Computational Statistics and Data Analysis 72 (2014) 147–157

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian test on equality of score parameters in the order restricted RC association model

Man-Suk Oh *

Department of Statistics, Ewha Womans University, Seo-dae-moon Gu, Seoul 120-750, Republic of Korea

ARTICLE INFO

Article history:
Received 16 November 2012
Received in revised form 24 September 2013
Accepted 12 November 2013
Available online 20 November 2013

Keywords:
Contingency tables
Model selection
Inequality constraints
Ordinal categorical variables
Markov chain Monte Carlo

ABSTRACT

In the RC association model for a two-way contingency table, it is often natural to impose order constraints on the score parameters of the row and column variables. In this article, a simple and efficient Bayesian model selection procedure is proposed that simultaneously compares all possible combinations of (in)equalities of successive score parameters in the order restricted RC association model. The method introduces normal latent variables into the model and uses a simple and accurate approximation to the likelihood function so that the full conditional posterior distributions of elements of the parameter are given as truncated normal distributions. The Gibbs sampling algorithm of Gelfand and Smith (1990) is employed to generate samples from the full model in which all the scores are strictly ordered, and posterior probabilities of all possible models are estimated by using the Gibbs output from the full model. A simulation study shows that the proposed method performs well in detecting the true model. Two real data sets are analyzed using the proposed method.

1. Introduction

In many application areas such as behavioral science, social science, and biomedical science, data are given as a two-way contingency table with ordered categories for both column and row variables (Clogg and Shidaheh, 1994; Agresti, 1999, 2001). A wide variety of association models have been introduced by many researchers, which reflect the ordinal nature of the variables (Goodman, 1979, 1981; Davis, 1988; Agresti, 2001). These models are more parsimonious and suitable in cases where the variables are ordinal.

The most commonly used association models for two-way contingency tables with ordinal row and column variables are Goodman's models. For the row and the column variables, scores μ_i's and ν_j's are assigned respectively and the interaction term is given by the product $\phi\mu_i\nu_j$ where ϕ is an unknown constant. When the scores are fixed for categories for both row and column variables, the model is called a Uniform association model or a Linear-by-Linear association model. When the scores of the column (row) variables are fixed but the scores of the row (column) variables are unknown, the model is called a Row (Column) effect association model. When both column and row scores are unknown, the model is called an RC association model.

In recent years statistical inference on RC and Row (Column) effect association models with score parameters subject to order constraints has received substantial interest since inference incorporating the order constraints yields reasonable score estimates and may improve power in testing. In particular, hypothesis test on the equality of successive scores under the order constraints attracts the interest of researchers and practitioners, since if some of the successive column or row scores are equal, the patterns of interaction in the table could be simplified (Kateri and Illiopoulos, 2003).

* Tel.: +82 2 3277 2374; fax: +82 2 3277 3607.
E-mail address: msoh@ewha.ac.kr.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.11.009

Estimation methods which yield estimated scores comprised with the given order restrictions have been proposed by Agresti et al. (1987), Ritov and Gilula (1991), and Galindo-Garre and Vermunt (2004, 2005). Bayesian inference on the association models with order restrictions are given by Iliopoulos et al. (2007), Hoijtink et al. (2008), Tarantola et al. (2008), Klugkist et al. (2010), and Iliopoulos et al. (2009).

In comparing models subject to restrictions on parameters, Bayesian approaches are preferred over classical approaches in many settings. The benefits of the Bayesian approach are (i) it does not rely on often complicated asymptotic distributions of test statistics, (ii) it can easily incorporate the given restrictions on parameters, (iii) it can compare all candidate models simultaneously, and (iv) it naturally gives penalties for more complex models so that it compromises between goodness-of-fit and model interpretability.

Recently Iliopoulos et al. (2009) proposed a Bayesian model comparison for the order restricted RC association model. They used a reversible jump Markov chain Monte Carlo (MCMC) to explore the model space and estimated the posterior model probabilities by the relative frequencies of visits to the models. The Metropolis–Hastings algorithm is used to generate samples of parameters in each model and indicators of equality/inequality of successive scores.

In this paper, we propose a simple and efficient Bayesian model selection procedure for the order restricted RC models. We introduce normal latent variables into the model and use a simple and accurate approximation to the Poisson distribution function, so that all the full conditional posterior distributions of elements of parameters are given as truncated normal distributions. The Gibbs sampling algorithm of Gelfand and Smith (1990) is employed to generate posterior samples from the full model in which all the scores are strictly ordered.

The Bayes factor of a candidate model with at least one equality of successive scores against the full model is represented as the ratio of the posterior and prior marginal density functions of a subset of the score parameters, from the Savage–Dickey density ratio (Dickey and Lieutz, 1970; Dickey, 1971, 1976). The marginal posterior densities are then estimated by using the full conditional posterior densities of the score parameters and the posterior samples from the full model, by adopting the method of Oh (1999). Given the Bayes factors of all candidate models, the posterior probabilities of all possible models can be obtained.

A key idea of the proposed method is derivation of the full conditional posterior density function of each score parameter in a closed form, with the aid of latent variables and an approximation to the likelihood. Given the convenient full conditional posterior distributions, sample generation and estimation of the Bayes factors can be conducted in an *automated* way. Another key feature of the proposed method is that it computes the Bayes factors, hence the posterior probabilities, of all possible models by using only one set of samples from the full model. Since it fits *only the full model*, the computational cost can be smaller in terms of algorithmic complexity and computation time, compared to methods which need to fit all candidate models, particularly when the number of candidate models is large.

This paper is organized as follows. The order restricted RC model is presented in Section 2. In Section 3, latent variables and an approximation to the likelihood function are described. Full conditional posterior distributions are derived in Section 4 and estimation of the Bayes factors is presented in Section 5. Results of a simulation study are given in Section 6. Two real data sets are analyzed by using the proposed method in Section 7. Conclusions are given in Section 8.

2. Model

Let $N_{ij},\ i = 1, \dots, r,\ j = 1, \dots, c$, be frequencies in an $r \times c$ contingency table of two ordinal variables. Assume that N_{ij} is obtained from a Poisson distribution with mean θ_{ij}, Poi(θ_{ij}), independently for each i and j. Assign parametric scores $(\mu_1, \dots, \mu_r)$ to the row variable X and $(\nu_1, \dots, \nu_c)$ to the column variable Y, and assume the multiplicative row–column (RC) model

$$\lambda_{ij} = \log \theta_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \phi\mu_i\nu_j, \quad i = 1, \dots, r,\ j = 1, \dots, c. \tag{1}$$

The parameter ϕ is a global measure of association and the adjacent category log odds ratio is given by

$$\log\left(\frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}\right) = \phi(\mu_{i+1} - \mu_i)(\nu_{j+1} - \nu_j)$$

(Goodman, 1979, 1981).

For identifiability of the parameters, certain constraints have to be imposed on them. The most commonly used constraints are sum-to-zero (STZ) and sum of squares to one (SSTO) constraints, given as

$$\sum_{i=1}^r \lambda_i^X = \sum_{j=1}^c \lambda_j^Y = \sum_{i=1}^r \mu_i = \sum_{j=1}^c \nu_j = 0,$$
$$\sum_{i=1}^r \mu_i^2 = \sum_{j=1}^c \nu_j^2 = 1.$$

However, STZ and SSTO constraints are difficult to implement in the MCMC procedure. See Kateri et al. (2005) and Iliopoulos et al. (2009) for more details. Thus, following Iliopoulos et al. (2009), we use the following identifiability constraints in this