Journal of Statistical Planning and Inference 157–158 (2015) 108–120

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

On species sampling sequences induced by residual allocation models

Abel Rodríguez[a], Fernando A. Quintana[b,*]

[a] *Department of Applied Mathematics and Statistics, University of California, Santa Cruz, USA*

[b] *Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile*

ARTICLE INFO

Article history:
Received 30 December 2013
Received in revised form 28 July 2014
Accepted 27 August 2014
Available online 8 September 2014

Keywords:
Exchangeable partition probability function
Generalized Dirichlet process
Probit-stick breaking process
Size-biased permutation

ABSTRACT

We discuss fully Bayesian inference in a class of species sampling models that are induced by residual allocation (sometimes called stick-breaking) priors on almost surely discrete random measures. This class provides a generalization of the well-known Ewens sampling formula that allows for additional flexibility while retaining computational tractability. In particular, the procedure is used to derive the exchangeable predictive probability functions associated with the generalized Dirichlet process of Hjort (2000) and the probit stick-breaking prior of Chung and Dunson (2009) and Rodriguez and Dunson (2011). The procedure is illustrated with applications to genetics and nonparametric mixture modeling.

1. Introduction

An exchangeable sequence of random variables $X_1, X_2, \ldots$ defined on a probability space $(\mathcal{X}, \mathcal{B})$ follows a species sampling model (SSM) if its joint distribution can be characterized by a sequence of predictive rules where $X_1 \sim G_0$ and

$$p(X_{n+1} \in B | X_1, \ldots, X_n) = \sum_{k=1}^{K^n} q_k^n(\mathbf{m}^n)\delta_{\tilde{X}_k}(B) + q_{K^n+1}^n(\mathbf{m}^n)G_0(B),$$

for some non-atomic measure G_0 on $(\mathcal{X}, \mathcal{B})$ and for all $B \in \mathcal{B}$. In the previous expression, δ_a denotes the degenerate probability measure placing probability 1 on a, $\mathbf{m}^n = (m_1^n, \ldots, m_{K^n}^n)$ with $m_k^n = \sum_{j=1}^n I(X_j = \tilde{X}_k)$ being the number of values among $X_1, \ldots, X_n$ that are equal to $\tilde{X}_k$, $K^n = \max\{k : m_k^n > 0\}$ being the number of unique values among $X_1, \ldots, X_n$, and q_k^n for $k \le n$ and $n = 1, 2, \ldots$ being a collection of functions of $\mathbf{m}^n$ (usually called the predictive probability functions, or PPFs) which for all n and $\mathbf{m}^n$ satisfy

$$q_k^n(\mathbf{m}^n) \ge 0, \qquad \sum_{k=1}^{K^n+1} q_k^n(\mathbf{m}^n) = 1.$$

The name species sampling model comes from the application of this type of models in ecology and genetics (Fisher et al., 1943; Ewens, 1972; Engen, 1978). Indeed, we could think of sequentially sampling individuals in an infinite population with an unknown (and potentially infinite) number of species. When the first individual is sampled, its species is assigned

[*] Corresponding author. Tel.: +56 2 354 4464; fax: +56 2 354 7229.
E-mail addresses: abel@ams.ucsc.edu (A. Rodríguez), quintana@mat.uc.cl (F.A. Quintana).

http://dx.doi.org/10.1016/j.jspi.2014.08.008
0378-3758/© 2014 Elsevier B.V. All rights reserved.

a random tag generated according to G_0. Subsequently, individual n either belongs to one of the previously observed K^n species with respective probabilities $q_1^n, \dots, q_{K^n}^n$, or to a new species (which is again given a random tag according to G_0) with probability $q_{K^n+1}^n$.

A well-defined set of PPFs implies a symmetric joint probability distribution $p(m_1^n, \dots, m_{K^n}^n)$ for the number of species and the sample sizes associated with each one of them, which can be obtained through a recursion where $p(1) = 1$,

$$p(m_1^n, \dots, m_k^n + 1, \dots, n_{K^n}) = q_k^n(m_1^n, \dots, m_{K^n}^n)p(m_1^n, \dots, m_k^n, \dots, m_{K^n}^n)$$

for $k \le K^n$, and

$$p(m_1^n, \dots, m_{K^n}^n, 1) = q_{K^n+1}^n(m_1^n, \dots, m_{K^n}^n)p(m_1^n, \dots, m_{K^n}^n)$$

(for example, see Pitman, 1995 and Lee et al., 2013). The function p is often called the exchangeable partition probability function (EPPF). Alternatively, this can be written as

$$q_k^n(m_1^n, \dots, m_{K^n}^n) = \begin{cases} \dfrac{p(m_1^n, \dots, m_k^n + 1, \dots, n_{K^n})}{p(m_1^n, \dots, m_k^n, \dots, m_{K^n}^n)} & k \le K^n, \\ \dfrac{p(m_1^n, \dots, m_{K^n}^n, 1)}{p(m_1^n, \dots, m_{K^n}^n)} & k = K^n + 1. \end{cases} \tag{1}$$

In principle, constructing EPPFs is a difficult task, as ensuring that $X_1, X_2, \dots$ is an exchangeable sequence implies quite strict conditions on the PPFs (for a discussion, see for example Pitman, 1995, Pitman, 1996b and Lee et al., 2013). As a consequence, the number of species sampling models available in the literature is rather small, with the most popular ones arguably being those associated with the Dirichlet process (DP) (Ferguson, 1973; Blackwell and MacQueen, 1973), the two-parameter Poisson–Dirichlet process (PDP) (Pitman, 1995), and the normalized inverse Gaussian measures (NIGM) (Lijoi et al., 2005).

The previous list suggests that there is a close link between the class of nonparametric priors on almost-surely discrete distributions often used in nonparametric Bayesian modeling and the class of SSMs. Indeed, a well known result due to Pitman (1996b) establishes that the de Finetti measure of any SSM can be written as

$$G(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{X_k^*}(\cdot) + RG_0 \tag{2}$$

for some sequence of positive random variables $\omega_1, \omega_2, \dots$ and R such that $1 - R = \sum_{k=1}^{\infty} \omega_k \le 1$ almost surely, $X_1^*, X_2^*, \dots$ is a random sample from a non-atomic G_0, and the sequences $\omega_1, \omega_2, \dots$ and $X_1^*, X_2^*, \dots$ are independent. The resulting SSM is termed *proper* if $R = 0$ with probability 1. This connection can be exploited to generate novel SSMs, in particular, we study the SSM induced by a class of residual allocation models, as well as the special cases associated with the generalized Dirichlet process (GDP) of Hjort (2000), and the probit stick-breaking processes (PSBP) (Rodriguez et al., 2009; Chung and Dunson, 2009; Rodriguez and Dunson, 2011). In addition to model construction, we discuss computational issues associated with Bayesian estimation in this class of models.

The remainder of the paper is organized as follows. Section 2 discusses the EPPF for a class of species sampling models, specifically, independent residual allocation models, giving a general expression and analyzing some special well-known particular cases. Section 3 applies the results to the probit stick-breaking priors of Chung and Dunson (2009) and Rodriguez and Dunson (2011) and to the generalized Dirichlet process of Hjort (2000). Section 4 discusses Bayesian inference for the parameter controlling the allocation distribution and the probability of discovery of a new species for the class of models considered here. Section 5 illustrates our methods using both simulated and real datasets. Final comments are given in Section 6. Appendix contains proofs of some auxiliary results.

2. Exchangeable partition probability functions and predictive probability functions derived from residual allocation models

A random probability measure G defined on a probability space $(\mathcal{X}, \mathcal{B})$ is said to follow an independent residual allocation prior if it can be represented as

$$G(\cdot) = \sum_{k=1}^{N} \omega_k \delta_{X_k^*}, \tag{3}$$

where $N \in \mathbb{N}$, $X_1^*, X_2^*, \dots$ is a sequence of independent and identically distributed realizations from some distribution G_0 and $\omega_k = z_k \prod_{\ell<k}\{1 - z_\ell\}$ where $z_k \sim H_k^\theta$ independently for all $k = 1, 2, \dots$ (with the convention $z_N = 1$ if $N < \infty$), and H_k^θ is a probability distribution on $[0, 1]$ indexed by the vector of parameters θ. Two of the best-known examples of residual allocation priors are the Dirichlet process (where $N = \infty$, $\theta = b$, and $z_k \sim \mathsf{Beta}(1, b)$ for some $b > 0$) and the Poisson–Dirichlet process (for which $\theta = (a, b)$, $z_k \sim \mathsf{Beta}(1 - a, b + ka)$ and either $N = \infty$, $0 \le a < 1$ and