Dirichlet process (HDP; Teh et al. (2006)), by introducing a Pitman–Yor process as common parent distribution. Let $\theta, \theta^* > 0$, $\alpha^* \in [0, 1)$ and $F^* \in \mathcal{P}$. Then $\boldsymbol{F} = (F_i)_{i=1}^m \sim \text{bHDP}(\theta, \alpha^*, \theta^*, F^*)$ if

$$\begin{aligned} F_i|\tilde{F} &\overset{\text{iid}}{\sim} \text{DP}(\theta, \tilde{F}) \qquad \text{for } i = 1, \dots, m; \\ \tilde{F} &\sim \text{PY}(\alpha^*, \theta^*, F^*). \end{aligned} \tag{2}$$

When $\alpha^* = 0$ we recover the HDP. Here we consider the bHDP as latent nonparametric prior within a multivariate Gaussian mixture model. Let $\boldsymbol{X}_i = (X_{i,1}, \dots, X_{i,n_i})$ be a sample of size n_i on $\mathbb{X}$, $i = 1, \dots, m$. We define the *boosted* HDP *(Gaussian) mixture model* for $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_m)$ as

$$\begin{aligned} \boldsymbol{X}_i|\boldsymbol{F}, \boldsymbol{\Sigma} &\overset{\text{ind}}{\sim} p_{F_i,\Sigma_i}^{n_i} \qquad \text{for } i = 1, \dots, m; \\ (\boldsymbol{F}, \boldsymbol{\Sigma}) &\sim \text{bHDP}(\theta, \alpha^*, \theta^*, F^*) \times G^m, \end{aligned} \tag{3}$$

where $\boldsymbol{F} = (F_i)_{i=1}^m$ are probability measures on $\mathbb{R}^d$, $\boldsymbol{\Sigma} = (\Sigma_i)_{i=1}^m$ are covariance matrices and G is a probability measure on the space of $d \times d$ positive-definite real matrices.

3.2 Posterior contraction rates

The boosted HDP mixture model induces partially exchangeable observations (1) and thus to study its contraction rates we consider this general setup.

As it is customary in posterior asymptotic analysis, we assume that the data are generated from a fixed distribution. Specifically, we consider $\boldsymbol{X}_i \overset{\text{ind}}{\sim} P_{0,i}^{n_i}$ for $i = 1, \dots, m$, where $P_{0,i} \in \mathcal{P}$ denotes the true data generating distribution in group i. It follows that $\boldsymbol{X}$ is made of m independent groups of observations and within each group the data are iid. It is also assumed that each $P_{0,i}$ is dominated by the same measure as the realizations of $\tilde{P}_i$ and we denote by $p_{0,i}$ its density. For a fixed number of groups m, we analyze the properties of $\Pi(\cdot|\boldsymbol{X})$ for recovering $(P_{0,i})_{i=1}^m$ as the number of observations n_i in each group diverges to $+\infty$ with possibly different speed. Hence, we define $n = \min(n_1, \dots, n_m)$ and we examine the behavior of $\Pi(\,\cdot\,|\boldsymbol{X})$ as $n \to +\infty$.

To this end, we recall that every topology on the space of distributions $\mathcal{P}$ is inherited by the product space $\mathcal{P}^m$ through the product topology. If d is a baseline metric $\mathcal{P}$ defined on densities, one of the most notable classes of product metrics are the ℓ_s–metrics for $1 \leq s < +\infty$, that is

$$d_s((p_i)_{i=1}^m,(q_i)_{i=1}^m)=\left(\sum_{i=1}^m d(p_i,q_i)^s\right)^{\frac{1}{s}}. \tag{4}$$

Definition 1 *A sequence ϵ_n is a posterior contraction rate at $(p_{0,i})_{i=1}^m$ with respect to d_s if*

$$\Pi(d_s((p_i)_{i=1}^m,(p_{0,i})_{i=1}^m)\geq M_n\epsilon_n|\boldsymbol{X})\to 0$$

in $(\prod_{i=1}^m P_{0,i}^\infty)$–probability for every $M_n\to+\infty$, as $n\to+\infty$.

We determine contraction rates for true densities $p_{0,i}$ that satisfy one of the two alternative smoothness regimes detailed below. To this aim, for a multi-index $k=(k_1,\ldots,k_d)$ of nonnegative integers k_i define $k_+=\sum_{j=1}^d k_j$ and let $D^k=\partial^{k_+}/\partial x_1^{k_1}\cdots\partial x_d^{k_d}$ denote the mixed partial derivative operator. Also denote by $[-z,z]^d=\prod_{i=1}^d[-z,z]$ the d-dimensional cube of side $[-z,z]$.

Definition 2 *A density function p_0 on $\mathbb{R}^d$ is said to be*

- *supersmooth if there exist (F_0,Σ_0) such that $p_0=p_{F_0,\Sigma_0}$ and $1-F_0([-z,z]^d)\lesssim e^{-c_0z^{r_0}}$ for every $z>0$, with $c_0>0$ and $r_0\geq 2$.*
- *β-smooth if*
  1. *The mixed partial derivatives D^kp_0 of order up to $k_+\leq\lfloor\beta\rfloor$ satisfy*

$$|(D^kp_0)(x+y)-(D^kp_0)(x)|\leq L(x)e^{c_1\|y\|^2}\|y\|^{\beta-\lfloor\beta\rfloor},\quad k_+=\lfloor\beta\rfloor,\ x,y\in\mathbb{R}^d;$$
$$P_0\left(\frac{L+|D^kp_0|}{p_0}\right)^{(2\beta+\epsilon)/\beta}<\infty,\quad k_+\leq\lfloor\beta\rfloor;$$

     for some function $L:\ \mathbb{R}^d\to[0,\infty)$ and positive constants c_1,ϵ.
  2. *For every $\|x\|>a$, $p_0(x)\leq ce^{-b\|x\|^\tau}$, for some positive constants a,b,c,τ.*

These two smoothness regimes are standard in the analysis of contraction rates for Bayesian nonparametric models. In particular, a density is supersmooth when it is a Gaussian mixture with light tailed mixture distribution, whereas it is β-smooth when the derivatives are sufficiently regular and the tails are sufficiently light (Fan, 1991).

3.3 Main result

Our main result concerns the posterior contraction rates of the boosted HDP mixture model in the two smoothness regimes above by using the Hellinger distance as baseline metric. It is worth mentioning that the same rates hold for the total variation distance as well. We make some standard assumptions on the parameters F^* and G of the bHDP mixture model in (3) and ask for the existence of positive constants a_k,C_k,b_k and κ such that

$$\begin{aligned}
1-F^*([-z,z]^d)&\leq b_1e^{-C_1z^{a_1}}\quad\text{for sufficiently large } z>0,\\
G(\Sigma:\operatorname{eig}_d(\Sigma^{-1})\geq s)&\leq b_2e^{-C_2s^{a_2}}\quad\text{for sufficiently large } s>0,\\
G(\Sigma:\operatorname{eig}_1(\Sigma^{-1})\leq s)&\leq b_3s^{a_3}\quad\text{for sufficiently small } s>0,\\
G\Big(\bigcap_{1\leq j\leq d}\Big\{\Sigma:\ s_j<\operatorname{eig}_j(\Sigma^{-1})<s_j(1+t)\Big\}\Big)&\geq b_4s_1^{a_4}t^{a_5}e^{-C_3s_d^{\kappa/2}},
\end{aligned} \tag{5}$$

where the last claim holds for any $0 < s_1 \leq \cdots \leq s_d$ and $0 < t < 1$. For example, these conditions hold with $\kappa = 2$ when G is equal to an inverse Wishart distribution with positive-definite scale matrix or when G is supported on the diagonal matrices with iid eigenvalues from a distribution with polynomial tail at 0 and an exponential tail at infinity.

Theorem 3 *Let $\Pi(\cdot|\boldsymbol{X})$ be the posterior of the bHDP mixture model such that (5) hold and the true density $p_{0,i}$ is supersmooth with tail parameter $r_{0,i}$ for $i = 1, \ldots, m$. Let $r_0 = \min(r_{0,1}, \ldots, r_{0,m})$ and assume that there exists $0 < \delta < (d + d/r_0 + 2)^{-1}$ such that $n_\vee \lesssim e^{n^\delta}$ for n large enough. Then the posterior convergence rate is $\epsilon_n = n^{-1/2}(\log n_\vee)^{(d+d/r_0+2)/2}$.*

A first thing to note is that, when $m = 1$, Theorem 3 provides the same rate of convergence of Dirichlet process mixtures, as stated for example in Ghosal and van der Vaart (2017, Theorem 9.9). We thus get the same rate of the Dirichlet process mixture for a new class of models for exchangeable observations, which correspond to the marginal sequences of the bHDP mixture model. This represents an interesting by-product since the marginal densities are not convoluted with a Dirichlet process, but rather with a mixture of Dirichlet processes, whose base probability is distributed according to a Pitman–Yor process.
More importantly, Theorem 3 provides contraction rates as the sample sizes n_i are allowed to grow at different speed. In particular, if the size of the largest group grows at a polynomial speed with respect to the smallest, i.e. $n_\vee \lesssim n^k$ for some $k > 0$, the contraction rate is parametric up to a logarithmic factor with respect to the cardinality of the smallest group, namely, $n^{-1/2}\log(n)^{(d+d/r_0+2)/2}$ is a contraction rate. However, when the growth becomes exponentially fast the contraction rate deteriorates progressively, becoming non–informative whenever $n_\vee \gtrsim e^{n^{1/(d+d/r_0+2)}}$. The twofold role of the tail parameter is remarkable: as r_0 increases, it makes both the contraction rate faster and the range of growth rates of $n_\vee$ wider. For the sake of illustration, consider two groups of observations having cardinality n_1 and n_2, respectively. Assume that $(\boldsymbol{X}_1, \boldsymbol{X}_2) \sim \text{Norm}(a_1, \Sigma_1)^{(n_1)} \times \text{Norm}(a_2, \Sigma_2)^{(n_2)}$, where $\text{Norm}(a, \Sigma)$ is a bivariate Gaussian distribution with mean vector a and covariance matrix Σ, so that $d = 2$ and $r_0 = +\infty$. If, with no loss of generality, we take n_1 and n_2 diverging such that $n_1 \lesssim n_2$, the posterior distribution of the densities corresponding to the bHDP mixture model contracts towards the vector of true Gaussian distributions with rate $\sqrt{\log(n_2)^4/n_1}$ whenever $n_2 \lesssim e^{{n_1}^{1/4}}$. On the contrary, when $n_2 \gtrsim e^{{n_1}^{1/4}}$, convergence to the truth is not ensured.
The following theorem presents the contraction rates in the ordinary smooth case.

Theorem 4 *Let $\Pi(\cdot|\boldsymbol{X})$ be the posterior of the bHDP mixture model such that (5) holds and the true density $p_{0,i}$ is β_i-smooth with tail parameter τ for $i = 1, \ldots, m$. Let $\beta = \min(\beta_1, \ldots, \beta_m)$ and assume that there exists $0 < a < 2\beta/d^*$ such that $n_\vee \lesssim n^{1+a}$ for n large enough. Then the posterior convergence rate is $\epsilon_n = (n_\vee/n)^{1/2}\, n_\vee^{-\beta/(2\beta+d^*)} \log(n_\vee)^t$, where $d^* = d \vee \kappa$ and $t > \beta(d^*/\tau + d^* + 1 + d^*/\beta)/(2\beta + d^*)$.*

In this case the posterior rate of Dirichlet process mixtures in Ghosal and van der Vaart (2017, Theorem 9.9) is recovered when $m = 1$ and $n_\vee \asymp n$. As soon as the sample sizes diverges at different speed, the contraction rates deteriorates by a factor proportional to the square root of $n_\vee/n$. This implies that, in contrast to the supersmooth case, a polynomial deviation in the sample sizes suffices to slow down the convergence, which is not guaranteed

whenever $n_\vee \geq n^{\frac{2\beta+d^*}{d^*}}$. Note that the rate in Theorem 3 is recovered when the smoothness level β diverges to infinity.

4. Posterior asymptotics for partially exchangeable models

In this section we extend Schwartz's theory to general partially exchangeable models (1), which is pivotal for obtaining the posterior convergence results on bHDP mixtures in Section 5. More specifically, in the following Theorem 7 we establish general conditions under which the convergence rate of the joint posterior $\Pi(\,\cdot\,|\boldsymbol{X})$ of $(\tilde{P}_1,\ldots,\tilde{P}_m)$ can be deduced from the marginal rates of $p_i|\boldsymbol{X_i}$ for $i=1,\ldots,m$. We stress that, whenever the cardinalities of the groups do not all grow at the same rate, this represents quite a delicate issue. In fact, the connection between marginal and joint convergence rates is far from trivial since, in general, posterior consistency for the marginal exchangeable sequences does not imply the one for the partially exchangeable ones. This is evident from the next example.

Example 1 We consider consistency with respect to the weak topology and set $m=2$. Let $\mathcal{U}_0=\{\mathrm{KL}(p_{0,1},p_1)<\tilde{\epsilon}\}\times\{\mathrm{KL}(p_{0,2},p_2)<\tilde{\epsilon}\}$ for some $\tilde{\epsilon}>0$. Let Π be a prior whose support is $\mathcal{U}_0^c$ and that satisfies

$$\int_{\{\mathrm{KL}(p_{0,1},p_1)<\epsilon\}\times\mathcal{P}}\Pi(dp_1,dp_2)>0,\qquad \int_{\mathcal{P}\times\{\mathrm{KL}(p_{0,2},p_2)<\epsilon\}}\Pi(dp_1,dp_2)>0, \tag{6}$$

for every $\epsilon>0$. Since $\Pi(\mathcal{U}_0)=0$, the posterior according to the partially exchangeable model (1) satisfies $\Pi(\mathcal{U}_0|\boldsymbol{X})=0$, $\mathcal{L}(\boldsymbol{X})$–almost surely, for every $n_1,n_2\in\mathbb{N}\setminus\{0\}$. Then $\Pi(\mathcal{U}_0|\boldsymbol{X})=0$ almost surely with respect to $P_{0,1}^{n_1}\times P_{0,2}^{n_2}$ and, thus, also in probability. As $\mathcal{U}_0$ is a neighborhood of $(P_{0,1},P_{0,2})$ according to the weak topology, Π is not consistent at $(P_{0,1},P_{0,2})$. However, (6) guarantees that the marginal random measures Π_1 and Π_2 satisfy the KL–property of Schwartz theorem for exchangeable sequences, ensuring marginal consistency (see e.g. Ghosal and van der Vaart (2017, Example 6.20)).

We assume that the product metric d_s in (4) is defined in terms of a metric d that satisfies the following basic testing assumption: given $p_0\in\mathcal{P}$, for every $n\in\mathbb{N},\epsilon>0$ and $p_1\in\mathcal{P}$ such that $d(p_0,p_1)>\epsilon$, there exists a test $\phi_n:\mathbb{X}^n\to[0,1]$ and some universal constants $K>0$ and $\xi\in(0,1)$ such that

$$P_0^n(\phi_n)\leq e^{-Kn\epsilon^2};\qquad \sup_{d(p,p_1)<\xi\epsilon}P^n(1-\phi_n)\leq e^{-Kn\epsilon^2}. \tag{7}$$

This standard requirement holds for the Hellinger distance with $K=1/8$ and $\xi=1/2$ (Le Cam, 1986). More generally, it holds for any metric $d\leq d_H$ that generates convex balls (cfr. Proposition D.8 in Ghosal and van der Vaart (2017)), including the total variation distance. It plays an important role in building marginal frequentist tests $\phi_{n_i}^i:\mathbb{X}^{n_i}\to\{0,1\}$ that separate the true distribution $P_{0,i}$ from the complement of any neighborhood with exponentially small error probabilities with respect to the number of observations. A crucial difference with respect to the exchangeable case is that we can not directly build a frequentist test $\phi:\mathbb{X}^{n_+}\to\{0,1\}$ that separates the true distributions $(P_{0,i})_{i=1}^m$ with exponentially bounded errors with respect to the total number of observations n_+, unless $n_i\asymp n_{i'}$ for

every i, i'. However, in Lemma 5 we manage to build an exponentially consistent sequence of tests with respect to the minimum group size n. Given tests $\{\phi^i\}_{i=1}^m$, we define the union–intersection test as

$$\phi(x_1, \dots, x_m) := \sum_{k=1}^{m} (-1)^{k-1} \sum_{I \in \mathcal{I}_{m,k}} \prod_{i \in I} \phi^i(x_i), \tag{8}$$

where $\mathcal{I}_{m,k} = \{I \subset \{1, \dots, m\} : |I| = k\}$. Indeed, if we restrict to $\phi^i = \mathbb{1}_{A_i^c}$, then the rejection region of ϕ is the union set $A = \bigcup_{i=1}^m A_i$, where with a small abuse of notation A_i also denotes its natural injection in the product space. Hence the null hypothesis of ϕ corresponds to the intersection of the null hypotheses of the tests ϕ^i, once injected in the product space.

Lemma 5 *Let $\mathbb{X}_1, \dots \mathbb{X}_m$ be Polish spaces and let $P_{0,i} \in \mathcal{P}_{\mathbb{X}_i}$ for $i = 1, \dots, m$. Given a neighborhood $\mathcal{U}_{0,i}$ of $P_{0,i}$ and a measurable subset $\mathcal{P}_i \subset \mathcal{P}_{\mathbb{X}_i}$, assume that there exists $\alpha_i, \beta_i > 0$ and a test $\phi^i : \mathbb{X}_i \to [0,1]$ such that*

$$P_{0,i}(\phi^i) < \alpha_i, \qquad \sup_{p \in \mathcal{P}_i \cap \mathcal{U}_{0,i}^c} P(1 - \phi^i) < \beta_i,$$

for $i = 1, \dots, m$. Then $\phi : \prod_{i=1}^m \mathbb{X}_i \to [0,1]$ in (8) is a test that satisfies

1. $\big(\prod_{i=1}^m P_{0,i}\big)(\phi) < m \max(\alpha_1, \dots, \alpha_m)$;
2. $\sup_{(P_i)_{i=1}^m \in \mathcal{P}^m \cap \mathcal{U}_0^c} \big(\prod_{i=1}^m P_i\big)(1 - \phi) < \max(\beta_1, \dots, \beta_m)$;

where $\mathcal{P}^m = \mathcal{P}_1 \times \cdots \times \mathcal{P}_m$ and $\mathcal{U}_0 = \mathcal{U}_{0,1} \times \cdots \times \mathcal{U}_{0,m}$.

We apply Lemma 5 to marginal frequentist tests $\phi_{n_i}^i : \mathbb{X}^{n_i} \to \{0,1\}$ with exponentially small error probabilities with respect to the number of observations n_i. The fact that ϕ is exponentially consistent with respect to n instead of n_+ leads to the need for a reinforced Kullback–Leibler condition. We define the *reinforced Kullback–Leibler variation neighborhood* as

$$\mathcal{V}_{0,\epsilon,n} = \left\{ \mathrm{KL}(p_{0,i}; p_i), V(p_{0,i}; p_i) \lesssim \frac{n}{n_i} \epsilon^2 \text{ for } i = 1, \dots m \right\}, \tag{9}$$

where $V(p;q) = P|\log(p/q) - \mathrm{KL}(p;q)|^2$ is the Kullback–Leibler variation. Note that (9) differs from the standard definition of Kullback–Leibler variation neighborhood $\mathcal{V}_{0,\epsilon} = \{\mathrm{KL}(p_{0,i}; p_i), V(p_{0,i}; p_i) \lesssim \epsilon^2 \text{ for } i = 1, \dots m\} \supset \mathcal{V}_{0,\epsilon,n}$, as it introduces an explicit dependence on the cardinality of the samples so to shrink each component $\{\mathrm{KL}(p_{0,i}; p_i), V(p_{0,i}; p_i) \lesssim \epsilon^2\}$ of the neighborhood proportionally to the ratio between $n = \min(n_1, \dots, n_m)$ and n_i. We added a subscript n in the notation $\mathcal{V}_{0,\epsilon,n}$, though technically it also depends on the whole vector $(n_1, \dots, n_m)$. We observe that when $n_i \asymp n_{i'}$ for every $i \neq i'$, $\mathcal{V}_{0,\epsilon,n}$ and $\mathcal{V}_{0,\epsilon}$ coincide, otherwise $\mathcal{V}_{0,\epsilon,n} \supset \mathcal{V}_{0,\epsilon}$. The need of a reinforced Kullback–Leibler neighborhood is strictly linked with the test ϕ being exponentially consistent with respect to n instead of n_+. Indeed, one can show that in such case Lemma 6 holds, whereas with the standard Kullback–Leibler variation one would only retain a lower bound in n_+; see Ghosal and van der Vaart (2017, Lemma 8.10).

Lemma 6 *Consider the general partially exchangeable model* (1) *with* Π *supported on dominated distributions. Then for any* $\epsilon, D > 0$ *and* n *sufficiently large, we have*

$$\int \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{p_i}{p_{0,i}}(X_{i,j})\, d\Pi(p_1, \ldots, p_m) \geq \Pi(\mathcal{V}_{0,\epsilon,n}) e^{-m(D+1)\epsilon^2 n}, \tag{10}$$

with $\left(\prod_{i=1}^{m} P_{0,i}^{\infty}\right)$*–probability at least* $1 - (mD^2\epsilon^2 n)^{-1}$.

Theorem 7 *Given a distance* d *that satisfies the basic testing assumption* (7)*, suppose that there exist* $\mathcal{P}_n \subset \mathcal{P}$ *and a constant* $C > 0$*, such that for* $\bar{\epsilon}_n \leq \epsilon_n$ *sequences of real numbers such that* $n\bar{\epsilon}_n^2 \to +\infty$*, the following hold for sufficiently large* n*:*

(7.a) $\Pi(\mathcal{V}_{0,\bar{\epsilon}_n,n}) \geq e^{-Cn\bar{\epsilon}_n^2}$;

(7.b) $\log(\mathscr{N}(\xi\epsilon_n, \mathcal{P}_n, d)) \leq n\epsilon_n^2$;

(7.c) $\Pi_i(\mathcal{P}_n^c) \leq e^{-(C+2m+1)n\bar{\epsilon}_n^2}$ *for* $i = 1, \ldots, m$.

Then ϵ_n *is a posterior rate of contraction at* $(p_{0,i})_{i=1}^m$ *with respect to* d_s*, for every* $s \geq 1$.

The proof of Theorem 7 is postponed to Section 7.3. Here we provide some intuition on the role of the conditions and their relation with the exchangeable case. The basic testing assumption (7) and conditions (7.b)–(7.c) are standard when building marginal frequentist tests $\phi_{n_i}^i : \mathbb{X}^{n_i} \to \{0, 1\}$ that separate the true distribution $P_{0,i}$ from the complement of any neighborhood with exponentially small errors with respect to the number of observations. In the statement of Theorem 7 we considered the same sieve $(\mathcal{P}_n)_{n\geq 1}$ for every marginal distribution Π_i for simplicity, since in most common frameworks, including the bHDP in (2), $\Pi_i = \Pi_{i'}$ for every i, i'. However, we point out that the result may be generalized to account for different subsets $\mathcal{P}_{n,i} \subset \mathcal{P}$, as showed in Section 7.3, which is of particular interest when the marginal exchangeable models with respect to Π_i require different sieves. Condition (7.a) on the reinforced Kullback–Leibler variation neighborhood is needed because one can not directly build a frequentist test $\phi : \mathbb{X}^{n_+} \to \{0, 1\}$ that separates the true distributions $(P_{0,i})_{i=1}^m$ with exponentially bounded small error probabilities with respect to the total number of observations n_+, unless $n_i \asymp n_{i'}$ for every i, i'.

5. Derivation of posterior contraction rates

Leveraging on Theorem 7, in this section we derive the contraction rates of Section 3.3 for multivariate bHDP Gaussian mixtures towards the true vector of densities $(p_{0,i})_{i=1}^m$. The conditions of Theorem 7 are divided into two separate blocks: (i) estimate of the prior mass of Kullback–Leibler variation neighborhood (4.a); (ii) high mass and low entropy sieves for the marginal exchangeable models (4.b; 4.c). Proposition 8 deals with (i), whereas Proposition 9 deals with (ii).

Proposition 8 *Let* $(\boldsymbol{F}, \boldsymbol{\Sigma}) \sim bHDP(\theta, \alpha^*, \theta^*, F^*) \times G^m$ *as in* (3) *and* $\mathcal{V}_{0,\epsilon,n}$ *be the Kullback–Leibler variation neighborhood defined in* (9)*. Then there exists* $C > 0$ *such that, provided that* $\bar{\epsilon}_n \to 0$*,* $\Pi(\mathcal{V}_{0,\bar{\epsilon}_n,n}) \geq e^{-Cn\bar{\epsilon}_n^2}$ *for sufficiently large* n*, where* $\bar{\epsilon}_n$ *is given by*

(i) $n^{-1/2}\log(n_\vee)^{(d+d/r_0+1)/2}$ *if* $p_{0,i}=p_{F_{0,i},\Sigma_{0,i}}$ *is supersmooth with* $\Sigma_{0,i}$ *in the support of* G *for each* $i=1,\dots,m$,

(ii) $(n_\vee/n)^{1/2}\, n_\vee^{-\beta/(2\beta+d^*)}\log(n_\vee)^{t_0}$, *if* $p_{0,i}$ *is* $\beta_{0,i}$*-smooth for each* $i=1,\dots,m$ *and* G *satisfies* (5), *where* $d^*=d\vee\kappa$ *and* $t_0=\beta(d^*/\tau+d^*+1+d^*/\beta)/(2\beta+d^*)$.

The next proposition provides a sieve that satisfies the desired conditions, in the same spirit of Shen et al. (2013). We point out that this holds for hierarchical models with conditionally Dirichlet marginals and mean measure $F^*(A)=\mathbb{E}(\tilde F(A))$ in general, regardless of the prior on $\tilde F$. First of all we define some relevant quantities:

$$\mathcal{F}_{N,a}=\Big\{\sum_{j=1}^{+\infty}\omega_j\delta_{z_j}:\ \sum_{j=N+1}^{+\infty}\omega_j<\epsilon_n^2, z_1,\dots,z_N\in[-a,a]^d\Big\};$$
$$\mathcal{S}_{\sigma,M}=\{\Sigma:\ \sigma_n^2\le \mathrm{eig}_1(\Sigma)\le \mathrm{eig}_d(\Sigma)<\sigma_n^2(1+\epsilon_n^2)^M\};$$
$$N_n=\frac{Kn\bar\epsilon_n^2}{\log(n\bar\epsilon_n^2)};\quad n\epsilon_n^2=KN_n\log n;\quad a_n^{a_1}=n\epsilon_n^2;\quad \sigma_n^{-2a_2}=n\epsilon_n^2;\quad M_n=n,$$

for $K>0$ and a_1 and a_2 defined in (5).

Proposition 9 *Let* $(\boldsymbol{F},\boldsymbol{\Sigma})\sim bHDP(\theta,\alpha^*,\theta^*)\times G^m$ *as in* (3) *such that conditions* (5) *hold. Define* $\mathcal{P}_n=\{p_{F_n,\Sigma_n}: F\in\mathcal{F}_{N_n,a_n},\Sigma\in\mathcal{S}_{\sigma_n,M_n}\}$ *and* $\bar\epsilon_n$ *as in (i) or (ii) of Proposition 8. If* $\bar\epsilon_n\to 0$, *then* $\log(\mathscr{N}(\epsilon_n,\mathcal{P}_n,d))\le n\epsilon_n^2$ *and for every* $C>0$ *there exists* $K>0$ *such that* $\Pi_i(\mathcal{P}_n^c)\le e^{-Cn\bar\epsilon_n^2}$ *for every* $i=1,\dots,m$.

Putting together Proposition 8 and Proposition 9, we obtain the following proof of Theorem 3 in the supersmooth case. Let $\bar\epsilon_n^2=n^{-1}\log(n_\vee)^{d+d/r_0+1}$ and $\epsilon_n^2=n^{-1}\log(n_\vee)^{d+d/r_0+2}$. In order for $\bar\epsilon_n,\epsilon_n\to 0$ as $n\to+\infty$, we ask that $n_\vee\lesssim e^{n^\delta}$ for some $0<\delta<(d+d/r_0+2)^{-1}$. Condition 1 on the reinforced Kullback–Leibler variation holds by Proposition 8 (i). Consider now $\mathcal{P}_n$ as in Proposition 9 and denote with $\tilde\epsilon_n$ the value of ϵ_n therein. For n large enough, $\tilde\epsilon_n\lesssim\epsilon_n$, so that (7.b) holds by Proposition 9. Finally, $\Pi_i(\mathcal{P}_n^c)\le e^{-Cn\bar\epsilon_n^2}$ for every $i=1,\dots,m$ and $C>0$. In particular, (7.c) holds as well. The proof of Theorem 4 in the ordinary smooth case follows similar arguments.

6. Future developments

In this paper we have laid the groundwork for the analysis of the frequentist properties of models involving dependent random probability measures, such as the popular HDP and the more general class of bHDP. In principle the same techniques can be used for the entire class of hierarchical PY mixture models, where both the child and the parent distribution are PYs instead than DPs. However, in order to treat this class we first need an exhaustive asymptotic theory for the exchangeable Pitman–Yor mixture model, which is currently missing in the multivariate scenario. The sieve proposed in Shen et al. (2013) for the multivariate Dirichlet process mixtures is inherently dependent on fast decreasing weights, leaving the Pitman–Yor case currently unresolved. Still, we may use Theorem 7 to derive contraction rates in this class of models by adding more restrictive assumptions, such as real-valued observations (Scricciolo, 2014) or multivariate distributions with compact support.

7. Proofs

7.1 Proof of Lemma 5

First we prove that ϕ is a test, i.e. a measurable function between 0 and 1. Sum and product of measurable functions are indeed measurable. For every $x_i \in \mathbb{X}_i$, $\phi^i(x_i) \in [0,1]$. We may assume that there exist independent events $\{A_i\}_{i=1}^m$ and a probability measure $\mathbb{P}$ such that $\mathbb{P}(A_i) = \phi^i(x_i)$ for $i = 1, \dots, m$. Then $\phi(x_1, \dots, x_m) = \mathbb{P}(\cup_{i=1}^m A_i)$, which is clearly between 0 and 1. To prove 1. we reason in a similar way. We observe that $P_{0,i}(\phi^i) \in [0,1]$ and consider independent events $\{A_i\}_{i=1}^m$ such that $\mathbb{P}(A_i) = P_{0,i}(\phi^i)$. Then $\mathbb{P}(\cup_{i=1}^m A_i) \leq \sum_{i=1}^m \mathbb{P}(A_i) \leq m \max(\mathbb{P}(A_1), \dots, \mathbb{P}(A_m)) < m \max(\alpha_1, \dots, \alpha_m)$. Finally, to prove 2. we consider independent events $\{A_i\}_{i=1}^m$ such that $\mathbb{P}(A_i) = P_i(\phi^i)$. Then $(\prod_{i=1}^m P_i)(1-\phi) = \mathbb{P}((\cup_{i=1}^m A_i)^c) = \mathbb{P}(\cap_{i=1}^m A_i^c)$. Since $(p_i)_{i=1}^m \in \mathcal{U}_0^c$, there exists $i' \in \{1, \dots, m\}$ such that $p_{i'} \in \mathcal{U}_{0,i'}$. Hence $\mathbb{P}(\cap_{i=1}^m A_i^c) \leq \mathbb{P}(A_{i'}^c) < \beta_{i'}$. We conclude by observing that $\beta_{i'} \leq \max(\beta_1, \dots, \beta_m)$.

7.2 Proof of Lemma 6

Define $d\Pi_{0,\epsilon}(p_1, \dots, p_m) \propto \mathbb{1}_{\mathcal{V}_{0,\epsilon,n}}(p_1, \dots, p_m)\, d\Pi(p_1, \dots, p_m)$ the restriction of Π to $\mathcal{V}_{0,\epsilon,n}$. The logarithm of the left hand side of (10) is bounded from below by

$$\log(\Pi(\mathcal{V}_{0,\epsilon,n})) + \log\left(\int \prod_{i=1}^m \prod_{j=1}^{n_i} \frac{p_i}{p_{0,i}}(X_{i,j})\, \Pi_{0,\epsilon}(p_1, \dots, p_m) \right).$$

Thus by Jensen's inequality the probability of the complement of the event in (10) is smaller or equal to the probability of

$$\int \log\left(\prod_{i=1}^m \prod_{j=1}^{n_i} \frac{p_i}{p_{0,i}}(X_{i,j}) \right) d\Pi_{0,\epsilon}(p_1, \dots, p_m) \leq -m(D+1)\epsilon^2 n.$$

Define $Z_i = \int \log(\prod_{j=1}^{n_i} p_i(X_{i,j}) p_{0,i}(X_{i,j})^{-1})\, d\Pi_{0,\epsilon}(p_1, \dots, p_m)$. Then the expected value $\mathbb{E}(Z_i) = -n_i \mathrm{KL}(p_i; p_{0,i}) > -\epsilon^2 n$ because of the definition of $\mathcal{V}_{0,\epsilon,n}$. Thus the probability of the complement of the event in (10) is smaller or equal to the one of

$$\left\{ \sum_{i=1}^m Z_i - \mathbb{E}\left(\sum_{i=1}^m Z_i \right) \leq -mD\epsilon^2 n \right\},$$

which, by the triangular inequality and Markov's inequality, admits $(mD\epsilon^2 n)^{-2} \sum_{i=1}^m \mathbb{E}(|Z_i - \mathbb{E}(Z_i)|^2)$ as an upper bound. The Marcinkiewicz–Zygmund inequality guarantees that $\mathbb{E}(|Z_i - \mathbb{E}(Z_i)|^2) \leq n_i V(p_{0,i}; p_i)$, which is smaller or equal to $\epsilon^2 n$ by definition of $\mathcal{V}_{0,\epsilon,n}$.

7.3 Proof of Theorem 7

We prove a more general version of Theorem 7, which accounts for potentially different sieves for each marginal distribution.

Theorem 10 *Given a distance d that satisfies the basic testing assumption* (7)*, suppose that there exist $\mathcal{P}_i = \mathcal{P}_{n,i} \subset \mathcal{P}$ and a constant $C > 0$, such that for $\bar{\epsilon}_n \leq \epsilon_n$ sequences of real numbers such that $n\bar{\epsilon}_n^2 \to +\infty$, the following hold for sufficiently large n:*

1. $\Pi(\mathcal{V}_{0,\bar{\epsilon}_n,n}) \geq e^{-Cn\bar{\epsilon}_n^2}$;
2. $\log(\mathscr{N}(\xi\epsilon_n, \mathcal{P}_i, d)) \leq n_i\epsilon_n^2$ *for* $i = 1, \ldots, m$;
3. $\Pi_i(\mathcal{P}_i^c) \leq e^{-(C+2m+1)n\bar{\epsilon}_n^2}$ *for* $i = 1, \ldots, m$.

Then ϵ_n *is a posterior rate of contraction at* $(p_{0,i})_{i=1}^m$ *with respect to* d_s, *for every* $s \geq 1$.

Proof Let $B_n = \{(p_i)_{i=1}^m \in \mathcal{P}^m : d_s((p_i)_{i=1}^m, (p_{0,i})_{i=1}^m) > M\epsilon_n\}$. Since L_1-convergence implies convergence in probability, we shall prove that $\mathbb{E}(\Pi(B_n|\boldsymbol{X})) \to 0$ as $n \to +\infty$. We observe that $B_n \subseteq \mathcal{U}_0^c$, where $\mathcal{U}_0 = \mathcal{U}_{0,1} \times \cdots \times \mathcal{U}_{0,m}$ and $\mathcal{U}_{0,i} = \{p : d(p, p_{0,i}) \leq m^{-1/s}M\epsilon_n\})$. Moreover, by applying Theorem D.5 in Ghosal and van der Vaart (2017) with $\epsilon = m^{-1/s}M\epsilon_n$ and $j = 1$, and by condition 2, the basic testing assumption (7) entails that there exist $K > 0$ and a test ϕ_{n_i} with error probabilities

$$P_{0,i}^{n_i}(\phi_{n_i}) \leq e^{\epsilon_n^2 n_i}\frac{e^{-Kn_i m^{-2/s}M^2\epsilon_n^2}}{1 - e^{-Kn_i m^{-2/s}M^2\epsilon_n^2}};$$
$$\sup_{p \in \mathcal{P}_i \cap \mathcal{U}_{0,i}^c} P^{n_i}(1 - \phi_{n_i}) \leq e^{-Kn_i m^{-2/s}M^2\epsilon_n^2}.$$

For $M^2 > K^{-1}m^{2/s}$, both error probabilities tend to zero. We observe that any constant larger than 1 would suffice. By Lemma 5 there exists a test ϕ that satisfies

$$\left(\prod_{i=1}^m P_{0,i}^{n_i}\right)(\phi) \leq me^{\epsilon_n^2 n}\frac{e^{-Knm^{-2/s}M^2\epsilon_n^2/2}}{1 - e^{-Knm^{-2/s}M^2\epsilon_n^2}};$$
$$\sup_{(p_i)_{i=1}^m \in (\mathcal{P}_1 \times \cdots \times \mathcal{P}_m) \cap \mathcal{U}_0^c}\left(\prod_{i=1}^m P_i^{n_i}\right)(1 - \phi) \leq e^{-Knm^{-2/s}M^2\epsilon_n^2}.$$

Let $A_n = \{\int \prod_{i=1}^m \prod_{j=1}^{n_i} p_i(X_{i,j})p_{0,i}(X_{i,j})^{-1}d\Pi(p_1, \ldots, p_m) \geq e^{(C+2m)\bar{\epsilon}_n^2 n}\}$. By Bayes' formula, the posterior probability of B_n is bounded above by

$$\phi + \mathbb{1}_{A_n^c} + e^{(C+2m)\bar{\epsilon}_n^2 n}\int_{B_n}\prod_{i=1}^m\prod_{j=1}^{n_i}\frac{p_i(X_{i,j})}{p_{0,i}(X_{i,j})}d\Pi(p_1, \ldots, p_m)(1 - \phi).$$

The expected value of the first term goes to zero by the previous argument. The one of the second term goes to zero by Lemma 6 with $D = 1$ and condition 1. By leveraging $B_n \subset \mathcal{U}_0^c$, the expected value of the third term is bounded above by

$$e^{(C+2m)\bar{\epsilon}_n^2 n}\big(\Pi((\mathcal{P}_1 \times \cdots \times \mathcal{P}_m)^c) + e^{-m^{-2/s}KM^2\epsilon_n^2 n}\big).$$

Since $\Pi((\mathcal{P}_1 \times \cdots \times \mathcal{P}_m)^c) \leq \sum_{i=1}^m \Pi_i(\mathcal{P}_i^c)$, $e^{(C+2m)\bar{\epsilon}_n^2 n}\Pi((\mathcal{P}_1 \times \cdots \times \mathcal{P}_m)^c) \leq me^{-n\bar{\epsilon}_n^2}$ by condition 3. We conclude by taking M large enough such that $m^{-2/s}KM^2 > C + 2m$. ■

7.4 Proof of Proposition 8

We first recall a known property of the Dirichlet distribution (Lemma 11) and prove an upper bound for the mixed moments of the Pitman–Yor process (Lemma 12). Lemma 11 can be easily deduced from the proof of Lemma 6.1 in Ghosal et al. (2000). It relies on the density of the Dirichlet distribution being bounded from below whenever the parameters are smaller than one. Let us consider a random vector $(X_1, \ldots, X_{k-1})$ taking values in the simplex $\mathbb{S}_{k-1}$ and, for any $(u_1, \ldots, u_{k-1}) \in \mathbb{S}^{k-1}$, we set $u_k = 1 - \sum_{i=1}^{k-1} u_i$. Then for any $\epsilon \leq k^{-1}$, one has

$$\mathbb{P}\bigg(\sum_{i=1}^{k} |X_i - u_i| \leq 2\epsilon, \min_{1\leq i\leq k} X_i > \frac{\epsilon^2}{2} \bigg) \geq \mathbb{P}\big(\max_{1\leq i\leq k-1} |X_i - u_i| \leq \epsilon^2 \big),$$

where, without loss of generality, we further assumed $u_k = \max_{i=1,\ldots,k} u_i$. This can be used to prove the following result. Let $\mathbf{1}_k$ denote the k–dimensional vector of all ones.

Lemma 11 *Let* $(Y_{1,1}, \ldots, Y_{1,k}, \ldots, Y_{N,1}, \ldots, Y_{N,k}) \sim \mathrm{Dir}(\gamma_1 \mathbf{1}_k, \ldots, \gamma_N \mathbf{1}_k)$ *with* $\gamma_i \leq 1$. *Then for every* $\epsilon \leq (kN)^{-1}$ *and* $\{u_{i,j}\}$ *s.t.* $\sum_{i=1}^{N} \sum_{j=1}^{k} u_{i,j} = 1$,

$$\mathbb{P}\bigg(\sum_{i=1}^{N} \sum_{j=1}^{k} |Y_{i,j} - u_{i,j}| \leq 2\epsilon, \min_{i,j} Y_{i,j} > \frac{\epsilon^2}{2} \bigg) \geq \Gamma\bigg(k \sum_{i=1}^{N} \gamma_i \bigg) \epsilon^{2(kN-1)} \prod_{i=1}^{N} \gamma_i^k,$$

where Γ *indicates the gamma function.*

Lemma 12 yields an upper bound on the mixed moments of the Pitman–Yor process. The proof relies on the relationship between the Pitman–Yor process and the stable completely random measure, together with some convenient tools for evaluating the mixed moments of normalized random measures with independent increments, as first developed in James et al. (2006). For this reason in model (2) we focused on $\theta^* > 0$.

Lemma 12 *Let* $\tilde{P} \sim \mathrm{PY}(\alpha, \theta, \mathrm{F})$ *with* $\alpha, \theta > 0$ *and* $F \in \mathcal{P}$, *where* $\mathbb{X}$ *is a Polish space. Then for any* $A_1, \ldots, A_k$ *pairwise disjoint Borel sets on* $\mathbb{X}$, *and* $n_1, \ldots, n_k \in \mathbb{N}$,

$$\mathbb{E}(\tilde{P}(A_1)^{n_1} \cdot \cdots \cdot \tilde{P}(A_k)^{n_k}) \geq \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta/\alpha+1)} \frac{\Gamma(k+\theta/\alpha)}{\Gamma(n_+ + \theta)} \prod_{i=1}^{k} \alpha F(A_i),$$

where $n_+ = n_1 + \cdots + n_k$ *and* Γ *is the gamma function.*

Proof Let $\mathbb{P}_{\alpha,F}$ be the law of an α–stable completely random measure with base measure F on the space of boundedly finite measures $\mathcal{M}_{\mathbb{X}}$ on $\mathbb{X}$. We let $\mathbb{P}_{\alpha,\theta,F}$ be another probability measure on $\mathcal{M}_{\mathbb{X}}$ that is absolutely continuous with respect to $\mathbb{P}_{\alpha,F}$ and with Radon–Nikodym derivative

$$\frac{d\mathbb{P}_{\alpha,\theta,F}}{d\mathbb{P}_{\alpha,F}}(\mu) = \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)} \mu^{-\theta}(\mathbb{X}).$$

As shown in Pitman and Yor (1997), the Pitman–Yor process $\mathrm{PY}(\alpha, \theta, F)$ may be obtained by normalizing a random measure $\tilde{\mu} \sim \mathbb{P}_{\alpha,\theta,F}$, i.e.

$$\tilde{P} = \frac{\tilde{\mu}(\cdot)}{\tilde{\mu}(\mathbb{X})} \sim \mathrm{PY}(\alpha, \theta, F).$$

This relationship between the Pitman–Yor process and the stable completely random measure may be conveniently used to derive the mixed moments, as shown in Canale et al. (2017). In particular, $\mathbb{E}(\tilde{P}(A_1)^{n_1} \cdot \cdots \cdot \tilde{P}(A_k)^{n_k})$ is equal to

$$\frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)} \frac{1}{\Gamma(n_+ + \theta)} \int_0^{+\infty} u^{n_+ + \theta - 1} e^{-u^\alpha} \prod_{i=1}^{k} \sum_{\ell=1}^{n_i} F(A_i)^\ell \xi_{n_i,\ell}(u)\, du, \tag{11}$$

where $\xi_{n,\ell}$ is defined as

$$\xi_{n,\ell}(u) = \frac{\alpha^\ell}{u^{n-\ell\alpha}\ell!} \sum_{\boldsymbol{q}} \binom{n}{q_1 \cdots q_\ell} \prod_{r=1}^{\ell} (1-\sigma)_{q_r - 1},$$

where $(\cdot)_q$ indicates the Pochammer function and the sum is over all vectors $\boldsymbol{q} = (q_1, \ldots, q_\ell)$ of positive integers such that $q_1 + \cdots + q_\ell = n$. We observe that $\sum_{\ell=1}^{n_i} F(A_i)^\ell \xi_{n_i,\ell}(u) \geq F(A_i)\xi_{n_i,1}(u) = F(A_i)\alpha u^{\alpha - n_i}$. Thus a lower bound to (11) is given by

$$\frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)} \frac{1}{\Gamma(n+\theta)} \left(\prod_{i=1}^{k} \alpha F(A_i) \right) \int_0^{+\infty} u^{\theta + k\alpha - 1} e^{-u^\alpha}\, du.$$

We conclude by observing that the integral in the last expression is equal to $\alpha^{-1}\Gamma(k+\theta/\alpha)$. ∎

We proceed now to the proof of assertion (i) of Proposition 8. Following Proposition 9.14 of Ghosal and van der Vaart (2017) for sufficiently small $\epsilon > 0$ and each $i = 1, \ldots, m$ there exist $N_i \lesssim \log_-(\epsilon)^{d + d/r_{0,i}}$, disjoint subsets $\{U_{i,j}\}_{j=1}^{N_i} \subseteq \mathbb{R}^d$ of diameter of the order ϵ^2 and weights $\{\omega_{i,j}\}_{j=1}^{N_i} \in \mathbb{S}_{N_i - 1}$ such that,

$$\{(F, \Sigma) : d_H(p_{0,i}, p_{F,\Sigma}) \lesssim \epsilon\} \supset B_{i,\epsilon} := \left\{ (F, \Sigma) : \sum\nolimits_{j=1}^{N_i} |F(U_{i,j}) - \omega_{i,j}| \leq \epsilon^2, \|\Sigma - \Sigma_{0,i}\| \leq \epsilon \right\}$$

Note that $N_i \lesssim \log_-(\epsilon)^{d + d/r_0}$ for $r_0 = \min(r_{0,1}, \ldots, r_{0,m})$. Moreover, without loss of generality, we can force the sets U_{ij} pertaining to different groups to be either disjoint or equal. This is needed later so to leverage on the aggregation properties of the Dirichlet process in the application of Lemma 11 and Lemma 12. With a slight abuse of notation, let $B_\epsilon = \bigcap_{i=1}^m B_{i,\epsilon}$ be the subset of $\left((F_i)_{i=1}^m, (\Sigma_i)_{i=1}^m\right)$ such that each (F_i, Σ_i) belong to $B_{i,\epsilon}$ defined in the display above. Conditionally on $\tilde{F}$, $F_1, \cdots, F_m$ are independent. Thus,

$$\mathbb{P}(B_\epsilon) = \mathbb{E}\left(\prod_{i=1}^{m} \mathbb{P}\left(\sum_{j=1}^{N_i} |F_i(U_{i,j}) - \omega_{i,j}| \leq \epsilon^2 \Big| \tilde{F} \right) \right) \prod_{i=1}^{m} \mathbb{P}(\|\Sigma_i - \Sigma_{0,i}\| \leq \epsilon).$$

Since G has continuous and positive density on its support and $\Sigma_{0,i}$ belongs to the support of G, $\mathbb{P}(\|\Sigma_i - \Sigma_{0,i}\| \leq \epsilon) \gtrsim \epsilon^q$, where q depends on the dimension of the support of G. Next, let $U_{0,i} = \mathbb{R}^d \setminus (\cup_{j=1}^{N_i} U_{i,j})$, so that $F_i(U_{0,i}), \ldots, F_i(U_{N_i,i})|\tilde{F} \sim \mathrm{Dir}(\theta\tilde{F}(U_{0,i}), \ldots, \theta\tilde{F}(U_{N_i,i}))$. Let $\underline{\theta} := \lceil \theta \rceil$ and let $\eta = \theta\underline{\theta}^{-1} \leq 1$. The aggregation properties of the Dirichlet distribution guarantee that, conditionally on $\tilde{F}$, $F_i(U_{i,j}) = \sum_{h=1}^{\underline{\theta}} Y_{i,j,h}$, where

$$((Y_{i,1,h})_{h=1}^{\underline{\theta}}, \ldots, (Y_{i,N_i,h})_{h=1}^{\underline{\theta}}) \sim \mathrm{Dir}(\eta\tilde{F}(U_{i,1})\mathbf{1}_{\underline{\theta}}, \ldots, \eta\tilde{F}(U_{i,N_i})\mathbf{1}_{\underline{\theta}}),$$

with $\mathbf{1}_{\underline{\theta}}$ denoting the $\underline{\theta}$–dimensional vector of all ones. Define $\omega_{i,0} = 0$. Then $\sum_{j=1}^{N_i} |F_i(U_{i,j}) - \omega_{i,j}| \leq \sum_{j=0}^{N_i} \sum_{h=1}^{\underline{\theta}} |Y_{i,j,h} - \omega_{i,j}\underline{\theta}^{-1}|$. Lemma 11 thus guarantees that

$$\Pi\Big(\sum_{j=1}^{N_i} |F_i(U_{i,j}) - \omega_{i,j}| \leq \epsilon^2 \Big| \tilde{F}\Big) \geq \Gamma(\theta)(\epsilon/\sqrt{2})^{2(\underline{\theta}(N_i+1)-1)}\eta^{N_i+1} \prod_{j=0}^{N_i} \tilde{F}(U_{i,j})^{\underline{\theta}}.$$

We observe that $\prod_{i=1}^{m} \Gamma(\theta)(\epsilon/\sqrt{2})^{2(\underline{\theta}(N_i+1)-1)}\eta^{N_i+1} \gtrsim e^{-c_1 \log_-(\epsilon)^{d+d/r_0+1}}$ for some $c_1 > 0$. In order to conclude that

$$\mathbb{P}(B_\epsilon) \gtrsim e^{-c \log_-(\epsilon)^{d+d/r_0+1}} \epsilon^{mq} \tag{12}$$

for some constant $c > 0$, we show next that

$$\mathbb{E}(\prod_{i=1}^{m} \prod_{j=0}^{N_i} \tilde{F}(U_{i,j})^{\underline{\theta}}) \gtrsim e^{c_2 \log_-(\epsilon)^{d+d/r_0+1}}$$

for some $c_2 > 0$. We indicate by $\{U_h : h = 1, \ldots, N\}$ the set of pairwise disjoint neighborhoods and by $k_h = |\{i : U_h = U_{i,j} \text{ for some } j\}|$ the number of groups containing a copy of U_h, so that $k_1 + \cdots + k_h = N_1 + \cdots + N_m =: N_+$. Define $U_0 = \mathbb{R}^d \setminus (\cup_{i=1}^{m} \cup_{j=1}^{N_i} U_{i,j})$, so that $(U_h)_{h=0}^{N}$ forms a partition of $\mathbb{R}^d$ and set $k_0 = m$. Since $U_0 \subseteq U_{0,i}$ for $i = 1, \ldots, m$,

$$\mathbb{E}\Big(\prod_{i=1}^{m} \prod_{j=0}^{N_i} \tilde{F}(U_{i,j})^{\underline{\theta}}\Big) \geq \mathbb{E}\Big(\prod_{h=0}^{N} \tilde{F}(U_h)^{k_h\underline{\theta}}\Big).$$

In order to compute the expected value on the right, we distinguish two ranges for the parameters. We first deal with the case $\alpha^* = 0$, so that $(\tilde{F}(U_h))_{h=1}^{N}$ has Dirichlet distribution on $\mathbb{R}^d$. Thus, by known properties of the Dirichlet distribution,

$$\mathbb{E}\Big(\prod_{h=0}^{N} \tilde{F}(U_h)^{k_h\underline{\theta}}\Big) = \frac{\prod_{h=0}^{N}(\theta^* F^*(U_h))^{[k_h\underline{\theta}]}}{(\theta^*)^{[(m+N_+)\underline{\theta}]}}, \tag{13}$$

where $\beta^{[n]} = \Gamma(\beta+n)/\Gamma(\beta)$ is the ascending factorial for $\beta > 0$ and $n \in \mathbb{N}$. Let $N_0 := \underline{\theta}(m + N_+)$ and note that $N_0 \lesssim \log_-(\epsilon)^{d+d/r_0}$. Since F^* is continuous and positive, $\theta^* F^*(U_i) \gtrsim \epsilon^2$ for $i = 1, \ldots, N$. Moreover, since $b^k \leq b^{[k]} \leq (b + k - 1)^k$, for ϵ sufficiently small the right side of (13) is greater than or equal to a constant multiplied by

$$\left(\frac{\epsilon^2}{\theta^* + N_0 - 1}\right)^{N_0} \geq \epsilon^{3N_0} \geq e^{-3 \log_-(\epsilon)^{d+d/r_0+1}}.$$

When $\alpha^* > 0$, the expression of the mixed moments is available thanks to the relationship between the Pitman–Yor process and the stable completely random measure. By Lemma 12,

$$\mathbb{E}\left(\tilde{F}(U_0)^{m\underline{\theta}} \prod_{h=1}^{N} \tilde{F}(U_h)^{k_h\underline{\theta}}\right) \gtrsim \frac{\Gamma(N+\theta^*/\alpha^*)}{\Gamma(N_0+\theta^*)} \prod_{h=0}^{N} \alpha^* F^*(U_h).$$

As $\epsilon \to 0$, $N \asymp N_0$. Thus, since $\alpha^* F^*(U_i) \gtrsim \epsilon^2$ for $i = 1, \ldots, N$, an upper bound for the right side of the previous expression is given by $\epsilon^{2N} = e^{-2\log_-(\epsilon)^{d+d/r_0+1}}$.

We proceed next to use the prior mass bound (12) for the reinforced Kullback–Leibler variation neighborhood $\mathcal{V}_{0,\epsilon,n}$ in (9). First note that $\{d_H(p_{0,i}, p_{F_i,\Sigma_i}) \lesssim \epsilon \text{ for } i = 1, \ldots, m\} \supseteq B_\epsilon$. Moreover, reasoning as in Proposition 9.14 in Ghosal and van der Vaart (2017), for each $i = 1, \ldots, m$, $\mathrm{KL}(p_{0,i}; p_{F_i,\Sigma_i})$ and $\mathrm{V}(p_{0,i}; p_{F_i,\Sigma_i})$ are bounded above by a multiple of $d_H^2(p_{0,i}, p_{F,\Sigma})(\log_-(d_H(p_{0,i}, p_{F,\Sigma}))^2$. Hence

$$\mathcal{V}_{0,\epsilon,n} \supseteq \left\{d_H(p_{0,i}, p_{F_i,\Sigma_i}) \log_-(d_H(p_{0,i}, p_{F_i,\Sigma_i}) \lesssim \sqrt{n/n_\vee}\,\epsilon \text{ for } i = 1, \ldots m\right\},$$

where we used $n_i \le n_\vee$ for each i. Consider now a sequence $\bar{\epsilon}_n \to 0$, to be determined later, that depends on n and $n_\vee$ such that $\sqrt{n/n_\vee}\,\bar{\epsilon}_n \to 0$. The function $f(x) = x\log_-(x)$ is strictly monotonic near zero with inverse f^{-1} satisfying $f^{-1}(x) \asymp x/\log_-(x)$ as $x \downarrow 0$. In fact

$$f\big(x/\log_-(x)\big) = \frac{x}{\log_-(x)}(\log_-(x) + \log(\log_-(x))) = x + o(x).$$

It follows that a lower bound on the prior mass of $\mathcal{V}_{0,\bar{\epsilon}_n,n}$ is obtained from the lower bound $e^{-c\log_-(\epsilon)^{d+d/r_0+1}}\epsilon^{mq}$ in (12) on the prior mass of B_ϵ upon replacing ϵ in the definition of B_ϵ with $\sqrt{n/n_\vee}\,\bar{\epsilon}_n/\log_-(\sqrt{n/n_\vee}\,\bar{\epsilon}_n)$. Since $\sqrt{n/n_\vee}\,\bar{\epsilon}_n \to 0$, the lower bound simplifies to $\mathbb{P}(\mathcal{V}_{0,\bar{\epsilon}_n,n}) \gtrsim e^{-C\log_-(\sqrt{n/n_\vee}\,\bar{\epsilon}_n)^{d+d/r_0+1}}$. Hence assertion (i) boils down to find $\bar{\epsilon}_n$ such that $\sqrt{n/n_\vee}\,\bar{\epsilon}_n \to 0$ and

$$e^{-c\log_-(\sqrt{n/n_\vee}\,\bar{\epsilon}_n)^{d+d/r_0+1}} = e^{-Cn\bar{\epsilon}_n^2},$$

for some $C > 0$. It is easy to show that $\bar{\epsilon}_n = n^{-1/2}\log(n_\vee)^{(d+d/r_0+1)/2}$ satisfies these requirements whenever $n_\vee \lesssim e^{n^\delta}$ for n large enough and $\delta < (d + d/r_0 + 2)^{-1}$.

As for assertion (ii), following Proposition 9.14 of Ghosal and van der Vaart (2017) for sufficiently small $\sigma, \epsilon > 0$ and each $i = 1, \ldots, m$, there exist $N_i \lesssim \log_-(\sigma)^{d/\tau}\sigma^{-d}(\log_-\epsilon)^d$, disjoint subsets $\{U_{i,j}\}_{j=1}^{N_i} \subseteq \mathbb{R}^d$ of diameter of the order $\sigma\epsilon^2$ and weights $\{\omega_{i,j}\}_{j=1}^{N_i} \in \mathbb{S}_{N_i-1}$ such that,

$$\begin{aligned}&\{(F,\Sigma) :\, d_H(p_{0,i}, p_{F,\Sigma}) \lesssim \sigma^{\beta_i} + \epsilon\}\\ &\supset B_{i,\epsilon,\sigma} := \Big\{(F,\Sigma) :\, \textstyle\sum_{j=1}^{N_i} |F(U_{i,j}) - \omega_{i,j}| \le \epsilon^2, \min_{1\le j\le N_i} F(U_{i,j}) \ge \epsilon^4, 1 \le \sigma^2 \mathrm{eig}(\Sigma^{-1}) \le 1 + \sigma^{\beta_i}\Big\}.\end{aligned}$$

Similarly to before, we can force the sets U_{ij} pertaining to different groups to be either disjoint or equal and use Lemma 11 and Lemma 12 to establish, under prior condition (5), that

$$\begin{aligned}\mathbb{P}(B_{\epsilon,\sigma}) &\gtrsim e^{-c_1\log_-(\sigma)^{d/\tau}\sigma^{-d}(\log_-\epsilon)^{d+1}} \prod_{i=1}^{m} \sigma^{-2a_4}\sigma^{2\beta_i a_5} e^{-C_3\sigma^{-\kappa}}\\ &\gtrsim e^{-c_1\log_-(\sigma)^{d/\tau}\sigma^{-d}(\log_-\epsilon)^{d+1} - c_2\sigma^{-\kappa}},\end{aligned} \tag{14}$$

for some $c_1, c_2 > 0$, where $B_{\epsilon,\sigma} = \bigcap_{i=1}^m B_{i,\epsilon,\sigma}$. We omit details. Following arguments used in the proof of Proposition 9.14 of Ghosal and van der Vaart (2017), we can establish that for $(F, \Sigma) \in B_{i,\epsilon,\sigma}$,

$$\mathrm{KL}(p_{0,i}; p_{F_i,\Sigma_i}), \mathrm{V}(p_{0,i}; p_{F_i,\Sigma_i}) \leq c\, d_H^2(p_{0,i}; p_{F_i,\Sigma_i})(\log_-(\epsilon^4/\sigma^d))^2 + o(\sigma^{2\beta_i})$$

for some $c > 0$, provided that ϵ^4/σ^d is sufficiently small. Hence the prior mass of the neighborhood $\mathcal{V}_{0,\bar\epsilon_n,n}$ is bounded below by the prior mass of the set $B_{\epsilon,\sigma}$ if ϵ and σ in the definition of $B_{\epsilon,\sigma}$ are chosen so that

$$(\sigma^\beta + \epsilon)^2(\log_-(\epsilon^4/\sigma^d))^2 \lesssim \frac{n}{n_\vee}\bar\epsilon_n^2, \quad \epsilon^4/\sigma^d = O(1), \tag{15}$$

where we used $\beta = \min(\beta_1, \ldots, \beta_m)$. As before $\bar\epsilon_n$ is a sequence to be determined later that depends on n and $n_\vee$ such that $\sqrt{n/n_\vee}\,\bar\epsilon_n \to 0$ as $n \to \infty$. The prior mass of $\mathcal{V}_{0,\bar\epsilon_n,n}$ is bounded below by $e^{-Cn\bar\epsilon_n^2}$ for some $C > 0$ if

$$\log_-(\sigma)^{d/\tau}\sigma^{-d}(\log_-\epsilon)^{d+1} + \sigma^\kappa \leq n\bar\epsilon_n^2. \tag{16}$$

Take $\epsilon^4 \asymp \sigma^d \wedge \sigma^{2\beta}$ so that (15) reduces to $\sigma^{2\beta}(\log_-\sigma)^2 \lesssim \frac{n}{n_\vee}\bar\epsilon_n^2$, and (16) reduces to

$$\log_-(\sigma)^{d/\tau+d+1}\sigma^{-d} + \sigma^\kappa \lesssim n\bar\epsilon_n^2. \tag{17}$$

Inequality (15) is satisfied by taking $\sigma = \sigma_n$ depending on n and $n_\vee$ such that

$$\sigma_n^\beta = \sqrt{\frac{n}{n_\vee}}\bar\epsilon_n \Big/ \log_-\left(\sqrt{\frac{n}{n_\vee}}\bar\epsilon_n\right).$$

Before substituting σ with σ_n in (17), let restrict the search for $\bar\epsilon_n$ to

$$\bar\epsilon_n = \sqrt{\frac{n_\vee}{n}}n_\vee^{-\gamma}(\log n_\vee)^{t_0},$$

by analogy to the supersmooth case. Here $\gamma \in (0, 1/2)$ and $t_0 > 0$ are constants to be determined later. Note that $\log_-(\sqrt{n/n_\vee}\,\bar\epsilon_n) \asymp \log n_\vee$, while the right hand side of (17) becomes

$$n\bar\epsilon_n^2 = n_\vee^{1-2\gamma}(\log n_\vee)^{2t_0}.$$

Let $d \geq \kappa$. Substituting $\sigma_n^\beta = \sqrt{n/n_\vee}\bar\epsilon_n/\log n_\vee = n_\vee^{-\gamma}(\log n_\vee)^{t_0-1}$ into the left hand side of (17) we get that, up to a constant, the leading term is

$$(\log n_\vee)^{d/\tau+d+1}n_\vee^{\gamma\frac{d}{\beta}}(\log n_\vee)^{(1-t_0)\frac{d}{\beta}}.$$

Equating the last two displays we get

$$\begin{cases} 1 - 2\gamma = \gamma\dfrac{d}{\beta} & \Longrightarrow \gamma = \dfrac{\beta}{2\beta + d} \\ 2t_0 = \dfrac{d}{\tau} + d + 1 + (1 - t_0)\dfrac{d}{\beta} & \Longrightarrow t_0 = \dfrac{\beta}{2\beta + d}\Big(\dfrac{d}{\tau} + d + 1 + \dfrac{d}{\beta}\Big), \end{cases}$$

as desired. The case $d < \kappa$ can be treated similarly. It is easy to show that $\bar\epsilon_n \to 0$ whenever $n_\vee \lesssim n^{1+a}$ for some $0 < a < 2\beta/d$. The case $d < \kappa$ can be treated similarly. The proof is then complete.

7.5 Proof of Proposition 9

In order to prove that $\log \mathscr{N}(\epsilon_n, \mathcal{P}_n, d_H) \leq n\epsilon_n^2$, we show that there exist constants C_1, C_2 not depending on n such that $\log \mathscr{N}(C_1\epsilon_n, \mathcal{P}_n, d_H) \leq C_2 n\epsilon_n^2$. The constants C_1, C_2 can be included in the rate by defining ϵ_n as $\epsilon_n \max(\sqrt{C_1}, C_2)$. By Lemma 9.15 in Ghosal and van der Vaart (2017), there exists a large constant A such that $\log \mathscr{N}(A\epsilon_n, \mathcal{P}_n, d_H)$ is less than or equal to a constant multiplied by

$$N_n \log\left(\frac{5}{\epsilon_n^2}\right) + dN_n \log\left(\frac{3a_n}{\sigma_n \epsilon_n^2}\right) + d^2 \log\left(\frac{5}{\epsilon_n^2}\right) + M_n d^2 \log(1+\epsilon_n^2) + d\log(M_n).$$

Here $\epsilon_n = K^2\bar{\epsilon}_n^2 \log n / \log(n\bar{\epsilon}_n^2)$ for $\bar{\epsilon}_n$ as in (i) or (ii) of Proposition 8, so that $\epsilon_n \asymp \bar{\epsilon}_n \log n$ in the supersmooth case and $\epsilon_n \geq \sqrt{K}\bar{\epsilon}_n$ in the ordinary smooth case. We show next that all summands are bounded from above by $n\epsilon_n^2$ up to a constant. First of all we observe that $n\epsilon_n^2 > \log n_\vee > \log n$ for sufficiently large n. Thus the last term is less than or equal to $dn\epsilon_n^2$. The fourth one is easily bounded by $d^2 n\epsilon_n^2$. Moreover, $\epsilon_n^{-2} < n$ implies $\log(\epsilon_n^{-2}) < \log n$ for n large enough. Thus, the third term is bounded by $d^2\epsilon_n^2 n$. As for the first term, we observe that $N_n/n\epsilon_n^2 = (K\log n)^{-1}$. Since $\log(\epsilon_n^{-2}) < \log n$, we have that $N_n \log(\epsilon_n^{-2})/n\epsilon_n^2 = O(1)$ for large n. As for the second term, it remains to show that $N_n \log(a_n/\sigma_n)/n\epsilon_n^2 = O(1)$. This follows from $a_n/\sigma_n = (n\epsilon_n^2)^{\frac{1}{a_1}+\frac{1}{2a_2}} \leq n^{\frac{1}{a_1}+\frac{1}{2a_2}}$.

We now prove that for every $C > 0$ there exists $K > 0$ such that $\Pi_i(\mathcal{P}_n^c) \geq e^{-Cn\bar{\epsilon}_n^2}$. We observe that $\Pi_i(\mathcal{P}_n^c) \leq \Pi(F_i \in \mathcal{F}_{N_n,a_n}^c) + \Pi(\Sigma_i \in \mathcal{S}_{\sigma_n,M_n}^c)$ and $\Pi(F_i \in \mathcal{F}_{N_n,a_n}^c) = \mathbb{E}(\Pi(F_i \in \mathcal{F}_{N_n,a_n}^c|\tilde{F}))$. Since $F_i|\tilde{F}$ is distributed as a Dirichlet process, by Proposition 2 in Shen et al. (2013), this is bounded from above by

$$\mathbb{E}\left(\left(\frac{2e\theta \log_- \epsilon_n}{N_n}\right)^{N_n} + N_n(1-\tilde{F}([-a_n,a_n]^d)\right),$$

which is equal to $(2e\theta \log_-(\epsilon_n)N_n^{-1})^{N_n} + N_n(1-F^*([-a_n,a_n]^d)$. On the other hand, $G(\mathcal{S}_{\sigma_n,M_n}^c) \leq G(\text{eig}_1 \geq \sigma_n^2(1+\epsilon_n^2)) + G(\text{eig}_d \leq \sigma_n^2)$. Putting these together, $\Pi_i(\mathcal{P}_n^c)$ is bounded from above by

$$\left(\frac{2e\theta \log_- \epsilon_n}{N_n}\right)^{N_n} + N_n e^{-C_1 a_n^{a_1}} + b_2 e^{-C_2/\sigma_n^{2a_2}} + b_3 \sigma_n^{-2a_3}(1+\epsilon_n^2)^{-a_3 M_n}.$$

The second and third summand are easily bounded by $e^{-C'n\epsilon_n^2} \leq e^{-KC'n\bar{\epsilon}_n^2}$, for some constant C'. In the last summand $(1+\epsilon_n^2)^{-a_3 M_n} \leq e^{-a_3 n\epsilon_n^2/2}$ by using $1+x \leq e^x$. As for the first summand, we first observe that $\log_- \epsilon_n/N_n = \log_- \epsilon_n \log(n\bar{\epsilon}_n^2)/n\bar{\epsilon}_n^2$. In the supersmooth case $\log_- \epsilon_n \log(n\bar{\epsilon}_n^2) \leq (\log n)^2 = (n\bar{\epsilon}_n^2)^\delta$ for some $0 < \delta < 1$, so $\log_- \epsilon_n/N_n \leq (n\bar{\epsilon}_n^2)^{-(1-\delta)}$. Thus for n sufficiently large,

$$(\log_- \epsilon_n/N_n)^{N_n} \leq e^{-Kn\bar{\epsilon}_n^2 \frac{(1-\delta)\log(n\bar{\epsilon}_n^2)}{\log(n\bar{\epsilon}_n^2)}} = e^{-K(1-\delta)n\bar{\epsilon}_n^2}.$$

In the ordinary smooth case, we use $\log_- \epsilon_n/N_n = K\log_- \epsilon_n \log n/n\epsilon_n^2$. The latter is bounded above by $K(\log n)^2/n\epsilon_n^2$ which in turn is bounded above by $n_\vee^{-\delta}$ for some constant $\delta > 0$. Hence, for some $\delta_1 > 0$,

$$(\log_- \epsilon_n/N_n)^{N_n} \leq e^{-N_n\delta \log n_\vee} \leq e^{-N_n\delta \log n} \leq e^{-K\delta_1 n\bar{\epsilon}_n^2}.$$

In both cases, by taking K large enough, we thus derive the desired upper bound.

Acknowledgments

The authors are grateful to the Editor and the anonymous Referees for insightful comments and remarks, which led to a substantial improvement of the manuscript.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

F. Camerlenghi, A. Lijoi, and I. Prünster. Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis*, 156:18–28, 2017.

F. Camerlenghi, A. Lijoi, P. Orbanz, and I. Prünster. Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92, 2019.

A. Canale and P. De Blasi. Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404, 2017.

A. Canale, A. Lijoi, B. Nipoti, and I. Prünster. On the Pitman-Yor process with spike and slab base measure. *Biometrika*, 104(3):681–697, 2017.

P. J. Cowans. Information retrieval using hierarchical Dirichlet processes. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 564–565, 2004.

L. T. Elliott, M. De Iorio, S. Favaro, K. Adhikari, and Y. W. Teh. Modeling population structure under hierarchical Dirichlet processes. *Bayesian Anal.*, 14(2):313–339, 2019.

J. Fan. On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems. *The Annals of Statistics*, 19(3):1257 – 1272, 1991.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

N. J. Foti and S. A. Williamson. A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):359–371, 2013.

S. Ghosal and A. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263, 2001.

S. Ghosal and A. van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017. ISBN 9780521878265.

S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27(1):143–158, 1999.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

T. Griffiths, K. Canini, A. Sanborn, and D. Navarro. Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

T. S. Haines and T. Xiang. Delta-dual hierarchical Dirichlet processes: A pragmatic abnormal behaviour detector. In *2011 International Conference on Computer Vision*, pages 2198–2205, 2011.

N. L. Hjort, C. C. Holmes, P. Müller, and S. G. Walker. *Bayesian nonparametrics*. Cambridge University Press, 2010.

L. F. James, A. Lijoi, and I. Prünster. Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, 33(1):105–120, 2006.

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer, 1986.

A. Y. Lo. On a class of bayesian nonparametric estimates: I. density estimates. *Ann. Statist.*, 12(1):351–357, 1984.

T. Nakamura, T. Nagai, and N. Iwahashi. Multimodal categorization by hierarchical Dirichlet process. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1520–1525, 2011.

X. Nguyen. Borrowing strengh in hierarchical bayes: Posterior concentration of the dirichlet base measure. *Bernoulli*, 22(3):1535–1571, 2016.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25(2):855–900, 1997.

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

L. Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

C. Scricciolo. Adaptive Bayesian density estimation in l^p -metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal.*, 9(2):475–520, 2014.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

W. Shen, S. T. Tokdar, and S. Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77: 291–330, 2008.

T. Taniguchi, R. Yoshino, and T. Takano. Multimodal hierarchical Dirichlet process-based active perception by a robot. *Frontiers in Neurorobotics*, 12:22, 2018.

Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*, pages 158–207. Cambridge University Press, 2010.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

S. T. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, 68:90–110, 2006.

Y. Wei and X. Nguyen. Convergence of de Finetti's mixing measure in latent structure models for observed exchangeable sequences. *The Annals of Statistics*, 2021. forthcoming.

Y. Wu and S. Ghosal. The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101(10):2411–2419, 2010.

E. P. Xing, M. I. Jordan, and R. Sharan. Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology*, 14(3):267–284, 2007.