On the Pitman–Yor process with spike and slab prior specification

Antonio Canale[1], Antonio Lijoi[2], Bernardo Nipoti[3] and Igor Prünster [4]

[1] Department of Statistical Sciences, University of Padova, Italy and Collegio Carlo Alberto, Moncalieri, Italy; e-mail: canale@stat.unipd.it

[2] Department of Decision Sciences and BIDSA, Bocconi University, Milano, Italy and Collegio Carlo Alberto, Moncalieri, Italy; e–mail: antonio.lijoi@unibocconi.it

[3] School of Computer Science and Statistics, Trinity College, Ireland; e–mail: nipotib@tcd.ie

[4] Department of Decision Sciences, BIDSA and IGIER, Bocconi University, Milano, Italy; e–mail: igor@unibocconi.it

Abstract

For the most popular discrete nonparametric models, beyond the Dirichlet process, the prior guess at the shape of the data generating distribution, also known as base measure, is assumed to be diffuse. Such a specification greatly simplifies the derivation of analytical results allowing for a straightforward implementation of Bayesian nonparametric inferential procedures. However, in several applied problems the available prior information leads naturally to incorporate an atom into the base measure and one is essentially left with the Dirichlet process as the only tractable choice for the prior. In this paper we fill this gap by considering the Pitman–Yor process featuring an atom in its base measure. We derive computable expressions for the distribution of the induced random partitions and for the predictive distributions. These findings allow us to devise an effective generalized Pólya urn Gibbs sampler. Applications to density estimation, clustering and curve estimation, with both simulated and real data, serve as an illustration of our results and allow comparisons with existing methodology. In particular, we tackle a functional data analysis problem concerning basal body temperature curves.

Keywords: Bayesian Nonparametrics; functional data; Pitman–Yor process; predictive distribution; random partition; spike and slab base measure.

1 Introduction

The specification of two-component mixture priors represents the most popular choice in Bayesian variable selection and when investigating sparsity phenomena. Such mixtures are commonly referred to as spike and slab priors according to a terminology that was introduced in Mitchell & Beauchamp (1988), who use a mixture whose components are a degenerate distribution at 0, referred to as spike, and a diffuse distribution, referred to as slab. The seminal contribution of George & McCulloch (1993), where a mixture of two normal distributions with zero mean and different variances is considered, originated a huge literature on the topic. Further developments, along with an insightful discussion of connections with frequentist procedures, can be found in Ishwaran & Rao (2005).

The present paper investigates the use of a spike and slab prior specification for Bayesian nonparametric inference on the clustering structure featured by the data. Among several possible motivating applications we consider a functional data analysis problem, where the data represent the basal body temperature (bbt) curves of women in their reproductive age. The daily bbt of a healthy woman during the menstrual cycle is known to follow a distinctive biphasic trajectory, which can be described by a specific parametric function of time as

$$f^*(t) = a + b\,\frac{\exp\{ct\}}{1+\exp\{ct\}} \tag{1}$$

and admits a clear clinical interpretation (see § 4.2). Nonetheless, a number of unhealthy women may display a far more irregular functional form that does not preserve the nice S–shape yielded by (1). It is then natural to think of these functional data as being generated "on average" by a mixture probability distribution having a spike at the functional form in (1) and a diffuse component that accommodates for irregular bbt behaviour. See Scarpa & Dunson (2009). In our fully nonparametric framework this idea translates into the use of a nonparametric prior $\tilde{P}$ whose base measure is a convex linear combination of a point mass at the function f^* in (1) and of a diffuse distribution P^* on a suitable set of functions, i.e. $E(\tilde{P}) = \zeta\,\delta_{f^*} + (1-\zeta)\,P^*$. Introducing an atom, corresponding to the regular S-shape, in the base measure allows us to embed useful prior information while maintaining the natural flexibility of the nonparametric approach, which is needed to model the potentially very irregular

shape of unhealthy women. Motivated by different applications, with real-valued data, Dunson et al. (2008), MacLehose et al. (2007), Yang (2012) and Barcella et al. (2016) conveniently adopted a Dirichlet process (DP) with base measure featuring an atom at 0: this allows them to simultaneously perform clustering and variable selection. In fact, an atom at 0 represents a natural way to incorporate the belief that some coefficients might be null with positive probability in the prior. The same construction is used in Suarez & Ghosal (2016) to model wavelet coefficients of functional data so to induce sparsity. Applications to multiple testing problems can be found in Bogdan et al. (2008) and in Kim et al. (2009). Among other contributions proposing testing procedures based on a DP whose base measure is a two-components mixture, we mention Guindani et al. (2009) and Do et al. (2005). When using the DP, the presence of the atom in the base measure does not impact the structure of the predictive distributions because of its conjugacy. Indeed, the predictive distribution can be determined as a linear functional of the posterior distribution, which is still the distribution of a DP regardless of the presence of atoms in the base measure P_0. However, when $\tilde{P}$ is not a DP an atom in P_0 considerably changes the posterior structure of the process and induces some challenging technical issues that need to be addressed in order to perform Bayesian inference.

In this work we investigate the distributional properties of the probably most popular generalization of the DP, namely the Pitman–Yor process (Perman et al., 1992; Pitman & Yor, 1997). We show that, even when an atom is included in the base measure, the process still preserves a considerable degree of analytical tractability. We derive explicit expressions for the associated exchangeable partition probability function (EPPF), the predictive distributions and the distribution of the a priori number of distinct values K_n in an n-sample $X^{(n)} = (X_1, \ldots, X_n)$. These expressions represent the building block of a generalized Blackwell McQueen Pólya urn scheme. The resulting algorithm is then used to carry out an extensive study involving both scalar and functional data. This empirical analysis uncovers some interesting features of the models we are considering and allows useful comparisons with possible alternatives available in the literature. First we assess the different inferential behaviour of Dirichlet and the Pitman–Yor process based models, when the base measure has an atomic component. Our findings show that, somehow similarly to what happens in the case of a diffuse base measure (Lijoi et al., 2007; Jara et al., 2010; De Blasi et al., 2015), models based on the Pitman–Yor pro-

cess are more flexible and more robust with respect to prior misspecifications on the clustering structure of the data. Moreover, we compare the Pitman–Yor process, with spike and slab base measure, with an alternative two-component mixture model defined as a linear combination of an atomic component and a Pitman-Yor process with diffuse base measure, in the spirit of Scarpa & Dunson (2009). Finally, we draw a comparison between models whose base measure is diffuse with models having a fixed atom in the base measure as in (5). The convenience of an atomic component in the base measure, to reflect prior information, is already pointed out in existing literature on the DP for the case of scalar data. Here, instead, we consider functional data in the more general Pitman–Yor setup and evaluate the potential gain in terms of inferential performance. An atom in the base measure defined on some functional space turns out to be greatly beneficial in terms of classification of functions.

2 Some preliminaries on random partitions

Since our goal is to study the clustering structure of the data from a Bayesian nonparametric standpoint, it is natural to consider a discrete random probability measure $\tilde{P}$ and to look at the exchangeable random partition associated to $\tilde{P}$. Assume the data $X_i \,|\, \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$, for $i = 1, \ldots, n$, take values in some space $\mathbb{X}$ and

$$\tilde{P} = \sum_{j \geq 1} \tilde{p}_j \, \delta_{Z_j} \tag{2}$$

is a discrete random probability measure such that $\sum_{j \geq 1} \tilde{p}_j = 1$, almost surely, and the Z_j's are independent and identically distributed $\mathbb{X}$-valued random elements with common distribution P_0. Due to the discreteness of $\tilde{P}$, the n-sample $X^{(n)} = (X_1, \ldots, X_n)$ induces a partition, say Ψ_n, of $[n] = \{1, \ldots, n\}$ such that i and j are in the same partition set when $X_i = X_j$. The corresponding probability distribution $\text{pr}(\Psi_n = \{C_1, \ldots, C_k\})$, where C_j for $j = 1, \ldots, k$ are the unique cluster labels, for any $k \leq n$, is also known as EPPF. See Pitman (1995). Once the EPPF is available, one can determine the predictive distributions associated to the exchangeable sequence $(X_i)_{i \geq 1}$. If the sample $X^{(n)}$ displays k distinct values $x_1^*, \ldots, x_k^*$, then

$$\text{pr}\Big(X_{n+1} \in \mathrm{d}x \mid X^{(n)}\Big) = w_{k,n}^{(0)} \, P_0(\mathrm{d}x) + \sum_{j=1}^{k} w_{k,n}^{(j)} \, \delta_{x_j^*}(\mathrm{d}x), \tag{3}$$

where $E(\tilde{P}) = P_0$ and the weights $\{w_{k,n}^{(j)} : j = 0, 1, \ldots, k\}$ can be expressed in terms of the underlying EPPF. Closed form expressions for predictive distributions in (3) are available for broad classes of discrete random probability measures under the crucial assumption of P_0 being diffuse. See, e.g., Pitman (2003); Lijoi et al. (2005, 2007); James et al. (2009). Beyond the DP, the literature on instances, where the assumption of diffuseness of P_0 is relaxed, is limited and essentially confined to theoretical investigations with no actual implementation. James et al. (2006) consider the class of homogeneous normalized random measures and study the predictive distribution for grouped data when the base measures has an atomic component. Their work sheds light on the technical problems arising when considering an atomic component in the base measure. A related result, confined to the Dirichlet case, can be found in Regazzini (1978). On the other hand, Sangalli (2006) studies the predictive distribution of Poisson–Kingman models when the base measure has an atomic component. Although in line of principle the results we present in this work could be derived from the more general but rather involved expressions in James et al. (2006) and Sangalli (2006), we opted to present a direct derivation that is less cumbersome and, importantly, better illustrates the learning mechanism corresponding to such random measures.

We first recall the definition of the Pitman–Yor process and introduce some notation used throughout. Let $\tilde{P}$ be as in (2) and assume that $(\tilde{p}_j)_{j\geq 1}$ and $(Z_j)_{j\geq 1}$ are independent. Then $\tilde{P}$ is a Pitman–Yor process, $\tilde{P} \sim \mathrm{PY}(\sigma, \theta; P_0)$ with $\sigma \in [0, 1)$ and $\theta > -\sigma$, if the $\tilde{p}_i$'s are constructed according to the following stick-breaking procedure (Perman et al., 1992): $\tilde{p}_1 = V_1$, $\tilde{p}_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$, for $j \geq 2$, and $(V_i)_{i\geq 1}$ is a sequence of independent random variables with $V_i \sim \mathrm{Beta}(1 - \sigma, \theta + i\sigma)$. For a diffuse P_0 the corresponding EPPF equals

$$\mathrm{pr}(\Psi_n = (C_1, \ldots, C_k)) = \Phi_k^{(n)}(n_1, \ldots, n_k; \sigma, \theta) = \frac{\prod_{j=1}^{k-1}(\theta + j\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1} \tag{4}$$

where $(a)_n = \Gamma(a + n)/\Gamma(a)$, for any integer $n \geq 0$, $n_j = \mathrm{card}(C_j)$ are positive integers such that $\sum_{i=1}^{k} n_i = n$. See Pitman (1995). However, if one assumes

$$P_0 = \zeta\, \delta_{x_0} + (1 - \zeta)\, P^*, \tag{5}$$

for some $x_0 \in \mathbb{X}$ and diffuse probability measure P^* on $\mathbb{X}$, then (4) no longer holds true.

Before stating the main results in the next section, we highlight a key difference between $\tilde{P} \sim \text{PY}(\sigma, \theta; P_0)$, with P_0 as in (5) and the alternative spike and slab prior specification

$$\tilde{Q} = \zeta\, \delta_{x_0} + (1 - \zeta)\, \tilde{Q}^*, \tag{6}$$

where $\tilde{Q}^* \sim \text{PY}(\sigma, \theta; P^*)$ and P^* is diffuse as in (5). Henceforth, we shall refer to $\tilde{P} \sim \text{PY}(\sigma, \theta; P_0)$, with P_0 as in (5) as the inner spike and slab model. Similarly, $\tilde{Q}$ as in (6) will be referred to as outer spike and slab model. It is worth noting that the model with and outer spike and slab (6) has been used in Scarpa & Dunson (2009) for the special case of $\tilde{Q}^*$ being a DP. Both processes share the same two–components mixture centering, since it is apparent that $E(\tilde{Q}) = E(\tilde{P}) = P_0$.

Remark 1 The inner and outer spike and slab models yield structurally different priors. An interesting comparison can be made when $\sigma = 0$, which implies that both $\tilde{Q}^*$ and $\tilde{P}$ are Dirichlet processes. If one sets $\zeta \sim \text{Beta}(1, \theta)$, $\tilde{Q}$ can be represented as $\sum_{j \geq 0} \pi_j\, \delta_{Y_j}$ with the random probability masses π_j admitting the same stick–breaking representation characterizing the weights of a DP. Nonetheless, $\tilde{Q}$ is not a DP since the location associated to the first stick–breaking weight, namely Y_0, equals x_0 and therefore the random variables of the sequence $(Y_j)_{j \geq 0}$ are not independent and identically distributed. The substantial difference between the two models can be further appreciated and somehow quantified through the next Proposition, which shows that the variabilities of $\tilde{P}$ and $\tilde{Q}$ around the shared mean, P_0, are different.

Proposition 1 *If $f : \mathbb{X} \to \mathbb{R}$ is any function such that $\int f^2 \, \mathrm{d}P^* < \infty$, then*

$$\text{var}\Big\{ \int f \, \mathrm{d}\tilde{P} \Big\} - \text{var}\Big\{ \int f \, \mathrm{d}\tilde{Q} \Big\} = \zeta(1 - \zeta) \frac{1 - \sigma}{\theta + 1} \int \{f(x_0) - f\}^2 \, \mathrm{d}P^* \geq 0. \tag{7}$$

Hence, the prior uncertainty associated to a Pitman–Yor process with a spike and slab base measure is larger than the uncertainty induced by an outer spike and slab model. In this sense, our fully nonparametric prior is less informative and provides more flexibility than the one used in Scarpa & Dunson (2009). A simple illustration of this finding may be given by choosing f as an indicator function. If $f = \mathbb{1}_{[0,t]}$, for some $t > 0$, one obtains the random

survival functions

$$\tilde{S}_{\tilde{P}}(t) := 1 - \int_0^\infty \mathbb{1}_{[0,t]}(x) \mathrm{d}\tilde{P}(x) \quad \text{and} \quad \tilde{S}_{\tilde{Q}}(t) := 1 - \int_0^\infty \mathbb{1}_{[0,t]}(x) \mathrm{d}\tilde{Q}(x),$$

defined as functionals of the inner and the outer spike and slab models $\tilde{P}$ and $\tilde{Q}$, respectively. By setting $x_0 = 0$, $\tilde{S}_{\tilde{P}}$ and $\tilde{S}_{\tilde{Q}}$ can be conveniently used as nonparametric prior distributions assigning positive probability to the event occurring at time $t = 0$, which in reliability applications may be interpreted as the failure of an item during its production. Let P^* be any diffuse probability measure on $\mathbb{R}^+$, and S^* denote the corresponding survival function. It is straightforward to show that both models have the same prior guess $E\{\tilde{S}_{\tilde{P}}(t)\} = E\{\tilde{S}_{\tilde{Q}}(t)\} = (1-\zeta)S^*(t)$. A direct application of Proposition 1 implies that, for every $t \geq 0$,

$$\mathrm{var}\left\{\tilde{S}_{\tilde{P}}(t)\right\} - \mathrm{var}\left\{\tilde{S}_{\tilde{Q}}(t)\right\} = \zeta(1-\zeta)\frac{1-\sigma}{\theta+1}S^*(t),$$

thus indicating that the random survival function based on the inner spike and slab model $\tilde{P}$ is less concentrated around the prior guess than the one based on the outer spike and slab model $\tilde{Q}$.

3 Pitman–Yor process with spike and slab base measure

The following result concerns a Pitman–Yor process having a point mass in its base measure and provides a closed form expression for its EPPF, denoted as $\Pi_k^{(n)}(n_1, \ldots, n_k)$. The expression is given in terms of generalized factorial coefficients $\mathscr{C}(n_j, i; \sigma) = \frac{1}{i!}\sum_{r=0}^{i}(-1)^r\binom{i}{r}(-r\sigma)_{n_j}$; see Charalambides (2005) for an exhaustive account on their properties. In this section we assume $\sigma \in (0,1)$ and note that the Dirichlet case is obtained by suitably taking the limit for $\sigma \to 0$.

Theorem 1 *The EPPF induced by* $\tilde{P} \sim PY(\sigma, \theta; P_0)$, *where* $P_0 = \zeta\,\delta_{x_0} + (1-\zeta)P^*$ *as in* (5), *is*

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = (1-\zeta)^k \, \frac{\prod_{j=1}^{k-1}(\theta+j\sigma)}{(\theta+1)_{n-1}} \, \prod_{j=1}^{k}(1-\sigma)_{n_j-1}$$
$$+ (1-\zeta)^{k-1} \sum_{j=1}^{k} \frac{\prod_{r=1}^{k-2}(\theta+r\sigma)}{(\theta+1)_{n-1}} \prod_{\ell\neq j}(1-\sigma)_{n_\ell-1} \sum_{i=1}^{n_j} \zeta^i \left(\frac{\theta}{\sigma}+k-1\right)_i \mathscr{C}(n_j,i;\sigma). \quad (8)$$

A simple rearrangement of (8) yields a nice probabilistic interpretation of the result. To this end, recall that the posterior distribution of a $\text{PY}(\sigma,\theta;P_0)$, conditional on a sample of size $n-n_j$ featuring $k-1$ distinct values $x_1^*,\ldots,x_{k-1}^*$, all different from the fixed atom x_0, is equal to the law of

$$\sum_{i=1}^{k-1} \pi_{i,j}\,\delta_{x_i^*} + \pi_{k,j}\tilde{P}_{k-1} \quad (9)$$

with $\tilde{P}_{k-1} \sim \text{PY}(\sigma,\theta+(k-1)\sigma;P_0)$, $(\pi_{1,j},\ldots,\pi_{k-1,j})$ having a $(k-1)$-variate Dirichlet distribution with parameters $(n_1-\sigma,\ldots,n_{j-1}-\sigma,n_{j+1}-\sigma,\ldots,n_k-\sigma;\theta+(k-1)\sigma)$ and $\pi_{k,j}=1-\sum_{i=1}^{k-1}\pi_{i,j}$. Moreover, $(\pi_{1,j},\ldots,\pi_{k-1,j})$ and $\tilde{P}_{k-1}$ are independent. Also note that the distribution of the number of distinct values $K_n \in \{1,\ldots,n\}$, in a sample of size n from a $\text{PY}(\sigma,\theta;P_0)$ process, depends on the parameters (σ,θ,ζ) and for this reason we will use the notation $\text{pr}\{K_n=k;\,(\sigma,\theta,\zeta)\}$ to identify it. In particular, $\zeta=0$ corresponds to a diffuse base measure leading to

$$\text{pr}\{K_n=k;(\sigma,\theta,0)\} = \frac{\prod_{r=1}^{k-1}(\theta+r\sigma)}{(\theta+1)_{n-1}} \, \frac{\mathscr{C}(n,k;\sigma)}{\sigma^k}.$$

Simple algebra, then, leads to the following result.

Corollary 1 *The EPPF of $\tilde{P}\sim PY(\sigma,\theta;P_0)$ with P_0 as in (5) can be represented as*

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = (1-\zeta)^k\,\Phi_k^{(n)}(n_1,\ldots,n_k;\sigma,\theta) + (1-\zeta)^{k-1}\sum_{j=1}^{k}\frac{(\theta+(k-1)\sigma)_{n_j}}{(\theta+n-n_j)_{n_j}}$$
$$\times\ \Phi_{k-1}^{(n-n_j)}(n_1,\ldots,n_{j-1},n_{j+1},\ldots,n_k;\sigma,\theta)\sum_{i=1}^{n_j}\zeta^i\, pr\{K_{n_j}=i;\,(\sigma,\theta+(k-1)\sigma,0)\}, \quad (10)$$

with $\Phi_k^{(n)}(n_1,\ldots,n_k;\sigma,\theta)$ defined as in (4).

The first summand on the right-hand side of (10) corresponds to the case where none of the k partition sets is identified by x_0, its probability being $(1-\zeta)^k$. The second summand corresponds to the case where one of the partition sets is at x_0. The probabilistic interpretation works as follows. For any $j = 1, \ldots, k$: (i) with probability equal to $\Phi_{k-1}^{(n-n_j)}(n_1, \ldots, n_{j-1}, n_{j+1}, \ldots, n_k; \sigma, \theta)$ a partition of $n - n_j$ observations into $k-1$ groups is generated through the diffuse component of the base measure; (ii) conditional on the $k-1$ clusters generated by $n-n_j$ observations through the diffuse component, $\{\theta+(k-1)\sigma\}_{n_j}/(\theta+n-n_j)_{n_j}$ is the probability that the remaining n_j observations are generated by $\tilde{P}_{k-1}$ in (9), which is the only component containing x_0; (iii) conditional on having n_j observations generated by $\tilde{P}_{k-1}$ and equal to x_0, i of them are from the base measure and, if we label them as if they generate separate clusters, the remaining $n_j - i$ are assigned to any of these i labeled groups. In other terms, according to (iii), it is as if the n_j observations are further split into i "fictitious" sub-clusters all identified by x_0.

Having derived a closed form expression for the EPPF, it is now possible to obtain the distribution of the number of distinct values K_n in $X^{(n)}$, for any vector of parameters (σ, θ, ζ), with $\zeta \in [0, 1]$.

Theorem 2 *If $X_i \,|\, \tilde{P} \overset{iid}{\sim} \tilde{P}$, for $i = 1, \ldots, n$, and $\tilde{P} \sim PY(\sigma, \theta; P_0)$, with P_0 as in (5), the probability distribution of the number of distinct values K_n in $X^{(n)}$ equals*

$$\begin{aligned}\operatorname{pr}\{K_n = k;\, (\sigma,\theta,\zeta)\} &= (1-\zeta)^k \operatorname{pr}\{K_n = k;\, (\sigma,\theta,0)\} \\ &+ (1-\zeta)^{k-1} \sum_{r=1}^{n-k+1} \binom{n}{r} \frac{(\theta+(k-1)\sigma)_r}{(\theta+n-r)_r} \operatorname{pr}\{K_{n-r} = k-1;\, (\sigma,\theta,0)\} \\ &\times \sum_{i=1}^{r} \zeta^i \operatorname{pr}\{K_r = i;\, (\sigma,\theta+(k-1)\sigma,0)\}. \quad (11)\end{aligned}$$

The predictive distributions associated to the exchangeable sequence $(X_i)_{i\geq 1}$ directed by $\tilde{P} \sim \mathrm{PY}(\sigma, \theta; P_0)$, with P_0 as in (5), can also be readily obtained from the corresponding EPPF. Suppose the observed sample $X^{(n)}$ displays k distinct values $x_1^*, \ldots, x_k^*$ with respective frequencies $n_1, \ldots, n_k$.

Theorem 3 *Let* $X_i \,|\, \tilde{P} \overset{iid}{\sim} \tilde{P}$, *for* $i = 1, \ldots, n$, *and* $\tilde{P} \sim PY(\sigma, \theta; P_0)$, *with* P_0 *as in* (5). *The corresponding predictive distribution is given by:*

(i) *if* $x_0 \notin \{x_1^*, \ldots, x_k^*\}$

$$\text{pr}(X_{n+1} \in A \mid X^{(n)}) = \frac{\theta + k\sigma}{\theta + n} P_0(A) + \frac{1}{\theta + n} \sum_{j=1}^{k} (n_j - \sigma)\, \delta_{x_j^*}(A);$$

(ii) *if* $x_0 = x_j^*$, *for some* $j = 1, \ldots, k$,

$$\begin{aligned}\text{pr}(X_{n+1} \in A \mid X^{(n)}) = (1-\zeta)\, \frac{\theta + (k-1)\sigma}{\theta + n}\, \frac{\sum_{i=1}^{n_j} \zeta^i\, \mathscr{C}(n_j, i; \sigma) \left(\frac{\theta}{\sigma} + k\right)_i}{\sum_{i=1}^{n_j} \zeta^i\, \mathscr{C}(n_j, i; \sigma) \left(\frac{\theta}{\sigma} + k - 1\right)_i} P^*(A) \\ + \frac{1}{\theta + n}\, \frac{\sum_{i=1}^{n_j+1} \zeta^i\, \mathscr{C}(n_j + 1, i; \sigma) \left(\frac{\theta}{\sigma} + k - 1\right)_i}{\sum_{i=1}^{n_j} \zeta^i\, \mathscr{C}(n_j, i; \sigma) \left(\frac{\theta}{\sigma} + k - 1\right)_i} \delta_{x_j^*}(A) + \frac{1}{\theta + n} \sum_{\ell \neq j} (n_\ell - \sigma)\, \delta_{x_\ell^*}(A).\end{aligned}$$

Note that the predictive distribution for case (i) coincides with the well-known ones of the Pitman–Yor case with diffuse base measure. Moreover, if $\zeta = 0$ in (5), the predictive distribution reduces to that of case (i), as required. Analogously it is easy to see that with $\zeta = 0$ in (5) the EPPF in Theorem 1 reduces to that of the non-atomic case in (4). Theorem 3 provides the basic ingredients for devising the Pólya urn type algorithm that will be used in § 4 and whose details are provided in the Appendix.

4 Illustration

4.1 Synthetic data

The previous results pave the way to a straightforward implementation of the inner spike and slab nonparametric model to a number of applications of interest. In §4.2 we will focus on a real data application concerning the analysis of functional data. However, we need to further investigate distributional properties of the inner spike and slab model. This will be achieved through the use of synthetic data in the present section.

For $\tilde{P} \sim \text{PY}(\sigma, \theta; P_0)$ with P_0 a diffuse probability measure, there is an extensive literature aimed at investigating the effects of σ on posterior inferences for the Pitman–Yor process and allied nonparametric priors (see Lijoi et al., 2007; Jara et al., 2010; De Blasi et al., 2015). Here, we aim at understanding whether these features are preserved when one allows the base measure P_0 of the Pitman–Yor process to have an atom. To this end we perform a simulation study with $R = 100$ replicated samples of size $n = 50, 100$ so to mimic a quality control application. In these context a given random element X is supposed to have a precise nominal value and the goal of the analysis is to assess whether the nominal value is plausible or not on the basis of an observed random sample $X^{(n)}$ of X. We assume that the measurements of X_i, with $i = 1, \ldots, n$, are taken with a measuring instrument with known precision. Data are simulated from a location-scale mixture of Gaussian kernels, i.e.

$$g_0(X) = \sum_{h=1}^{5} \pi_h \phi(X; m_h, t_h^{-1}),$$

where $\phi(.; m, t^{-1})$ is the normal density with mean m and precision t. We further set $m_1 = 0$ and t_1^{-1} =0.04 as the nominal value of X and the variance of the measuring instrument, respectively. The values of the remaining parameters are reported in the Appendix. Data are analyzed assuming

$$X_i \mid (\mu_i, \tau_i) \overset{\text{ind}}{\sim} N(\mu_i, \tau_i^{-1}), \quad (\mu_i, \tau_i) \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \quad \tilde{P} \mid \zeta \sim \text{PY}(\sigma, \theta; P_0).$$

Given the prior information on the nominal value and on the precision of the measuring instrument, the base measure P_0 is specified so to assign positive mass to the pair (m_1, t_1) and thus is a mixture of a point mass and a diffuse density, and precisely

$$P_0 = \zeta \delta_{(m_1, t_1)} + (1 - \zeta) P^*,$$

where P^* is normal-gamma. In order to also learn the proportion of observations that can be suitably modeled by the spike in (m_1, t_1), we further assume that ζ has a uniform prior between zero and one. The analysis is repeated with different choices of σ and θ. Specifically we take $\sigma \in \{0, 0.25, 0.5, 0.75\}$ and we fix θ using the results of Theorem 2 so to have a prior expected number of mixture components equal to 3 or 15 thus corresponding, respectively,

$E(K_n)$	σ	$n = 50$	$n = 100$
3	0	3.14	3.04
	0.25	3.40	3.35
	0.50	4.13	3.94
	0.75	5.71	4.68
15	0	11.99	11.33
	0.25	11.49	10.21
	0.50	10.50	8.07
	0.75	8.42	5.58

Table 1: *Posterior number of mixture components for location-scale mixture simulation experiment.*

to an under- and over-estimation of the true number of components. Details on the values of θ and on the Markov chain Monte Carlo sampling algorithm employed to sample from the posterior distribution of the parameters are reported in the Appendix. The results, displayed in Table 1, are coherent with the findings in the case of nonparametric mixtures with non-atomic base measures (Lijoi et al., 2007; De Blasi et al., 2015). Specifically, for larger values of σ, the estimated number of mixture components is closer to the true value. This nicely showcases the effectiveness of the additional model flexibility conveyed by σ in overcoming possible prior misspecifications. The numerical estimates are reported in Table 1 with the largest Monte Carlo standard errors being equal to 0.75 and 1.25 for the first and last four rows, respectively.

The second simulation experiment compares the inner and outer spike and slab models in terms of estimation of the proportion of observations allocated to the spike component. A straightforward application of Proposition 1 with $f = \mathbb{1}_{\{x_0\}}$, shows that the variance of the random mass assigned by the inner model to the atom x_0 is larger than the variance of the corresponding mass assigned by the outer model, the difference being equal to $\zeta(1-\zeta)(1-\sigma)/(\theta+1)$. This difference suggests that the inner spike and slab model should provide more robust posterior inference on the proportion of observations allocated to the spike, when ζ is

	σ	$n = 50$	$n = 100$
Inner	0	0.43	0.41
	0.25	0.42	0.40
	0.50	0.41	0.39
	0.75	0.42	0.39
Outer	0	0.50	0.49
	0.25	0.49	0.49
	0.50	0.49	0.49
	0.75	0.49	0.49

Table 2: *Posterior proportion of subjects allocated to the spike for location-scale mixture simulation experiment.*

fixed and its value misspecified. In order to check this we consider exactly the same simulated data as in the first experiment and keep $x_0 = (m_1, t_1)$. For both the inner and the outer models, we fix $\zeta = 0.8$ instead of assigning it a uniform prior. Given that the true value is 0.4, this amounts to a strong misspecification. For every value of $\sigma \in \{0{,}0.25{,}0.5{,}0.75\}$ we set the parameter θ in both models so that the prior expected number of components is equal to 5, the true number of components in our simulations. Details on how we set θ are reported in the Appendix. The results displayed in Table 2 show that the inner spike and slab model is clearly superior than the outer model in overcoming possible prior misspecifications. The largest Monte Carlo standard error in Table 2 is 0.08.

The third simulation experiment aims at highlighting the benefit of the inclusion of the spike in the base measure when there is supporting prior information. In fact, one might be tempted to think that the flexibility of the Pitman–Yor process alone is enough to detect the spike and assign sufficient posterior mass to it. Our simulation study shows that this is not the case. We simulate $R = 50$ datasets that mimic the characteristics of the bbt functional data of our motivating application. The daily bbt of a healthy woman is known to follow a distinctive biphasic trajectory that can be described, in simplified terms, as a function of time

t as

$$f(t) = \frac{\mathrm{e}^t}{1+\mathrm{e}^t}. \tag{12}$$

Unhealthy women, however, tend to exhibit far more irregular curves' shapes. For each dataset, we simulate $n = 50$ functional data from the following data generating process

$$X_{it} \mid f_i \overset{\text{ind}}{\sim} N(f_i(t), \sigma^2), \quad f_i \overset{\text{iid}}{\sim} P, \quad P = \sum_{j=1}^{5} \delta_{f_j^*} \pi_j,$$

where f_1^* is exactly (12) and $\pi_1 = 0.4$. The remaining curves f_j^*, for $j = 2, \ldots, 5$, and values of the parameters are reported in the Appendix. Data are analyzed assuming

$$X_{it} \mid f_i \overset{\text{ind}}{\sim} N(f_i(t), \sigma^2), \quad f_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \quad \tilde{P} \mid \zeta \sim \text{PY}(\sigma, \theta; P_0)$$

with two different specifications for P_0, which will be assumed as being either a mixture of a point mass at (12) and a diffuse measure over the space of functions, or a plain non-atomic measure. For both choices of P_0 we set $\sigma = 0.5$ and fix θ so to have the same prior expected number of mixture components. Additional details on the prior specification and posterior computation are reported in the Appendix.

The results highlight the benefits of including the spike in the prior specification. Ignoring prior information concerning a prevalent functional form for the data and consequently using a diffuse base measure, leads to a significant worsening of the inferences. This can be deduced, for example, from the posterior clustering structure and, in particular, from the binary classification of a subject into a cluster with or without biphasic shape. For the model with spike and slab P_0 this corresponds to checking if a subject belongs to the cluster represented by the fixed atom. For the model with diffuse P_0, we label as "biphasic" the cluster in which the majority of the data coming from (12) are clustered. The numerical results are reported in Table 3, with the largest Monte Carlo error for the global accuracy being equal to 0.07. The significantly better performance in terms of accuracy and false positives rates provides clear evidence in favor of the spike and slab specification of P_0.

Another appealing inferential implication of the spike and slab base measure specification is that the subject specific posterior functional means are more precise for the subjects coming from (12). Figure 1 displays the estimated functional mean and 90% pointwise posterior

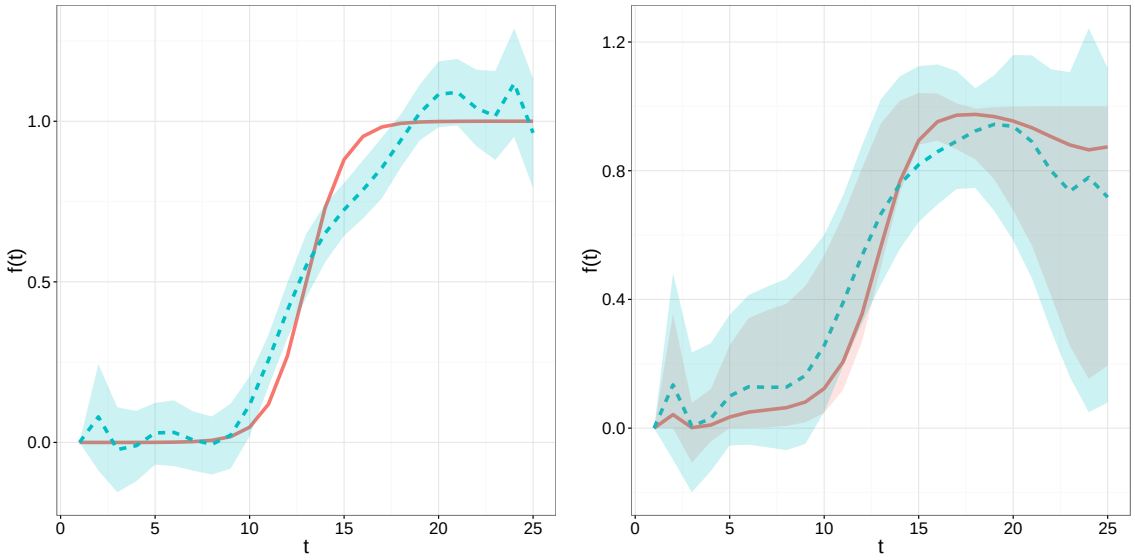
	Spike and slab base measure	Diffuse base measure
Accuracy	0.834	0.747
False Positive	0.305	0.557
False Negative	0.072	0.049

Table 3: *Confusion matrix for the third simulation experiment.*

credible bands for two subjects having true mean equal to (12). The functional mean and limits of the credible bands are estimated with the empirical mean and empirical quantiles of order 0.05, 0.95, determined through the Markov chain Monte Carlo iterations. The plots refer to one of the $R = 50$ datasets, though qualitatively similar results can be found in almost any replicate. Panel (a) concerns a subject classified in the true cluster with the spike and slab model more than 99.9% of the Markov chain Monte Carlo iterations. In such a case, as the cluster's shape is not estimated but fixed, there is no credible band around the continuous line. In contrast, for the model without spike the curve's shape clearly cannot coincide with (12) since it is estimated from the data and it is worth noting that this estimate is erratic on the left and right part of the domain. Panel (b) concerns a borderline subject classified as biphasic in the 85% of the Markov chain Monte Carlo iterations for the spike and slab model and only in the 60% of the iterations for the non-atomic model. This leads to wider credible bands in both cases.

4.2 Basal body temperature functional data

As previously mentioned our motivating application is concerned with functional data analysis. Specifically we study a dataset on daily measurements of bbt consisting of 1118 non-conception cycles from $n = 157$ women in the Verona center of the Colombo & Masarotto (2000) study. As shown in panel (a) of Figure 2, the bbt curve trajectory over time of healthy women in reproductive age follows a biphasic trajectory that can be described by (12) or, more in general,



(a) (b)

Figure 1: *Posterior functional means for two subjects having true mean equal to* (12)*: continuous red lines correspond to the model spike and slab base measure, whereas dashed cyan lines to the model with non-atomic base measure. Shaded areas depict the posterior pointwise 90% credible bands. panel (a) corresponds to a subject clearly belonging to* (12)*, panel (b) to a "borderline" case.*

by the parametric function of time t

$$f(t;\tau_1,\tau_2,\lambda,\omega) = \tau_1 + \tau_2 \left(\frac{\exp\{\frac{t-\lambda}{\omega}\}}{1+\exp\{\frac{t-\lambda}{\omega}\}} \right). \tag{13}$$

The representation in (13) is particularly convenient, since the parameters have a clear clinical interpretation. For example τ_1 represents the value of hypothermia during the follicular phase of the cycle, λ can be interpreted as the moment of ovulation, $(\tau_1 + \tau_2)$ is the level that the bbt reaches after the sharp increase, controlled by ω, which happens just before the ovulation. In contrast, unhealthy women tend to have different curve's shapes as shown, for example, in panel (b) of Figure 2.

The nonparametric model with spike and slab base measure perfectly fits the present setup: it allows to assign prior positive mass to curves with the peculiar healthy women's shape, and at the same time, to account for abnormal deviations from this standard shape via an extremely flexible nonparametric functional data mixture model. The same dataset has been previously analyzed by Scarpa & Dunson (2009), with similar goals but a rather different approach: as formalized in (6), they rely on a parametric model with possible nonparametric contaminations, rather than a fully nonparametric model with an informative prior specification.

Let n_{ij} denote the duration of cycle $j = 1, \ldots, n_i$ of woman $i = 1, \ldots, N$. For every $t = 1, 2, \ldots, n_{ij}$, the bbt $X_{ij}(t)$ is observed. We assume that the measurements $X_{ij}(t)$ can be modelled as

$$X_{ij}(t) = \tau_{1ij} + \tau_{2ij} f_{ij}\left(\tfrac{t-\lambda_{ij}}{\omega_{ij}}\right) + \epsilon_{ij}(t), \tag{14}$$

where $\epsilon_{ij}(t)$ are independent measurement errors modeled as $\epsilon_{ij}(t) \sim N(0, \sigma^2)$, and f_{ij} is a smooth random function with prior

$$f_{ij} \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \quad \tilde{P} \mid \zeta \sim \text{PY}(\theta, \sigma; P_0).$$

where P_0 has a spike and slab structure of the type

$$P_0 = \zeta \delta_{f_0} + (1 - \zeta) P^*,$$

$f_0(t) = e^t/(1 + e^t)$ represents the biphasic curve and P^* is a non-atomic probability measure on a function space. The almost sure discreteness of the Pitman–Yor process induces ties, with positive probability, among the f_{ij}'s. We denote these atoms by f_h^* for $h = 1, \ldots, k$.

As probability measure on the function space we consider the prior induced by a B-spline basis expansion, namely

$$g \sim P^*, \quad g(t) = B(t)^T \beta, \quad \beta \sim MVN(\beta_0, \Sigma_0),$$

with $B(.)$ denoting the B-splines basis, $MVN(m, V)$ the multivariate normal distribution with mean vector m and variance matrix V, and β a finite vector of basis coefficients. The Bayesian

specification of the model is then completed by eliciting prior distributions for all the remaining parameters that we assume independent. We let

$$\begin{aligned} (\tau_{1ij}, \tau_{2ij}) &\sim N(\alpha_i, \Omega), & \alpha_i &\sim N(\alpha_0, R) \\ \lambda_{ij} &\sim U(b_{ij}+10, b_{ij}+20), & \omega_{ij} &\sim \text{Ga}(1/2, 1), \\ 1/\sigma^2 &\sim \text{Ga}(1/2, 1/2), & \zeta &\sim U(0, 1), \end{aligned} \tag{15}$$

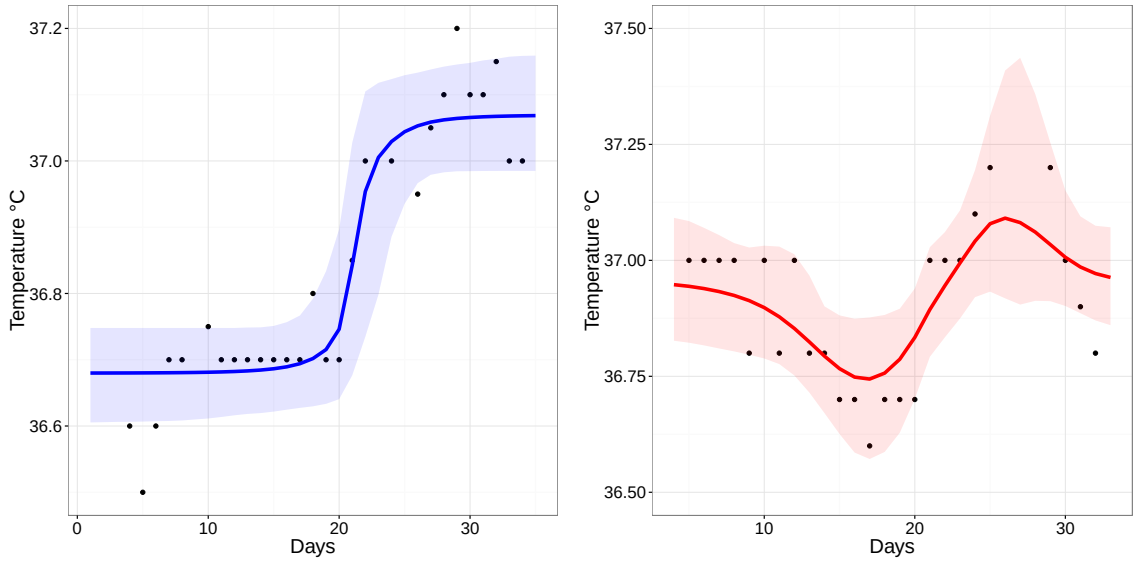
where b_{ij} denotes the first day after bleeding for cycle i of woman j, $U(a, b)$ denotes the uniform distribution over (a, b) and $\text{Ga}(c, d)$ stands for the gamma distribution with expected value c/d. For simplicity Ω, R, and Σ_0 are identity matrices while α_0 and β_0 are vectors of suitable dimensions of zeroes. Note that the specifications in (15) allow to model within- and between-woman heterogeneity thanks to the presence of the woman specific parameters α_i. The parameters of the Pitman–Yor process are set equal to $\theta = 1$ and σ =0.25, while a uniform prior on ζ, the prior proportion of cycles belonging to the parametric atom, is assumed in order to allow the model to learn this feature from the data.

Posterior sampling is performed with the Gibbs sampler described in the Appendix. Its derivation, for the parametric part, is straightforward and follows standard results on linear regression and spline interpolation. For the nonparametric part, the sampler is obtained by using the results of § 3. We perform our analysis by running the algorithm for 8,000 iterations and discarding the first 3,000 as burn in. Convergence was assessed by visual inspection of the traceplots which provided no evidence against it.

The posterior probability of being allocated to the biphasic component f_0 was greater than 50% for 94.09% of the cycles under study. The posterior mean of ζ is 0.9283 with 95% quantile based posterior credible interval equal to (0.9097, 0.9450).

Panel (a) of Figure 2 displays the pointwise posterior mean and 95% credible bands for a biphasic cycle of a healthy woman. For this observation, as for all observations falling in the biphasic cluster, we can perform inference on important features such as, for instance, the day of ovulation and the level of the low and high plateau for the first cycle. The corresponding posterior distributions are depicted in Figure 3.

The cycles that do not fit the biphasic pattern are clustered in separate groups by our model. More specifically, the posterior median number of clusters is equal to 4 with first and



(a) (b)

Figure 2: *Panels (a) and (b) report the bbt data for two cycles along with pointwise posterior means and 95% credible bands.*

third quartiles equal to 4 and 5, respectively. These are potentially abnormal or related to unhealthy women. Panel (b) of Figure 2 shows an example.

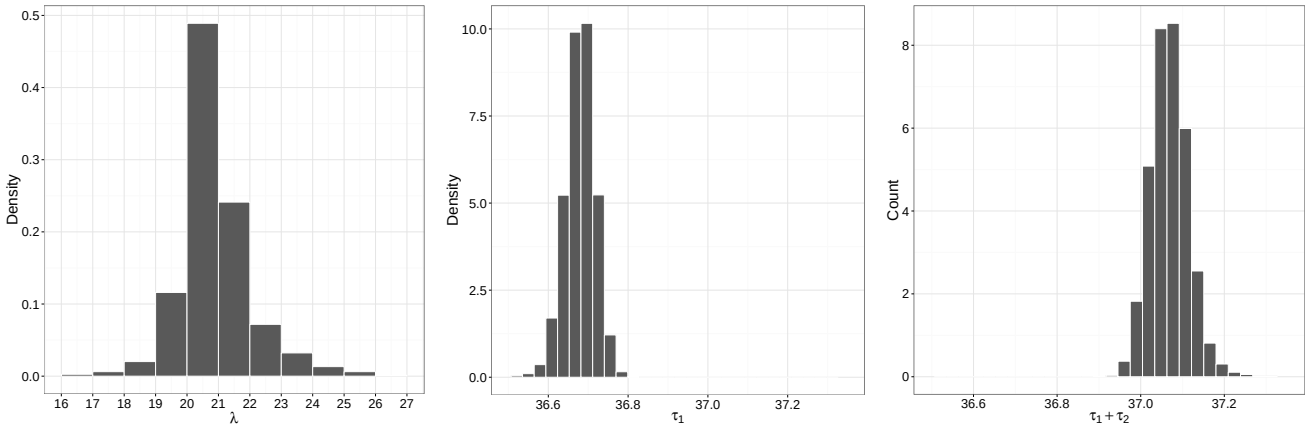
Appendix A: Proofs

A.1 Proof of Proposition 1

One may proceed along the same lines as in Proposition 1 in James et al. (2006) and show that for a Pitman–Yor process $\tilde{H}$ with parameters (σ, θ) and any type of base measure H_0, i.e. diffuse or atomic or combinations thereof, one has

$$\operatorname{var}\Big\{\int f \,\mathrm{d}\tilde{H}\Big\} = \frac{1-\sigma}{\theta+1}\Big\{\int f^2 \,\mathrm{d}H_0 - \Big(\int f \,\mathrm{d}H_0\Big)^2\Big\}.$$

Specializing this for $\tilde{P}$ and $\tilde{Q}$ as in the statement yields (7). □



(a) (b) (c)

Figure 3: *Panels (a), (b), and (c) display the estimated posterior distributions of the day of ovulation, the level of the low and high plateau for the cycle in the left panel of Figure 2.*

A.2 Proof of Theorem 1

In order to prove the result we resort to an alternative construction of the Pitman–Yor process that makes use of completely random measures and is more convenient when the goal is to derive distributional properties. See Lijoi & Prünster (2010) for a review of nonparametric priors using completely random measures as unifying concept. To this end, recall that a completely random measure is a random measure $\tilde{\mu}$ on $\mathbb{X}$ such that, for any collection of pairwise disjoint subsets $A_1, \ldots, A_k$ of $\mathbb{X}$, and $k \geq 1$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_k)$ are mutually independent. For homogeneous and (almost surely) finite completely random measures without fixed points of discontinuity, which are of interest here, the Laplace functional is of the form

$$E\left\{\mathrm{e}^{-\int_{\mathbb{X}} f(x)\,\tilde{\mu}(\mathrm{d}x)}\right\} = \exp\left\{-\int_{\mathbb{R}^+\times\mathbb{X}} \left(1-\mathrm{e}^{-sf(x)}\right)\rho(s)\,\mathrm{d}s\,cP_0(\mathrm{d}x)\right\} \tag{16}$$

for any $f : \mathbb{X} \to \mathbb{R}^+$ with $\rho(s)\,cP_0(\mathrm{d}x)$ the Lévy intensity characterizing $\tilde{\mu}$. The σ-stable completely random measure (Kingman, 1975) is identified by setting $\rho(s) = \sigma s^{-1-\sigma}/\Gamma(1-\sigma)$, for some $\sigma \in (0,1)$ and let $\mathbb{P}_\sigma$ denote its probability distribution. The construction of the

Pitman–Yor process, due to Pitman & Yor (1997), is then as follows. For any $\theta \geq 0$, introduce another probability measure $\mathbb{P}_{\sigma,\theta}$, which is absolutely continuous with respect to $\mathbb{P}_\sigma$ and such that

$$\frac{\mathrm{d}\mathbb{P}_{\sigma,\theta}}{\mathrm{d}\mathbb{P}_\sigma}(m) = \frac{\Gamma(\theta+1)}{\Gamma(\theta/\sigma+1)}\, m^{-\theta}(\mathbb{X}). \tag{17}$$

The resulting random measure $\tilde{\mu}_{\sigma,\theta}$ with distribution $\mathbb{P}_{\sigma,\theta}$ is almost surely discrete while not completely random. Moreover, $\tilde{P} = \tilde{\mu}_{\sigma,\theta}/\tilde{\mu}_{\sigma,\theta}(\mathbb{X})$ is a Pitman–Yor process $\tilde{P} \sim \mathrm{PY}(\sigma,\theta;P_0)$.

Given this, the proof amounts to determining

$$E\Big\{ \int_{\mathbb{X}^k} \tilde{P}^{n_1}(\mathrm{d}x_1)\, .s\, \tilde{P}^{n_k}(\mathrm{d}x_k) \Big\} \tag{18}$$

for any k-tuple of positive integers $n_1,\ldots,n_k$ such that $\sum_{i=1}^k n_i = n$, and integrating variables such that $x_1 \neq .s \neq x_k$. By virtue of Fubini's theorem and the definition of the Pitman–Yor process, (18) equals

$$\frac{\Gamma(\theta+1)}{\Gamma(\theta/\sigma+1)}\,\frac{1}{\Gamma(\theta+n)} \int_{\mathbb{X}^k}\int_0^\infty u^{\theta+n-1} E\Big\{ \mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\mathbb{X})} \prod_{j=1}^k \tilde{\mu}_{\sigma,0}^{n_j}(\mathrm{d}x_j) \Big\}\,\mathrm{d}u, \tag{19}$$

where $\tilde{\mu}_{\sigma,0}$ denotes the σ-stable completely random measure. Let us focus on the determination of $E\{\prod_{j=1}^k \tilde{P}^{n_j}(\mathrm{d}x_j)\}$, i.e. the inner integral in (19). If none of the x_j's equals x_0, only the diffuse component P^* of P_0 contributes to the integral and the integrand boils down to the known expression of the Pitman–Yor process with diffuse base measure, i.e.

$$\begin{aligned} E\Big\{ \prod_{j=1}^k \tilde{P}^{n_j}(\mathrm{d}x_j) \Big\} \cong \Big\{ \prod_{j=1}^k (1-\zeta) P^*(\mathrm{d}x_j) \Big\}\, \frac{\sigma^k}{(\theta+1)_{n-1}\,\Gamma(\theta/\sigma+1)} \\ \times \int_0^\infty u^{\theta+n-1}\,\mathrm{e}^{-u^\sigma} \Big\{ \prod_{j=1}^k \frac{1}{\Gamma(1-\sigma)} \int_0^\infty s^{n_j-\sigma-1}\,\mathrm{e}^{-us} \Big\}\,\mathrm{d}u \end{aligned} \tag{20}$$

where each $\mathrm{d}x_j$ stands for an infinitesimal neighbourhood around x_j. Hence, (20) is a first-order approximation of $E\{\prod_{j=1}^k \tilde{P}^{n_j}(\mathrm{d}x_j)\}$ and note that the higher order terms vanish when computing the integral over $\mathbb{X}^k$ in (19). The right-hand side of (20) can be rewritten as

$$\Big\{ \prod_{j=1}^k P^*(\mathrm{d}x_j) \Big\}\, \frac{(1-\zeta)^k \sigma^k \Big\{ \prod_{j=1}^k (1-\sigma)_{n_j-1} \Big\}}{(\theta+1)_{n-1}\,\Gamma(\theta/\sigma+1)} \int_0^\infty u^{\theta+k\sigma-1}\,\mathrm{e}^{-u^\sigma}\,\mathrm{d}u$$

$$= \Big\{ \prod_{j=1}^{k} P^*(\mathrm{d}x_j) \Big\} \, \frac{(1-\zeta)^k \prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \, \prod_{j=1}^{k} (1-\sigma)_{n_j-1}$$

On the other hand, if $x_0 = x_j$ for some $j \in \{1, \ldots, k\}$, the expected value in the integral in (19) equals

$$E\left\{ \mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\{x_0\})} \, \tilde{\mu}_{\sigma,0}^{n_j}(\{x_0\}) \right\} E\Big\{ \mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\mathbb{X}\setminus\{x_0\})} \prod_{\ell \neq j} \tilde{\mu}_{\sigma,0}^{n_\ell}(\mathrm{d}x_\ell) \Big\} \tag{21}$$

where the factorization follows from the definition of completely random measure. The second factor on the right-hand side of (21) can be easily evaluated since $x_\ell \neq x_0$ for any $\ell \neq j$ and it, thus, involves only the diffuse component P^* of P_0, i.e.

$$E\Big\{ \mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\mathbb{X}\setminus\{x_0\})} \prod_{\ell \neq j} \tilde{\mu}_{\sigma,0}^{n_\ell}(\mathrm{d}x_\ell) \Big\} \cong \Big\{ \prod_{\ell \neq j} P^*(\mathrm{d}x_\ell) \Big\}$$
$$\times \; (1-\zeta)^{k-1} \, \mathrm{e}^{-(1-\zeta)\psi(u)} \, u^{(k-1)\sigma - n + n_j - 1} \sigma^{k-1} \prod_{\ell \neq j} (1-\sigma)_{n_\ell - 1}$$

and the above approximation is to be interpreted as the one given in (20). As for the first factor, one has

$$E\Big\{ \mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\{x_0\})} \, \tilde{\mu}_{\sigma,0}^{n_j}(\{x_0\}) \Big\} = (-1)^{n_j} \frac{\mathrm{d}^{n_j}}{\mathrm{d}u^{n_j}} \mathrm{e}^{-\zeta\,\psi(u)} = \mathrm{e}^{-\zeta\,\psi(u)} \sum_{i=1}^{n_j} \zeta^i \, \xi_{n_j,i}(u)$$

where $\psi(u) = \int_0^\infty (1 - \mathrm{e}^{-us}) \, \rho(s) \, \mathrm{d}s$ and for any $n \geq 1$

$$\xi_{n,i}(u) = \frac{1}{i!} \sum_{j=0}^{i} (-1)^{n-j} \binom{i}{j} \psi^{i-j}(u) \, \frac{\mathrm{d}^n}{\mathrm{d}u^n} \, \psi^j(u).$$

Since $\tilde{\mu}_{\sigma,0}$ has intensity $\sigma s^{-1-\sigma} \, P_0(\mathrm{d}x)/\Gamma(1-\sigma)$, then $\psi(u) = u^\sigma$ and

$$\xi_{n,i}(u) = u^{i\sigma - n} \, \frac{1}{i!} \sum_{j=0}^{i} \binom{i}{j} (-1)^j (-j\sigma)_n = u^{i\sigma-n} \, \mathscr{C}(n, i; \sigma).$$

Hence

$$E\Big\{ \mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\{x_0\})} \, \tilde{\mu}_{\sigma,0}^{n_j}(\{x_0\}) \Big\} = \mathrm{e}^{-\zeta\,u^\sigma} \sum_{i=1}^{n_j} \zeta^i \, u^{i\sigma - n_j} \, \mathscr{C}(n_j, i; \sigma). \tag{22}$$

To sum up, the integrand in (19) is a linear combination of the case where $x_0 \notin \{x_1, \ldots, x_k\}$ and the case where $x_0 = x_j$, for $j = 1, \ldots, k$ and it can be represented as follows

$$E\Big\{\prod_{j=1}^{k} \tilde{P}^{n_j}(\mathrm{d}x_j)\Big\} \cong \Big\{1 - \sum_{j=1}^{k} \delta_{x_0}(\mathrm{d}x_j)\Big\}\Big\{\prod_{j=1}^{k} P^*(\mathrm{d}x_j)\Big\} \frac{(1-\zeta)^k \prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1}$$
$$+ \sum_{j=1}^{k} \delta_{x_0}(\mathrm{d}x_j) \frac{\Big\{\prod_{\ell \neq j} P^*(\mathrm{d}x_\ell)\Big\}(1-\zeta)^{k-1}\sigma^{k-1}}{(\theta+1)_{n-1}\,\Gamma(\theta/\sigma+1)} \Big\{\prod_{\ell \neq j}(1-\sigma)_{n_\ell-1}\Big\}$$
$$\times \sum_{i=1}^{n_j} \zeta^i \mathscr{C}(n_j, i; \sigma) \int_0^\infty u^{\theta+(k-1+i)\sigma-1}\, \mathrm{e}^{-u^\sigma}\, \mathrm{d}u$$

which, as before, is a first-order approximation with vanishing higher order terms, and equals

$$\Big\{1 - \sum_{j=1}^{k} \delta_{x_0}(\mathrm{d}x_j)\Big\}\Big\{\prod_{j=1}^{k} P^*(\mathrm{d}x_j)\Big\} \frac{(1-\zeta)^k \prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1}$$
$$+ \sum_{j=1}^{k} \delta_{x_0}(\mathrm{d}x_j) \frac{\Big\{\prod_{\ell \neq j} P^*(\mathrm{d}x_\ell)\Big\}(1-\zeta)^{k-1}\sigma^{k-2}}{(\theta+1)_{n-1}\,\Gamma(\theta/\sigma+1)} \Big\{\prod_{\ell \neq j}(1-\sigma)_{n_\ell-1}\Big\}$$
$$\times \sum_{i=1}^{n_j} \zeta^i \mathscr{C}(n_j, i; \sigma)\, \Gamma\Big(\frac{\theta}{\sigma} + k - 1 + i\Big)$$

If we plug this expression in (19), simple algebra yields (8). □

A.3 Proof of Theorem 2

This follows from (10) in Corollary 1 and the fact that

$$\mathrm{pr}\{K_n = k;\, (\sigma, \theta, 0)\} = \frac{1}{k!} \sum_{\Delta_{k,n}} \binom{n}{n_1 \,.s\, n_k} \Phi_k^{(n)}(n_1, \ldots, n_k; \sigma, \theta)$$

where $\Delta_{k,n}$ is the set of all vectors of positive integers $(n_1, \ldots, n_k)$ such that $\sum_{i=1}^{k} n_i = n$. □

A.4 Proof of Theorem 3

Recall that the weights of the predictive distribution in (3), may be determined as follows

$$w_{k,n}^{(0)} = \frac{\Pi_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1)}{\Pi_k^{(n)}(n_1, \ldots, n_k)}, \qquad w_{k,j}^{(j)} = \frac{\Pi_k^{(n+1)}(n_1, \ldots, n_j + 1, \ldots, n_k)}{\Pi_k^{(n)}(n_1, \ldots, n_k)}.$$

In view of Theorem 1, if $x_0 \notin \{x_1^*, \ldots, x_k^*\}$, then only the first summand on the right-hand side of (8) is involved in the determination of $w_{k,n}^{(0)}$ and $w_{k,n}^{(j)}$, for $j = 1, \ldots, n$. It is clear, now, that (i) follows and, as expected, it equals the predictive distribution one would have had if P_0 were diffuse. On the other hand, if $x_0 = x_j^*$ for some $j = 1, \ldots, k$, then the second summand on the right-hand side of (8) determines the predictive weights and simple algebra yields (ii). □

Appendix B: Model specifications and sampling schemes

The R code used in the paper is available at the github repository
`github.com/tonycanale/PitmanYorSpikeAndSlab/`

B.1 Blackwell–MacQueen Pólya urn scheme

Before detailing in the next section the specific algorithms we resorted to in our experiments on simulated and real data, we stress that their main ingredient is represented by the predictive distributions derived in Theorem 3. These can be used to tailor the general Blackwell–MacQueen Pólya urn scheme for $\tilde{P} \sim \mathrm{PY}(\sigma, \theta; P_0)$ with a spike and slab base measure P_0 reported below. Let $X_i \,|\, \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$, for $i = 1, \ldots, n$. We assume that the distinct values of X_i are $x_0^*, x_1^*, \ldots, x_k^*$, where x_0^* represents the atom in the base measure (5). If the distinct values do not contain the atom, the algorithm below simplifies to a standard Blackwell–MacQueen Pólya urn scheme. Let furthermore $\mathrm{pr}(X_i \mid X_{\setminus i})$ be the probability of X_i conditionally on all the remaining quantities, $k_{\setminus i}$ be the number of distinct values of x_j^* labelled from 0 to $k_{\setminus i} - 1$ and n_j be the number of observations equal to x_j^*. Then the induced Blackwell–MacQueen Pólya

urn scheme is obtained sampling X_i for $i = 1, \ldots, n$, from a multinomial with cell probabilities

$$\begin{aligned}
\text{pr}(X_i = x_0^* \mid X_{-i}) &\propto \frac{1}{\theta + n - 1} \frac{\sum_{l=1}^{n_0+1} \zeta^l \mathscr{C}(n_0 + 1, l; \sigma)(\theta/\sigma + k_{\setminus i} - 1)_l}{\sum_{l=1}^{n_0} \zeta^l \mathscr{C}(n_0, l; \sigma)(\theta/\sigma + k_{\setminus i} - 1)_l}, \\
\text{pr}(X_i = x_j^* \mid X_{-i}) &\propto \frac{(n_j - \sigma)}{\theta + n - 1}, \quad \text{for } j = 1, \ldots k_{\setminus i} - 1, \\
\text{pr}(X_i = k_{\setminus i} \mid X_{-i}) &\propto (1 - \zeta) \frac{\theta + (k_{\setminus i} - 1)\sigma}{\theta + n - 1} \frac{\sum_{l=1}^{n_0} \zeta^l \, \mathscr{C}(n_0, l; \sigma) \left(\theta/\sigma + k_{\setminus i}\right)_l}{\sum_{l=1}^{n_0} \zeta^i \, \mathscr{C}(n_0, l; \sigma) \left(\theta/\sigma + k_{\setminus i} - 1\right)_l}.
\end{aligned}$$

B.2 Details on the inner spike and slab location-scale mixture

We now focus attention on the specific examples developed in §4 of the manuscript. The simulated scalar data of § 4.1 are generated from the following location-scale mixture of Gaussian

$$0.4\phi(0, 0.2) + 0.1\phi(-3.5, 1) + 0.1\phi(3.5, 1) + 0.2\phi(1, 0.8) + 0.2\phi(-1, 0.8).$$

In the first simulation experiment, data are analyzed assuming the inner model with the base measure

$$P_0 = \zeta \delta_{(m_1, t_1)} + (1 - \zeta) P^*,$$

where P^* is a prior over $\mathbb{R} \times \mathbb{R}^*$, namely

$$P^*(\mathrm{d}\mu, \mathrm{d}\tau) = \phi(\mu; \mu_0, \kappa\tau^{-1}) \times \text{Ga}(\tau; a, b) \, \mathrm{d}\mu \, \mathrm{d}\tau, \tag{23}$$

where $\mu_0 = 0$, $a = 0.5$, $b = 2$, and κ is set equal to the sample variance of the data. Note that the latter is parametrized in terms of precision $\tau = 1/\sigma^2$. The prior on ζ is uniform, $\zeta \sim U(0, 1)$. The analysis is repeated with different choices of σ and θ reported in Table 4 obtained using equation (11).

B.3 Details on the Gibbs sampler for the inner spike and slab location-scale mixture

Given the above prior specification, in order to perform Markov chain Monte Carlo sampling from the posterior distribution of the parameters, we use the Gibbs sampler composed by the following steps.

$E(K_n)$	σ	θ	
		$n = 50$	$n = 100$
3	0	0.72	0.60
	0.25	0.13	0.03
	0.5	-0.35	-0.40
	0.75	-0.71	-0.73
15	0	16.43	9.27
	0.25	10.25	4.89
	0.5	4.63	1.43
	0.75	0.39	-0.44

Table 4: *Prior parameters for the simulation experiment.*

1. Let $S_1, \ldots, S_n$ be the current cluster allocation, with $S_j = 0$ if X_j is allocated to the cluster of the spike. For $i = 1, \ldots, n$ let $k_{\setminus i}$ be the number of distinct values of S_j labeled from 0 to $k_{\setminus i} - 1$ and n_h is the number of observations belonging to cluster h. Then allocate the i-th observation to the cluster of the spike, if already occupied, with probability proportional to

$$\text{pr}(S_i = 0 \mid -) \propto \frac{1}{\theta + n - 1} \frac{\sum_{l=1}^{n_0+1} \zeta^l \mathscr{C}(n_0 + 1, l; \sigma)(\theta/\sigma + k_{\setminus i} - 1)_l}{\sum_{l=1}^{n_0} \zeta^l \mathscr{C}(n_0, l; \sigma)(\theta/\sigma + k_{\setminus i} - 1)_l} \phi(X_i; \mu_0, \tau_0^{-1}),$$

to one of the existing clusters, different from the spike, with probability proportional to

$$\text{pr}(S_i = h \mid -) \propto \frac{(n_h - \sigma)}{\theta + n - 1} \phi(X_i; \mu_h^*, \tau_h^{*-1}), \quad \text{for } h = 1, \ldots k_{\setminus i} - 1$$

and finally to a new cluster with probability proportional to

$$\text{pr}(S_i = k_{\setminus i} \mid -) \propto (1 - \zeta) \frac{\theta + (k_{\setminus i} - 1)\sigma}{\theta + n - 1} \frac{\sum_{l=1}^{n_0} \zeta^l \, \mathscr{C}(n_0, l; \sigma) \left(\theta/\sigma + k_{\setminus i}\right)_l}{\sum_{l=1}^{n_0} \zeta^i \, \mathscr{C}(n_0, l; \sigma) \left(\theta/\sigma + k_{\setminus i} - 1\right)_l} \times \phi(X_i; \mu_*, \tau_*^{-1}),$$

where (μ_*, τ_*) are new drawn from P^*.

2. Update (μ_h^*, τ_h^*) from its conditional posterior

$$(\mu_h^*, \tau_h^*) \sim N(\hat{\mu}_h, \hat{\kappa}_h \tau_h^{*-1}) \text{Ga}(\hat{a}_{\tau_h}, \hat{b}_{\tau_h})$$

with

$$\begin{aligned}
\hat{\kappa}_h &= (\kappa^{-1} + n_h)^{-1}, \\
\hat{\mu}_h &= \hat{\kappa}_h(\kappa^{-1}\mu_0 + n_h \bar{y}_h), \\
\hat{a}_{\tau_h} &= a_\tau + \frac{n_h}{2}, \\
\hat{b}_{\tau_h} &= b_\tau + \frac{1}{2}\left(\sum_{i:S_i=h} (X_i - \bar{X}_h)^2 + \frac{n_h}{1+\kappa n_h}(\bar{X}_h - \mu_0)^2 \right).
\end{aligned}$$

3. Update $\zeta \sim \text{Beta}(1 + n_0, 1 + n - n_0)$.

B.4 Details on the outer spike and slab location-scale mixture

In the second simulation experiment, we compare the inner and outer models. For the latter the mixing distribution is defined as

$$\tilde{Q} = \zeta\, \delta_{(m_1,t_1)} + (1-\zeta)\, \tilde{Q}^*,$$

where $\tilde{Q}^* \sim \text{PY}(\sigma, \theta; P^*)$ and P^* is equal to (23). The analysis is carried out for different choices of σ and θ as reported in Table 5. The specific values are set so to have the prior expected number of components equal to 5 and are determined by using (11) for the inner model and the following result for the outer model.

Proposition 2 *Let K_n be the number of distinct values in an exchangeable sampe $X^{(n)}$ from the outer spike and slab model (6). Then*

$$E(K_n) = 1 - (1-\zeta)^n - \frac{\theta}{\sigma} + \frac{\theta}{\sigma}\frac{(\theta+\sigma)_n}{\theta_n}\, {}_2F_1(-n, -\sigma; 1-n-\theta-\sigma; \zeta),$$

where ${}_2F_1$ denotes the Gaussian hypergeometric function.

Proof. Denote by n_0 the number of observations in $X^{(n)}$ that coincide with the atom x_0. Then we have

$$E\left(K_n\right) = \sum_{j=0}^{n} \binom{n}{j} \zeta^j (1-\zeta)^{n-j} \, E\left(K_n \mid n_0 = j\right).$$

If K_n' is the number of distinct values in a sample of size n from an exchangeable sequence governed by $\tilde{Q}^*$, one has

$$\begin{aligned} E\left(K_n \mid n_0 = j\right) &= \{1 - \delta_0(\{j\})\} + E(K'_{n-n_0} \mid n_0 = j) \\ &= \{1 - \delta_0(\{j\})\} + \frac{\theta}{\sigma} \left\{ \frac{(\theta+\sigma)_{n-j}}{\theta_{n-j}} - 1 \right\}. \end{aligned}$$

for any $j = 0, 1, \ldots, n$. Thus we have

$$\begin{aligned} E\left(K_n\right) &= \sum_{j=1}^{n} \binom{n}{j} \zeta^j (1-\zeta)^{n-j} + \frac{\theta}{\sigma} \sum_{j=0}^{n} \binom{n}{j} \zeta^j (1-\zeta)^{n-j} \left\{ \frac{(\theta+\sigma)_{n-j}}{\theta_{n-j}} - 1 \right\} \\ &= 1 - (1-\zeta)^n - \frac{\theta}{\sigma} + \frac{\theta}{\sigma} \sum_{j=0}^{n} \binom{n}{j} \zeta^j (1-\zeta)^{n-j} \frac{(\theta+\sigma)_{n-j}}{\theta_{n-j}} \\ &= 1 - (1-\zeta)^n - \frac{\theta}{\sigma} + \frac{\theta}{\sigma} \frac{(\theta+\sigma)_n}{\theta_n} \, {}_2F_1\left(-n, -\sigma; 1-n-\theta-\sigma; \zeta\right). \end{aligned}$$

□

B.5 Details on the Gibbs sampler for the outer spike and slab location-scale mixture

Given the above prior specification, to perform Markov chain Monte Carlo sampling from the posterior distribution of the parameters under the outer spike and slab location-scale mixture model, we use a Gibbs sampler composed by the following steps.

1. Let $S_1, \ldots, S_n$ be the current cluster allocation, with $S_j = 0$ if X_j is allocated to the cluster of the spike. For $i = 1, \ldots, n$ let $k_{\setminus i}$ be the number of distinct values of S_j labeled from 0 to $k_{\setminus i} - 1$ and n_h is the number of observations belonging to cluster h.

Model	σ	θ	
		$n = 50$	$n = 100$
Inner	0	11.86	7.24
	0.25	7.11	3.66
	0.5	2.90	0.91
	0.75	-0.04	-0.52
Outer	0	2.03	1.22
	0.25	1.07	0.46
	0.5	0.19	-0.17
	0.75	-0.52	-0.66

Table 5: *Prior parameters for the simulation experiment assuming* $E(K_n) = 5$.

Then allocate the i-th observation to the cluster of the spike, if already occupied, with probability proportional to

$$\text{pr}(S_i = 0 \mid -) \propto \zeta \phi(X_i; \mu_0, \tau_0^{-1}),$$

to one of the existing clusters, different from the spike, with probability proportional to

$$\text{pr}(S_i = h \mid -) \propto (1-\zeta)\frac{n_h - \sigma}{\theta + n - n_0 - 1}\phi(X_i; \mu_h^*, \tau_h^{*-1}), \quad \text{for } h = 1, \dots k_{\setminus i} - 1$$

and finally to a new cluster with probability proportional to

$$\text{pr}(S_i = k_{\setminus i} \mid -) \propto (1-\zeta)\frac{\theta + (k_{\setminus i} - 1)\sigma}{\theta + n - n_0 - 1}\phi(X_i; \mu_*, \tau_*^{-1}),$$

where (μ_*, τ_*) are new drawn from P^*.

2. Update (μ_h^*, τ_h^*) from its conditional posterior

$$(\mu_h^*, \tau_h^*) \sim \text{N}(\hat{\mu_h}, \hat{\kappa}_h \tau_h^{*-1})\text{Ga}(\hat{a}_{\tau_h}, \hat{b}_{\tau_h})$$

with

- $\hat{\kappa}_h = (\kappa^{-1} + n_h)^{-1}$,
- $\hat{\mu}_h = \hat{\kappa}_h(\kappa^{-1}\mu_0 + n_h\bar{y}_h)$,
- $\hat{a}_{\tau_h} = a_\tau + n_h/2$,
- $\hat{b}_{\tau_h} = b_\tau + 1/2(\sum_{i:S_i=h}(X_i - \bar{X}_h)^2 + n_h/(1+\kappa n_h)(\bar{X}_h - \mu_0)^2)$.

B.6 Details on the functional data simulation

The functional data of § 4.1 are generated on an equi-spaced grid of $T = 25$ points adding independent random normal noises with fixed variance $\sigma^2 = 0.25$ to the random functional means sampled from

$$P = \sum_{j=1}^{5} \delta_{f_j^*} \pi_j,$$

where the five f_j^* are reported in Figure 4 and

$$\pi = (\pi_1, \ldots, \pi_5) = (0.4, 0.2, 0.2, 0.1, 0.1).$$

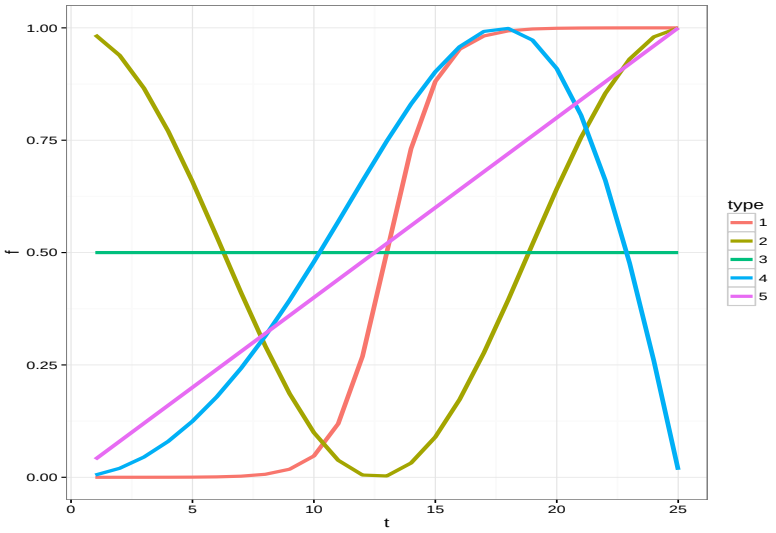


Figure 4: Functional means for the functional data simulation experiment.

Data are analysed assuming $P \sim \text{PY}(\sigma, \theta; P_0)$ with two different choices for P_0: (i) a mixture of a point mass on (12) and a diffuse measure over the space of functions, (ii) a diffuse non-atomic base measure. In both cases the diffuse measure is induced by a B-spline basis expansion, namely

$$f(t) = B(t)^T \beta, \quad \beta \sim MVN(\beta_0, \Sigma_0),$$

where $B(.)$ denotes the B-splines basis and β a finite vector of basis coefficients. We specify the B-splines basis assuming a fixed set of knots at 2, 5, 9, 13, 17, 21, 24. For simplicity, Σ_0 is an identity matrix and β_0 is a vector of zeroes. In both cases $\sigma = 0.5$ while $\theta = 1$ and $\theta = 0.178$ for first and second prior, respectively.

B.7 Details on the Gibbs sampler for functional data simulation

Given the above prior specification, the Gibbs sampler is composed by the following steps.

1. Let $S_1, \ldots, S_n$ be the current cluster allocation, with $S_j = 0$ if the corresponding observation is allocated to the cluster of the spike. For $i = 1, \ldots, n$ let $k_{\setminus i}$ be the number of distinct values of S_j labeled from 0 to $k_{\setminus i} - 1$ and n_h is the number of observations belonging to cluster h. Then allocate the i-th observation to the cluster of the spike, if already occupied, with probability proportional to

$$\text{pr}(S_i = 0 \mid -) \propto \frac{1}{\theta + n - 1} \frac{\sum_{l=1}^{n_0+1} \zeta^l \mathscr{C}(n_0 + 1, l; \sigma)(\theta/\sigma + k_{\setminus i} - 1)_l}{\sum_{l=1}^{n_0} \zeta^l \mathscr{C}(n_0, l; \sigma)(\theta/\sigma + k_{\setminus i} - 1)_l} \prod_{t=1}^{T} \phi(X_{it}; f_0(t), \sigma^2),$$

   to one of the existing clusters, different from the spike, with probability proportional to

$$\text{pr}(S_i = h \mid -) \propto \frac{n_h - \sigma}{\theta + n - 1} \prod_{t=1}^{T} \phi(X_{it}; f_h^*(t), \sigma^2), \quad \text{for } h = 1, \ldots k_{\setminus i} - 1$$

   and finally to a new cluster with probability proportional to

$$\text{pr}(S_i = k_{\setminus i} \mid -) \propto (1 - \zeta) \frac{\theta + (k_{\setminus i} - 1)\sigma}{\theta + n - 1} \frac{\sum_{l=1}^{n_0} \zeta^l \mathscr{C}(n_0, l; \sigma) \left(\theta/\sigma + k_{\setminus i}\right)_l}{\sum_{l=1}^{n_0} \zeta^l \mathscr{C}(n_0, l; \sigma) \left(\theta/\sigma + k_{\setminus i} - 1\right)_l} \times \prod_{t=1}^{T} \phi(X_{it}; f_*(t), \sigma^2),$$

where f_* is a new draw from the base measure.

2. Update the cluster baseline functions from the multivariate normal with covariance matrix and mean

$$V_{\beta_h} = \left(\Sigma_0^{-1} + \frac{n_h}{\sigma^2} B^T B\right)^{-1} \qquad m_{\beta_h} = V_{\beta_h} \left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2} \sum_{S_i=h} B^T X_i\right).$$

3. Update σ^2 form the conjugate inverse-gamma distribution

$$1/\sigma^2 \sim \text{Ga}\left(a + \frac{nT}{2}, b + \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_i(t) - f_i(t))^2\right).$$

4. Update $\zeta \sim \text{Beta}(1 + n_0, 1 + n - n_0)$.

B.8 Computational details on § 4.2

The Gibbs sampler used in § 4.2 is composed by the following steps.

1. For each cycle i of woman j, conditionally on X_{ij} and on λ_{ij}, ω_{ij}, and n_{ij}, the model can be written as simple linear model, that is

$$X_{ij}(t) = Z_{ij}\theta + \epsilon_{ij}(t)$$

where

$$Z_{ij} = \begin{pmatrix} 1 & z_{ij}(1) \\ 1 & z_{ij}(2) \\ \vdots & \vdots \\ 1 & z_{ij}(n_{ij}) \end{pmatrix}, \qquad z_{ij}(t) = f_{ij}\left(\frac{t - \lambda_{ij}}{\omega_{ij}}\right),$$

meaning that it can be seen as a standard linear regression for each pair (i, j). Hence the full conditional distribution for τ_{1ij} and τ_{2ij} is $(\tau_{1ij}, \tau_{2ij})^T \sim N(a_1, V_1)$, where

$$V_1 = (\Omega^{-1} + \sigma^{-2} Z_{ij}^T Z_{ij})^{-1} \qquad a_1 = V_1(\Omega^{-1}\alpha_i + \sigma^{-2} Z_{ij}^T X_{ij}).$$

2. For each cycle i of woman j, conditionally on X_{ij} and on τ_{1ij} e τ_{2ij}, the model can be written as

$$X_{ij}(t) = \tilde{f}_{ij}\left(\frac{t-\lambda_{ij}}{\omega_{ij}}\right) + \epsilon_{ij}(t)$$

where $\tilde{f}_{ij} = \tau_{1ij} + \tau_{2ij} f_{ij}$. We then proceed with the following two steps.

   - Update the value of λ_{ij} using direct sampling from the posterior. Given the uniform prior and that the days are discrete, the full conditional posterior is simply a multinomial with probabilities proportional to the likelihood function.

   - Update ω_{ij} via Metropolis–Hastings sampling.

3. For each $i = 1, \ldots, n$, sample the woman specific mean $\alpha_i \sim N(a_2, V_2)$, where

$$V_2 = (R + n_i\Omega^{-1})^{-1}, \qquad a_2 = V_2(R\alpha + \Omega^{-1}\sum_{j=1}^{n_i}(\tau_{1ij}, \tau_{2ij})^T)$$

   and n_i is the total number of cycles for woman i.

4. Update the cluster allocation via Pólya urn sampling. Specifically let $S_1, \ldots, S_n$ be the current cluster allocation, with $S_j = 0$ if the corresponding observation is allocated to the cluster of the spike. For $i = 1, \ldots, n$ let $k_{\setminus i}$ be the number of distinct values of S_j labeled from 0 to $k_{\setminus i} - 1$ and n_h is the number of observations belonging to cluster h. Then allocate the i-th observation to the cluster of the spike, if already occupied, with probability proportional to

$$\text{pr}(S_{ij} = 0 \mid -) \propto \frac{1}{\theta + n - 1} \frac{\sum_{l=1}^{n_0+1} \zeta^l \mathscr{C}(n_0+1, l; \sigma)(\theta/\sigma + k_{\setminus i} - 1)_l}{\sum_{l=1}^{n_0} \zeta^l \mathscr{C}(n_0, l; \sigma)(\theta/\sigma + k_{\setminus i} - 1)_l} \prod_{t=1}^{T} \phi(X_{it}; f_0(t), \sigma^2),$$

   to one of the existing clusters, different from the spike, with probability proportional to

$$\text{pr}(S_{ij} = h \mid -) \propto \frac{n_h - \sigma}{\theta + n - 1} \prod_{t=1}^{T} \phi(X_{it}; f_h^*(t), \sigma^2), \quad \text{for } h = 1, \ldots k_{\setminus i} - 1,$$

   and finally to a new cluster with probability proportional to

$$\text{pr}(S_{ij} = k_{\setminus i} \mid -) \propto (1-\zeta)\frac{\theta + (k_{\setminus i} - 1)\sigma}{\theta + n - 1} \frac{\sum_{l=1}^{n_0} \zeta^l \mathscr{C}(n_0, l; \sigma) \left(\theta/\sigma + k_{\setminus i}\right)_l}{\sum_{l=1}^{n_0} \zeta^l \mathscr{C}(n_0, l; \sigma) \left(\theta/\sigma + k_{\setminus i} - 1\right)_l} \times \prod_{t=1}^{T} \phi(X_{it}; f_*(t), \sigma^2),$$

where f_* is a new draw from the base measure.

5. Update the cluster baseline functions f_h^* for $h = 1, \ldots, k-1$ from the multivariate normal with covariance matrix and mean

$$V_{\beta_h} = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{S_{ij}=h} \tau_{2ij}^2 B_{ij}^T B_{ij} \right)^{-1},$$

$$m_{\beta_h} = V_{\beta_h} \left(\Sigma_0^{-1} \beta_0 + \frac{1}{\sigma^2} \sum_{S_{ij}=h} B_{ij}^T (X_{ij} - \tau_{1ij}) \right),$$

where $B_{ij} = B((t - \lambda_{ij})/\omega_{ij})$.

6. Update σ^2 form the conjugate inverse-gamma distribution

$$1/\sigma^2 \sim \text{Ga}\left(\frac{1}{2} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} n_{ij}, \frac{1}{2} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{t=1}^{n_{ij}} (X_{ij}(t) - f_{ij}(t))^2 \right).$$

7. Update $\zeta \sim \text{Beta}(1 + n_0, 1 + n - n_0)$.

Acknowledgements

A. Lijoi and I. Prünster are supported by the European Research Council (ERC) through StG "N-BNP" 306406. A. Canale is supported by grant CPDA154381/15 from the University of Padova, Italy.

References

Barcella, W., De Iorio, M., Baio, G. & Malone-Lee, J. (2016). Variable selection in covariate dependent random partition models: an application to urinary tract infection. *Stat. Med.* **35**, 1373–1389.

Bogdan, M., Ghosh, J. K. & Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, vol. 1 of *Inst. Math. Stat. Collect.* Inst. Math. Statist., Beachwood, OH, pp. 211–230.

Charalambides, C. A. (2005). *Combinatorial methods in discrete distributions*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Colombo, B. & Masarotto, G. (2000). Daily fecundability: first results from a new data base. *Demographic research* **3**, N. 5.

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. Pattern Anal. Mach. Intell.* **37**, 212–229.

Do, K.-A., Müller, P. & Tang, F. (2005). A Bayesian mixture model for differential gene expression. *J. Roy. Statist. Soc. Ser. C* **54**, 627–644.

Dunson, D. B., Herring, A. H. & Engel, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *J. Amer. Statist. Assoc.* **103**, 534–546.

George, E. I. & McCulloch, R. E. (1993). Bayesian variable selection via gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

Guindani, M., Müller, P. & Zhang, S. (2009). A Bayesian discovery procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**, 905–925.

Ishwaran, H. & Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Statist.* **33**, 730–773.

James, L. F., Lijoi, A. & Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.* **33**, 105–120.

James, L. F., Lijoi, A. & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36**, 76–97.

Jara, A., Lesaffre, E., De Iorio, M., Quintana, F. et al. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics* **4**, 2126–2149.

Kim, S., Dahl, D. B. & Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Anal.* **4**, 707–732.

Kingman, J. F. C. (1975). Random discrete distribution. *J. Roy. Statist. Soc. Ser. B* **37**, 1–22.

Lijoi, A., Mena, R. H. & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.* **100**, 1278–1291.

Lijoi, A., Mena, R. H. & Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 715–740.

Lijoi, A. & Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian nonparametrics*, Camb. Ser. Stat. Probab. Math. Cambridge Univ. Press, Cambridge, pp. 80–136.

MacLehose, R. F., Dunson, D. B., Herring, A. H. & Hoppin, J. A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology* **18**, 199–207.

Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83**, 1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors.

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

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102**, 145–158.

Pitman, J. (2003). Poisson-Kingman partitions. In *Statistics and science: a Festschrift for Terry Speed*, vol. 40 of *IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Beachwood, OH, pp. 1–34.

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

Regazzini, E. (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilita. *Giornale dell'Istituto Italiano degli Attuari* **41**, 77–89.

Sangalli, L. M. (2006). Some developments of the normalized random measures with independent increments. *Sankhyā* **68**, 461–487.

Scarpa, B. & Dunson, D. B. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics* **65**, 772–780.

Suarez, A. J. & Ghosal, S. (2016). Bayesian clustering of functional data using local features. *Bayesian Anal.* **11**, 71–98.

Yang, M. (2012). Bayesian variable selection for logistic mixed model with nonparametric random effects. *Comput. Statist. Data Anal.* **56**, 2663–2674.