Density Estimation Via Hierarchies of Nonparametric Priors

Federico Camerlenghi * Antonio Lijoi † Igor Prünster ‡

Abstract
In Bayesian Nonparametrics partial exchangeability is a useful assumption tailored for heterogeneous, though related, groups of observations. Recent contributions in Bayesian literature have focused on the construction of dependent nonparametric priors to accommodate for partially exchangeable sequences of observations. In the present paper we concentrate on vectors of hierarchical Pitman-Yor processes, in which the dependence is created by choosing a common random base measure for each group of observations. These hierarchical processes are then used to define dependent hierarchical mixtures. We finally apply the model to estimate densities arising from multiple groups of observations by performing a suitable Gibbs sampling algorithm.

Key Words: Bayesian nonparametrics, partial exchangeability, hierarchical process, Pitman-Yor process, density estimation, mixture model.

1. Introduction

Bayesian inference, as well as statistical inference in general, can be carried out if a certain number of *analogous* observations are available. Exchangeability reflects this idea of analogy or symmetry of the data in some applications of interest. We remind that a sequence of observations $(\theta_i)_{i\geq 1}$ is exchangeable iff, for any $n \geq 1$, the distribution of the vector $(\theta_1, \ldots, \theta_n)$ is invariant under a permutation of its components. By virtue of de Finetti's representation theorem, such an assumption can be conveniently rephrased as conditional independence and identical distribution of the $\theta_i's$, more precisely

$$\begin{aligned} \theta_i|\tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \quad i \in \mathbb{N} \\ \tilde{p} &\sim Q \end{aligned}$$

where $\tilde{p}$ is a random probability measure with distribution Q, called the de Finetti measure of the sequence and working as a prior distribution to carry out posterior inference. The most famous Bayesian nonparametric prior is certainly the distribution of the Dirichlet process introduced by Ferguson (1973). The Pitman-Yor process has been the first generalization of the Dirichlet process (see Pitman and Yor (1997)); other important contributions for the construction of random probability measures have been proposed by Regazzini, Lijoi and Prünster (2003) and De Blasi et al. (2015).
However, as pointed out by de Finetti (1938) himself, exchangeability could be a quite restrictive assumption when data are affected by some sort of heterogeneity, e.g. in multiple related studies. In these situations exchangeability can be considered only a limiting case and one should resort to more general dependence structures, which are still analytically tractable. Some dependent nonparametric priors have been recently proposed in Bayesian literature, and many of these rely on the notion of partial exchangeability. To provide a formal definition of such a fundamental assumption, suppose that Θ is a Polish space equipped

*Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy. E-mail: federico.camerlenghi@unimib.it
†Department of Decision Sciences, BIDSA and IGIER, Bocconi University, via Röntgen 1, 20136 Milano, Italy. E-mail: antonio.lijoi@unibocconi.it
‡Department of Decision Sciences, BIDSA and IGIER, Bocconi University, via Röntgen 1, 20136 Milano, Italy. E-mail: igor@unibocconi.it

with its Borel σ-algebra $\mathscr{T}$. Consider d sequences of observations $\theta^{(i)} := (\theta_{i,j})_{j\geq 1}$, for $i = 1, \ldots, d$, defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\Theta, \mathscr{T})$. They are partially exchangeable if and only if the distribution of $(\theta^{(1)}, \ldots, \theta^{(d)})$ coincides with the one of $(\pi_1\theta^{(1)}, \ldots, \pi_d\theta^{(d)})$, where $\pi_i\theta^{(i)} = (\theta_{i,\pi_i(j)})_{j\geq 1}$, for all the d-tuples $(\pi_1, \ldots, \pi_d)$ of finite permutations on $\mathbb{N}^d$. The de Finetti representation theorem provides with an equivalent formulation of such a notion, more precisely $(\theta_{i,j})_{j\geq 1}$, for $i = 1, \ldots, d$, are partially exchangeable if and only if there exists a vector of random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_d)$, such that

$$\begin{aligned} (\theta_{1,j_1}, \ldots, \theta_{d,j_d}) \,|\, (\tilde{p}_1, \ldots, \tilde{p}_d) &\overset{\text{iid}}{\sim} \tilde{p}_1 \times \cdots \times \tilde{p}_d \qquad (j_1, \ldots, j_d) \in \mathbb{N}^d \\ (\tilde{p}_1, \ldots, \tilde{p}_d) &\sim Q_d. \end{aligned} \tag{1}$$

Denoted by P_Θ the space of all probability measures on Θ, which is assumed to be endowed with the corresponding Borel σ-field $\mathscr{P}_\Theta$, then Q_d is a probability law on the space $(\mathsf{P}_\Theta^d, \mathscr{P}_\Theta^d)$. The definition of Q_d, or equivalently of the dependence across $\tilde{p}_1, \ldots, \tilde{p}_d$, has been recently addressed in the Bayesian nonparametric literature, starting from the contribution of MacEachern (1999, 2000). In the present paper we focus on hierarchical random probability measures, developing a suitable model to face Bayesian density estimation when data come from different, though related, sources of randomness. The idea of hierarchical priors has been first introduced in Teh et al. (2006) for the Dirichlet process case. Other contributions in this direction include Gasthaus and Teh (2010); Teh and Jordan (2010); Wood et al. (2011); Nguyen (2016), a complete distribution theory for hierarchical processes has been developed in Camerlenghi et al. (2018a), see also Camerlenghi, Lijoi and Prünster (2018b) for the exchangeable case. The structure of the present paper is as follows. In Section 2 we recall the definition of hierarchical Pitman-Yor processes, which will be used to define a vector of random dependent densities. The MCMC algorithm for density estimation will be presented in Section 3, then we conclude the paper with some numerical illustrations.

2. Hierarchies of Pitman-Yor processes

In the sequel we focus on hierarchies of the Pitman-Yor (PY) process (see Pitman and Yor (1997)). It is worth to remind that a Pitman-Yor random probability measure $\tilde{p}$ is characterized by two parameters (c, σ) and a base measure P_0. The admissible parameter values are $c > -\sigma$ and $\sigma \in (0, 1)$ or $c = m|\sigma|$ and $\sigma < 0$ for some $m \in \mathbb{N}$. For our purposes, we assume $\sigma \in (0, 1)$ and $c > 0$. There are many ways to construct $\tilde{p}$, and the simplest one is based on a stick-breaking procedure. More precisely $\tilde{p}$ is a discrete random probability measure $\tilde{p} = \sum_{j\geq 1} \tilde{\pi}_j \delta_{Z_j}$ such that

$$\tilde{\pi}_1 = V_1, \quad \tilde{\pi}_j = V_j \prod_{i=1}^{j-1} (1 - V_i) \text{ for } j \geq 2,$$

where the $(Z_j)_{j\geq 1}$'s are i.i.d. random variables taking values in $(\Theta, \mathscr{T})$, with common distribution P_0, and the V_i's are independent Beta random variables with parameters $(c + i\sigma, 1 - \sigma)$. In addition the sequences $(V_i)_{i\geq 1}$ and $(Z_i)_{i\geq 1}$ are assumed to be independent. To fix the notation, we will write $\tilde{p} \sim \text{PY}(\sigma, c; P_0)$.

We are now ready to define a vector of hierarchical Pitman-Yor processes, randomizing the base measure referring to each $\tilde{p}_i$ of the vector $(\tilde{p}_1, \ldots, \tilde{p}_d)$ in (1). More precisely we

say that Q_d in (1) is the distribution of a Hierarchical Pitman-Yor Process (HPYP) if

$$\begin{aligned} \tilde{p}_i|\tilde{p}_0 &\overset{\text{ind}}{\sim} \text{PY}(\sigma_i, c_i; \tilde{p}_0) \quad i = 1, \ldots, d \\ \tilde{p}_0 &\sim \text{PY}(\sigma_0, c_0; P_0) \end{aligned} \tag{2}$$

being $\sigma_i, \sigma_0 \in (0, 1)$, $c_i, c_0 > 0$, for any $i = 1, \ldots, d$, and P_0 is a non-atomic probability measure on $(\Theta, \mathscr{T})$. Note that the use of the same base measure $\tilde{p}_0$ for each group of observations generates dependence across the diverse random probability measures. It is worth to underline that the random probability measures $\tilde{p}_1, \ldots, \tilde{p}_d$ in (2) are almost surely discrete, since they are PY processes conditionally on $\tilde{p}_0$. For the definition of a broader class of hierarchial priors, having almost surely discrete realizations, we refer to Camerlenghi et al. (2018a).

2.1 Hierarchical mixture models

It is now easy to employ the vector $(\tilde{p}_1, \ldots, \tilde{p}_d)$ in (2), to define a corresponding vector of random dependent densities $(\tilde{f}_1, \ldots, \tilde{f}_d)$, by putting $\tilde{f}_i(x) := \int_\Theta h(x; \theta)\tilde{p}_i(\mathrm{d}\theta)$, where $h(\,\cdot\,;\,\cdot\,)$ is a kernel function. More precisely we assume to be provided with a sample $\boldsymbol{X}_i := (X_{i,1}, \cdots, X_{i,n_i})$ for population i, where $i = 1, \ldots, d$, and that the variables $X_{i,j}$'s take values in a Polish space $\mathbb{X}$, equipped with the Borel σ-field $\mathscr{X}$. The vector $\boldsymbol{X}_i$ is associated with the corresponding vector of latent variables $\boldsymbol{\theta}_i := (\theta_{i,1}, \cdots, \theta_{i,n_i})$, for any $i = 1, \ldots, d$. We can summarize the hierarchical structure as follows:

$$\begin{aligned} (X_{1,j_1}, \ldots, X_{d,j_d})|(\theta_{1,j_1}, \ldots, \theta_{d,j_d}) &\overset{\text{ind}}{\sim} h(\,\cdot\,;\theta_{1,j_1}) \times \cdots \times h(\,\cdot\,;\theta_{d,j_d}) \\ (\theta_{1,j_1}, \ldots, \theta_{d,j_d})|(\tilde{p}_1, \ldots, \tilde{p}_d) &\overset{\text{ind}}{\sim} \tilde{p}_1 \times \cdots \times \tilde{p}_d \end{aligned} \tag{3}$$

for any $j_i \in \{1, \ldots, n_i\}$, $i = 1, \ldots, d$, and $(\tilde{p}_1, \ldots, \tilde{p}_d)$ is supposed to be a vector of hierarchical Pitman-Yor processes. In order to fix the notation we define $\boldsymbol{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$ and besides $\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$.

The discreteness of the different random probability measures in (3) allows for latent ties across the different vectors of latent variables $\boldsymbol{\theta}_i$, for $i = 1, \ldots, d$. Such ties induce a latent random partition within each group and across the diverse groups of observations, whose distributional properties has been carefully investigated in Camerlenghi et al. (2018a). In the following we make use of these results to determine the joint distribution of the variables in the model (3). We will assume that the latent variables $\boldsymbol{\theta}$ may display k distinct values, denoted here as $\theta_1^*, \ldots, \theta_k^*$. Moreover, $\boldsymbol{n}_i := (n_{i,1}, \ldots, n_{i,k})$, for $i = 1, \ldots, d$, is the vector of frequency counts in the i-th vector $\boldsymbol{\theta}_i$, namely $n_{i,j} \geq 0$ is the number of elements of the i-th vector $\boldsymbol{\theta}_i$ that coincide with the j-th distinct values; we further set $\bar{n}_{\bullet j} := \sum_{i=1}^d n_{i,j}$ the total number of observations coinciding with the j-th distinct value. We obviously have that $\sum_{j=1}^k n_{i,j} = n_i$ for any $i = 1, \ldots, d$, and $n_{i,j} = 0$ means that the j-th distinct has not been recorded in $\boldsymbol{\theta}_i$. The induced partition structure, known as *partially Exchangeable Partition Probability Function* (pEPPF), may be easily interpreted in terms of the Chinese Restaurant Franchise (CRF) representation (see Teh et al. (2006)). According to this metaphor, $\boldsymbol{\theta}_i$ identifies the i-th Chinese restaurant in a franchise of d restaurants, all sharing the same menu. $\boldsymbol{\theta}_i$ are the dishes' labels that have been selected by the n_i customers seated in the i-th restaurant. People seating at the same table eat the same dish, and the same dish can be served at different tables within the same restaurant or across different restaurants. Accordingly, $n_{i,j} \geq 0$ is the number of customers in restaurant i eating dish j, for $i = 1, \ldots, d$ and $j = 1, \ldots, k$.

As discussed in Camerlenghi et al. (2018a) the evaluation of the partition structure in full

generality is a difficult task, hence, in order to obtain a more tractable expression of the pEPPF, one needs to introduce suitable latent variables $\boldsymbol{T}_i = (T_{i,1}, \ldots, T_{i,n_i})$, for each restaurant i, which represent the tables' labels where the people are seated at and are collected in the vector $\boldsymbol{T} := (\boldsymbol{T}_1, \ldots, \boldsymbol{T}_d)$. The latent tables determine a refinement of the partition induced by data, whereby the $n_{i,j}$ customers eating dish j in restaurant i may be partitioned into $\ell_{i,j} \in \{1, \ldots, n_{i,j}\}$ distinct tables, the t-th of which has $q_{i,j,t}$ customers, for $t = 1, \ldots, \ell_{i,j}$. Hence we have that $n_{i,j} = \sum_{t=1}^{\ell_{i,j}} q_{i,j,t}$. We further introduce the compact notation for the vectors of counts $\boldsymbol{\ell}_i := (\ell_{i,1}, \ldots, \ell_{i,k})$ and $\boldsymbol{q}_{i,j} := (q_{i,j,1}, \ldots, q_{i,j,\ell_{i,j}})$, while $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_d)$ denotes the overall tables frequencies, whereas $\bar{\ell}_{\bullet j} = \sum_{i=1}^{d} \ell_{i,j}$, $\bar{\ell}_{i\bullet} = \sum_{j=1}^{k} \ell_{i,j}$ denote the number of tables serving dish j and the overall number of tables, respectively, in restaurant i. We introduce the set $C_{i,j} := \{r \in \{1, \ldots, n_i\} : \theta_{i,r} = \theta_j^*\}$ collecting the indexes of observations from population i which coincides with the j-th distinct value θ_j^*; we finally denote by $T^*_{i,j,1}, \ldots, T^*_{i,j,\ell_{i,j}}$ the $\ell_{i,j}$ distinct tables' labels in restaurant i serving dish j. Using the representation of the partition structure of $\boldsymbol{\theta}$ derived in Camerlenghi et al. (2018a), one can determine the joint distribution of $(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta})$, when $(\tilde{p}_1, \ldots, \tilde{p}_d)$ is a vector of hierarchical Pitman-Yor processes:

$$\begin{aligned}\mathscr{L}(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}) = \prod_{i=1}^{d} \prod_{j=1}^{k} \prod_{r \in C_{i,j}} h(x_{i,r}; \theta_j^*) \mathrm{d}x_{i,r} \frac{\prod_{r=1}^{k-1}(c_0 + r\sigma_0)}{(c_0+1)_{|\boldsymbol{\ell}|-1}} \prod_{t=1}^{k} (1-\sigma_0)_{\bar{\ell}_{\bullet t}-1} \\ \times \prod_{i=1}^{d} \left(\frac{\prod_{r=1}^{\bar{\ell}_{i\bullet}-1}(c_i + r\sigma_i)}{(c_i+1)_{n_i-1}} \prod_{v=1}^{k} \prod_{t=1}^{\ell_{i,v}} (1-\sigma_i)_{q_{i,v,t}-1} Q_0(\mathrm{d}t_{i,v,t}) \right) \prod_{j=1}^{k} P_0(\mathrm{d}\theta_j^*).\end{aligned} \tag{4}$$

where $(a)_n := \Gamma(a+n)/\Gamma(a)$ denotes the ascending factorial, with the proviso $(a)_{-1} \equiv 1$, and Q_0 is a non-atomic probability measure on the space of tables' labels.

3. Algorithm

On the basis of the joint distribution (4), in this section we are able to devise a Gibbs sampler algorithm to estimate the posterior expected values of the random dependent densities $\tilde{f}_i$, for $i = 1, \ldots, d$. In particular we propose a suitable extension of the algorithm in Escobar and West (1995), which is valid for an arbitrary number d of populations. As in Escobar and West (1995), we assume that P_0 is a normal/Inverse-Gamma distribution, i.e. $P_0(\mathrm{d}M, \mathrm{d}V) = P_{0,1}(\mathrm{d}V) P_{0,2}(\mathrm{d}M|V)$, where $P_{0,1}$ is an Inverse-Gamma with parameters (s_0, S_0), and $P_{0,2}$ is Gaussian with mean m and variance τV. The hyperpriors are chosen of the type $\tau^{-1} \sim \mathrm{Gam}(w/2, W/2)$ and $m \sim N(a, A)$, for some real parameters $w, W, A > 0$ and $a \in \mathbb{R}$. In the simulation studies we will set $(w, W) = (1, 100)$, $(a, A) = (\bar{\boldsymbol{X}}, 2)$. The parameters (c_i, σ_i), for $i = 0, \ldots, d$, are assumed to be independent random variables, in particular we suppose that $c_i \sim \mathrm{Gam}(1, 1)$ and $\sigma_i \sim \mathrm{U}(0, 1)$ *a priori* for any $i = 0, \ldots, d$. For the sake of notational simplicity we further set $\boldsymbol{c} := (c_0, c_1, \ldots, c_d)$, $\boldsymbol{\sigma} := (\sigma_0, \sigma_1, \ldots, \sigma_d)$ and $\boldsymbol{\Delta} := (\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \tau, m, \boldsymbol{c}, \boldsymbol{\sigma})$; we finally write $\boldsymbol{\Delta}_{-v}$ to denote all the variables but v. The basic steps of the algorithm are described below in details.

Update the couples of dishes and tables

Fix, $i = 1, \ldots, d$, we use the notation v^{-r} to indicate the value of a random variable v after the removal of the couple $(\theta_{i,r}, T_{i,r})$. The full conditional distribution of the couple

$(\theta_{i,r}, T_{i,r})$, for $r = 1, \ldots, n_i$ and $i = 1, \ldots, d$, boils down to

$$\mathbb{P}(\theta_{i,r} \in \mathrm{d}\theta, T_{i,r} \in \mathrm{d}t | \boldsymbol{\Delta}_{-(\theta_{i,r}, T_{i,r})}) = w_{i,0} P^*_{i,r}(\mathrm{d}\theta) Q_0(\mathrm{d}t) + \sum_{h=1}^{k^{-r}} w_{i,h} \delta_{\{\theta^{*,-r}_h\}}(\mathrm{d}\theta) Q_0(\mathrm{d}t) + \sum_{h=1}^{k^{-r}} \sum_{\kappa=1}^{\ell^{-r}_{i,h}} w_{i,h,\kappa} \delta_{\{\theta^{*,-r}_h\}}(\mathrm{d}\theta) \delta_{\{T^{*,-r}_{i,h,\kappa}\}}(\mathrm{d}t)$$

where $T^{*,-r}_{i,h,1}, \ldots, T^{*,-r}_{i,h,\ell^{-r}_{i,h}}$ are the distinct tables' labels at the first restaurant where the h-th dish is served, after the removal of $T_{i,r}$, with $P^*_{i,r}(\mathrm{d}\theta) = h(x_{i,r}; \theta) P_0(\mathrm{d}\theta) / \int_\Theta h(x_{i,r}; \theta) P_0(\mathrm{d}\theta)$, while

$$w_{i,0} \propto \frac{(c_0 + k^{-r}\sigma_0)(c_i + \bar{\ell}^{-r}_{i\bullet}\sigma_i)}{(c_i + n_i - 1)(c_0 + |\boldsymbol{\ell}^{-r}|)} \int_\Theta h(x_{i,r}; \theta) P_0(\mathrm{d}\theta)$$

and, for any $h = 1, \ldots, k^{-r}$ and $\kappa = 1, \ldots, \ell^{-r}_{i,h}$

$$w_{i,h} \propto \frac{(\bar{\ell}^{-r}_{\bullet h} - \sigma_0)(c_i + \bar{\ell}^{-r}_{i\bullet}\sigma_i)}{(c_i + n_i - 1)(c_0 + |\boldsymbol{\ell}^{-r}|)} h(x_{i,r}; \theta^{*,-r}_h),$$

$$w_{i,h,\kappa} \propto \frac{(q_{i,h,\kappa} - \sigma_i)}{(c_i + n_i - 1)} h(x_{i,r}; \theta^{*,-r}_{i,h}) \mathbb{1}_{\{n^{-r}_{i,h} > 0\}}.$$

Update the parameters

The updating of τ, m, σ_i and c_i, for $i = 0, 1, \ldots, d$, is based on their full conditional distributions, that we report here. It is easy to see that

$$\mathscr{L}(c_0 | \boldsymbol{\Delta}_{-c_0}) \propto \frac{\prod_{r=1}^{k-1}(c_0 + r\sigma_0)}{(c_0 + 1)_{|\boldsymbol{\ell}|-1}} \kappa_{c_0}(\mathrm{d}c_0)$$

$$\mathscr{L}(c_i | \boldsymbol{\Delta}_{-c_i}) \propto \frac{\prod_{r=1}^{\bar{\ell}_{i\bullet}-1}(c_i + r\sigma_i)}{(c_i + 1)_{n_i - 1}} \kappa_{c_i}(\mathrm{d}c_i) \quad \text{for } i = 1, \ldots, d$$

$$\mathscr{L}(\sigma_0 | \boldsymbol{\Delta}_{-\sigma_0}) \propto \prod_{r=1}^{k-1}(c_0 + r\sigma_0) \prod_{t=1}^{k}(1 - \sigma_0)_{\bar{\ell}_{\bullet t} - 1} \kappa_{\sigma_0}(\mathrm{d}\sigma_0)$$

$$\mathscr{L}(\sigma_i | \boldsymbol{\Delta}_{-\sigma_i}) \propto \prod_{r=1}^{\bar{\ell}_{i\bullet}-1}(c_i + r\sigma_i) \prod_{v=1}^{k} \prod_{t=1}^{\ell_{i,v}}(1 - \sigma_i)_{q_{i,v,t} - 1} \kappa_{\sigma_i}(\mathrm{d}\sigma_i) \quad \text{for } i = 1, \ldots, d$$

where $\kappa_{c_i}(\cdot)$ and $\kappa_{\sigma_i}(\cdot)$ are the prior distributions for c_i and σ_i, $i = 0, 1, \ldots, d$, respectively. It is apparent that these parameters have to be updated through a Metropolis-Hastings algorithm.
As for τ and m we get

$$\mathscr{L}(\tau | \boldsymbol{\Delta}_{-\tau}) \sim \mathrm{IG}\Big(\frac{w}{2} + \frac{k}{2}, \frac{W}{2} + \sum_{j=1}^{k} \frac{(M^*_j - m)^2}{2V^*_j}\Big),$$

$$\mathscr{L}(m | \boldsymbol{\Delta}_{-m}) \sim \mathrm{N}\Big(\frac{R}{D}, \frac{1}{D}\Big),$$

where

$$R = \frac{a}{A} + \sum_{j=1}^{k} \frac{M^*_j}{\tau V^*_j}, \quad D = \frac{1}{A} + \sum_{j=1}^{k} \frac{1}{\tau V^*_j}.$$

Acceleration step

It is well known that the Pólya urn sampler tends to mix slowly, since the probability of sampling a new value is lower than the probability of sampling an already observed one. In order to avoid this problem one can speed up the algorithm resampling the distinct values at the end of every iteration. Observe that the full conditional distribution for θ_j^* is given by

$$\mathscr{L}(\theta_j^*|\boldsymbol{\Delta}_{-\theta_j^*}) \propto \prod_{i=1}^{d} \prod_{r \in C_{i,j}} h(x_{i,r}; \theta_j^*) P_0(\mathrm{d}\theta_j^*),$$

hence, putting

$$S' = S_0 + \frac{\sum_{i=1}^{d} \sum_{r \in C_{i,j}} x_{i,r}^2}{2} + \frac{m^2 \bar{n}_{\bullet j} - \sum_{i=1}^{d} \sum_{r \in C_{i,j}} x_{i,r} (2m + \tau \sum_{i=1}^{d} \sum_{r \in C_{i,j}} x_{i,r})}{2(\tau \bar{n}_{\bullet j} + 1)},$$

we get

$$\begin{aligned} V_j^* &\sim \mathrm{IG}\Big(s_0 + \frac{\bar{n}_{\bullet j}}{2}, S'\Big), \\ M_j^* | V_j^* &\sim \mathrm{N}\Big(\frac{m + \tau \sum_{i=1}^{d} \sum_{r \in C_{i,j}} x_{i,r}}{\tau \bar{n}_{\bullet j} + 1}, \frac{\tau V_j^*}{\tau \bar{n}_{\bullet j} + 1}\Big). \end{aligned}$$

4. Illustrations: multiple populations

For the sake of illustration, we apply the algorithm described in the previous section and collect the MCMC outputs to estimate the posterior expected values of the random dependent densities in $d = 6$ populations. The simulation study is based on 20.000 iterations after a burn-in period of 20.000 iterations. We have considered six simulated datasets of observations $X_{i,1}, \ldots, X_{i,n_i} \overset{\text{iid}}{\sim} X_i$, being $n_i = 200$ for any $i = 1, \ldots, 6$, where the X_i's are specified as follows

$$\begin{aligned} X_1 &\sim 0.5\mathrm{N}(6, 0.6) + 0.25\mathrm{N}(10, 0.6) + 0.25\mathrm{N}(15, 0.6) \\ X_2 &\sim 0.25\mathrm{N}(10, 0.6) + 0.5\mathrm{N}(15, 0.6) + 0.25\mathrm{N}(20, 0.6) \\ X_3 &\sim 0.5\mathrm{N}(6, 0.6) + 0.5\mathrm{N}(10, 0.6) \\ X_4 &\sim 0.2\mathrm{N}(0, 0.6) + 0.4\mathrm{N}(3, 0.6) + 0.2\mathrm{N}(15, 0.6) + 0.2\mathrm{N}(20, 0.6) \\ X_5 &\sim 0.5\mathrm{N}(0, 0.6) + 0.5\mathrm{N}(15, 0.6) \\ X_6 &\sim 0.25\mathrm{N}(0, 0.6) + 0.25\mathrm{N}(6, 0.6) + 0.25\mathrm{N}(10, 0.6) + 0.25\mathrm{N}(15, 0.6). \end{aligned}$$

We define the symmetric matrix A, which is a 6×6 matrix whose generic element $a_{i,j}$ counts the number of components shared by the two mixtures generating population i and j:

$$A = \begin{pmatrix} 3 & 2 & 2 & 1 & 1 & 3 \\ 2 & 3 & 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 0 & 0 & 2 \\ 1 & 2 & 0 & 4 & 2 & 2 \\ 1 & 1 & 0 & 2 & 2 & 2 \\ 3 & 2 & 2 & 2 & 2 & 4 \end{pmatrix}.$$

The matrix A has a lot of non-zero entries, it is then apparent that a lot of components are shared across the different mixtures, hence the partially exchangeable framework is the most appropriate one to model these data.

The estimated densities are quite accurate and are reported in Figure 1. Besides Figure 2 shows the estimated posterior distribution of the number of components for each mixture: the distributions are highly concentrated around the true value in all the cases. Finally, we study the estimated number of shared components for the different couples of mixtures, to this end let us denote by $K_{i,j}$ the random number of components shared by the two samples from population i and j, with the convention $K_{i,i} = K_i$. On the basis of the MCMC output, we can approximate the matrix P, whose generic element is defined as $p_{i,j} = \mathbb{P}(K_{i,j} = a_{i,j}|\boldsymbol{X})$, i.e. the posterior probability that $K_{i,j}$ equals the true value $a_{i,j}$. The MCMC output allows us to estimate P, indeed we have

$$P = \begin{pmatrix} 0.48 & 0.86 & 0.89 & 0.93 & 0.94 & 0.82 \\ 0.86 & 0.56 & 0.91 & 0.92 & 0.95 & 0.86 \\ 0.89 & 0.91 & 0.78 & 0.97 & 0.99 & 0.90 \\ 0.93 & 0.92 & 0.97 & 0.45 & 0.87 & 0.83 \\ 0.94 & 0.95 & 0.99 & 0.87 & 0.73 & 0.89 \\ 0.82 & 0.86 & 0.90 & 0.83 & 0.89 & 0.32 \end{pmatrix}.$$

One can easily realize that the model recognizes the right number of shared components for every couple of populations, indeed the off-diagonal elements of P are very close to 1.

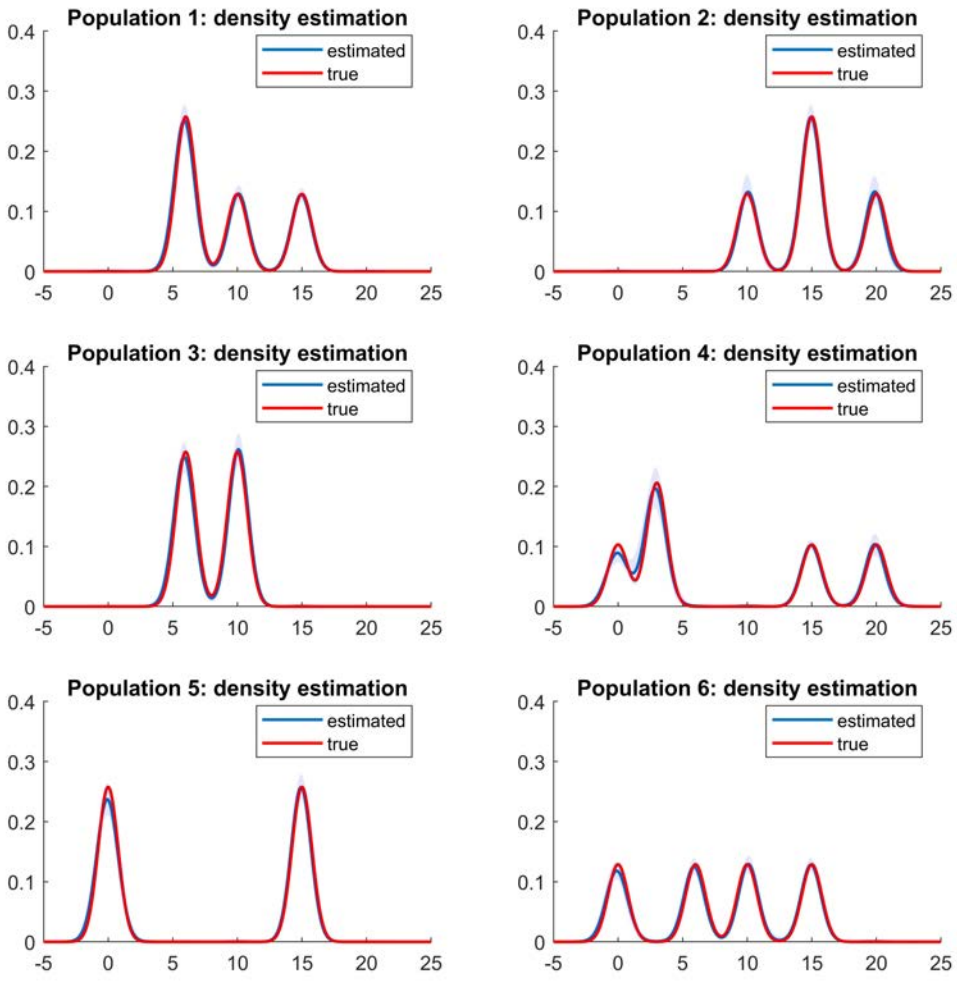

Figure 1: Estimated densities (blue) and true densities (red) for the six populations; the estimated credible intervals are shaded.

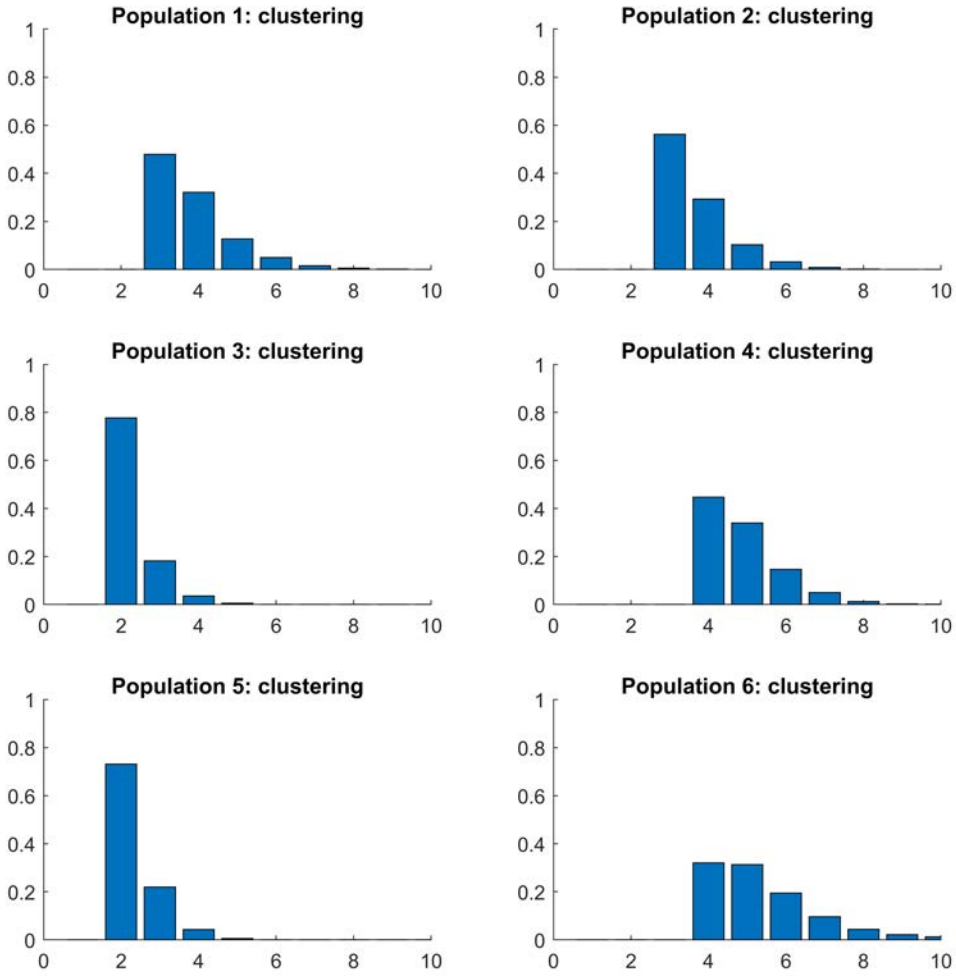

Figure 2: Posterior distribution of the number of mixture components for each population.

5. Acknowledgements

This work is supported by MIUR through the PRIN Project 2015SNS29B.

References

de Finetti, B. (1938). "Sur la condition d'equivalence partielle", *Actualités scentiques et industrielles*, 5–18.

Camerlenghi, F., Lijoi, A., Orbanz, P. and Prünster, I. (2018). "Distribution theory for hierarchical processes", *Ann. Statist.*, in press.

Camerlenghi, F., Lijoi, A. and Prünster, I. (2017). "Bayesian prediction with multiple–sample information", *J. Multivariate Anal.*, 156, 18–28.

Camerlenghi, F., Lijoi, A. and Prünster, I. (2018). "Bayesian nonparametric inference beyond the Gibbs–type framework", *Scand. J. Stat.*, doi: 10.1111/sjos.12334

De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Ruggiero, M. and Prünster, I. (2015). "Are Gibbs–type priors the most natural generalization of the Dirichlet process?", *IEEE Trans. Pattern Anal. Mach. Intell.*, 37, 212–229.

Escobar, M.D. and West, M. (1995). "Bayesian density estimation and inference using mixtures", *J. Amer. Statist. Assoc.*, 90, 577–588.

Ferguson, T.S. (1973). "A Bayesian analysis of some nonparametric problems", *Ann. Statist.*, 1, 209–230.

Gasthaus, J. and Teh, Y.W. (2010). "Improvements to the sequence memoizer", *Advances in Neuronal Information Processing Systems*, 23.

MacEachern, S.N. (1999). "Dependent nonparametric processes", in *ASA Proceedings of the SBSS*, Alexandria: American Statistical Association, pp. 50-55.

MacEachern, S.N. (2000). "Dependent Dirichlet processes", *Technical Report*, Department of Statistics, Ohio State University.

Nguyen, X. (2016). "Borrowing strength in hierarchical Bayes: convergence of the Dirichlet base measure", *Bernoulli*, 22, 1535–1571.

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

Regazzini, E., Lijoi, A. and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments", *Ann. Statist.*, 31, 560–585.

Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006). "Hierarchical Dirichlet processes", *J. Amer. Statist. Assoc.*, 101, 1566–1581.

Teh, Y.W. and Jordan, M.I. (2010). "Hierarchical Bayesian nonparametric models with applications", in *Bayesian Nonparametrics*, Cambridge Univ. Press, Cambridge, pp. 158-207.

Wood, F., Gasthaus, J., Archambeau, C., James, L.F. and Teh, Y.W. (2011). "The sequence memoizer", *Communications ACM*, 54, 91–98.