On some distributional properties of hierarchical processes

Federico Camerlenghi * Antonio Lijoi † Igor Prünster ‡

Abstract
Vectors of hierarchical random probability measures are popular tools in Bayesian nonparametrics. They may be used as priors whenever partial exchangeability is assumed at the level of either the observations or of some latent variables involved in the model. The first contribution in this direction can be found in Teh et al. (2006), who introduced the hierarchical Dirichlet process. Recently, Camerlenghi et al. (2017) have developed a general distribution theory for hierarchical processes, which includes the derivation of the partition structure, the posterior distribution and the prediction rules. The present paper is a review of these theoretical findings for vectors of hierarchies of Pitman–Yor processes.

Key Words: Bayesian Nonparametrics, hierarchical processes, partial exchangeablity, Pitman–Yor process, partition structure, posterior distribution.

1. Introduction

Exchangeability is quite standard an assumption in the Bayesian literature. Recall that a sequence of observations is exchangeable if the order in which the observations are recorded is irrelevant and this is tantamount to saying that the distribution of the sequence is invariant under the group of all finitary permutation. Less formally, such an assumption corresponds to a notion of homogeneity among all observations that may not hold true in several applications where data are affected by some source of heterogeneity. An obvious example emerges when data are generated by different, though related, such as in clinical trials, multicenter studies, change–point problems, and so on. In such situations one needs to resort to a more general dependence structure, and *partial exchangeability* often is a natural fit (see de Finetti (1938)). Roughly speaking, partial exchangeablity corresponds to assuming that the whole population splits into a certain number of subpopulations which are exchangeable in their own right.
More precisely, assume that $\mathbb{X}$ is a Polish space equipped with its Borel σ–field $\mathscr{X}$. Consider d sequences of observations $\boldsymbol{X}_i := (X_{i,j})_{j\geq 1}$, for $i = 1,\dots,d$, defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathbb{X}, \mathscr{X})$. They are partially exchangeable if and only if $(\boldsymbol{X}_1,\dots,\boldsymbol{X}_d) \stackrel{\mathrm{d}}{=} (\pi_1\boldsymbol{X}_1,\dots,\pi_d\boldsymbol{X}_d)$, where $\pi_i\boldsymbol{X}_i = (X_{i,\pi_i(j)})_{j\geq 1}$ and $\pi_1,\dots,\pi_d$ are finite permutations on $\mathbb{N}$. A representation theorem by B. de Finetti states that the sequences $\boldsymbol{X}_1,\dots,\boldsymbol{X}_d$ are partially exchangeable if and only if there exists a vector of random probability measures $(\tilde p_1,\dots,\tilde p_d)$, such that

$$
\begin{aligned}
(X_{1,j_1},\dots,X_{d,j_d}) \,|\, (\tilde p_1,\dots,\tilde p_d) &\stackrel{\text{iid}}{\sim} \tilde p_1\times\cdots\times\tilde p_d \qquad (j_1,\dots,j_d)\in\mathbb{N}^d\\
(\tilde p_1,\dots,\tilde p_d) &\sim Q_d. \qquad (1)
\end{aligned}
$$

*Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, 40126 Bologna, Italy. E-mail: federico.camerlenghi@unibo.it
†Department of Decision Sciences, BIDSA, 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

Having denoted by $\mathsf{P}_{\mathbb{X}}$ the space of all probability measures on $\mathbb{X}$, which is assumed to be endowed with the corresponding Borel σ–field $\mathscr{P}_{\mathbb{X}}$, then Q_d is a probability law on the space $(\mathsf{P}_{\mathbb{X}}^d, \mathscr{P}_{\mathbb{X}}^d)$.
The case where $d = 1$ corresponds to exchangeability and it has been extensively studied in the literature, starting from the seminal contribution of Ferguson (1973), who introduced the Dirichlet process. Subsequently, a wide variety of priors Q_1 have been introduced and investigated to model exchangeable observations, beyond the Dirichlet process. A popular generalization of the Dirichlet prior is the Pitman–Yor process (see Perman, Pitman and Yor (1992) and Pitman and Yor (1997)). Other general and flexible class of priors suited for the exchangeable case have been obtained by means of suitable transformations of completely random measures and are surveyed in Lijoi and Prünster (2010). Besides the random partition structure, posterior distributions and prediction rules induced by these nonparametric priors have been widely investigated and well understood.
On the other hand, when $d \geq 2$, the definition and investigation of Q_d in (1) is still the subject of a lively and fastly evolving literature. Starting from the seminal contribution of MacEachern (1999, 2000), the construction of dependent random probability measures has become a popular subject in the Bayesian nonparametric community. Although this line of research has been pursued by many authors in Bayesian nonparametrics, the derivation of theoretical properties in such a context is a hard task.
In the present paper we focus on a special class of dependent random probability measures, namely hierarchical Pitman–Yor processes. The hierarchical construction has been first introduced in Teh et al. (2006) for the Dirichlet process case, with other subsequent contributions in this direction that include Gasthaus and Teh (2010); Teh and Jordan (2010); Wood et al. (2011) and Nguyen (2016), among others. Hierarchical priors have proved to be effective tools in several applied areas, most notably in topic modeling problems, though the investigation of some of their most relevant distributional properties has been quite limited so far. Recent works by Camerlenghi et al. (2017) and Camerlenghi, Lijoi and Prünster (2017a) have successfully tackled the analytical hurdles related to hierarchical constructions in a partially exchangeable setting, deriving closed form expressions for the partition probability functions, the posterior distributions and the prediction rules when the $\tilde{p}_i$'s are suitable transformations of completely random measures. The present paper summarizes the most relevant findings corresponding to the Pitman–Yor process case. The proofs are not reported here, the interested reader may refer to Camerlenghi et al. (2017) for all the technical details and a more extensive treatment of the subject.

2. Hierarchical Pitman–Yor processes

For the reader's convenience, we remind that a Pitman–Yor (PY) process with parameters $\sigma \in (0,1), \vartheta > 0$ and base measure P_0 is a discrete random probability measure $\tilde{p} \equiv \sum_{j\geq 1} \tilde{\pi}_j \delta_{Z_j}$ 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 $(\mathbb{X}, \mathscr{X})$, with common distribution P_0, and the V_i's are independent Beta random variables with parameters $(\vartheta + i\sigma, 1-\sigma)$. Moreover, the sequences $(V_i)_{i\geq 1}$ and $(Z_i)_{i\geq 1}$ are independent. We will write $\tilde{p} \sim \mathrm{PY}(\sigma, \vartheta; P_0)$. The Dirichlet process may be recovered as a limiting case, when $\sigma \to 0$.

Having recalled the stick–breaking construction of the Pitman–Yor process, we now introduce hierarchies of Pitman–Yor random probability measures. The main idea behind the definition of hierarchical priors relies in randomizing the base measure of each random probability $\tilde{p}_i$ 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

$$\begin{aligned} \tilde{p}_i | \tilde{p}_0 &\overset{\text{iid}}{\sim} \mathrm{PY}(\sigma, \vartheta; \tilde{p}_0) \quad i = 1, \ldots, d \\ \tilde{p}_0 &\sim \mathrm{PY}(\sigma_0, \vartheta_0; P_0) \end{aligned} \tag{2}$$

being $\sigma, \sigma_0 \in (0,1)$, $\vartheta, \vartheta_0 > 0$ and P_0 is a non–atomic probability measure on $(\mathbb{X}, \mathscr{X})$. Note that the randomness of the base measure $\tilde{p}_0$ enables dependence across the different groups of observations. 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 $\tilde{p}_0$.
To get some intuition on the dependence structure induced by the hierarchical construction, one may first look for the correlation structure of the model, which is described by the following:

Theorem 1. *Consider a vector of random probability measures as in (2). Then, for any $A \in \mathscr{X}$ and $i \neq j$*

$$\mathrm{corr}(\tilde{p}_i(A), \tilde{p}_j(A)) = \left\{ 1 + \frac{1-\sigma}{1-\sigma_0} \, \frac{\vartheta_0 + \sigma_0}{\vartheta + 1} \right\}^{-1}. \tag{3}$$

We underline that the correlation (3) between different $\tilde{p}_i$'s is always positive and that this formula does not depend on the choice of the set A.
In the what follows we overview the relevant theoretical results concerning the HPYP, with a particular emphasis on the partition structure (Section 2.1), the posterior characterization of $(\tilde{p}_1, \ldots, \tilde{p}_d)$ (Section 2.3) and the distribution of the number of clusters out a sample of size n for the whole population (Section 2.2).

2.1 Random partition induced by HPYP

Assume to be provided with a sample $\boldsymbol{X}^{(n_i)} := (X_{i,1}, \ldots, X_{i,n_i})$ of size n_i for each population $i = 1, \ldots, d$, and denote by $n := \sum_{i=1}^d n_i$ the total number of observations. The discreteness of the different random probability measures in (2) allows for ties within the same sample and across different samples. Such ties induce a random partition and one is naturally led to determine its probability distribution, termed *partially exchangeable partition probability function* (pEPPF), which have been derived in Camerlenghi et al. (2017). In order to formalize the notion of pEPPF, suppose that he d samples $\{\boldsymbol{X}^{(n_i)} : i = 1, \ldots, d\}$ display k distinct values. 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 sample, namely $n_{i,j} \geq 0$ is the number of elements of the i–th sample that coincide with the j–th distinct among the k that have been overall recorded. 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 value does not appear in the sample $\boldsymbol{X}^{(n_i)}$. Having fixed the notation, the pEPPF is defined as follows

$$\Pi_k^{(n)}(\boldsymbol{n}_1,\dots,\boldsymbol{n}_d) = \mathbb{E}\int_{\mathbb{X}^k}\prod_{j=1}^k \tilde p_1^{n_{1,j}}(\mathrm{d}x_j)\cdots \tilde p_d^{n_{d,j}}(\mathrm{d}x_j). \tag{4}$$

In the case of hierarchical processes, the partition structure (4) may be interpreted in terms of the Chinese Restaurant Franchise (CRF) representation first introduced in Teh et al. (2006). According to this metaphor, $\boldsymbol{X}_i$ identifies the i–th Chinese restaurant in a franchise of d restaurants, all sharing the same menu. The samples $\boldsymbol{X}^{(n_i)}$ 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 $i=1,\dots,d$ and $j=1,\dots,k$.
The observed dishes' labels induce a random partition of $\{1,\dots,n\}$ that is characterized through (4). However, the evaluation of $\Pi_k^{(n)}$ is too difficult and, in order to obtain a tractable expression, one needs to introduce suitable latent variables $\boldsymbol{T}^{(n_i)} = (T_{i,1},\dots,T_{i,n_i})$, for each restaurant i, which represent the tables' labels where the people are seated at. 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 $\ell_{i,j}\in\{1,\dots,n_{i,j}\}$ distinct tables, the t–th of which has $q_{i,j,t}$ customers, for $t=1,\dots,\ell_{i,j}$. 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 $\boldsymbol{\ell}_i := (\ell_{i,1},\dots,\ell_{i,k})$ and $\boldsymbol{q}_{i,j} := (q_{i,j,1},\dots,q_{i,j,\ell_{i,j}})$, while $\boldsymbol{\ell} = (\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_d)$ and $\boldsymbol{q}$ denotes the overall tables frequencies. Finally $\bar\ell_{\bullet j} = \sum_{i=1}^d \ell_{i,j}$, $\bar\ell_{i\bullet} = \sum_{j=1}^k \ell_{i,j}$ denote the number of tables serving dish j and the overall number of tables, respectively, in restaurant i. The following representation of the pEPPF has been established in Camerlenghi et al. (2017):

Theorem 2. *Let $\{(X_{i,j})_{j\geq 1} : i=1,\dots,d\}$ be partially exchangeable as in (1), with Q_d characterized by*

$$\tilde p_i \,|\, \tilde p_0 \overset{\text{iid}}{\sim} \mathrm{PY}(\sigma,\vartheta;\tilde p_0) \quad (i=1,\dots,d), \qquad \tilde p_0 \sim \mathrm{PY}(\sigma_0,\vartheta_0;P_0)$$

Then

$$\begin{aligned}\Pi_k^{(n)}(\boldsymbol{n}_1,\dots,\boldsymbol{n}_d) = \sum_{\boldsymbol{\ell}}\sum_{\boldsymbol{q}} C(\boldsymbol{n}_1,\dots\boldsymbol{n}_d;\boldsymbol{\ell},\boldsymbol{q}) \frac{\prod_{r=1}^{k-1}(\vartheta_0+r\sigma_0)}{(\vartheta_0+1)_{|\boldsymbol{\ell}|-1}} \prod_{j=1}^k (1-\sigma_0)_{\bar\ell_{\bullet j}-1} \\ \times \prod_{i=1}^d \frac{\prod_{r=1}^{\ell_{i\bullet}-1}(\vartheta+r\sigma)}{(\vartheta+1)_{n_i-1}} \prod_{j=1}^k\prod_{t=1}^{\ell_{i,j}} (1-\sigma)_{q_{i,j,t}-1}\end{aligned} \tag{5}$$

where $(a)_n = \Gamma(a+n)/\Gamma(a)$ denotes the ascending factorial, with the convention $(a)_{-1}\equiv 1$, and we have set

$$C(\boldsymbol{n}_1,\dots\boldsymbol{n}_d;\boldsymbol{\ell},\boldsymbol{q}) := \prod_{i=1}^d\prod_{j=1}^k \frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1},\dots,q_{i,j,\ell_{i,j}}}.$$

As one may easily realize, the introduction of latent tables allows to get rid of the two sums in (5) as well as the related coefficient $C(\boldsymbol{n}_1,\dots\boldsymbol{n}_d;\boldsymbol{\ell},\boldsymbol{q})$, hence obtaining an expression of the pEPPF that may be more easily implemented (see e.g. Camerlenghi, Lijoi and Prünster (2017a)). More specifically, we would like to emphasize that the augmented version of the pEPPF given by

$$\frac{\prod_{r=1}^{k-1}(\vartheta_0+r\sigma_0)}{(\vartheta_0+1)_{|\boldsymbol{\ell}|-1}}\prod_{j=1}^{k}(1-\sigma_0)_{\bar{\ell}_{\bullet j}-1}\prod_{i=1}^{d}\frac{\prod_{r=1}^{\ell_{i\bullet}-1}(\vartheta+r\sigma)}{(\vartheta+1)_{n_i-1}}\prod_{j=1}^{k}\prod_{t=1}^{\ell_{i,j}}(1-\sigma)_{q_{i,j,t}-1}$$

describes the partition structure induced by the tables and dishes, while (5) corresponds to its marginalization over all the possible configurations of tables.

2.2 Distribution of the number of clusters

Having derived the partition structure of hierarchical Pitman–Yor processes, one is naturally led to determine the distribution of K_n, namely the random number of distinct values out of n partially exchangeable observations. In order to do this, let us define:

i) the number of distinct values K'_{i,n_i} in $\boldsymbol{T}^{(n_i)}=(T_{i,1},\dots,T_{i,n_i})$ for any $i=1,\dots,d$;

ii) $K_{0,t}$, the number of distinct values out of t exchangeable observations generated from $\tilde{p}_0$.

These quantities may be explained in terms of the Chinese restaurant franchise metaphor, indeed K'_{i,n_i} represents the number of distinct tables in restaurant i, while $K_{0,t}$ is the number of distinct dishes in the overall franchise. Camerlenghi et al. (2017) have derived the distribution of K_n:

Theorem 3. *Assume that K_n is the number of distinct values out of d partially exchangeable samples $\{\boldsymbol{X}^{(n_i)}: i=1,\dots,d\}$ governed by a vector of hierarchical Pitman–Yor processes, i.e. $\tilde{p}_i|\tilde{p}_0 \overset{\text{iid}}{\sim} \mathrm{PY}(\sigma,\vartheta;\tilde{p}_0)$ and $\tilde{p}_0\sim \mathrm{PY}(\sigma_0,\vartheta_0;P_0)$. Then, for any $k=1\dots,n$ one has*

$$\mathbb{P}[K_n=k]=\sum_{t=k}^{n}\mathbb{P}[K_{0,t}=k]\,\mathbb{P}\Big[\sum_{i=1}^{d}K'_{i,n_i}=t\Big]. \tag{6}$$

In the sequel we denote by $\mathscr{C}(n,k;\sigma)$ the generalized factorial coefficient, which is defined as

$$(\sigma t)_n=\sum_{k=1}^{n}\mathscr{C}(n,k;\sigma)(t)_k.$$

See also Charalambides (2005). Since we are working with Pitman–Yor processes, the expression in (6) may also be expressed as

$$\mathbb{P}[K_N=k]=\sum_{t=k}^{N}\frac{\prod_{r=1}^{k-1}(\vartheta_0+r\sigma_0)}{(\vartheta_0+1)_{t-1}}\frac{\mathscr{C}(t,k;\sigma_0)}{\sigma_0^k}\times\sum_{\{(\zeta_1,\dots,\zeta_d)\in\Delta_{d,t}\}}\prod_{i=1}^{d}\frac{\prod_{r=1}^{\zeta_i-1}(\vartheta+r\sigma)}{(\vartheta+1)_{n_i-1}}\frac{\mathscr{C}(n_i,\zeta_i;\sigma)}{\sigma^{\zeta_i}}$$

where $\Delta_{d,t}=\{(r_1,\dots,r_d): r_i\ge 1,\ \sum_{i=1}^{d}r_i=t\}$. In addition, from the representation (6), one may immediately deduce that

$$K_n\overset{d}{=}K_{0,K'_{1,n_1}+\dots+K'_{d,n_d}},$$

whose validity is useful to understand the asymptotic behavior of K_n. In particular Camerlenghi et al. (2017) have proven the following asymptotic result for K_n:

Theorem 4. *Suppose K_n is the number of distinct values in the d partially exchangeable samples $\{\boldsymbol{X}^{(n_i)} : i = 1, \dots, d\}$ governed by a vector of hierarchical Pitman–Yor processes. Furthermore, let $n_1 = \cdots = n_d = n/d$. Then*

$$\lim_{n \to +\infty} \frac{K_n}{n^{\sigma\,\sigma_0}} = Z$$

almost surely, where Z is some positive random variable.

2.3 Posterior characterization of the HPYP

To conclude our theoretical analysis of hierarchical Pitman–Yor processes, we focus on the posterior distribution of $(\tilde{p}_1, \dots, \tilde{p}_d)$ given the observations $\boldsymbol{X}$ and the latent tables $\boldsymbol{T}$ of the whole franchise. Such a characterization has been first derived in Camerlenghi et al. (2017) and no results were available before that. We underline the importance of such a characterization, indeed it allows to make inference on non–linear functionals of the hierarchical random probability measures.
To fix the notation, we set $k_i := \#\{j : n_{i,j} \geq 1\}$ and we agree on the fact that the Dirichlet distribution having parameters $(n_{i,1} - \ell_{i,1}\sigma, \dots, n_{i,k} - \ell_{i,k}\sigma, \vartheta + k_i\sigma)$ is defined on the k_i–dimensional simplex, after the removal of the parameters with $n_{i,j} = 0$. Below we report the posterior characterization proved by Camerlenghi et al. (2017).

Theorem 5. *The posterior distribution of $\tilde{p}_0$, conditional on $(\boldsymbol{X}, \boldsymbol{T})$, equals the distribution of the random probability measure*

$$\sum_{j=1}^{k} W_j \delta_{X_j^*} + W_{k+1}\, \tilde{p}_{0,k} \tag{7}$$

where $(W_1, \dots, W_k)$ is a k–variate Dirichlet random vector with parameters $(\bar{\ell}_{\bullet 1} - \sigma_0, \dots, \bar{\ell}_{\bullet k} - \sigma_0, \vartheta_0 + k\sigma_0)$, $W_{k+1} = 1 - \sum_{i=1}^{k} W_i$ and $\tilde{p}_{0,k} \sim \mathrm{PY}(\sigma_0, \vartheta_0 + k\sigma_0; P_0)$. Moreover, the posterior distribution of each $\tilde{p}_i$, conditional on $(\tilde{p}_0, \boldsymbol{X}, \boldsymbol{T})$, equals the distribution of the random probability measure

$$\sum_{j=1}^{k} W_{i,j}\, \delta_{X_j^*} + W_{i,k+1}\, \tilde{p}_{i,k} \tag{8}$$

where $(W_{i,1}, \dots, W_{i,k})$ is a k–variate Dirichlet random vector with parameters $(n_{i,1} - \ell_{i,1}\sigma, \dots, n_{i,k} - \ell_{i,k}\sigma, \vartheta + k_i\sigma)$, $W_{i,k+1} = 1 - \sum_{j=1}^{k} W_{i,j}$ and $\tilde{p}_{i,k} \,|\, \tilde{p}_0 \overset{\mathrm{ind}}{\sim} \mathrm{PY}(\sigma, \vartheta + k_i\sigma; \tilde{p}_0)$.

The posterior representations (7)–(8) resemble the ones derived by Pitman (1996) for the exchangeable case and feature the quasi–conjugacy property of the Pitman–Yor process.

3. Concluding remarks

In the present paper we have summarized the results proved by Camerlenghi et al. (2017) for hierarchical Pitman–Yor random probability measures, which are useful

dependent nonparametric priors to model partially exchangeable data. The theoretical results presented in Section 2 concern the partition structure (Section 2.1), the distribution of the number of distinct values out of n partially exchangeable observations (Section 2.2) and the posterior characterization of the hierarchical Pitman–Yor process (Section 2.3). All the proofs are based on the representation of the $\tilde{p}_i$'s as suitable transformations of completely random measures. We further underline that similar results, with respect to those presented here, hold true when the $\tilde{p}_i$'s are hierarchical normalized completely random measure: we refer to Camerlenghi et al. (2017) for all the details and proofs.
Our theoretical findings are of paramount importance to devise suitable sampling schemes in many applied problems. See, e.g., Camerlenghi, Lijoi and Prünster (2017a) for applications of these results to prediction in species sampling problems. Even if the interest towards hierarchical priors first arose in a partially exchangeable framework, one may employ hierarchical processes to model the *a priori* opinion in presence of exchangeable data: refer to Camerlenghi, Lijoi and Prünster (2017b) for theoretical investigations and noteworthy applications.

Acknowledgements

A. Lijoi and I. Prünster are partially supported by the European Research Council (ERC), StG "N-BNP" 306406 and by MIUR, PRIN Project 2015SNS29B.

References

Camerlenghi, F., Lijoi, A., Orbanz, P. and Prünster, I. (2017). Distribution theory for hierarchical processes. *Submitted*.

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

Camerlenghi, F., Lijoi, A. and Prünster, I. (2017b). Bayesian nonparametric inference beyond the Gibbs–type framework. *Submitted*.

Charalambides, C.A. (2005). *Combinatorial methods in discrete distributions*. Hoboken, NJ: Wiley.

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

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**.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, pp. 80–136. Cambridge University Press, Cambridge.

MacEachern, S.N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the SBSS*. Alexandria: American Statistical Association, 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.

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

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (T.S. Ferguson, L.S. Shapley and J.B. MacQueen, Eds.). IMS Lecture Notes Monogr. Ser., Vol. **30**. IMS, Hayward, 245–267.

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*, pp. 158-207, Cambridge Univ. Press, Cambridge.

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