Biometrika (2017), **104**, 3, *pp.* 681–697
Printed in Great Britain

doi: 10.1093/biomet/asx041
Advance Access publication 3 August 2017

On the Pitman–Yor process with spike and slab base measure

By A. CANALE

Department of Statistical Sciences, University of Padua, Via C. Battisti 241, 35121 Padua, Italy
canale@stat.unipd.it

A. LIJOI

Department of Decision Sciences, Bocconi University, via Röntgen 1, 20136 Milan, Italy
lijoi@unibocconi.it

B. NIPOTI

School of Computer Science and Statistics, Trinity College, College Green, Dublin 2, Ireland
nipotib@tcd.ie

and I. PRÜNSTER

Department of Decision Sciences, Bocconi University, via Röntgen 1, 20136 Milan, Italy
igor@unibocconi.it

Summary

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 the 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 the incorporation of an atom into the base measure, and then the Dirichlet process is essentially the only tractable choice for the prior. In this paper we fill this gap by considering the Pitman–Yor process with 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.

Some key words: Bayesian nonparametric inference; Functional data; Pitman–Yor process; Predictive distribution; Random partition; Spike and slab base measure.

1. Introduction

Two-component mixture priors represent the most popular choice in Bayesian variable selection and when investigating sparsity phenomena. Such mixtures are commonly referred to as spike and slab priors, a terminology that was introduced in Mitchell & Beauchamp (1988), who use a mixture whose components are a degenerate distribution at zero, referred to as a spike, and a diffuse distribution, referred to as a 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 amount of literature. 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 curves of women. The daily basal body temperature 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. However, unhealthy women may display a far more irregular functional form that does not preserve the 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 irregular basal body temperature 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 a diffuse distribution P^* on a suitable set of functions, i.e., $E(\tilde{P}) = \zeta\,\delta_{f^*} + (1-\zeta)\,P^*$. Introducing an atom in the base measure, corresponding to the regular S-shape, 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 female body temperature curves. Motivated by different applications, with real-valued data, Dunson et al. (2008), MacLehose et al. (2007), Yang (2012) and Barcella et al. (2016) adopted a Dirichlet process with base measure featuring an atom at zero: this allows them to simultaneously perform clustering and variable selection. In fact, an atom at zero 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 as 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 Dirichlet process, whose base measure is a two-component mixture, we mention Guindani et al. (2009) and Do et al. (2005). When using the Dirichlet process 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 Dirichlet process regardless of the presence of atoms in the base measure P_0. However, when $\tilde{P}$ is not a Dirichlet process, 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 perhaps the most popular generalization of the Dirichlet process, 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 considerable analytical tractability. We derive explicit expressions for the associated exchangeable partition probability function, 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 blocks of a generalized Blackwell–McQueen–Pólya urn scheme. The resulting algorithm is then used to perform an extensive study involving both scalar and functional data. This empirical analysis uncovers some interesting features of the model and allows useful comparisons with alternatives. First we assess the different inferential behaviour of Dirichlet and Pitman–Yor process-based models, when the base measure has an atomic component. Our findings show that, somewhat similar 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 process are more flexible and more robust with respect to prior misspecification of 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 and 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, has already been pointed out in existing literature on the Dirichlet process for the case of scalar data. Here, instead, we consider functional data in the more general Pitman–Yor set-up 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 with $\tilde{P}$. Assume that the data $X_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$, for $i = 1, \ldots, n$, take values in some space $\mathbb{X}$, where

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

is a discrete random probability measure such that $\sum_{j \geqslant 1} \tilde{p}_j = 1$ almost surely, with the Z_j being independent and identically distributed $\mathbb{X}$-valued random elements having 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 $k \leqslant n$, is also known as the exchangeable partition probability function. See Pitman (1995). Once the exchangeable partition probability function is available, one can determine the predictive distributions associated with the exchangeable sequence $(X_i)_{i \geqslant 1}$. If the sample $X^{(n)}$ contains k distinct values $x_1^*, \ldots, x_k^*$, then

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

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 exchangeable partition probability function. Closed-form expressions for predictive distributions in (3) are available for broad classes of discrete random probability measures under the crucial assumption that P_0 is diffuse (Pitman, 2003; Lijoi et al., 2005, 2007; James et al., 2009).

Beyond the Dirichlet process, the literature on instances where the assumption of diffuseness of P_0 can be relaxed is essentially confined to theoretical investigations with no actual implementation. James et al. (2006) consider homogeneous normalized random measures and study the predictive distribution for grouped data when the base measure 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 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 $(\tilde p_j)_{j\geqslant 1}$ and $(Z_j)_{j\geqslant 1}$ 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 $\tilde p_i$ 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 $j\geqslant 2$, and $(V_i)_{i\geqslant 1}$ is a sequence of independent random variables with $V_i\sim \mathrm{Be}(1-\sigma,\theta+i\sigma)$. For a diffuse P_0 the corresponding exchangeable partition probability function 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 $n\geqslant 0$ and $n_j=\mathrm{card}(C_j)$ 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) is no longer true.

Before stating the main results in the next section, we highlight a key difference between $\tilde P\sim\mathrm{PY}(\sigma,\theta;P_0)$ 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 $\tilde Q^*\sim\mathrm{PY}(\sigma,\theta;P^*)$ and P^* is diffuse as in (5). Henceforth, we shall refer to $\tilde P\sim\mathrm{PY}(\sigma,\theta;P_0)$ with P_0 as in (5) as the inner spike and slab model. Similarly, $\tilde Q$ as in (6) will be referred to as the outer spike and slab model. The model with an outer spike and slab (6) has been used in Scarpa & Dunson (2009) when $\tilde Q^*$ is a Dirichlet process. Both processes share the same two-component mixture centring, 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 $\zeta\sim\mathrm{Be}(1,\theta)$, $\tilde Q$ can be represented as $\sum_{j\geqslant 0}\pi_j\,\delta_{Y_j}$ with the random probability masses π_j admitting the same stick-breaking representation characterizing the weights of a Dirichlet process. Nonetheless, $\tilde Q$ is not a Dirichlet process since the location associated with the first stick-breaking weight, namely Y_0, equals x_0 and therefore the random variables of the sequence $(Y_j)_{j\geqslant 0}$ are not independent and identically distributed. The substantial difference between the two models can be further appreciated and in some sense quantified through Proposition 1, 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*

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

Hence, the prior uncertainty associated with 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 that used in Scarpa & Dunson (2009). A simple illustration of this may be obtained by choosing f to be an indicator function. If $f = \mathbb{1}_{[0,t]}$ for some $t > 0$, one obtains the random survival functions

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

defined as functionals of the inner and 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 let 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 $t \geqslant 0$,

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

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 that 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 exchangeable partition probability function, denoted by $\Pi_k^{(n)}(n_1, \ldots, n_k)$. The expression is given in terms of generalized factorial coefficients $\mathscr{C}(n_j, i; \sigma) = (i!)^{-1}\sum_{r=0}^{i}(-1)^r\, i!/\{r!(i-r)!\}\,(-r\sigma)_{n_j}$; see Charalambides (2005). In this section we assume that $\sigma \in (0,1)$; the Dirichlet case is obtained by taking the limit $\sigma \to 0$.

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

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

A simple rearrangement of (8) yields a nice probabilistic interpretation of the result. Recall that the posterior distribution of a $\mathrm{PY}(\sigma, \theta; P_0)$, 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

$$\sum_{i=1}^{k-1}\pi_{i,j}\,\delta_{x_i^*}+\pi_{k,j}\tilde P_{k-1} \tag{9}$$

with $\tilde P_{k-1}\sim \mathrm{PY}\{\sigma,\theta+(k-1)\sigma;P_0\}$, $(\pi_{1,j},\ldots,\pi_{k-1,j})$ 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. The distribution of the number of distinct values $K_n\in\{1,\ldots,n\}$ in a sample of size n from a $\mathrm{PY}(\sigma,\theta;P_0)$ process depends on the parameters (σ,θ,ζ), so we will use the notation $\mathrm{pr}\{K_n=k;\ (\sigma,\theta,\zeta)\}$ for it. In particular, $\zeta=0$ corresponds to a diffuse base measure leading to

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

Simple algebra leads to the following result.

Corollary 1. *The exchangeable partition probability function of $\tilde P\sim \mathrm{PY}(\sigma,\theta;P_0)$ with P_0 as in (5) can be represented as*

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

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 is as follows. For any $j=1,\ldots,k$: (i) with probability $\Phi_{k-1}^{(n-n_j)}(n_1,\ldots,n_{j-1},n_{j+1},\ldots,n_k;\sigma,\theta)$ 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 labelled groups. In other terms, according to (iii), it is as if the n_j observations are further split into i fictitious subclusters all identified by x_0.

Having derived a closed-form expression for the exchangeable partition probability function, we can 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 $X_i \mid \tilde P \overset{\text{iid}}{\sim} \tilde P$ for $i = 1,\ldots,n$ and $\tilde P \sim \mathrm{PY}(\sigma,\theta;P_0)$ with P_0 as in (5), the probability distribution of the number of distinct values K_n in $X^{(n)}$ equals*

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

The predictive distributions associated with the exchangeable sequence $(X_i)_{i\geqslant 1}$ 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 exchangeable partition probability function. 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 $X_i \mid \tilde P \overset{\text{iid}}{\sim} \tilde P$ for $i = 1,\ldots,n$ and $\tilde P \sim \mathrm{PY}(\sigma,\theta;P_0)$ with P_0 as in (5). The corresponding predictive distribution is*

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

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

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

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

The predictive distribution for case (i) coincides with that 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 exchangeable partition probability function in Theorem 1 reduces to that of the nonatomic 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 Supplementary Material.

4. Illustration

4·1. *Synthetic data*

The previous results pave the way to a straightforward implementation of the inner spike and slab nonparametric model. For $\tilde P \sim \mathrm{PY}(\sigma,\theta;P_0)$ with P_0 a diffuse probability measure, there is

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

$E(K_n)$	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

extensive literature on the effects of σ on posterior inferences for the Pitman–Yor process and allied nonparametric priors (Lijoi et al., 2007; Jara et al., 2010; De Blasi et al., 2015). Here, we aim to understand 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 $n = 50, 100$ to mimic a quality control application. In this 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 $X_1, \ldots, X_n$ 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 $\phi(\cdot; m, t^{-1})$ is the normal density with mean m and precision t. We further set $m_1 = 0$ and $t_1^{-1} = 0 \cdot 04$ as the nominal value of X and the variance of the measuring instrument, respectively. Data are analysed 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 assigns positive mass to the pair (m_1, t_1) and thus is a mixture of a point mass and a diffuse density,

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

where P^* is normal-gamma. In order to also learn the proportion of observations that can be suitably modelled by the spike at (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 $\sigma \in \{0, 0{\cdot}25, 0{\cdot}5, 0{\cdot}75\}$ and we fix θ, using the results of Theorem 2, to have a prior expected number of mixture components equal to 3 or 15 thus corresponding, respectively, to underestimation and overestimation 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 Supplementary Material. The results, displayed in Table 1, are consistent with the findings in the case of nonparametric mixtures with nonatomic

Table 2. *Posterior proportion of subjects allocated to the spike for the location-scale mixture simulation experiment. The largest Monte Carlo standard error is* 0·08

	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

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 conferred 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 exceeds the variance of the corresponding mass assigned by the outer model by $\zeta(1-\zeta)(1-\sigma)/(\theta+1)$. 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 fixed and its value misspecified. In order to check this we consider the same simulated data as in the first experiment and keep $x_0 = (m_1, t_1)$. For both inner and outer models, we fix $\zeta = 0{\cdot}8$ 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 $\sigma \in \{0, 0{\cdot}25, 0{\cdot}5, 0{\cdot}75\}$ we set the parameter θ in both models so that the prior expected number of components is equal to five, the true number of components in our simulations. Details on how we set θ are reported in the Supplementary Material. The results displayed in Table 2 show that the inner spike and slab model is clearly better than the outer model in overcoming prior misspecifications.

A third simulation experiment aims to highlight the benefit of including the spike in the base measure when there is supporting prior information. 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 basal body temperature functional data of our motivating application. The daily basal body temperature 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,

$$f(t) = \frac{\mathrm{e}^t}{1+\mathrm{e}^t}. \tag{11}$$

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

$$X_{it} \mid f_i \overset{\text{ind}}{\sim} N\{f_i(t), \sigma^2\}, \quad f_i \overset{\text{iid}}{\sim} P, \quad P = \sum_{j=1}^{5} \delta_{f_j^*}\pi_j, \tag{12}$$

Table 3. *Confusion matrix for the third simulation experiment. The largest Monte Carlo error for the global accuracy is equal to* 0·07

	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

where f_1^* is (11) and $\pi_1 = 0{\cdot}4$. The remaining curves f_j^*, for $j = 2, \ldots, 5$, and values of the parameters are reported in the Supplementary Material. Data are analysed 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 (11) and a diffuse measure over the space of functions, or a plain nonatomic measure. For both choices of P_0 we set $\sigma = 0{\cdot}5$ and fix θ so as to have the same prior expected number of mixture components. Additional details on the prior specification and posterior computation are reported in the Supplementary Material.

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 from (11) are clustered. The numerical results are reported in Table 3. The better performance provides clear evidence in favour 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 (11). Figure 1 displays the estimated functional mean and 90% pointwise posterior credible bands for two subjects having true mean (11). The functional mean and limits of the credible bands are estimated with the empirical mean and the 0·05 and 0·95 quantiles from 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.

Figure 1(a) concerns a subject classified in the true cluster with the spike and slab model for 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 cannot coincide with (11) since it is estimated from the data, and it is worth noting that this estimate is erratic on the left and right parts of the domain. Figure 1(b) concerns a borderline subject classified as biphasic in 85% of the Markov chain Monte Carlo iterations for the spike and slab model and in only 60% of the iterations for the nonatomic model. This leads to wider credible bands in both cases.

4·2. *Basal body temperature functional data*

We study a dataset on daily measurements of basal body temperature, consisting of 1118 non-conception cycles from $n = 157$ women in the Verona centre of the Colombo & Masarotto (2000) study. As shown in Fig. 2(a), the basal body temperature curve trajectory over time of healthy

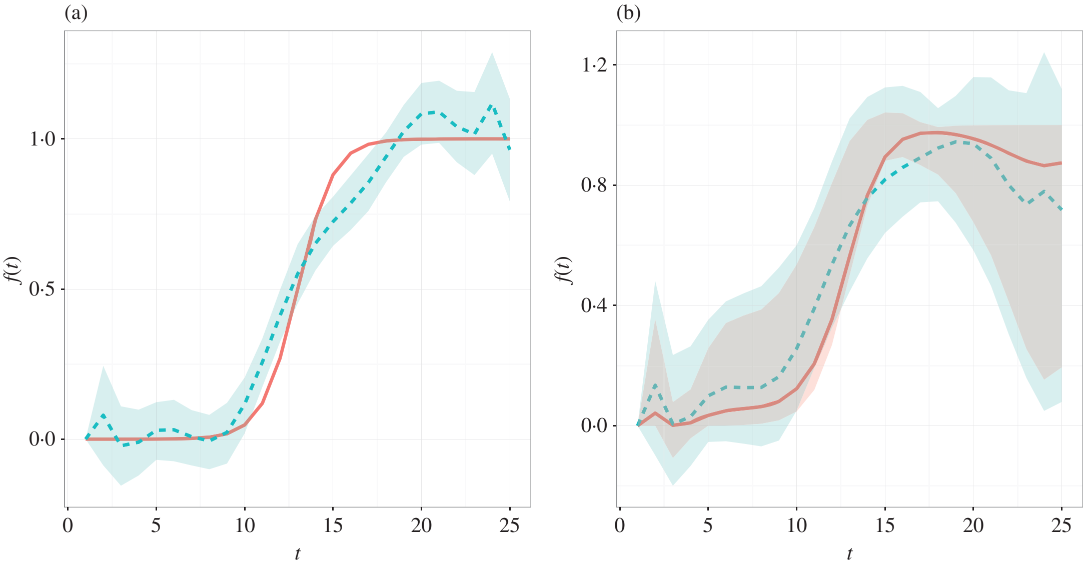


Fig. 1. Posterior functional means, conditional on simulated data from (12), for two observations having true mean equal to (11): solid lines correspond to the model with spike and slab base measure, and dashed lines to the model with nonatomic base measure. Shaded areas depict the posterior pointwise 90% credible bands. Panel (a) corresponds to a subject clearly belonging to (11), panel (b) to a borderline case.

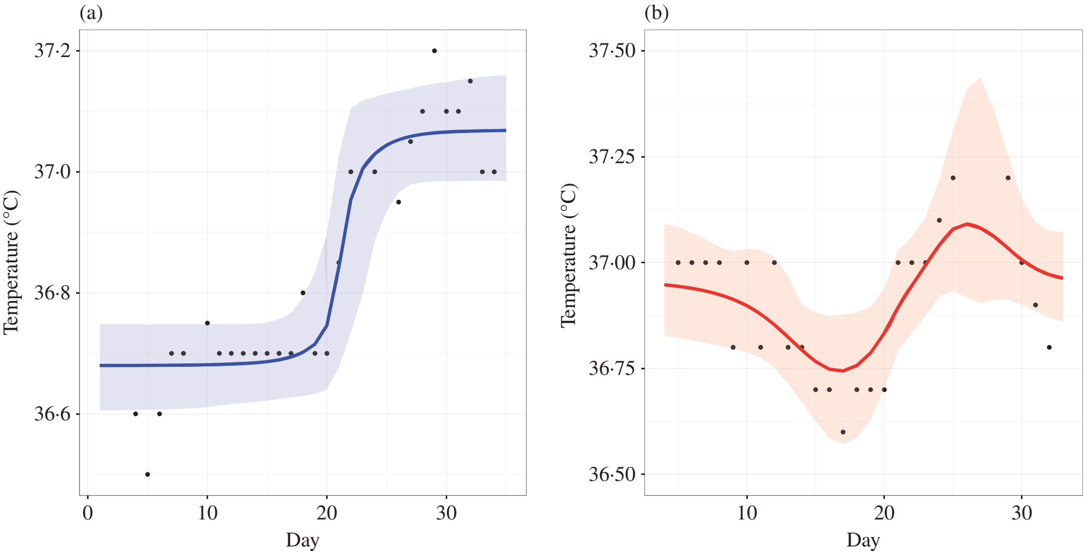


Fig. 2. Panels (a) and (b) show the basal body temperature data for two cycles along with pointwise posterior means and 95% credible bands.

women of reproductive age follows a biphasic trajectory that can be described by (11) or, more generally, by the following parametric function of time $t \geqslant 0$:

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

where $\tau_1, \tau_2, \lambda \geqslant 0$ and $\omega > 0$. 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 basal body temperature reaches after the sharp increase, controlled by ω, which happens just before the ovulation. In contrast, unhealthy women tend to have different curve shapes as shown, for example, in Fig. 2(b).

The nonparametric model with spike and slab base measure perfectly fits the present set-up: it allows one to assign prior positive mass to curves with the typical healthy women's shape, and also to account for abnormal deviations from this standard shape via an extremely flexible nonparametric functional data mixture model. The same dataset was analysed by Scarpa & Dunson (2009), with similar goals by means of the outer model.

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

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

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

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

where P_0 has a spike and slab structure of the type

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

with $f_0(t) = \mathrm{e}^t/(1 + \mathrm{e}^t)$ representing the biphasic curve and P^* being a nonatomic probability measure on a function space. The almost sure discreteness of the Pitman–Yor process induces ties among the f_{ij}, with positive probability. We denote these atoms by f_h^* for $h = 1, \ldots, k$.

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

$$g \sim P^*, \quad g(t) = B(t)^{\mathrm{T}} \beta, \quad \beta \sim N_p(\beta_0, \Sigma_0),$$

with $B(\cdot)$ denoting the B-spline basis, $N_p(m, V)$ the multivariate normal distribution of suitable dimension p with mean vector m and variance matrix V, β a finite vector of basis coefficients, β_0 a p-dimensional vector of zeros and Σ_0 the p-dimensional identity matrix. The Bayesian specification of the model is then completed by eliciting prior distributions for all the remaining parameters, which we assume to be independent. We let

$$\begin{aligned} (\tau_{1ij}, \tau_{2ij}) &\sim N_2(\alpha_i, \Omega), & \alpha_i &\sim N_2(\alpha_0, R), \\ \lambda_{ij} &\sim U(b_{ij} + 10, b_{ij} + 20), & \omega_{ij} &\sim \mathrm{Ga}(1/2, 1), \\ 1/\sigma^2 &\sim \mathrm{Ga}(1/2, 1/2), & \zeta &\sim U(0, 1), \end{aligned} \tag{14}$$

where b_{ij} denotes the first day after menstruation for cycle i of woman j, $U(a, b)$ denotes the uniform distribution over (a, b) and $\mathrm{Ga}(c, d)$ stands for the gamma distribution with expected value c/d. For simplicity Ω and R are identity matrices while α_0 is a bidimensional vector of zeros. The specifications in (14) allow us 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 $\theta = 1$ and $\sigma = 0{\cdot}25$, 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 Supplementary Material. For the parametric part, its derivation 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 run the algorithm for 8000 iterations and discard the first 3000 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 a 95% quantile based posterior credible interval equal to (0·9097, 0·9450).

Figure 2(a) 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 features such as the day of ovulation and the levels of the low and high plateaus.

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

Acknowledgement

The authors are grateful to an associate editor and two referees for helpful comments and suggestions. A. Lijoi and I. Prünster are also affiliated with the Bocconi Institute for Data Science and Analytics, Milano, Italy, and supported by the European Research Council. A. Canale and A. Lijoi are also affiliated with the Collegio Carlo Alberto, Moncalieri, Italy.

Supplementary material

Supplementary material available at *Biometrika* online contains further details on the model specifications and the derivation of the computational schemes employed in § 4.

Appendix

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

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

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

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 a unifying concept. 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 $k \geqslant 1$, 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

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

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

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

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

$$E\left\{\int_{\mathbb{X}^k} \tilde{P}^{n_1}(\mathrm{d}x_1)\cdots\tilde{P}^{n_k}(\mathrm{d}x_k)\right\} \tag{A1}$$

for any k-tuple of positive integers $n_1, \ldots, n_k$ such that $\sum_{i=1}^k n_i = n$ and integration variables such that $x_1 \neq \cdots \neq x_k$. By virtue of Fubini's theorem and the definition of the Pitman–Yor process, (A1) equals

$$\frac{\Gamma(\theta+1)}{\Gamma(\theta/\sigma+1)}\,\frac{1}{\Gamma(\theta+n)}\int_{\mathbb{X}^k}\int_0^\infty u^{\theta+n-1}E\left\{\mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\mathbb{X})}\prod_{j=1}^k\tilde{\mu}_{\sigma,0}^{n_j}(\mathrm{d}x_j)\right\}\mathrm{d}u, \tag{A2}$$

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 (A2). If none of the x_j equal 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.,

$$E\left\{\prod_{j=1}^k\tilde{P}^{n_j}(\mathrm{d}x_j)\right\} \cong \left\{\prod_{j=1}^k(1-\zeta)P^*(\mathrm{d}x_j)\right\}\frac{\sigma^k}{(\theta+1)_{n-1}\,\Gamma(\theta/\sigma+1)}$$
$$\times\int_0^\infty u^{\theta+n-1}\,\mathrm{e}^{-u^\sigma}\left\{\prod_{j=1}^k\frac{1}{\Gamma(1-\sigma)}\int_0^\infty s^{n_j-\sigma-1}\,\mathrm{e}^{-us}\right\}\mathrm{d}u, \tag{A3}$$

where each $\mathrm{d}x_j$ stands for an infinitesimal neighbourhood around x_j. Hence, (A3) is a first-order approximation of $E\{\prod_{j=1}^k\tilde{P}^{n_j}(\mathrm{d}x_j)\}$ and the higher-order terms vanish when computing the integral over $\mathbb{X}^k$ in (A2).

The right-hand side of (A3) can be rewritten as

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

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

On the other hand, if $x_0 = x_j$ for some $j \in \{1,\ldots,k\}$, the expected value in the integral in (A2) equals

$$E\left\{\mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\{x_0\})}\,\tilde{\mu}^{n_j}_{\sigma,0}(\{x_0\})\right\} E\left\{\mathrm{e}^{-u\tilde{\mu}_{\sigma,0}(\mathbb{X}\setminus\{x_0\})}\prod_{\ell\neq j}\tilde{\mu}^{n_\ell}_{\sigma,0}(\mathrm{d}x_\ell)\right\}, \tag{A4}$$

where the factorization follows from the definition of completely random measure. The second factor on the right-hand side of (A4) can be easily evaluated since $x_\ell \neq x_0$ for any $\ell \neq j$, and thus it involves only the diffuse component P^* of P_0, i.e.,

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

and the above approximation is to be interpreted as the one given in (A3). As for the first factor, one has

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

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

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

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

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

To sum up, the integrand in (A2) 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:

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

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

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

If we insert this expression into (A2), simple algebra yields (8).

Proof of Theorem 2

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

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

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

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 now clear 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).

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. *Statist. Med.* **35**, 1373–89.

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*, N. Balakrishnan, E. A. Peña & M. J. Silvapulle, eds., vol. 1 of *Institute of Mathematical Statistics Collections* Beachwood, Ohio: Institute of Mathematical Statistics, pp. 211–30.

Charalambides, C. A. (2005). *Combinatorial Methods in Discrete Distributions*. Hoboken, New Jersey: Wiley.

Colombo, B. & Masarotto, G. (2000). Daily fecundability: First results from a new data base. *Demographic Res.* **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. Pat. Anal. Mach. Intel.* **37**, 212–29.

Do, K.-A., Müller, P. & Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Appl. Statist.* **54**, 627–44.

Dunson, D. B., Herring, A. H. & Engel, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *J. Am. Statist. Assoc.* **103**, 534–46.

George, E. I. & McCulloch, R. E. (1993). Bayesian variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–9.

Guindani, M., Müller, P. & Zhang, S. (2009). A Bayesian discovery procedure. *J. R. Statist. Soc.* B **71**, 905–25.

Ishwaran, H. & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33**, 730–73.

James, L. F., Lijoi, A. & Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.* **33**, 105–20.

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

Jara, A., Lesaffre, E., De Iorio, M. & Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Ann. Appl. Statist.* **4**, 2126–49.

Kim, S., Dahl, D. B. & Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Anal.* **4**, 707–32.

Kingman, J. F. C. (1975). Random discrete distributions. *J. R. Statist. Soc.* B **37**, 1–22.

Lijoi, A., Mena, R. H. & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Am. Statist. Assoc.* **100**, 1278–91.

Lijoi, A., Mena, R. H. & Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Staistt. Soc.* B **69**, 715–40.

Lijoi, A. & Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, N. L. Hjort, C. Holmes, P. Müller & S. G. Walker, eds. Cambridge: Cambridge University Press, 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. Am. Statist. Assoc.* **83**, 1023–36. 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. *Prob. Theory Rel. Fields* **92**, 21–39.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Prob. Theory Rel. Fields* **102**, 145–58.

Pitman, J. (2003). Poisson–Kingman partitions. In *Statistics and Science: a Festschrift for Terry Speed*, D. R. Goldstein, ed., vol. 40 of *IMS Lecture Notes Monograph Series*. Beachwood, OH: Institute of Mathematical Statistics, pp. 1–34.

Pitman, J. & Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Prob.* **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**, 461–87.

Scarpa, B. & Dunson, D. B. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics* **65**, 772–80.

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. *Comp. Statist. Data Anal.* **56**, 2663–74.

[*Received on* 11 *November* 2016. *Editorial decision on* 10 *June* 2017]

Biometrika (2017), *pp.* 1–7
Printed in Great Britain

Supplementary material for "On the Pitman-Yor process with spike and slab base measure"

BY A. CANALE

Department of Statistical Sciences, University of Padua, Via C. Battisti 241, 35121 Padua, Italy,
canale@stat.unipd.it

A. LIJOI

Department of Decision Sciences, Bocconi University, via Röntgen 1, 20136 Milan, Italy
lijoi@unibocconi.it

B. NIPOTI

School of Computer Science and Statistics, Trinity College, College Green, Dublin 2, Ireland
nipotib@tcd.ie

AND I. PRÜNSTER

Department of Decision Sciences, Bocconi University, via Röntgen 1, 20136 Milan, Italy
igor@unibocconi.it

SUMMARY

This supplementary material contains further details on the model specification and the derivation of the computational schemes employed in §4 of the main paper. The R code used in the paper is available at the github repository `github.com/tonycanale/PitmanYorSpikeAndSlab/`.

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 $X_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$, 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

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

$$\mathrm{pr}(X_i = k_{\setminus i} \mid X_{-i}) \propto (1-\zeta)\frac{\theta + (k_{\setminus i}-1)\sigma}{\theta+n-1}\,\frac{\sum_{l=1}^{n_0}\zeta^l\,\mathscr{C}(n_0,l;\sigma)\left(\theta/\sigma + k_{\setminus i}\right)_l}{\sum_{l=1}^{n_0}\zeta^i\,\mathscr{C}(n_0,l;\sigma)\left(\theta/\sigma + k_{\setminus i}-1\right)_l}.$$

DETAILS ON § 4·1

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

$$P^*(\mathrm{d}\mu,\mathrm{d}\tau) = \phi(\mu;\mu_0,\kappa\tau^{-1})\times \mathrm{Ga}(\tau;a,b)\,\mathrm{d}\mu\,\mathrm{d}\tau, \tag{A8}$$

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 A1 obtained using equation (11).

Table A1: Prior parameters for the simulation experiment

		θ	
$E(K_n)$	σ	$n=50$	$n=100$
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

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.

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

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

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

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

and finally to a new cluster with probability proportional to

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

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

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

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

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

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

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 $\tilde{Q}^* \sim \mathrm{PY}(\sigma, \theta; P^*)$ and P^* is equal to (A8). The analysis is carried out for different choices of σ and θ as reported in Table A2. 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 A1. *Let K_n be the number of distinct values in an exchangeable sample $X^{(n)}$ from the outer spike and slab model* (6)*. Then*

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

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

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

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

Table A2: Prior parameters for the simulation experiment assuming $E(K_n) = 5$

		θ	
Model	σ	$n = 50$	$n = 100$
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

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 $k_{\backslash i}$ be the number of distinct values of S_j labeled from 0 to $k_{\backslash i} - 1$ 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

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

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

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

   and finally to a new cluster with probability proportional to

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

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

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

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

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

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 A1 and $\pi = (\pi_1, \ldots, \pi_5) = (0.4, 0.2, 0.2, 0.1, 0.1)$.

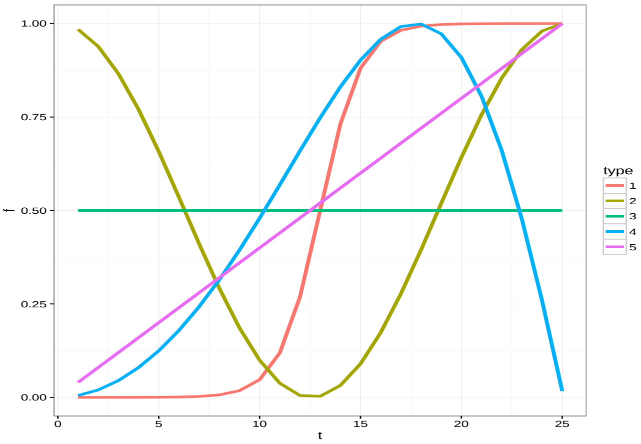


Fig. A1: Functional means for the functional data simulation experiment.

Data are analysed assuming $P \sim \mathrm{PY}(\sigma, \theta; P_0)$ 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

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

where $B(\cdot)$ denotes the cubic B-splines basis and β a finite vector of basis coefficients of size $p = 3 +$ number of knots. 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.

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

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

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

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

and finally to a new cluster with probability proportional to

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

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

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

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

DETAILS ON § 4·2

Computational details

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 $i = 1, \dots, n$, 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 $S_1, \dots, S_n$ be the current cluster allocation, with $S_j = 0$ if the corresponding observation is allocated to the cluster of the spike. For $i = 1, \dots, n$ 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

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

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

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

   and finally to a new cluster with probability proportional to

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

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

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

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

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

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

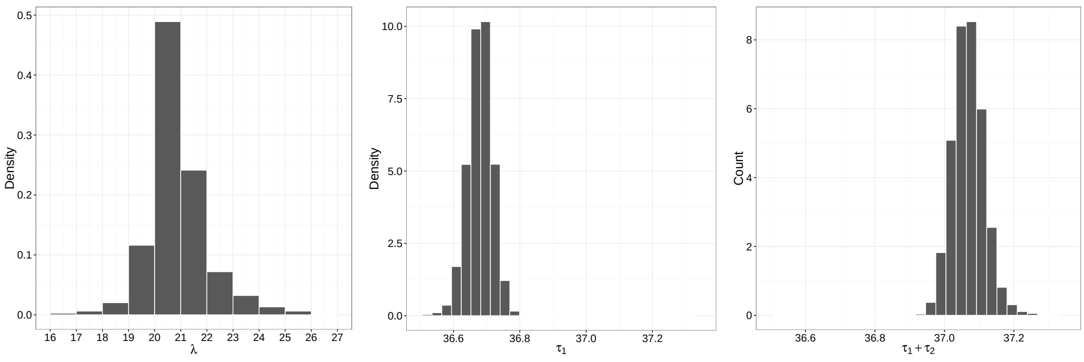
$$1/\sigma^2 \sim \mathrm{Ga} \left[\frac{1}{2} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} n_{ij}, \frac{1}{2} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{t=1}^{n_{ij}} \{X_{ij}(t) - f_{ij}(t)\}^2 \right].$$

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

Additional plots

Figure A2 reports 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.

[*Received November* 2016. *Revised April* 2017]



(a) (b) (c)

Fig. A2: 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.