Journal of Multivariate Analysis 131 (2014) 79–98

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Posterior analysis of rare variants in Gibbs-type species sampling models

Oriana Cesari [a], Stefano Favaro [b,*,1], Bernardo Nipoti [b,1]

[a] Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy

[b] University of Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy

ARTICLE INFO

Article history:
Received 21 November 2012
Available online 26 June 2014

AMS 2000 subject classifications:
60G57
62G05
62F15

Keywords:
Bayesian nonparametric inference
Asymptotic credible intervals
Exchangeable random partition
Gibbs-type random probability measure
Index of diversity
Sampling formula
Species sampling problem
Rare variant
Two parameter Poisson–Dirichlet process

ABSTRACT

Species sampling problems have a long history in ecological and biological studies and a number of statistical issues, including the evaluation of species richness, are still to be addressed. In this paper, motivated by Bayesian nonparametric inference for species sampling problems, we consider the practically important and technically challenging issue of developing a comprehensive posterior analysis of the so-called rare variants, namely those species with frequency less than or equal to a given abundance threshold. In particular, by adopting a Gibbs-type prior, we provide an explicit expression for the posterior joint distribution of the frequency counts of the rare variants, and we investigate some of its statistical properties. The proposed results are illustrated by means of two novel applications to a benchmark genomic dataset.

1. Introduction

Suppose that statistical units drawn from a population are representative of different species. Their labels are denoted by $\hat{X}_i$ and their respective proportions in the population by $\tilde{p}_i$, for $i \geq 1$. Therefore, models for species sampling problems can be usefully embedded in the framework of discrete random probability measures, $\tilde{P} = \sum_{i\geq 1} \tilde{p}_i \delta_{\hat{X}_i}$, where δ_a denotes the point mass at a. Discrete random probability measures emerge as remarkable tools for theoretical and applied analysis in, e.g., population genetics, ecology, genomics, mathematical physics, machine learning. The most celebrated example of discrete random probability measure is the Dirichlet process introduced by Ferguson [14] and whose random masses $\tilde{p}_i$ are obtained either by normalizing the jumps of a Gamma completely random measure or by means of a stick-breaking procedure. This process has been also popularized under the name of (one parameter) Poisson–Dirichlet process and characterized in terms of the distribution of its ranked random masses by Kingman [22]. The reader is referred to Lijoi and Prünster [28] for an up-to-date review of classes of discrete random probability measures generalizing the Dirichlet process.

* Corresponding author.
E-mail addresses: oriana.cesari@carloalberto.org (O. Cesari), stefano.favaro@unito.it (S. Favaro), bernardo.nipoti@unito.it (B. Nipoti).
[1] Also affiliated to Collegio Carlo Alberto, Moncalieri, Italy.

http://dx.doi.org/10.1016/j.jmva.2014.06.017
0047-259X/

In this paper our attention will be focused on statistical issues related to species sampling problems: these will be addressed by a Bayesian nonparametric approach. We consider data from a population whose species composition is directed by a discrete random probability measure $\tilde{P}$ with distribution Π, i.e.

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

for any $n \geq 1$. According to de Finetti's representation theorem, $(X_i)_{i\geq 1}$ is exchangeable and Π takes on the interpretation of a prior distribution over the composition of the population. Since $\tilde{P}$ is discrete, we expect ties in a sample $(X_1, \ldots, X_n)$ from $\tilde{P}$. Precisely we expect $K_n \leq n$ distinct observations, or species, with frequencies $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$ such that $\sum_{1\leq i\leq K_n} N_i = n$. Accordingly, the sample induces a random partition of $\{1, \ldots, n\}$, in the sense that any index $i \neq j$ belongs to the same partition set if and only if $X_i = X_j$. We denote by $p_j^{(n)}(n_1, \ldots, n_j)$ the function corresponding to the probability of any particular partition of $\{1, \ldots, n\}$ having $K_n = j$ blocks with frequencies $\mathbf{N}_n = (n_1, \ldots, n_j)$. This function is known as the exchangeable partition probability function (EPPF), a concept introduced in Pitman [34] as a development of earlier results in Kingman [23]. See Pitman [36] for a comprehensive account on exchangeable random partitions.

Under the framework (1), with $\tilde{P}$ being in the class of the Gibbs-type random probability measures by Pitman [35], Lijoi et al. [25] introduced a novel Bayesian nonparametric methodology for making inferences on quantities related to an additional unobserved sample $(X_{n+1}, \ldots, X_{n+m})$ from $\tilde{P}$, given an observed sample $(X_1, \ldots, X_n)$. A particularly important example is represented by the estimation of the number of new species that will be observed in the additional sample. See Lijoi et al. [29], Favaro et al. [12], Favaro et al. [11] and Bacallado et al. [1] for estimators of other features related to species richness under Gibbs-type priors. This class of priors stands out for both mathematical tractability and flexibility. Indeed, apart from the Dirichlet process, the class of the Gibbs-type random probability measures includes as special cases the two parameter Poisson–Dirichlet process, also known as Pitman–Yor process, and the normalized generalized Gamma process. We refer to Perman et al. [33], Pitman and Yor [37] and Ishwaran and James [18] for details on the two parameter Poisson–Dirichlet process, and to James [19], Pitman [35], Lijoi et al. [27] and James [20] for details on the normalized generalized Gamma process. Gibbs-type priors also stand out for being particularly suited in the context of inferential problems with a large unknown number of species, which typically occur in several genomic applications. See, e.g., Lijoi et al. [26], Guindani et al. [16] and De Blasi et al. [5].

Motivated by the goal of performing Bayesian nonparametric inference for species sampling problems, in this paper we develop a comprehensive posterior analysis of the so-called rare variants, namely the species with frequency less than or equal to a given abundance threshold τ. Ecological and biological literature have always devoted special attention to rare variants. In ecology, for instance, conservation of biodiversity represents a fundamental theme and it can be formalized in terms of the number of species whose frequency is greater than a specified threshold; indeed, any form of management on a sustained basis requires a certain number of sufficiently abundant species, the so-called breeding stock. See, e.g., Usher [39] and Magurran [31] for detailed surveys on measurements of biodiversity, conservation of populations, commonness and rarity of species. On the other hand in genetics one is typically interested in the number of individuals with rare genes, the reasons being that rare genes of a specific type may be associated with a deleterious disease. See, e.g., Elandt-Johnson [7] and Laird and Lange [24] for a detailed account on the role of rare variants in genetics.

Under the statistical framework (1), with $\tilde{P}$ being a Dirichlet process, Joyce and Tavaré [21] first investigated the prior distribution of the rare variants, namely the joint distribution of the frequency counts of the rare variants induced by an initial sample $(X_1, \ldots, X_n)$ from $\tilde{P}$. In particular they mainly focused on the study of the asymptotic behavior, for a large sample size n, of such a prior distribution. In this paper we derive the prior distribution of the rare variants under the more general assumption of $\tilde{P}$ being a Gibbs-type random probability measure. Furthermore, following ideas set forth in Lijoi et al. [25], we derive and investigate the posterior distribution of the rare variants. Such a posterior distribution corresponds to the conditional joint distribution of the frequency counts of the rare variants induced by an additional sample $(X_{n+1}, \ldots, X_{n+m})$, given $(X_1, \ldots, X_n)$. Precisely, this is as a suitable convolution of: (i) the joint posterior distribution of the new rare variants that are generated from the additional sample and do not coincide with rare variants already detected in the initial sample; (ii) the joint posterior distribution of the old rare variants that arise by updating, via the additional sample, the rare variants already detected in the initial sample. Our distributional results are derived by generalizing some of the combinatorial techniques originally developed in Favaro et al. [12], where special cases of the results in this paper have been presented. After submitting the first version of this paper we learnt that the posterior distribution of the rare variants have been recently obtained independently, and by means of different techniques, in Cerquetti [2]. For additional distributional results on rare variants we refer to the M.Sc. Thesis of Cesari [3], from which the main contributions of the present papers are taken.

Our prior and posterior distributional results admit several applications, not necessarily related to the study of the rare variants. In this paper we focus on two representative applications, which will be illustrated under the assumption of a two parameter Poisson–Dirichlet prior. Firstly, we devise a novel methodology to approximately quantify the uncertainty of a Bayesian nonparametric estimator for the number of rare species. This estimator has been recently introduced in Favaro et al. [12] and the problem of evaluating its accuracy is of great importance in several applied contexts. To this end, we