Computational Statistics and Data Analysis 56 (2012) 4067–4080

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian inference through encompassing priors and importance sampling for a class of marginal models for categorical data

Francesco Bartolucci [a], Luisa Scaccia [b,*], Alessio Farcomeni [c]

[a] Department of Economics, Finance and Statistics, University of Perugia, Italy

[b] Department of Economic and Financial Institutions, University of Macerata, Italy

[c] Department of Public Health and Infectious Diseases, Sapienza - University of Rome, Italy

ARTICLE INFO

Article history:
Received 7 November 2011
Received in revised form 10 April 2012
Accepted 11 April 2012
Available online 21 April 2012

Keywords:
Bayes factor
Generalized logits
Inequality constraints
Marginal likelihood
Positive association

ABSTRACT

A Bayesian approach is developed for selecting the model that is most supported by the data within a class of marginal models for categorical variables, which are formulated through equality and/or inequality constraints on generalized logits (local, global, continuation, or reverse continuation), generalized log-odds ratios, and similar higher-order interactions. For each constrained model, the prior distribution of the model parameters is specified following the encompassing prior approach. Then, model selection is performed by using Bayes factors estimated through an importance sampling method. The approach is illustrated by three applications based on different datasets, which also include explanatory variables. In connection with one of these examples, a sensitivity analysis to the prior specification is also performed.

1. Introduction

Even if log-linear models are frequently used for the analysis of contingency tables, they do not allow to express, and consequently test, several hypotheses that are usually of interest, mainly because lower order interactions do not refer to the marginal distributions to which they seem to refer. This motivated McCullagh and Nelder (1989, Chapter 6) to introduce a class of models in which the joint distribution of a set of categorical variables is parametrized through the highest log-linear interaction within each possible marginal distribution. Several other models have been proposed following the original idea of McCullagh and Nelder (1989) (see Glonek and McCullagh, 1995; Glonek, 1996; Colombi and Forcina, 2001; Bergsma and Rudas, 2002; Bartolucci et al., 2007).

In this paper, we deal with a flexible class of models in which: (i) the parameters of the saturated model are given by *generalized* logits, in the sense of Douglas et al. (1990), for each univariate marginal distribution, generalized log-odds ratios for each bivariate marginal distribution, and similar interactions for each higher-order marginal distribution; (ii) any constrained model may be formulated through linear equality and inequality constraints on such parameters. In this way, we may express several hypotheses, which are of special interest in the presence of ordinal variables (see Bartolucci et al., 2001; Colombi and Forcina, 2001), as for instance, that: (i) the marginal distribution of one variable is stochastically larger than that of another variable, provided that these have the same categories; (ii) a certain type of positive association between a pair of variables holds; (iii) the marginal distribution of one variable is stochastically increasing with respect to the level of an explanatory variable.

* Correspondence to: Department of Economic and Financial Institutions, University of Macerata, Via Crescimbeni 20, 62100 Macerata, Italy. Tel.: +39 0733 258 3242; fax: +39 0733 258 3205.
E-mail address: scaccia@unimc.it (L. Scaccia).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.04.006

For the above class of models, we develop Bayesian inference and, in particular, a model selection strategy based on the Bayes factor (see Jeffreys, 1935, 1961; Kass and Raftery, 1995), which is defined as the ratio between the marginal likelihoods of two competing models. For the marginal models considered in this paper, the use of the Bayes factor, compared to classical strategies based on the likelihood ratio test, has different advantages. First of all, using the Bayes factor allows for an easy comparison of models parametrized through different types of logit, which would be otherwise cumbersome using a likelihood ratio test. Moreover, since the Bayes factor is computed as the ratio between two marginal likelihoods, the presence of nuisance parameters does not affect the inferential results. Notice that this is a relevant problem in categorical data analysis when studying the association between two variables and the problem also exists when testing for a certain type of positive association using a likelihood ratio test; for a discussion on this point, see Bartolucci et al. (2001) and Dardanoni and Forcina (1998). The distribution of this test statistic depends, in fact, on the nuisance parameters and a possible solution is the conditional approach, which, however, results in a multivariate generalized hypergeometric distribution for the test statistic. Such a distribution is almost intractable whenever the frequencies or the dimension of the table are moderately large, since computing the probability of observing a certain table requires enumerating all the possible tables with the same margins. The strategy based on Monte Carlo maximum likelihood (Bartolucci and Scaccia, 2004) for making inference on the parameters of the model under these constraints helps in overcoming the intractability problem but is still computationally intensive. On the other hand, a drawback of the Bayesian approach proposed here is the need to specify a prior distribution on the parameters, that does not exist within the likelihood ratio approach.

While the decision theoretic approach leads us to select the model with largest marginal likelihood, we can also use the Bayes factor as a measure of evidence. In order to assess this evidence we refer to the Jeffreys (1961) scale, which gives the following guideline: a log Bayes factor below 0.5 indicates *poor* evidence, between 0.5 and 1 *substantial*, between 1 and 2 *strong*, and *decisive* evidence is provided by a log Bayes factor larger than 2. See also Kass and Raftery (1995).

Bayesian methods for the analysis of categorical data have been dealt with by several authors. For instance, Albert (1996, 1997) used the Bayes factor to test hypotheses such as independence, quasi-independence, symmetry, or constant association in two-way and three-way contingency tables. Dellaportas and Forster (1999) proposed a general framework for selecting a log-linear model through the Reversible Jump algorithm of Green (1995) under a multivariate Normal prior distribution on the parameters. In practice, both Albert (1996, 1997) and Dellaportas and Forster (1999) dealt with log-linear models obtained by imposing some linear equality constraints on the parameters of the saturated model; a particular case is the constraint that a subset of the parameters is equal to zero. Klugkist and Hoijtink (2007), Hoijtink et al. (2008), Klugkist et al. (2005a,b, 2010), and Wetzels et al. (2010) used, instead, the Bayes factor to compare competing models expressed through linear inequality and *about equality* constraints on the saturated model. Under their *encompassing prior* approach, the Bayes factor between a constrained model and the encompassing model reduces to the ratio of the probability that the constraints hold under the encompassing posterior distribution and the probability that they hold under the encompassing prior distribution. By encompassing model we mean a model whose parameter space includes that of any other model under consideration. Therefore, once the prior distribution has been specified on the encompassing model parameters (encompassing prior), it is automatically specified for each submodel.

The selection strategy we adopt for the class of models considered in this paper is related to the approach of Klugkist et al. (2010). We exploit their encompassing prior approach, which leads to a logically coherent assessment of prior and posterior model probabilities and parameter distributions, as well as an easy estimation of the Bayes factors. However, our work differs from that of Klugkist et al. (2010) mainly in three respects: (i) we consider a more general class of models for categorical data; (ii) we propose an importance sampling method to improve the efficiency of the Bayes factor estimates for models with very small prior and, possibly, posterior probabilities; (iii) we introduce an iterative algorithm to estimate the Bayes factor for models specified through about equality constraints, which does not require to sample from a constrained model parameter space.

The paper is organized as follows. In Section 2 we describe the class of models of interest. In Section 3 we review the encompassing prior approach and we deal with Bayesian model selection. In Section 4 we illustrate the proposed approach through three applications involving some datasets of interest in the categorical data analysis literature. Finally, we conclude with a brief discussion which is provided in Section 5.

2. Marginal models for categorical variables

In this section, we introduce the class of marginal models developed by McCullagh and Nelder (1989) and illustrate the parametrization based on generalized logits and log-odds ratios; for an overview see also Bergsma et al. (2009). Then, we show how hypotheses of interest may be expressed through linear equality and inequality constraints imposed on the parameters of the saturated model.

2.1. Preliminaries

Let $\boldsymbol{A} = (A_1, \dots, A_q)$ be a vector of q categorical variables and $\{1, \dots, m_i\}$ be the support of A_i, $i = 1, \dots, q$. Also let $r = \prod_i m_i$ be the number of possible configurations of $\boldsymbol{A}$ and let $\boldsymbol{\pi}$ be the r-dimensional column vector of the joint probabilities $\pi_{\boldsymbol{a}} = p(\boldsymbol{A} = \boldsymbol{a})$ arranged in lexicographical order. Suppose, for instance, that there are two categorical variables