Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis 50 (2006) 659 – 674

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

An improved collapsed Gibbs sampler for Dirichlet process mixing models

Lynn Kuo[a,*], Tae Young Yang[b]

[a]*Department of Statistics, University of Connecticut, 215 Glenbrook Road, Storrs 062694120, USA*
[b]*Department of Mathematics, Myongji University, Yongin, Kyonggi 449-728, Republic of Korea*

Received 2 December 2003; received in revised form 25 September 2004; accepted 28 September 2004
Available online 27 October 2004

Abstract

We study a nonparametric Bayesian formulation based on the Dirichlet process mixing models for the frequency counts in a two-way contingency table. The formulation provides a model averaging framework for a cluster analysis in the contingency table, because it specifies various partitions of the subjects according to their classification probabilities. We develop a new Gibbs sampler that improves upon the current collapsed Gibbs sampler by blocking and reducing the number of classification probabilities to be updated using the clustering configuration. The performance of the new Gibbs sampler is compared to the existing one. We apply the new method to two data sets; one is on testing for homogeneity in the contingency table, the other is on determining whether or not the residue frequency is independent of the position in a protein binding site from a data set of DNA sequences.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Cluster analysis; Contingency table; DNA sequences; Gibbs sampling; Homogeneity; Mixtures of Dirichlet processes

1. Introduction

Suppose we observe frequency counts n_{ij} in a two-way $k \times l$ contingency table. The row (count) vectors $\boldsymbol{n}_i = (n_{i1}, \ldots, n_{il})$ for $i = 1, \ldots, k$ are assumed to be independent; each has a multinomial distribution with sample size $n_{i+} = \sum_{j=1}^{l} n_{ij}$ and the classification probability $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{il})$. Note p_{ij} denotes the probability of a subject in the ith group

[*] Corresponding author. Tel.: +1 8604862951; fax: +1 8604864113.
E-mail addresses: lynn@stat.uconn.edu (L. Kuo), tyang@wh.myongji.ac.kr (T.Y. Yang).

doi:10.1016/j.csda.2004.09.013

being classified into the jth category. We have $\sum_{j=1}^{l} p_{ij} = 1$ for all i, where $p_{ij} > 0$ for each i and j.

To provide a nonparametric Bayesian formulation for the classification probabilities, we assume the probability vectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k$ are taken as a random sample from an unknown distribution G on the l dimensional simplex, where G follows a Dirichlet process (Ferguson, 1973) with two parameters G_0 and c suitably chosen or unknown but with a further hierarchical structure. Here G_0, the "location" parameter, represents our prior belief on the mean of the unknown measure G, and $c > 0$, a concentration parameter, represents our strength of this prior belief on G.

This nonparametric Bayesian formulation based on mixtures of Dirichlet processes (Antoniak, 1974) provides a natural framework for cluster analysis, because it specifies a discrete mixture distribution for all possible partitions of the k units with each cluster containing units with exactly the same classification probability. Note that we use the term "k units" instead of "k groups" to avoid confusion, because one may use the term a "group" to denote a cluster of units having the same classification probability.

This hierarchical formulation can be considered as a special case of the Bayes empirical Bayes problem developed by Lo (1984) and Kuo (1986a, b), where Dirichlet process mixing (DPM) over normal and binomial models are the main focus. This paper considers the Gibbs sampler for the DPM over multinomial models with data from contingency tables. Quintana (1998) develops a sequential importance sampling algorithm for the same problem. The Gibbs sampler developed in this paper can be applied to contingency tables with large dimensions, in addition to small and moderate. However, we expect the efficiency will decrease relative to the increase of the dimension.

Gibbs samplers for the DPM over various models including linear and generalized linear models have been developed in several papers. They include Doss (1994), Escobar (1994), MacEachern (1994, 1998), West et al. (1994), Escobar and West (1995, 1998), Bush and MacEachern (1996), Müller et al. (1996), Kuo and Mallick (1997), Mukhopadhyay and Gelfand (1997), MacEachern and Müller (1998), and Neal (2000). Ishwaran and James (2001, 2003) consider Gibbs samplers for a general class of stick-breaking priors that include the Dirichlet process prior. The sequential importance sampler for the DPM over the binomial model has been developed by Liu (1996) and MacEachern et al. (1999).

Tests of homogeneity in contingency tables refer to testing the hypothesis $\mathrm{H}_0 : \boldsymbol{p}_1 = \cdots = \boldsymbol{p}_k$ (all classification probabilities are equal componentwise) against the alternative that at least one unit has a different classification probability (at least one component is different). So we evaluate the posterior probability of the null hypothesis. If it is large, we would accept the null hypothesis. In our formulation, this translates into evaluating the posterior probability that there is only one cluster among all units.

To apply the Gibbs sampler (referred to as the standard) developed by Escobar (1994) and Escobar and West (1995) directly to our multinomial DPM model with a continuous distribution G_0, we will sample each parameter $\boldsymbol{p}_i$ sequentially from a discrete-continuous mixture distribution, where the discrete components share the same support as the other $\boldsymbol{p}$'s. This results in a "sticky" phenomenon in the Gibbs sampler where the values of the $\boldsymbol{p}$'s repeat a number of times without being updated. To avoid the stickiness phenomenon, we adopt the proposal in MacEachern (1994) where a collapsed Gibbs sampler for the DPM normal model was considered. In addition to changing models from normal to multinomial