Journal of Statistical Planning and Inference 166 (2015) 14–23

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# Bayesian nonparametric inference for shared species richness in multiple populations

Sergio Bacallado [a], Stefano Favaro [b,1], Lorenzo Trippa [c,*,2]

[a] Department of Statistics, Stanford University, United States
[b] Department of Economics and Statistics, University of Torino, Italy
[c] Department of Biostatistics, Harvard University, United States

ARTICLE INFO

*Article history:*
Received 3 March 2014
Accepted 14 March 2014
Available online 9 July 2014

*Keywords:*
Bayesian nonparametrics
Ewens–Pitman sampling model
Partial exchangeability
Shared species richness
Species sampling problem
Two parameter Poisson–Dirichlet process
Urn sampling scheme

ABSTRACT

We introduce the branching Ewens–Pitman sampling model for dependent species sequences. The model defines random probability measures having marginally two-parameter Poisson–Dirichlet process distributions. These random measures are associated with the nodes of a binary tree which describes the strength of dependence of the resulting random partitions. We discuss Bayesian analysis under the introduced model and provide algorithms for posterior inference. The model is applied to evaluate similarities across microbial populations in the human esophagus.



## 1. Introduction

Classical species sampling problems are associated to situations where an experimenter is sampling from a population of individuals belonging to different species with unknown proportions. Species labels are denoted by $(X_i^*)_{i\geq 0}$ and their respective proportions in the population by $(q_i)_{i\geq 0}$. Given the information yielded by an initial sample, most of the statistical issues to be faced are related to the species richness, which can be quantified in different ways. For example, given an initial observed sample of size $n$, species richness might be measured by the estimated number of new distinct species, or distinct species with a certain frequency of interest, to be detected in an additional sample of size $m$. It can be alternatively evaluated by the probability of discovering a new species at the $(n+m+1)$-th draw, which yields the discovery rate as a function of $m$. Important contributions in this direction date back to the seminal papers by Good (1953), Good and Toulmin (1956) and Efron and Thisted (1976). These problems, along with many others in the field, have originated a rich and consolidated body of literature in which most of the models and techniques adopted rely on a frequentist approach, either parametric or nonparametric. The reader is referred to the comprehensive and stimulating reviews by Bunge and Fitzpatrick (1993) and Chao (2005).

* Corresponding author.
*E-mail addresses:* sergio.bacallado@gmail.it (S. Bacallado), stefano.favaro@unito.it (S. Favaro), ltrippa@jimmy.harvard.edu (L. Trippa).
1 Also affiliated to Collegio Carlo Alberto, Moncalieri, Torino.
2 Also affiliated to Dana–Farber Cancer Institute, Boston, Massachusetts, USA.

http://dx.doi.org/10.1016/j.jspi.2014.03.010
0378-3758/

A Bayesian nonparametric approach to species sampling problems has been proposed by Lijoi et al. (2007a). This approach is based on the randomization of the unknown species proportions $q_i$'s. Specifically, the data are assumed from an exchangeable sequence $(X_i)_{i\geq 0}$ directed by a discrete random probability measure $Q = \sum_{i\geq 0} q_i \delta_{X_i^*}$, where the nonnegative weights $q_i$'s sum up to one almost surely, and they are independent from the locations $X_i^*$'s. Hence, by de Finetti's representation theorem for exchangeable random variables,

$$\begin{aligned} X_i \mid Q &\overset{\text{iid}}{\sim} Q \quad i = 1, \ldots, n \\ Q &\sim \Pi \end{aligned} \tag{1}$$

for any $n \geq 1$, with $\Pi$ playing the role of the prior for Bayesian inference. Under the model (1), with $\Pi$ being in the class of Gibbs-type priors introduced by Gnedin and Pitman (2005), Lijoi et al. (2007a) derived estimators for some quantities of interest related to an additional unobserved sample $(X_{n+1}, \ldots, X_{n+m})$ conditionally on an observed initial sample $(X_1, \ldots, X_n)$. See, e.g., Lijoi et al. (2008), Favaro et al. (2012), Favaro et al. (2013) and Bacallado et al. (in press) for theoretical developments, and Lijoi et al. (2007b) for practitioner oriented illustrations. Other contributions to species sampling problems adopting a Bayesian approach can be found in, e.g., Barger and Bunge (2010), Christen and Nakamura (2003), Guindani et al. (2014), Trippa and Favaro (2001), Zhang and Stern (2009) and Bacallado et al. (2013).

In this paper we consider species sampling problems associated to situations where an experimenter is sampling from a finite collection of populations sharing species. For instance, in ecology these populations could represent candidate sites for conservation or restoration, or areas at different latitudes or elevation above the sea level, or the same area at two different times. In genetics multiple populations arise with complementary DNA libraries from different tissues. Given the information yielded by a collection of initial samples, one from each of the populations of interest, the main issue consists in comparing populations. A common approach to compare populations is to measure the extent of similarity, or dissimilarity, by means of the so-called overlap indexes. See, e.g., Gower (1985), Pielou (1975), Pielou (1977) and Ludwig and Reynolds (1988). Most of the commonly used measures of similarity are functions of the number of shared distinct species. Hence, an estimator of such a quantity plays a crucial role in comparing populations and forms the basis to construct various types of overlap indexes. So far estimators of the shared species richness have been derived under the frequentist nonparametric framework. See, e.g., Chao et al. (2000), Chao et al. (2006), Pan et al. (2009), Chao and Lin (2012), Yue et al. (2001), Mao and Lindsay (2004) and Yue and Clayton (2012). Here we present the first Bayesian nonparametric approach to the problem of estimating the shared species richness of a finite collection of populations.

Our approach is an extension, from a single to multiple populations, of the approach in Lijoi et al. (2007a). Let $\mathcal{Z}$ be a set of distinct symbols. We consider a finite number $p > 1$ of populations, each one indexed by a distinct element of a $\mathcal{Z}$. Moreover, let $\{(X_{z_i}^*)_{i\geq 0}, z \in \mathcal{Z}\}$ be an array of species labels, one for each of the populations of interest, and let $\{(q_{z_i})_{i\geq 0}, z \in \mathcal{Z}\}$ be their respective unknown proportions. Since populations may share species, we need to assume shared labels among the $p$ sequences $\{(X_{z_i}^*)_{i\geq 0}, z \in \mathcal{Z}\}$. As in Lijoi et al. (2007a), our approach is based on the randomization of the $q_{z_i}$'s. Specifically, the data are assumed from a partially exchangeable array $\{(X_{z_i})_{i\geq 0}, z \in \mathcal{Z}\}$ directed by a dependent discrete random probability measure $\mathbf{Q} = \{Q(z), z \in \mathcal{Z}\}$, where $Q(z) = \sum_{i\geq 0} q_{z_i} \delta_{X_{z_i}^*}$. Hence, by de Finetti's representation theorem for partially exchangeable random variables,

$$\begin{aligned} X_{z_i} \mid \mathbf{Q} &\overset{\text{iid}}{\sim} Q(z) \quad i = 1, \ldots, n_z \quad z \in Z \\ \mathbf{Q} &\sim \Pi, \end{aligned}$$

for any $n_z \geq 1$. In order to specify the distribution of $\{(X_{z_i})_{i\geq 0}, z \in \mathcal{Z}\}$, and hence the prior $\Pi$, we resort to an idea in Muliere et al. (2005). Specifically, let $T$ be a directed binary tree and let $(z_i)_{i\geq 0}$ be sequence of vertices along the $z$-th branch of $T$. We assume $\{(X_{z_i})_{i\geq 0}, z \in \mathcal{Z}\}$ to be an array of random variables indexed by the vertices of $T$ in such a way that: (i) along the $z$th branch of $T$ the sequence $(X_{z_i})_{i\geq 0}$ is exchangeable, for any $z \in \mathcal{Z}$; (ii) the de Finetti measure of the exchangeable sequence $(X_{z_i})_{i\geq 0}$ coincides with the distribution of the two parameter Poisson–Dirichlet process introduced in Perman et al. (1992), for any $z \in \mathcal{Z}$. The resulting $\{(X_{z_i})_{i\geq 0}, z \in \mathcal{Z}\}$ is a collection of $p$ dependent exchangeable sequences with the same marginal distribution, and the dependence is directed by the geometry of $T$. In particular, according to this construction, the number of shared distinct species among populations depends on the number of common vertices among the branches of $T$.

Interest in species sampling models has been ignited recently by a burgeoning community of biologists studying the human microbiome. These studies aim to characterize the diversity of microbial populations in our body, especially in the gut, and understand the effect of environmental factors such as diet and antibiotics on them. Microbial imbalances have been recently implicated in a range of autoimmune disorders, such as diabetes and rheumatoid arthritis, as well as obesity and even cognitive disorders, such as autism, schizophrenia and depression. See Cho and Blaser (2012). This field of research uses diverse technologies, from broad-range amplification of ribosomal DNA followed by high-throughput sequencing, to more sophisticated deep-sequencing techniques, which have in common that they produce discrete species sampling data. The analysis pipeline involves methods to estimate the underlying distribution of species in each population, exploratory approaches to contrast different populations, and hypothesis tests that establish a difference between populations. See, for instance, McMurdie and Holmes (2013). The difference between two microbial distributions is summarized by various beta-diversity functions, such as the Jensen–Shannon distance and the Bray–Curtis dissimilarity, or distances based on phylogeny such as UniFrac, a Wasserstein distance which uses a metric derived from a phylogenetic tree.