Statistics and Probability Letters 83 (2013) 1539–1546

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

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

On the Jeffreys prior for the multivariate Ewens distribution

Abel Rodríguez

University of California, Santa Cruz, Department of Applied Mathematics and Statistics, 1156 High Street, Mailstop SOE2, Santa Cruz, CA, 95064, United States

ARTICLE INFO

Article history:
Received 8 September 2012
Received in revised form 18 February 2013
Accepted 19 February 2013
Available online 28 February 2013

Keywords:
Multivariate Ewens distribution
Jeffreys prior
Dirichlet process
Ewens sampling formula

ABSTRACT

We derive the Jeffreys prior for the parameter of the Multivariate Ewens Distribution and study some of its properties. In particular, we show that this prior is proper and has no finite moments. We also investigate the impact of this prior on the a priori distribution of the number of species and the a priori probability of discovery of a new species, which are usually employed in subjective prior elicitation. The effect of the Jeffreys prior for posterior inference is illustrated using examples arising in the context of inference for species sampling models and Dirichlet process mixture models.

1. Introduction

The Multivariate Ewens Distribution (MED), also known as the Ewens Sampling Formula (ESF) (Ewens, 1972; Johnson et al., 1997), is a probability distribution on the partitions of the set $\{1, 2, \ldots, n\}$. It appears often in genetics as the distribution of the number of distinct alleles in a sample of size n drawn from an infinite idealized population, or as the limiting distribution of other, more general models (Kingman, 1978; Hoppe, 1987). More generally, the MED belongs to the class of species sampling models, which describe distributions for exchangeable random partitions (Aldous, 1985; Pitman, 1995; Lijoi et al., 2007). The MED also appears in the context of Bayesian nonparametric statistics, where it is related to the number of unique values in a random sample of size n taken from a random distribution that follows a Dirichlet process prior (Antoniak, 1974). Hence, in the context of Dirichlet process mixture models, the MED acts as the prior distribution on the number and size of the clusters imposed by the model.

This paper is concerned with Bayesian estimation and prediction under the MED model. Gamma priors, or mixtures thereof, are commonly used as priors in this context because of the existence of simple Gibbs sampling algorithms based on data augmentation (Escobar and West, 1995). Hyperparameters are elicited by either exploiting their link with the expected number of distinct alleles (Escobar and West, 1995) or, in the case of nonparametric mixture models, their link with the mean and variance of the observations (Walker and Mallick, 1997). Alternatively, Carota and Parmigiani (2002) and Griffin and Steel (2004) propose eliciting priors on the probabilities of new alleles, which in turn imply a prior on the parameter of interest. In either case, elicitation can be difficult because of the lack of relevant prior information in specific applications. To deal with the lack of prior information, numerous authors have used very dispersed Gamma priors. In this paper we derive the Jeffreys prior associated with the MED, show that this prior is proper, and investigate some of its properties. Because of its invariance to transformations, the Jeffreys prior provides a natural alternative to the priors discussed above without a substantial increase in computational complexity.

The remaining of the paper is organized as follows. Section 2 briefly reviews the Multivariate Ewens Distribution and derives the Jeffreys prior associated with its parameter. Section 3 discusses some of the properties of the Jeffreys prior. Section 4 presents some illustrations in the context of species sampling models and Dirichlet process mixture models. We conclude in Section 5 with some brief remarks and future directions.

E-mail address: abel@soe.ucsc.edu.

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.02.014

2. The Jeffreys prior for the multivariate Ewens distribution

Consider a partition of the set $\{1, 2, \dots, n\}$ into $K \le n$ subsets so that there are r_j subsets of size j, where $\sum_{j=1}^{n} r_j = K$, and $\sum_{j=1}^{n} jr_j = n$. For example, the n elements of the original set might represent individuals being sampled from an infinite population, while the subsets into which they are divided could be interpreted as the species to which these individuals belong. The Multivariate Ewens Distribution (MED) assigns such a partition a probability

$$p(K, r_1, \dots, r_n \mid \beta) = \frac{\Gamma(\beta)\Gamma(n+1)}{\Gamma(\beta+n)} \beta^K \prod_{j=1}^{n} \frac{1}{j^{r_j}\Gamma(r_j+1)}, \tag{1}$$

where $0 < \beta < \infty$ is a parameter controlling the shape of the distribution.

The partitions associated with the MED can alternatively be described in terms of a sequence of exchangeable indicators $\xi_1, \dots, \xi_n$ such that $\xi_i = k$ if the i individual in the population belongs to species k. Assuming that the species are labeled consecutively between 1 and K, and letting $m_k = \sum_{i=1}^{n} I(\xi_i = k)$ be the number of individuals in species k, then

$$p(\xi_1, \dots, \xi_n \mid \beta) = p(K, m_1, \dots, m_K \mid \beta) = \frac{\Gamma(\beta)}{\Gamma(\beta+n)} \beta^K \prod_{k=1}^{K} \Gamma(m_k). \tag{2}$$

From (2) we can compute the probability mass function associated with the species of a new individual

$$p(\xi_{n+1} = k \mid \xi_1, \dots, \xi_n, \beta) = \begin{cases} \dfrac{p(K, m_1, \dots, m_k+1, \dots, m_K \mid \beta)}{p(K, m_1, \dots, m_k, \dots, m_K \mid \beta)} = \dfrac{m_k}{\beta+n} & k \le K \\ \dfrac{p(K+1, m_1, \dots, m_k, \dots, m_K, 1 \mid \beta)}{p(K, m_1, \dots, m_k, \dots, m_K \mid \beta)} = \dfrac{\beta}{\beta+n} & k = K+1. \end{cases}$$

This sequence of predictive distributions is sometimes called the Chinese restaurant process.

We are interested in estimating the parameter β based on either an observed sample $\xi_1, \dots, \xi_n$, or the sufficient statistic K.

Lemma 1. *For $n \ge 2$, the Jeffreys prior associated with (1) and (2) is given by*

$$\pi_n^J(\beta) \propto \sqrt{\frac{1}{\beta} \sum_{j=1}^{n-1} \frac{j}{(\beta+j)^2}}. \tag{3}$$

Proof. By definition, $\pi_n^J(\beta) \propto |\mathcal{I}(\beta)|^{1/2}$ where $\mathcal{I}(\beta) = -\mathsf{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\beta^2} \log\{p(K, m_1, \dots, m_K \mid \beta)\}\right]$ is the Fisher information associated with β. Now,

$$-\mathsf{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\beta^2} \log\{p(K, m_1, \dots, m_K \mid \beta)\}\right] = -\psi'(\beta) + \psi'(\beta+n) + \frac{\mathsf{E}\{K\}}{\beta^2},$$

where ψ' denotes the trigamma function (Abramowitz and Stegun, 1965). Using the facts that $\mathsf{E}\{K\} = \sum_{j=0}^{n-1} \frac{\beta}{\beta+j}$ (e.g., see Antoniak, 1974) and $\psi'(\beta+n) = \psi'(\beta) - \sum_{j=0}^{n-1} \frac{1}{(\beta+j)^2}$ (e.g., see Abramowitz and Stegun, 1965) we get

$$\mathcal{I}_E(\beta) = -\sum_{j=0}^{n-1} \frac{1}{(\beta+j)^2} + \frac{1}{\beta^2} \sum_{j=0}^{n-1} \frac{\beta}{\beta+j} = \frac{1}{\beta} \sum_{j=0}^{n-1} \frac{j}{(\beta+j)^2} = \frac{1}{\beta} \sum_{j=1}^{n-1} \frac{j}{(\beta+j)^2},$$

which directly leads to (3). □

In particular, for $n = 2$ (the smallest sample size containing information about β), the Jeffreys prior on β corresponds to a standard Cauchy prior on $\nu = \beta^{1/2}$. Note that like other "non-informative" priors frequently used in Bayesian analysis (e.g., the Zellner–Siow prior for linear regression models, Zellner and Siow, 1980), the prior in (3) depends explicitly on the sample size n. Hence, any statistical procedure derived under this prior will depend on the stopping rule associated with the experiment.