Bayesian prediction with multiple-samples information *

Federico Camerlenghi[1,4], Antonio Lijoi[2,4] and Igor Prünster[3]

[1] Department of Statistics, University of Bologna, via delle Belle 41, 40126 Bologna, Italy.
[2] Department of Economics & Management, University of Pavia, via San Felice 5, 27100 Pavia, Italy.
[3]Department of Decision Sciences, BIDSA and IGIER, Bocconi University, via Röntgen 1, 20136 Milano, Italy.
[4] Collegio Carlo Alberto, Moncalieri, Italy.

Abstract

The prediction of future outcomes of a random phenomenon is typically based on a certain number of "analogous" observations from the past. When observations are generated by multiple samples, a natural notion of analogy is partial exchangeability and the problem of prediction can be effectively addressed in a Bayesian nonparametric setting. Instead of confining ourselves to the prediction of a single future experimental outcome, as in most treatments of the subject, we aim at predicting features of an unobserved additional sample of any size. We first provide a structural property of prediction rules induced by partially exchangeable arrays, without assuming any specific nonparametric prior. Then we focus on a general class of hierarchical random probability measures and devise a simulation algorithm to forecast the outcome of m future observations, for any $m \geq 1$. The theoretical result and the algorithm are illustrated by means of a real dataset, which also highlights the "borrowing strength" behavior across samples induced by the hierarchical specification.

Keywords: Bayesian Nonparametrics, Hierarchical processes, Partial exchangeability, Prediction, Pitman–Yor process, Species sampling models.
AMS Classification 2010: 62F15; 60G57; 62G05.

1 Introduction

A fundamental goal of statistics consists in predicting future outcomes of a certain experiment given *analogous observations* that have been recorded in the past. If the observed data are denoted as $X_1, \ldots, X_n$, one may be interested in predicting specific features related to a future sample $X_{n+1}, \ldots, X_{n+m}$ of size $m \geq 1$. Bruno de Finetti repeatedly emphasized in his writings the importance of prediction. For instance, in [8] he wrote "science cannot limit itself to theorize about accomplished facts but must foresee." Also in the philosophical debate prediction plays a dominant role. In the fundamental work of Rudolf Carnap [4], where he provides a taxonomy of the varieties of inductive inference he stresses that prediction is "the most important and fundamental kind of inductive inference." And, in fact, *singular predictive inference*, in which the additional

*A. Lijoi and I. Prünster are supported by the European Research Council (ERC) through StG "N-BNP" 306406. F. Camerlenghi is partially supported by an INdAM-GNAMPA Grant 2016.

future sample consists of just one individual (namely $m = 1$), represents only a special case of predictive inference. See [35] for a stimulating account.

Here, very much in de Finetti's spirit, we focus on prediction in its generality therefore not confining ourselves to singular predictive inference but considering m-step ahead prediction, for any $m \geq 1$. In order to face the problem of prediction, the observations $X_1, X_2, \ldots$ need to satisfy some symmetry condition that allows one to consider them as being *analogous*. In a Bayesian nonparametric context such a mathematical hypothesis corresponds to *exchangeability* (see [9]) and defines quite a general kind of dependence across data. A random infinite sequence $(X_n)_{n\geq 1}$ is exchangeable when its distribution is invariant under the group of all finitary permutations. This means that $(X_n)_{n\geq 1} \stackrel{\mathrm{d}}{=} (X_{\pi(n)})_{n\geq 1}$, where $\pi : \mathbb{N} \to \mathbb{N}$ is any permutation such that $\pi(j) = j$ for any $j \geq N$, for some $N \geq 1$, i.e., π permutes an arbitrary but finite number of indices of the sequence. Equivalently, homogeneity across data is coded through an invariance property with respect to the order with which the observations are recorded. Prediction within an exchangeable setting has become routine in a number of applications where data display a clustering structure such as, e.g., in ecology, genomics, linguistics, topic modeling, and analysis of network data. The probabilistic investigations that have had a more direct impact in the area are exhaustively accounted for in [31], whereas statistical contributions are reviewed in [7, 21, 26].

However, in a large variety of applications exchangeability is quite restrictive an assumption. An interesting example is where data are originated from d different, though related, experiments that identify d sequences $(X_{1,n})_{n\geq 1}, \ldots, (X_{d,n})_{n\geq 1}$. In such situations, the homogeneity assumption may reasonably hold within each experimental condition $(X_{i,n})_{n\geq 1}$, though not across different experimental conditions $(X_{i,n})_{n\geq 1}$ and $(X_{j,n})_{n\geq 1}$, where $i \neq j$. In these cases, one may rely on a more general and appropriate form of dependence such as *partial exchangeability*. See [10].

In this paper the problem of prediction within the general framework of partially exchangeable observations is addressed and some properties of the induced prediction rules are investigated. Section 2 provides a brief introduction to the goals pursued in the paper. The main result is contained in Section 3, where we show a structural predictive property of partial exchangeability, which holds true without any specific assumptions on the prior distribution. In Section 5 an algorithm to predict future observations is devised for the hierarchical Pitman–Yor process defined in Section 4. Finally we face the problem of prediction for species sampling problems as an illustration of the theoretical results. Moreover, we compare prediction corresponding to independent exchangeable samples with prediction in the partially exchangeable framework. The latter is discussed in the case of multiple samples and showcases a significant *borrowing strength* phenomenon.

2 Partial exchangeability and prediction

Suppose that $\mathbb{X}$ is a Polish space, associated with the possible outcomes of an experiment, and let $\mathscr{X}$ be the corresponding Borel σ-algebra. Moreover, $\mathsf{P}_{\mathbb{X}}$ is the set of all probability measures on $(\mathbb{X}, \mathscr{X})$, which is assumed to be endowed with the topology of weak convergence so that $\mathscr{P}_{\mathbb{X}}$ is its Borel σ-algebra. Consider, now, an array of d sequences of observations $\boldsymbol{X} = \{(X_{i,j})_{j\geq 1} : \, i = 1, \ldots, d\}$ defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathbb{X}, \mathscr{X})$. They are partially exchangeable if for any choice of finite permutations $\pi_1, \ldots, \pi_d$ of $\mathbb{N}$ one has

$$\{(X_{i,j})_{j\geq 1} : \, i = 1, \ldots, d\} \stackrel{\mathrm{d}}{=} \left\{(X_{i,\pi_i(j)})_{j\geq 1} : \, i = 1, \ldots, d\right\}.$$

The analogue of de Finetti's representation theorem for the partial exchangeable case states that $\boldsymbol{X}$ is partially exchangeable if and only if

$$\mathbb{P}\Big\{\bigcap_{i=1}^{d}(\boldsymbol{X}_i^{(n_i)} \in A_i)\Big\} = \int_{\mathsf{P}_{\mathbb{X}}^d}\prod_{i=1}^{d} p_i^{(n_i)}(A_i)\, Q_d(\mathrm{d}p_1, \ldots, \mathrm{d}p_d) \tag{1}$$

for any integer $n_i \geq 1$ and $A_i \in \mathscr{X}^{n_i}$, where $\boldsymbol{X}_i^{(n_i)} = (X_{i,1}, \ldots, X_{i,n_i})$ and $p^{(q)} = p \times \cdots \times p$ is the q-fold product measure on $\mathbb{X}^q$, for any $q \geq 1$. In (1) Q_d is termed the *de Finetti measure* of the sequence.

In a Bayesian framework, partial exchangeability is usually rephrased in terms of random probability measures; consider a vector of dependent random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_d)$ having law Q_d, then (1) may be expressed as:

$$\begin{gathered}(X_{1,j_1}, \ldots, X_{d,j_d}) \,|\, (\tilde{p}_1, \ldots, \tilde{p}_d) \overset{\text{iid}}{\sim} \tilde{p}_1 \times \cdots \times \tilde{p}_d, \quad (j_1, \ldots, j_d) \in \mathbb{N}^d \\ (\tilde{p}_1, \ldots, \tilde{p}_d) \overset{\text{iid}}{\sim} Q_d\end{gathered} \tag{2}$$

It is apparent that Q_d acts as a prior distribution on $\mathsf{P}_{\mathbb{X}}^d$ and is the starting point for the determination of posterior inferences in a nonparametric setting.

The first contribution to the Bayesian nonparametric literature in this direction can be traced back to [6]. Nonetheless, the research in the area has experienced a significant boost only more recently, inspired by the seminal papers of S.N. MacEachern [22, 23]. Stimulating accounts can be found in [16, 26, 27]. Among the most recent interesting contributions we mention [1, 14, 15, 17, 18, 25, 28, 29, 36].

The key in this setting is the specification of Q_d, as it defines the dependence structure among $\tilde{p}_1, \ldots, \tilde{p}_d$ and, hence, determines the predictive distributions across different samples. A large portion of the contributions currently available on this topic work on multivariate extensions of celebrated priors used for exchangeable sequences such as, e.g., the Dirichlet process or the Pitman–Yor process.

We now consider the problem of m-step prediction in a partially exchangeable setting and highlight an interesting property of predictive rules that holds true for any choice of Q_d and emerges as a structural property of partial exchangeability. To be more precise, if we suppose that for each sample $i \in \{1, \ldots, d\}$ we have $n_i \geq 1$ observations $X_{i,1}, \ldots, X_{i,n_i}$, we would like to predict specific features of an additional sample $\boldsymbol{X}_i^{(m_i|n_i)} = (X_{i,n_i+1}, \ldots, X_{i,n_i+m_i})$ of size m_i, for any $m_i \geq 1$. In view of (2), one then has to determine

$$\begin{aligned}\mathbb{P}\Big\{\bigcap_{i=1}^{d}(\boldsymbol{X}_i^{(m_i|n_i)} \in A_i) \,\Big|\, \boldsymbol{X}_1^{(n_1)}, \ldots, \boldsymbol{X}_d^{(n_d)}\Big\} \\ = \int_{\mathsf{P}_{\mathbb{X}}^d}\prod_{i=1}^{d} p_i^{(m_i)}(A_i)\, Q_d\Big(\mathrm{d}p_1, \ldots, \mathrm{d}p_d \,\Big|\, \boldsymbol{X}_1^{(n_1)}, \ldots, \boldsymbol{X}_d^{(n_d)}\Big)\end{aligned}$$

for any $A_i \in \mathscr{X}^{m_i}$ and $i \in \{1, \ldots, d\}$, where $Q_d(\,\cdot \mid \boldsymbol{X}_1^{(n_1)}, \ldots, \boldsymbol{X}_d^{(n_d)})$ denotes the posterior distribution of $(\tilde{p}_1, \ldots, \tilde{p}_d)$. It is worth noting that if Q_d is such that $(\tilde{p}_1, \ldots, \tilde{p}_d)$ are independent, i.e., $Q_d(C_1, \ldots, C_d) = \prod_{i=1}^{d} \Lambda_i(C_i)$, for any choice of $A_1, \ldots, A_d$ in $\mathscr{P}_{\mathbb{X}}$, and each Λ_i is a probability

measure on $\mathscr{P}_{\mathbb{X}}$, then independence is preserved also a posteriori and

$$\begin{aligned}\mathbb{P}\Big\{\bigcap_{i=1}^{d}(\boldsymbol{X}_i^{(m_i|n_i)}\in A_i)\,\Big|\,\boldsymbol{X}_1^{(n_1)},\ldots,\boldsymbol{X}_d^{(n_d)}\Big\}&=\prod_{i=1}^{d}\int_{\mathsf{P}_{\mathbb{X}}}p^{(m_i)}(A_i)\,\Lambda_i\Big(\mathrm{d}p\,\Big|\,\boldsymbol{X}_i^{(n_i)}\Big)\\&=\prod_{i=1}^{d}\mathbb{P}\Big(\boldsymbol{X}_i^{(m_i|n_i)}\in A_i\,\Big|\,\boldsymbol{X}_i^{(n_i)}\Big).\end{aligned}\tag{3}$$

In this case, sample i, for $i\neq j$, has no effect on the prediction of future outcomes for sample j. The other extreme situation corresponds to Q_d degenerating on the diagonal $\{p_1=\cdots=p_d\}$, which corresponds to exchangeability across all samples that is

$$\mathbb{P}\Big\{\bigcap_{i=1}^{d}(\boldsymbol{X}_i^{(m_i|n_i)}\in A_i)\,\Big|\,\boldsymbol{X}_1^{(n_1)},\ldots,\boldsymbol{X}_d^{(n_d)}\Big\}=\int_{\mathsf{P}_{\mathbb{X}}}\prod_{i=1}^{d}p^{(m_i)}(A_i)\,Q\Big(\mathrm{d}p\,\Big|\,\boldsymbol{X}_1^{(n_1)},\ldots,\boldsymbol{X}_d^{(n_d)}\Big).\tag{4}$$

Interest typically lies in specific features of $\boldsymbol{X}_i^{(m_i|n_i)}$, for $i=1,\ldots,d$, which can be described in terms of suitable summaries of the predictive distribution. If the realizations of $\tilde{p}_1,\ldots,\tilde{p}_d$ are discrete probability measures, as for many popular nonparametric priors, there may be ties among data within the same sample and across different samples (with positive probability). Hence, the outcomes of the additional m_i-sample for population i can be either "new" values or values that have been already observed in $\boldsymbol{X}_i^{(n_i)}$ and/or in another sample $\boldsymbol{X}_j^{(n_j)}$ for $j\neq i$.

A natural quantity one is then interested in predicting is the expected proportion of elements in $\boldsymbol{X}_i^{(m_i|n_i)}$ yielding "new" or "old" distinct values. In Section 3 we show that such a quantity is constant as the size of the additional sample m_i varies and coincides with the singular predictions, i.e., $\mathbb{P}(X_{i,n_i+1}=\text{"new"}|\boldsymbol{X}_1^{(n_1)},\ldots,\boldsymbol{X}_d^{(n_d)})$ and $\mathbb{P}(X_{i,n_i+1}=\text{"old"}|\boldsymbol{X}_1^{(n_1)},\ldots,\boldsymbol{X}_d^{(n_d)})$, respectively. Importantly, such a result holds true whatever the choice of the prior distribution Q_d and represents a direct consequence of the partial exchangeability assumption giving a neat indication of the interaction among the populations summarized by a singular prediction alone.

3 A structural predictive property of partial exchangeability

Within the framework laid out in Section 2 let us assume, without loss of generality, $d=2$ to simplify the notation. A key quantity of interest in an additional sample $\boldsymbol{X}_i^{(m_i|n_i)}$, for $i=1,2$, is the number of observations that coincide with specific subsets of the observed data $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$. Define

$$A_{h,k}=\{X_{1,1},\ldots,X_{1,n_1}\}^h\cap\{X_{2,1},\ldots,X_{2,n_2}\}^k,\quad(h,k)\in\{0,1\}^2\tag{5}$$

where $B^1=B$ and $B^0=B^c$. Then $A_{1,1}$ is the set of values shared by the two samples, $A_{1,0}$ and $A_{0,1}$ the set of values appearing in the first (second) sample but not in the second (first) and $A_{0,0}$ the set of potential new values. Hence, one might be interested in estimating the number (or the proportion) of elements in $\boldsymbol{X}_i^{(m_i|n_i)}$ that belong to any of the four sets that one can identify in (5). For sample $s\in\{1,2\}$, these are defined as

$$L_{s,m}^{0,1}=\sum_{r=1}^{m}\mathbb{1}_{A_{0,1}}(X_{s,n_s+r}),\quad L_{s,m}^{1,0}=\sum_{r=1}^{m}\mathbb{1}_{A_{1,0}}(X_{s,n_s+r}),$$

$$L_{s,m}^{1,1} = \sum_{r=1}^{m} \mathbb{1}_{A_{1,1}}(X_{s,n_s+r}), \quad L_{s,m}^{0,0} = \sum_{r=1}^{m} \mathbb{1}_{A_{0,0}}(X_{s,n_s+r}),$$

where $\mathbb{1}_A$ is the indicator function of set A. The following result provides the corresponding posterior expectations.

Theorem 1. *Suppose* $\boldsymbol{X} = \{(X_{i,j})_{j\geq 1} : i = 1, 2\}$ *is partially exchangeable according to* (1)*. Then, for each sample* $s = 1, 2$ *and each partition set of the possible outcomes* $A_{h,k}$ *with* $(h, k) \in \{0, 1\}^2$*, one has*

$$\mathbb{E}(L_{s,m}^{h,k} \mid \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)}) = m\, \mathbb{P}(X_{s,n_s+1} \in A_{h,k} \mid \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)}). \tag{6}$$

Proof. We shall focus on $L_{1,m}^{0,1}$ and the proof for the other quantities involved follows in a straightforward way. First of all, note that by virtue of the partial exchangeability assumption, one has

$$\mathbb{P}(X_{1,n_1+i} \in A_{0,1} \mid \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)}) = \mathbb{P}(X_{1,n_1+1} \in A_{0,1} \mid \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)})$$

for any $i \geq 1$. Such an identity in distribution, conditional on $\boldsymbol{X}^{(n_1)}$ and $\boldsymbol{X}^{(n_2)}$, yields

$$\begin{aligned}\mathbb{E}(L_{1,m}^{0,1} \mid \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)}) &= \sum_{i=1}^{m} \mathbb{E}\{\mathbb{1}_{A_{0,1}}(X_{1,n_1+i}) \mid \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)}\} \\ &= \sum_{i=1}^{m} \mathbb{P}(X_{1,n_1+i} \in A_{0,1} \mid \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)})\end{aligned}$$

and (6) follows. □

The result in (6) entails that the prediction over an additional sample $X_{1,n_1+1}, \ldots, X_{1,n_1+m}$ of size m is linear in m when evaluating the number of observations that coincide with: (i) any of the distinct values specific to $\boldsymbol{X}_2^{(n_2)}$ and not shared by $\boldsymbol{X}_1^{(n_1)}$ (i.e., $L_{1,m}^{0,1}$); (ii) any of the distinct values specific to $\boldsymbol{X}^{(n_1)}$ and not shared by $\boldsymbol{X}_2^{(n_2)}$ (i.e., $L_{1,m}^{1,0}$); (iii) any of the distinct values shared by $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ (i.e., $L_{1,m}^{1,1}$); (iv) new distinct values that have been observed neither in $\boldsymbol{X}_1^{(n_1)}$ nor in $\boldsymbol{X}_2^{(n_2)}$ (i.e., $L_{1,m}^{0,0}$). The same conclusion obviously holds true for $X_{2,n_2+1}, \ldots, X_{2,n_2+m}$.

Phrased in different terms, the expected proportion of values that will belong to each of the $A_{h,k}$'s is constant with respect to additional sampling. This result is deep and trivial at the same time. Indeed, if one thinks about the independent and identically distributed case with a single sample, one has a binomial experiment and hence linearity in the number of trials, which is the size of the additional sample. Since independence does not hold true in this setting, the quantities in Theorem 1 represent generalizations of the binomial experiment with a more general form of dependence and dimensionality, the latter being meant as the number of populations. And (6) implies that with partial exchangeability (*a fortiori* in the exchangeable case) the expected number of observations in an additional sample, which replicate specific subsets of distinct values in the basic samples, is still linear and the slope is the corresponding singular predictive probability, i.e., at step $n_s + 1$. Such a result is due to the conditional identity in distribution, given the observed samples, and is a deep structural property implied by the partial exchangeability assumption.

It is apparent that linearity holds true whatever the choice of Q_d in (1). On the other hand, the choice of Q_d affects the structure of the singular predictive probability in (6). Consider sample 1. The discreteness of Q_d implies that the probability of re-observing values of sample 1, i.e.,

$\mathbb{P}(X_{1,n_1+1} \in A_{1,k} \,|\, \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)})$ for $k = 0, 1$, is strictly positive. The specific structure of Q_d determines the presence and, in the affirmative case, the intensity of the "borrowing strength" between the two exchangeable sequences, which is clearly desirable from an inferential point of view. Being interested in studying the dependence among different samples, in the following we will focus on priors Q_d such that the probability $\mathbb{P}(X_{1,n_1+1} \in A_{0,1} \,|\, \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)})$ is strictly positive. This is equivalent to a positive probability of ties across samples and, consequently, $\mathbb{P}(X_{1,n_1+1} \in A_{0,1} \,|\, \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)})$ provides a sort of quantification of the "borrowing strength" phenomenon. A more subtle display emerges from the probability of re-observing values present in both samples being higher than were they observed in sample 1 alone. A number of dependent processes $(\tilde{p}_1, \tilde{p}_2)$ yield such a property. Here we focus on a class of hierarchical random probability measures that naturally lend themselves to investigating the phenomenon we have been hinting at.

4 Hierarchical Pitman–Yor processes

Hierarchical processes have been first considered [34], where the popular Hierarchical Dirichlet process (HDP) has been introduced. The hierarchical construction induces dependence among the random probability measures in (2) through the base measure, which is taken to be random instead of being deterministic. Recall that a Dirichlet process with parameter $\vartheta > 0$ and base measure P_0, $\mathscr{D}(\vartheta; P_0)$, can be defined by means of a stick-breaking procedure as a discrete random probability measure $\tilde{p} \equiv \sum_{j\geq 1} \tilde{\pi}_j \delta_{Z_j}$ with

$$\tilde{\pi}_1 = V_1, \quad \tilde{\pi}_j = V_j \prod_{i=1}^{j-1} (1 - V_i) \text{ for } j \geq 2, \tag{7}$$

$(Z_j)_{j\geq 1}$ i.i.d. random variables on $(\mathbb{X}, \mathscr{X})$ with common distribution P_0 and the V_i's i.i.d. Beta random variables with parameters $(\vartheta, 1)$, namely $V_i \overset{\text{iid}}{\sim} \mathcal{B}(\vartheta, 1)$. Moreover, the sequences $(V_i)_{i\geq 1}$ and $(Z_i)_{i\geq 1}$ are independent. The HDP is defined as

$$\begin{aligned} \tilde{p}_i | \tilde{p}_0 &\overset{\text{iid}}{\sim} \mathscr{D}(\vartheta; \tilde{p}_0), \quad i = 1, \ldots, d \\ \tilde{p}_0 &\sim \mathscr{D}(\vartheta_0; P_0) \end{aligned} \tag{8}$$

where P_0 is a fixed non-atomic probability measure on $(\mathbb{X}, \mathscr{X})$. Such a model has been highly successful in topic modeling applications for classification of documents in a corpus at different levels.

A natural extension is obtained by replacing the Dirichlet process in (8) with a Pitman–Yor process [32] leading to a Hierarchical Pitman–Yor process (HPYP); see, e.g., [11, 33]. Also the Pitman–Yor process admits a stick-breaking representation as (7) with $V_i \overset{\text{ind}}{\sim} \mathcal{B}(\vartheta + i\sigma, 1 - \sigma)$ for $\sigma \in (0, 1)$ and $\vartheta > 0$. Pitman–Yor processes are more flexible than Dirichlet processes in that the probability of sampling a new value depends on the number of distinct observations in the sample and not only on the total number of data. This allows for richer predictive structures. See [7] for details. More recently hierarchical processes have been investigated from an analytical viewpoint in [3], where the authors derived the partition structure, the prediction rules and the posterior distribution of $(\tilde{p}_1, \ldots, \tilde{p}_d)$ for hierarchical structures, when $\tilde{p}_i$ are normalized random measures. In such a general context HDP and HPYP follow as particular cases.

If $\text{PY}(\sigma, \vartheta; P_0)$ denotes a Pitman–Yor process with parameters $\sigma \in (0, 1)$, $\vartheta > 0$ and base

measure P_0, we consider the following prior distribution Q_d for the model in (2)

$$\begin{aligned} \tilde{p}_i|\tilde{p}_0 &\overset{\text{iid}}{\sim} \text{PY}(\sigma, \vartheta; \tilde{p}_0), \quad i = 1, \ldots, d \\ \tilde{p}_0 &\sim \text{PY}(\sigma_0, \vartheta_0; P_0) \end{aligned} \tag{9}$$

being $\sigma, \sigma_0 \in (0,1)$, $\vartheta, \vartheta_0 > 0$ and P_0 is a non-atomic probability measure on $(\mathbb{X}, \mathscr{X})$. The random probability measures $\tilde{p}_1, \ldots, \tilde{p}_d$ in (9) are almost surely discrete, thus allowing 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, also referred to as *partially exchangeable partition probability functions* (pEPPF). This is of paramount importance to carry out posterior inference.

The problem has been successfully addressed in [3] for a broad and general class of hierarchical priors. Here we specialize the result of [3] to the case of Pitman–Yor priors. Moreover, let $d = 2$ for the sake of illustration. The partition structure may be better explained in terms of a culinary metaphor, known as the Chinese Restaurant Franchise (CRF) representation and introduced by [34]. There are $d = 2$ restaurants sharing the same menu, the two samples $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ represent the dishes' labels eaten by the $n_1 + n_2$ customers of the overall franchise. Discreteness entails that the number of distinct dishes being tasted in the whole franchise is $k \in \{1, \ldots, n_1 + n_2\}$ and we let $X_1^*, \ldots, X_k^*$ denote their respective labels. We suppose that in restaurant i there are $n_{i,j} \geq 0$ customers eating dish j, and the frequencies are reported in the vector $\boldsymbol{n}_i \equiv (n_{i,1}, \ldots, n_{i,k})$, for $i = 1, 2$. Each table is served the same dish, chosen by the first seated customer, and the same dish can be served at different tables within the same restaurant or across different restaurants.

In order to obtain a more tractable expression of the pEPPF, one needs to introduce sets of latent variables $\boldsymbol{T}_i^{(n_i)} = (T_{i,1}, \ldots, T_{i,n_i})$, for $i = 1, 2$. Roughly speaking, and still in terms of the Chinese restaurant franchise, $T_{i,j}$ is the table where the jth customer is seated in restaurant i eating dish $X_{i,j}$. The introduction of latent tables corresponds to a refinement of the partition determined by data, in fact the $n_{i,j}$ customers eating dish j in restaurant i may be partitioned into $\ell_{i,j}$ tables, each one containing $q_{i,j,t}$ clients, for $t = 1, \ldots, \ell_{i,j}$. In particular we have that $n_{i,j} = \sum_{t=1}^{\ell_{i,j}} q_{i,j,t}$.

To fix the notation we write $\boldsymbol{\ell}_i \equiv (\ell_{i,1}, \ldots, \ell_{i,k})$, $\boldsymbol{q}_{i,j} \equiv (q_{i,j,1}, \ldots, q_{i,j,\ell_{i,j}})$, while $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$ and $\boldsymbol{q}$ denotes the overall tables frequencies, finally dots in the indexes denote that we are summing over that index, e.g., $\ell_{\bullet j} = \sum_{i=1}^{d} \ell_{i,j}$. If $\Phi_{k,0}^{(n)}$ and $\Phi_{k,i}^{(n)}$ denote the exchangeable partition probability functions (EPPFs) induced by $\tilde{p}_0$ and $\tilde{p}_i$, respectively, then

$$\Pi_k^{(n_1+n_2)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \boldsymbol{\ell}, \boldsymbol{q}) = \Phi_{k,0}^{(\ell_{\bullet\bullet})}(\ell_{\bullet 1}, \ldots, \ell_{\bullet k}) \prod_{i=1}^{2} \Phi_{\ell_{i\bullet},i}^{(n_i)}(\boldsymbol{q}_{i,1}, \ldots, \boldsymbol{q}_{i,k}). \tag{10}$$

See Theorem 1 in [3]. For the HPYP, (10) boils down to

$$\begin{aligned} &\Pi_k^{(n_1+n_2)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \boldsymbol{\ell}, \boldsymbol{q}) \\ &\quad = \frac{\prod_{r=1}^{k-1}(\theta_0 + r\sigma_0)}{(\theta_0+1)_{\ell_{\bullet\bullet}-1}} \prod_{t=1}^{k} (1-\sigma_0)_{\ell_{\bullet}-1} \frac{\prod_{r=1}^{\ell_{1\bullet}-1}(\theta + r\sigma)}{(\theta+1)_{n_1-1}} \prod_{v=1}^{k} \prod_{t=1}^{\ell_{1,v}} (1-\sigma)_{q_{1,v,t}-1} \\ &\qquad \times \frac{\prod_{r=1}^{\ell_{2\bullet}-1}(\theta + r\sigma)}{(\theta+1)_{n_2-1}} \prod_{v=1}^{k} \prod_{t=1}^{\ell_{2,v}} (1-\sigma)_{q_{2,v,t}-1} \end{aligned} \tag{11}$$

with the convention $(1-\sigma)_{-1} \equiv 1$. The closed form expression displayed in (11) is a fundamental tool to derive the full conditional distributions of the Gibbs sampler described in Section 5. As for the actual determination of $\Pi_k^{(n_1+n_2)}$, a proof can be found in [3].

5 Algorithm for predictions

Once the probability distribution of the underlying partially exchangeable random partition has been determined through (10), one can address the issue of predicting m future outcomes of a certain experiments as mentioned in Section 2. To be more specific, conditional on observed data $\boldsymbol{X}_i^{(n_i)}$, interest lies in predicting specific features of additional and unobserved samples $\boldsymbol{X}_i^{(m_i|n_i)}$, for $i = 1, 2$. Had one solely been interested in estimating $L_{s,m}^{i,j}$, in view of Theorem 1 it would have been enough to determine the single-step prediction ($m = 1$) and obtain the estimate for a general m by linearity. However, since we also aim at identifying highest posterior density (HPD) regions of $L_{s,m}^{i,j}$, a general m-step prediction algorithm is mandatory (of which $m = 1$ represents a simple special case). Furthermore, the simulation of realizations of $\boldsymbol{X}_i^{(m_i|n_i)}$, for $i = 1, 2$, is of interest if one is also willing to infer other quantities of interest in species sampling problems such as, e.g., the number of new species that will be observed or the so-called discovery probability. See [19]. Finally, note that despite the presentation concerns the multiple-samples case, obvious modifications allow one to devise an algorithm for the exchangeable case ($d = 1$).

Our goal is to generate samples $X_{1,n_1+1}, \ldots, X_{1,n_1+m_1}$ and $X_{2,n_2+1}, \ldots, X_{2,n_2+m_2}$, conditional on $\boldsymbol{X}^{(n_1)}$ and $\boldsymbol{X}^{(n_2)}$, for any two positive integers m_1 and m_2. In order to employ (11) one needs to introduce $n_1 + m_1 + n_2 + m_2$ latent variables $T_{1,1}, \ldots, T_{1,n_1+m_1}$, $T_{2,1}, \ldots, T_{2,n_2+m_2}$, which are the labels identifying the tables at which the different costumers are seated in the restaurants. If the additional $m = m_1 + m_2$ data induce j new distinct observations not included in $\boldsymbol{X}^{(n_1)}$ and $\boldsymbol{X}^{(n_2)}$, the determination of the full conditionals follows immediately from the pEPPF, which is easily deduced from (11). One finds

$$
\begin{aligned}
&\Pi_k^{(n_1+n_2+m)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \boldsymbol{\ell}, \boldsymbol{q}) \\
&\quad = \Phi_{k+j,0}^{(\ell_{\bullet\bullet})}(\ell_{\bullet 1}, \ldots, \ell_{\bullet k+j}) \prod_{i=1}^{2} \Phi_{\ell_{i\bullet},i}^{(n_i+m_i)}(\boldsymbol{q}_{i,1}, \ldots, \boldsymbol{q}_{i,k+j}) \\
&\quad = \frac{\prod_{r=1}^{k+j-1}(\theta_0 + r\sigma_0)}{(\theta_0+1)_{\ell_{\bullet\bullet}-1}} \prod_{t=1}^{k+j} (1-\sigma_0)_{\ell_{\bullet t}-1} \frac{\prod_{r=1}^{\ell_{1\bullet}-1}(\theta + r\sigma)}{(\theta+1)_{n_1+m_1-1}} \prod_{v=1}^{k+j} \prod_{t=1}^{\ell_{1,v}} (1-\sigma)_{q_{1,v,t}-1} \\
&\qquad \times \frac{\prod_{r=1}^{\ell_{2\bullet}-1}(\theta + r\sigma)}{(\theta+1)_{n_2+m_2-1}} \prod_{v=1}^{k+j} \prod_{t=1}^{\ell_{2,v}} (1-\sigma)_{q_{2,v,t}-1}.
\end{aligned} \tag{12}
$$

Based on (12) one can devise a Gibbs sampler that generates $(T_{i,1}, \ldots, T_{i,n_i})$, for $i = 1, 2$, and $(X_{i,n_i+r}, T_{i,n_i+r})$, for $r = 1, \ldots, m_i$ and $i = 1, 2$, from their respective full conditionals. Details are provided for $i = 1$, the case $i = 2$ being identical with the appropriate adaptations. If V is a variable that is a function of $(T_{i,1}, \ldots, T_{i,n_i+m_i})$ and of $(X_{i,n_i+1}, \ldots, X_{i,n_i+m_i})$, use V^{-r} to denote the generic value of the variable V after removal of $T_{i,r}$, for $r = 1, \ldots, n_i$, and of $(X_{i,r}, T_{i,r})$, for $r = n_i + 1, \ldots, n_i + m_i$.

(1) At $t = 0$, start from an initial configuration $X_{i,n_i+1}^{(0)}, \ldots, X_{i,n_i+m_i}^{(0)}$ and $T_{i,1}^{(0)}, \ldots, T_{i,n_i+m_i}^{(0)}$, for $i = 1, 2$.

(2) At iteration $t \geq 1$,

(2.a) With $X_{1,r} = X_h^*$ generate latent variables $T_{1,r}^{(t)}$, for $r = 1, \ldots, n_i$, from

$$\mathbb{P}(T_{1,r} = \text{“new”} | \cdots) \propto w_{h,r} \, \frac{(\theta + \ell_{1\bullet}^{-r}\sigma)}{(\ell_{\bullet\bullet}^{-r} + \theta_0)},$$

$$\mathbb{P}(T_{1,r} = T_{1,h,\kappa}^{*,-r} | \cdots) \propto (q_{1,h,\kappa}^{-r} - \sigma) \quad \text{for } \kappa = 1, \ldots, \ell_{1,h}^{-r},$$

where $w_{h,r} = \ell_{\bullet h}^{-r} - \sigma_0$ if $\ell_{\bullet h}^{-r} > 0$ and $w_{h,r} = 1$ otherwise. Moreover, $T_{1,h,1}^{*,-r}, \ldots, T_{1,h,\ell_{1,h}^{-r}}^{*,-r}$ are the tables at the first restaurant where the hth dish is served, after the removal of $T_{1,r}$.

(2.b) For $r = 1, \ldots, m_1$, generate $(X_{n_i+r}^{(t)}, T_{n_i+r}^{(t)})$ from the following predictive distributions

$$\mathbb{P}(X_{1,r} = \text{“new”}, T_{1,r} = \text{“new”} | \cdots) = \frac{(\theta_0 + (k + j^{-r})\sigma_0)}{(\theta + n_1 + m_1 - 1)} \frac{(\theta + \ell_{1\bullet}^{-r}\sigma)}{(\theta_0 + \ell_{\bullet\bullet}^{-r})}$$

while, for any $h = 1, \ldots, k + j^{-r}$ and $\kappa = 1, \ldots, \ell_{1,h}^{-r}$,

$$\mathbb{P}(X_{1,r} = X_h^{*,-r}, T_{1,r} = \text{“new”} | \cdots) = \frac{(\ell_{\bullet h}^{-r} - \sigma_0)}{(\theta + n_1 + m_1 - 1)} \frac{(\theta + \ell_{1\bullet}^{-r}\sigma)}{(\theta_0 + \ell_{\bullet\bullet}^{-r})},$$

$$\mathbb{P}(X_{1,r} = X_{1,h}^{*,-r}, T_{1,r} = T_{1,h,\kappa}^{*,-r} | \cdots) = \frac{q_{1,h,\kappa}^{-r} - \sigma}{\theta + n_1 + m_1 - 1}.$$

6 Illustrations

We are now ready to apply the algorithm devised in Section 5 to face prediction in species sampling problems. To this end, we assume that the observations originate from d different populations of individuals that can be grouped into classes identified by different types or species. One can think, for example, of data related to communities of plants or animals from different species in unknown proportions. In this case, $\tilde{p}_i$ in (2) is the distribution of the species in the ith community and similarities between communities motivate dependence across the $\tilde{p}_i$'s.

Here we focus on the analysis of genomic data known as Expressed Sequence Tags (ESTs). These are generated by partially sequencing randomly isolated gene transcripts that have been converted into cDNA. In very simplified terms, ESTs are tool for gene identification and an EST sample of size n consists of K_n distinct genes, with expression levels, i.e., frequencies, $N_1, \ldots, N_{K_n}$, where $N_1 + \cdots + N_{K_n} = n$. A large amount of literature, both frequentist and Bayesian, has been developed for addressing prediction problems related to exchangeable data in several application areas, most notably in Ecology, Biology and Genomics. ESTs represent an important instance of genomics application.

Given a basic observed sample of size n and a potential additional sample of size m various types of prediction problems can be addressed. For instance, one may consider estimation of the number of new genes arising in the additional sample of size m or the m-discovery probability, which is the probability of discovering a new gene at the $(n + m + 1)$th draw, without having observed the additional sample of size m. The frequentist approach dates back to the pioneering contributions of Good and Toulmin [12, 13] and has seen countless contributions since then. Among them we

mention [2, 5, 24, 30] and references therein. A Bayesian nonparametric approach to this type of prediction problems in the exchangeable setting was first proposed in [19] and developments to date are accounted for in [7]. A method for comparing and testing across different EST libraries is set forth in [20]. However, species' prediction problems within a rigorous partially exchangeable framework have not been considered in the literature yet.

Here we drop the exchangeability assumption and address prediction in the more general and realistic framework of multiple populations. We consider two different cDNA libraries of fruits of a *citrus clementina*, namely FlavFr1 and RindPdig24, which, for simplicity, we refer to as FRUIT 1 and FRUIT 2, respectively. The EST sample corresponding to FRUIT 1, $\boldsymbol{X}_1^{(n_1)}$, contains $n_1 = 1593$ ESTs with $K_{n_1} = 806$ distinct genes, whereas the sample corresponding to FRUIT 2, $\boldsymbol{X}_2^{(n_2)}$, is made of $n_2 = 900$ ESTs with $K_{n_2} = 687$ unique genes. Moreover, the two libraries share 183 distinct genes and, in particular, 520 and 317 ESTs of, respectively, FRUIT 1 and FRUIT 2 refer to these common genes. The details of the two EST samples and the sample obtained by merging the two are given in Table 1. These data are freely available at the website `http://www.ncbi.nlm.nih.gov/unigene/`.

Expression level	FRUIT 1	FRUIT 2	FRUITS
1	561	549	905
2	148	99	231
3	37	20	79
4	18	12	32
5	6	4	11
6	5		9
7	12	1	11
8	1	1	4
9	1		6
10	3	1	2
11	1		3
12	2		3
13			1
14	3		1
15	2		1
16	1		2
17			2
19	1		
20			1
22	1		
23	1		1
24			2
26			1
58	1		1
117	1		1
n	1593	900	2493
K_n	806	687	1310

Table 1: *Citrus clementina*: EST clustering profile of cDNA libraries of different fruits. FRUITS is FRUIT 1 + FRUIT 2.

Given EST data, the main inferential goal consists in prediction of the outcomes of additional sequencing, in our case from the two *clementina* libraries. More precisely, we focus on the number of genes coinciding with new values to be detected in an additional sample of size m, which can be distinguished into: (a) $L_{s,m}^{0,0}$ for $s = 1, 2$; (b) $L_{1,m}^{0,1}$ and $L_{2,m}^{1,0}$; (c) $L_{1,m}^{0} = L_{1,m}^{0,0} + L_{1,m}^{0,1}$ and

$L^0_{2,m} = L^{0,0}_{2,m} + L^{1,0}_{2,m}$. Recall that (a) and (b) were defined right before Theorem 1. Since closed form expressions for estimators of these quantities are not available under hierarchical nonparametric models (2), we approximate all the predictions by resorting to the algorithm described in Section 5. Indeed, at every iteration t, we generate the trajectory $X^{(t)}_{i,n_i+1}, \ldots, X^{(t)}_{i,n_i+m}$ in order to evaluate the quantities of interest. For example, we have

$$\hat{L}^0_{s,m} = \frac{1}{T} \sum_{t=1}^{T} \sum_{r=1}^{m} \mathbb{1}_{\{X_{s,1},\ldots,X_{s,n_s}\}^c}(X^{(t)}_{s,n_s+r})$$

for $s = 1, 2$ and $m \in \mathbb{N}$. Here we compare the predictions obtained in the simple exchangeable case, in which the quantities are estimated separately for the two datasets, with the results obtained in the partially exchangeable case. Note that the latter is a more natural choice, since the two libraries share a high number of genes, and the assumption of partial exchangeability triggers the *borrowing of strength* phenomenon across the libraries. The following numerical outputs are based on 10,000 iterations of the Gibbs sampler after 5,000 burn-in sweeps.

6.1 Independent exchangeable datasets

We first analyze the two libraries separately, which is equivalent to assuming independence of the $\tilde{p}_i$'s and the prediction rule takes on the form displayed in (3). The corresponding model specification is

$$\begin{aligned}
(X_{1,i}, X_{2,j}) \,|\, (\tilde{p}_1, \tilde{p}_2) &\overset{\text{iid}}{\sim} \tilde{p}_1 \times \tilde{p}_2, \\
(\tilde{p}_1, \tilde{p}_2) \,|\, (\tilde{p}_{1,0}, \tilde{p}_{2,0}) &\sim \mathrm{PY}(\sigma_1, \theta_1; \tilde{p}_{1,0}) \times \mathrm{PY}(\sigma_2, \theta_2; \tilde{p}_{2,0}), \\
(\tilde{p}_{1,0}, \tilde{p}_{2,0}) &\sim \mathrm{PY}(\sigma_{1,0}, \theta_{1,0}; P_0) \times \mathrm{PY}(\sigma_{2,0}, \theta_{2,0}; P_0)
\end{aligned}$$

and one can rely on a suitable adaptation of the algorithm devised in Section 5 in order to obtain approximation predictions. We also set independent non-informative priors for $(\sigma_{i,0}, \theta_{i,0})$ and (σ_i, θ_i), for $i = 1, 2$ given by

$$(\sigma_{i,0}, \sigma_i, \theta_{i,0}, \theta_i) \overset{\text{iid}}{\sim} \mathcal{U}(0,1) \times \mathcal{U}(0,1) \times \mathcal{G}(300, 5^{-1}) \times \mathcal{G}(300, 5^{-1}),$$

where $\mathcal{U}(0,1)$ stands for the uniform distribution on the interval $(0,1)$ and $\mathcal{G}(a,b)$ denotes the Gamma distribution with parameters (a,b); the values of $(\sigma_{i,0}, \sigma_i, \theta_{i,0}, \theta_i)$ are generated through a Metropolis–Hastings step. In other terms, the two samples are independent and inferences with data from one sample do not impact inferences concerning the other sample. Given that there are 183 shared observations, the independence assumption is quite restrictive but it serves as a useful exercise for drawing comparisons with the more appropriate partial exchangeability assumption. The estimators of $L^0_{s,m}$ for $s = 1, 2$ are summarized, for different sizes of the additional sample m, in Table 2. In accordance with Theorem 1, $\hat{L}_{1,m}$ increases linearly in m, with a slope which is larger for the second dataset FRUIT 2, consequence of a higher probability of sampling a new value at step $n+1$ for library 2. Finally, the posterior estimates for the parameters of the marginal PY processes are equal to

$$\begin{aligned}
(\hat{\theta}_{1,0}, \hat{\sigma}_{1,0}, \hat{\theta}_1, \hat{\sigma}_1) &= (1213.4, 0.4676, 1387.5, 0.0545), \\
(\hat{\theta}_{2,0}, \hat{\sigma}_{2,0}, \hat{\theta}_2, \hat{\sigma}_2) &= (1428.1, 0.2726, 1543.3, 0.6532).
\end{aligned} \tag{13}$$

	Citrus clementina: FRUIT 1		*Citrus clementina*: FRUIT 2	
m	$\hat{L}^0_{1,m}$	HPD (95%)	$\hat{L}^0_{2,m}$	HPD (95%)
200	68.21	(54, 83)	122	(106, 138)
400	136.21	(114, 159)	244	(219, 269)
600	204.28	(175, 235)	366	(331, 401)
800	272.34	(236, 310)	488	(444, 531)
1000	340.37	(297, 385)	610	(557, 662)
1200	408.48	(358, 461)	731	(670, 792)
1400	476.51	(419, 536)	853	(783, 924)
1600	544.71	(481, 611)	975	(897, 1054)
1800	612.74	(542, 687)	1097	(1011, 1185)
2000	680.83	(604, 760)	1219	(1124, 1314)

Table 2: Posterior expected number of new ESTs with corresponding 95% highest posterior density intervals for FRUIT 1 and FRUIT 2 in the independent exchangeable setting for the HPYP.

	Citrus clementina: FRUIT 1		*Citrus clementina*: FRUIT 2	
m	$\hat{L}^0_{1,m}$	HPD (95%)	$\hat{L}^0_{2,m}$	HPD (95%)
200	53.21	(41, 66)	85.32	(71, 100)
400	106.45	(87, 126)	170.70	(149, 192)
600	159.69	(136, 184)	256.04	(228, 284)
800	212.99	(185, 242)	341.37	(307, 376)
1000	266.13	(233, 300)	426.74	(387, 467)
1200	319.28	(282, 357)	512.13	(467, 558)
1400	372.45	(331, 414)	597.53	(547, 648)
1600	425.78	(380, 473)	682.86	(627, 739)
1800	479.01	(429, 530)	768.17	(708, 829)
2000	532.37	(479, 587)	853.59	(788, 920)

Table 3: Posterior expected number of new ESTs with corresponding 95% highest posterior density intervals for FRUIT 1 and FRUIT 2 in the independent exchangeable setting for the HDP.

It is useful to briefly compare the results of HPYP with the more familiar HDP, which arises by setting $\sigma_i = \sigma_{i,0} = 0$ for $i = 1, 2$. The estimated values of $L^0_{s,m}$ for $s = 1, 2$ for the HDP are reported in Table 3. Not surprisingly, given the findings in [19], the Dirichlet process leads to strong underestimation. Clearly, if the HDP were the appropriate model to use in this case, the estimates of the σ parameters for the two samples would have all been close to 0, whereas it is clear from (13) that they are not.

6.2 Partially exchangeable samples

The presence of 183 shared genes across the two libraries indicates that a more elaborate model accounting for dependence among the two samples is appropriate. The hierarchical structure (2), with $d = 2$ and $\tilde{p}_1$, $\tilde{p}_2$ and $\tilde{p}_0$ as in (9), seems ideally suited to account for interactions among the

two samples. This corresponds to assuming the data to be exchangeable within each library and conditionally independent across the two libraries. Hence, the number of new genes to be detected in the additional sample for each library depends also on the sample of the other library. Moreover, there is a single shared set of parameter values $(\theta_0, \sigma_0, \theta, \sigma)$ for which we set independent priors as follows

$$(\sigma_0, \sigma, \theta, \theta_0) \sim \mathcal{U}(0,1) \times \mathcal{U}(0,1) \times \mathcal{G}(300, 5^{-1}) \times \mathcal{G}(300, 5^{-1}).$$

In a similar fashion as in the exchangeable framework, these parameters are generated through a Metropolis–Hastings step embedded within the Gibbs sampler. The generalized Blackwell–MacQueen urn scheme in Section 5, then, yields the simulated trajectories that are used to approximate posterior inferences. The relevant numerical summaries arising from the algorithm are reproduced in Table 4. The estimates of $L^{0,1}_{1,m}$ and $L^{1,0}_{2,m}$ show how many of the ESTs become "shared" as the size of the additional sample increases. For instance, after $m = 2000$ additional sequencing, we predict that in FRUIT 1 we will detect 144.67 ESTs originally observed only in FRUIT 2. By comparing $L^{0,1}_{1,m}$ and $L^{1,0}_{2,m}$ it is apparent that the rate of detection for new values specific to the FRUIT 1 sample in library FRUIT 2 is faster than vice versa. Also the number of new genes not previously recorded in any of the two samples, $L^{0,0}_{1,m}$ and $L^{0,0}_{2,m}$, is larger when sampling additional genes for the FRUIT 2 library.

	Citrus clementina: FRUIT 1				*Citrus clementina*: FRUIT 2			
m	$\hat{L}^{0,0}_{1,m}$	$\hat{L}^{0,1}_{1,m}$	$\hat{L}^{0}_{1,m}$	$\hat{L}^{0}_{1,m}$–HPD	$\hat{L}^{0,0}_{2,m}$	$\hat{L}^{1,0}_{2,m}$	$\hat{L}^{0}_{2,m}$	$L^{0}_{2,m}$–HPD
200	67.65	14.45	82.09	(68, 97)	82.88	25.34	108.22	(93, 123)
400	135.28	28.89	164.17	(142, 186)	165.84	50.78	216.62	(193, 240)
600	202.84	43.37	246.20	(218, 275)	248.67	76.25	324.91	(294, 355)
800	270.52	57.86	328.38	(293, 364)	331.66	101.63	433.29	(396, 470)
1000	338.19	72.33	410.51	(369, 452)	414.57	127.08	541.65	(498, 585)
1200	405.82	86.80	492.61	(445, 540)	497.42	152.55	649.97	(600, 699)
1400	473.41	101.27	574.68	(521, 628)	580.21	177.97	758.19	(702, 814)
1600	541.03	115.76	656.79	(597, 716)	663.09	203.40	866.50	(805, 927)
1800	608.67	130.24	738.91	(675, 804)	745.92	228.88	974.80	(906, 1042)
2000	676.25	144.67	820.92	(750, 891)	828.90	254.24	1083.14	(1009, 1157)

Table 4: *Citrus clementina*: posterior expected number of new ESTs and 95% highest posterior density intervals for the two libraries of fruits in the partially exchangeable framework for the HPYP.

In accordance with Theorem 1, the posterior estimates of the quantities of interest turn out to be linear in m. Thanks to Theorem 1 we also obtain estimates of the following one-step predictions probabilities

$$\mathbb{P}(X_{1,n_1+1} \in A_{0,1} \,|\, \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)}) \approx 0.0723,$$
$$\mathbb{P}(X_{2,n_2+1} \in A_{1,0} \,|\, \boldsymbol{X}_1^{(n_1)}, \boldsymbol{X}_2^{(n_2)}) \approx 0.127.$$

Hence, the slope of the linear estimator $\hat{L}^{0,1}_{2,m}$ is higher than that of $\hat{L}^{1,0}_{1,m}$. This is also apparent from Figure 1.

It is also interesting to compare Table 2 with Table 4. The appropriate quantities to focus on are $L^{0}_{s,m}$ from which the desired phenomenon of borrowing of strength is apparent. This is

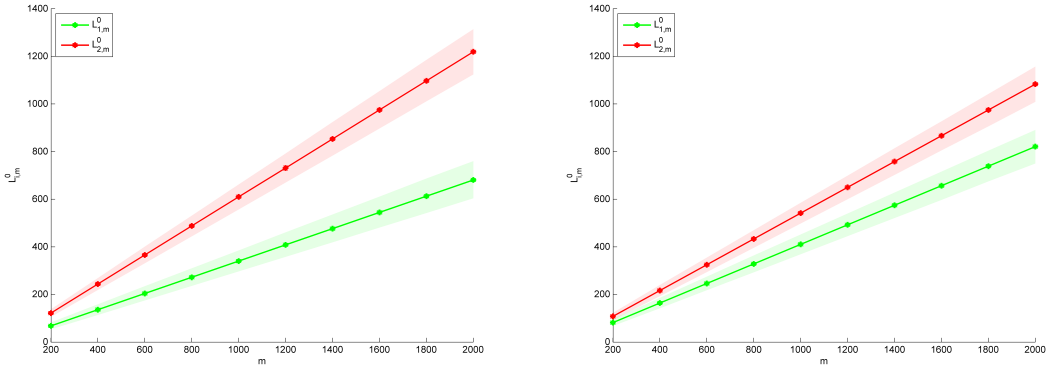


Figure 1: HPYP: total number of new ESTs $\hat{L}^0_{i,m}$ in the exchangeable (a) and partially exchangeable (b) settings as the size m of additional sample increases.

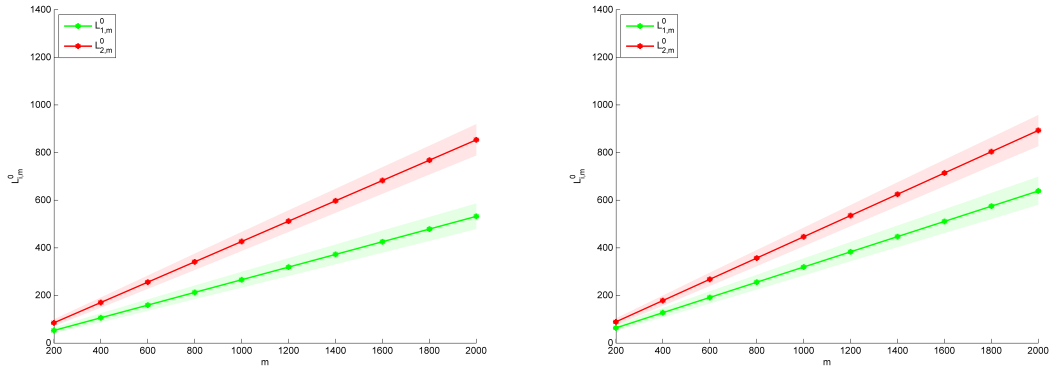
even more explicit in Figure 1, which depicts the posterior estimates of $L^0_{1,m}$ and $L^0_{2,m}$ as m increases both in the exchangeable (Figure 1(a)) and partially exchangeable (Figure 1(b)) settings. We may conclude that the discrepancies between $\hat{L}^0_{1,m}$ and $\hat{L}^0_{2,m}$ are much lower in the partially exchangeable case. Furthermore the 95% HPD intervals are significantly narrower for the partial exchangeable model, showing the beneficial influence of the borrowing of strength, which reduces the uncertainty about the estimates. Besides, the estimates of the model parameters equal

$$(\hat{\sigma}_0, \hat{\sigma}, \hat{\theta}, \hat{\theta}_0) = (0.3449, 0.5595, 1241.40, 1044.54). \tag{14}$$

Finally, we also consider the HDP case, which corresponds to $\sigma = \sigma_0 = 0$. The estimated quantities are reported in Table 5 and can be directly compared to those in Table 4 corresponding to the HPYP. Figure 2 displays the posterior estimates of $L^0_{1,m}$ and $L^0_{2,m}$, as m increases, for both the exchangeable (Figure 2(a)) and partially exchangeable (Figure 2(b)) settings. The previous considerations clearly apply also to the HDP. A first noteworthy, though not surprising, difference is that the rate of detection of new genes is much slower in the HDP case. Moreover, the shrinking phenomenon in the partially exchangeable setup is less evident for the HDP. This means that, besides the growth rate of new species being detected in additional samples, the key parameters σ and σ_0 have also a considerable effect on the intensity of the shrinkage phenomenon. Finally, one may argue as in Section 6.1 and note that there is no doubt about the HPYP yielding the better performance: if the HDP were the model to use, the posterior estimates of σ and σ_0 would have been close to 0, namely consistent with the HDP model, which is clearly not the case as the numerical values displyed in (14) illustrate.

	Citrus clementina: FRUIT 1				*Citrus clementina*: FRUIT 2			
m	$\hat{L}^{0,0}_{1,m}$	$\hat{L}^{0,1}_{1,m}$	$\hat{L}^{0}_{1,m}$	$\hat{L}^{0}_{1,m}$–HPD	$\hat{L}^{0,0}_{2,m}$	$\hat{L}^{1,0}_{2,m}$	$\hat{L}^{0}_{2,m}$	$L^{0}_{2,m}$–HPD
200	48.94	14.90	63.84	(51, 78)	62.54	26.83	89.37	(75, 104)
400	98.08	29.79	127.87	(108, 149)	125.10	53.66	178.76	(157, 200)
600	147.09	44.67	191.76	(166, 218)	187.46	80.53	267.00	(240, 296)
800	196.14	59.54	255.68	(225, 287)	249.81	107.34	357.17	(323, 391)
1000	245.13	74.41	319.55	(284, 356)	312.34	134.15	446.49	(407, 485)
1200	294.16	89.25	383.41	(344, 425)	374.73	160.99	535.71	(491, 580)
1400	343.22	104.04	447.26	(403, 494)	437.25	187.80	625.05	(575, 676)
1600	392.21	118.87	511.07	(462, 562)	499.90	214.54	714.43	(659, 770)
1800	441.33	133.72	575.05	(521, 631)	562.50	241.30	803.80	(744, 864)
2000	490.32	148.53	638.84	(581, 699)	625.06	268.09	893.14	(827, 958)

Table 5: *Citrus clementina*: posterior expected number of new ESTs and 95% highest posterior density intervals for the two libraries of fruits in the partially exchangeable framework for the HDP.



(a) (b)

Figure 2: HDP: total number of new ESTs $\hat{L}^0_{i,m}$ in the exchangeable (a) and partially exchangeable (b) settings as the size m of additional sample increases.

References

[1] Barrientos, A.F., Jara, A., and Quintana, F.A. (2016). Fully nonparametric regression for bounded data using dependent Bernstein polynomials. *J. American Statist. Assoc.*, doi: 10.1080/01621459.2016.1180987.

[2] Bunge, J., Willis, A. and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* **1**, 427–445.

[3] Camerlenghi, F., Lijoi, A., Orbanz, P. and Prünster, I. (2016). Distribution theory for hierarchical processes. *Submitted*.

[4] Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

[5] Chao, A. and Jost, L. (2012). Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. *Ecology* **93**, 2533–2547.

[6] Cifarelli, D. and Regazzini, E. (1978). Problemi statistici nonparametrici in condizioni di scambiabilità parziale. *Quaderni Istituto di Matematica Finanziaria, Università di Torino.*

[7] 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.

[8] de Finetti, B. (1931). Probabilismo. *Logos* **14**, 163–219. [Translated in *Erkenntnis* **31**, 169–223, 1989].

[9] de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, **7**, 1–68.

[10] de Finetti, B. (1938). Sur la condition d'équivalence partielle. *Actualités scientifiques et industrielles*, 5–18.

[11] Gasthaus, J. and Teh, Y.W. (2010). Improvements to the sequence memoizer. *Advances in Neuro Information Processing Systems* **23**. 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.

[12] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

[13] Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, **43**, 45–63.

[14] Griffin, J.E. and Leisen, F. (2016), Compound random measures and their use in Bayesian non-parametrics. *J. R. Stat. Soc. B.* doi:10.1111/rssb.12176

[15] Gutierrez, L., Mena, R.H. and Ruggiero, M. (2016). A time dependent Bayesian nonparametric model for air quality analysis *Computat. Statist. Data Anal.* **95**, 161–175.

[16] Hjort, N.L, Holmes, C., Müller, P., Walker, S.G. (Eds.) (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK.

[17] Huynh, V., Phung, D., Venkatesh, S., Nguyen, X., Hoffman, M. and Bui, H.H. (2016). Scalable nonparametric Bayesian multilevel clustering *Proceedings of the ICML 2014*.

[18] Jo, S., Lee, J., Müller, P., Quintana, F.A., and Trippa, L. (2016). Dependent species sampling models for spatial density estimation. *Bayesian Anal.*, in press.

[19] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering a new species *Biometrika*, **94**, 769–786.

[20] Lijoi, A, Mena, R.H. and Prünster, I. (2008). A Bayesian nonparametric approach for comparing clustering structures in EST libraries. *J. Comput. Biol.*, **15**, 1315–1327.

[21] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C. Müller, P., Walker, S.G. Eds.), pp. 80–136. Cambridge University Press, Cambridge.

[22] MacEachern, S.N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria: American Statistical Association, 50–55.

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

[24] Mao, C.X. (2004). Prediction of the conditional probability of discovering a new class. *J. Amer. Statist. Assoc.* **99**, 1108–1118.

[25] Mena, R.H. and Ruggiero, M. (2016) Dynamic density estimation with diffusive Dirichlet mixtures *Bernoulli* **22**, 901–926.

[26] Müller, P., Quintana, F.A., Jara, A., Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer, New York.

[27] Müller, P. and Quintana, F.A. (2004). Nonparametric Bayesian data analysis. *Statist. Science* **19**, 95–110.

[28] Nguyen, V., Phung, D., Nguyen, X., Venkatesh, S. and Bui, H.H. (2014). Bayesian nonparametric multilevel clustering with group-level contexts. *Proceedings of the ICML 2014*.

[29] Nguyen, X. (2016). Borrowing strength in hierarchical Bayes: posterior concentration of the Dirichlet base measure. *Bernoulli* **22**, 1535–1571.

[30] Orlitsky, A., Suresh, A.T. and Wu, Y. (2016). Optimal prediction of the number of unseen species. *PNAS*, **113**, 13283–13288.

[31] Pitman, J. (2006). *Combinatorial stochastic processes.* École d'été de probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875, Springer, New York.

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

[33] Teh, Y.W. (2006). A hierarchical Bayesian language model based on Pitman–Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 985–92. Morristown, NJ: Association for Computational Linguistics.

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

[35] Vickers, J. (2011). The Problem of Induction. In *The Stanford Encyclopedia of Philosophy* (Ed. E.N. Zalta).

[36] Zhu, W. and Leisen, F. (2015). A multivariate extension of a vector of two-parameter Poisson-Dirichlet processes. *J. Nonparam. Statist.* **27**, 89–105.