www.istatistikciler.org
İstatistikçiler Dergisi 2 (2009) 82-88

İstatistikçiler Dergisi

A Bayesian model averaging approach to the dumping severity data over hierarchical log-linear models

Haydar Demirhan
Hacettepe University
Department of Statistics
06800-Beytepe, Ankara, Türkiye
haydarde@hacettepe.edu.tr

Canan Hamurkaroğlu
Hacettepe University
Department of Statistics
06800-Beytep, Ankara, Türkiye
caca@hacettepe.edu.tr

Abstract

In this article, we consider a Bayesian model averaging approach for hierarchical log-linear models to analyse the "dumping severity" dataset from the Bayesian perspective. A wide range of log-linear models fits this dataset well when analyzed with classical approaches. However, only one log-linear model is determined to fit the dataset well by the Bayesian model averaging approach. Therefore, use of the BMA approach for this dataset is found to be more advantageous than classical model selection approaches.

Keywords: Gibbs sampling; Markov Chain Monte Carlo; Model selection; Log-linear modelling; Prior distribution; Full conditional posterior.

Özet

Ağrı şiddeti veri kümesi için hiyerarşik log-doğrusal modeller üzerinden bir Bayesci model ortalama yaklaşımı

Çalışmada hiyerarşik log-doğrusal modeller için Bayesci model ortalama yaklaşımı kullanılarak "ağrı şiddeti" veri kümesi Bayesci yaklaşım ile çözümlenmiştir. Veri kümesinin özelliği, birçok log-doğrusal modelin bu veri kümesi için uygun bulunmasıdır. Uygun bulunan modellerden hangisinin bu veri kümesi için en iyi model olduğu tartışmalıdır. Bayesci model ortalama yaklaşımı ile bu veri kümesi için sadece bir model uygun model olarak bulunmuştur. Bu nedenle, Bayesci model ortalama yaklaşımının bu veri kümesi için kullanımının klasik yaklaşımların kullanılmından daha üstün olduğu söylenebilir.

Anahtar sözcükler: *Gibbs öreklemesi algoritması; Markov zinciri Monte Carlo; Model seçimi; Log-doğrusal modelleme; Önsel dağılım; Tam koşullu sonsal dağılım.*

1. Introduction

Categorical data is collected and analyzed in many fields of scientific investigation. Collected categorical data includes an association structure. Appropriate exploration of this association structure is one of the major concerns of categorical data analysis. Log-linear models are widely used for this aim. Each log-linear model corresponds to an association structure and parameters of it give detailed information on the variables. Choosing the log-linear model that fits best the data is an important step in the log-linear analysis of categorical data. There are a huge number of approaches for model selection in the classical and Bayesian settings. Some of the methods are based on a series of significance tests and some of them include prior information, use Markov chain Monte Carlo methods (MCMC) and some are based on Bayes factors. Almost each method has its own difficulties or problems. The most important problem is on the inclusion of the model uncertainty in the model selection process. If a particular model is selected and inferences are conditionally based on the selected model, model uncertainty is ignored. When the model uncertainty is ignored, drawn inferences are valid only if the assumed model is exactly the correct model. Underestimation due to the model uncertainty can lead too risky false decisions [10]. This case is especially seen in the classical setting.

This difficulty can be overcome by using Bayesian approaches to the model selection. Bayesian model averaging (BMA) is one of these approaches [13, 17]. BMA is used to determine a particular model for model determination and parameter estimation, simultaneously. In the BMA approach, posterior distribution of considered quantity is obtained over the set of suitable models, and then they are weighted by their posterior model probabilities. Model uncertainty is included in the analysis by this way. The most common way of including the information, provided by different sources, is to use the average. From the Bayesian point of view, this averaging is applied in BMA such that posterior distribution of the quantity of interested is obtained over the set of suitable models, then they are weighted by their posterior model probabilities [13, 17].

Roberts [20] extended the model averaging idea of Leamer [13]. However, computational difficulties were a handicap for progress of model averaging idea. Draper [8] and Raftery [18] review the Bayesian model averaging and the cost of ignoring model uncertainty. Madigan and Raftery [16] give Occam's razor and Occam's window approaches to reduce the number of candidate models. Hoeting et al. [11] give a good tutorial for BMA. Demirhan and Hamurkaroglu [6] consider a BMA approach for hierarchical log-linear models, propose an approach for the calculation of posterior model probabilities, and apply the approach to a traffic accidents dataset.

In this article, we analyze a well known dumping severity dataset, which is given by Grizzle et al. [9], by using the approach given by Demirhan and Hamurkaroglu [6]. Results of this analysis were presented at the Applied Statistics 2009 Conference [7]. A wide range of log-linear models including simple independence model fit this dataset well, when analyzed with classical approaches [1]. However, only one log-linear model is determined to fit the dataset well by the BMA approach. Therefore, use of the BMA approach for this dataset is found to be more advantageous than classical model selection approaches.

Log-linear model notation and hierarchy principle are mentioned in Section 2. BMA for hierarchical log-linear models is outlined in Section 3. The analysis of "Dumping Severity" data of Grizzle et al. [9] by using BMA approach is given in Section 4. Conclusions are given in Section 5.

2. Notation and hierarchical log-linear models

Number of terms of a log-linear model increases by the increase in the number of categorical variables and standard notations become cumbersome. Instead, King and Brooks [12] give very flexible and practical notations, which are also used in this work.

Set of sources (variables constituting the contingency table of interest), where the data come from, is denoted by S. Number of elements of a set is denoted by $|\cdot|$, so each source is labelled such that $\mathrm{S} = \{\mathrm{S}_\lambda : \lambda = 1,\dots,|\mathrm{S}|\}$. Set of levels for source S_i is K_i, for $i = 1,\dots,|\mathrm{S}|$. Cells of a contingency table are represented by the set $\mathrm{K} = \mathrm{K}_1 \times \cdots \times \mathrm{K}_{|\mathrm{S}|}$, so the cells are indexed by $\mathbf{k} \in \mathrm{K}$. Expected cell counts and observed cell counts are denoted by $n_{\mathbf{k}}$ and $y_{\mathbf{k}}$ for $\mathbf{k} \in \mathrm{K}$, respectively. The set of subsets of S is defined by $\mathcal{P}(\mathrm{S}) = \{s : s \subseteq \mathrm{S}\}$. Then $m \subseteq \mathcal{P}(\mathrm{S})$ is used to represent a log-linear model, where m lists the log-linear terms presented in the model. Each element of the model, m is included in a set c such that $c \in \mathcal{P}(\mathrm{S})$. Constant term of the log-linear model is represented by $\varnothing \in \mathcal{P}(\mathrm{S})$. The set $\mathbf{M}^c = \left\{\mathbf{m}_1^c,\dots,\mathbf{m}_{|\mathbf{M}^c|}^c\right\}$ contains all possible combinations of the levels of sources included in c. In general, the highest level is not included by the elements of $\mathbf{M}^c$. The log-linear model vector for each $c \in \mathcal{P}(S)$ is $\boldsymbol{\beta}^c = \left(\beta^c_{\mathbf{m}_1}, \ldots, \beta^c_{\mathbf{m}_{|\mathbf{M}^c|}} \right)$. Thus the log-linear parameter vector for the model m is $\boldsymbol{\beta}_m = \left(\boldsymbol{\beta}^{\mathbf{c}_1}, \ldots, \boldsymbol{\beta}^{\mathbf{c}_{|m|}} \right)$. Design matrix or model matrix corresponding to the model $m \subseteq \mathcal{P}(S)$ is denoted by $\mathbf{X}_m$, columns of which include -1, 0, and 1 values when all variables are nominal. If there is an ordinal variable $\mathbf{X}_m$ includes -1, 0, 1 and score values. Score values are represented by $x^{c_i}_{m_j}$, $i = 1, \ldots, |m|$ and $j = 1, \ldots, |\mathbf{M}^{c_i}|$, where c_i corresponds to an ordinal variable. Using this design matrix and the parameter vector, the log-linear model is represented as follows:

$$\log \mathbf{n} = \mathbf{X}_m \boldsymbol{\beta}_m$$

More detailed notations for the elements of design matrix, order of parameters and cells, and examples are given in King and Brooks [12] and Demirhan [3, 4].

The family of hierarchical models is such that if any $\boldsymbol{\beta}^c$ term is not included in the model then all of its higher relatives must not be included in the model, and all of its lower order relatives must be in the model at the same time [2]. Hierarchy principle helps us to decrease the number of considered models.

3. Bayesian model averaging for hierarchical log-linear models

The probabilistic definition of BMA is given as follows. Let the quantity of interest, which would be a group of parameter, effect size, odds ratio, etc., be Δ, and D be data then

$$P(\Delta|D) = \sum_{m \subseteq \mathcal{P}(S)} P(\Delta|m, D) \frac{P(D|m)P(m)}{\sum_{m \subseteq \mathcal{P}(S)} P(D|m)P(m)} \tag{1}$$

where

$$P(D|m) = \int P(D|\boldsymbol{\beta}_m, m) P(\boldsymbol{\beta}_m|m) d\boldsymbol{\beta}_m . \tag{2}$$

Posterior mean of Δ is as follows:

$$E(\Delta|D) = \sum_{m \subseteq \mathcal{P}(S)} \hat{\Delta}_m P(m|D),$$

where $\hat{\Delta}_m = E(\Delta|m, D)$.

Although the application of BMA is simple, BMA has some difficulties, which makes the BMA unpopular. Dimension of the model space can be enormous, which prevents considering whole model space. The integral in (2) is in general hard to compute. Specification of $\mathcal{P}(S)$ is not clear, especially if there is prior information on some of the models of $\mathcal{P}(S)$. Choice of the model class over which to average is also problematic [8, 11, 16].

To reduce the dimension of model space, Occam's window and Occam's Razor principles of Madigan and Raftery [16] is used. A subset of whole model space, which is found by Occam's window, has the property that it includes models that are not predict data far less well than the best model. Complex models, which receive less support from the data than their counterparts are excluded by the Occam's Razor approach.

Number of log-linear models grows rapidly by the increase of the number of categorical variables. Occam's Window approach of Madigan and Raftery [16] provides a solution to this situation. Adapted form of the Occam's Window approach for hierarchical log-linear (HLL) models is given by Demirhan and Hamurkaroglu [6].

Another difficulty of the BMA, which is also valid for HLL models, is the computation of the integral (2). The integral is calculated by Demirhan and Hamurkaroglu [6], and $P(D|m)$ for HLL models is obtained as follows:

$$P(D|m) = \frac{\sqrt{\det(\mathbf{A})}(2\pi)^{-\dim(\boldsymbol{\beta}_m)/2}}{\sqrt{\det(\mathbf{V}_{\mathbf{b}_m})\det(\boldsymbol{\Sigma}_m)}} \exp\left\{-\frac{1}{2}\left[\mathbf{b}_m^T \mathbf{V}_{\mathbf{b}_m}^{-1}\mathbf{b}_m + \boldsymbol{\mu}_m^T \boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m - \mathbf{z}^T \mathbf{A}\mathbf{z}\right]\right\},$$

where $\mathbf{b}_m$ is the vector of maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}_m$, $\dim(\cdot)$ is the dimension of inner vector, $\mathbf{V}_{\mathbf{b}_m}^{-1}$ is the inverse of covariance matrix of MLEs, $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m^{-1}$ are prior mean vector and the inverse of prior covariance matrix of $\boldsymbol{\beta}_m$, respectively, and $\mathbf{A} = \mathbf{V}_{\mathbf{b}_m}^{-1} + \boldsymbol{\Sigma}_m^{-1}$ and $\mathbf{A}\mathbf{z} = \mathbf{V}_{\mathbf{b}_m}^{-1}\mathbf{b}_m + \boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m$.

To represent our degree of belief in the prior information, a prior distribution for $\boldsymbol{\Sigma}_m$ is specified in two stages. In the first stage, covariance matrix of the prior distribution is taken as, $\boldsymbol{\Sigma}_m = \alpha\mathbf{C}_m = \alpha c\mathbf{I}_m$, where $\mathbf{I}_m$ is the identity matrix dimension of $\dim(\boldsymbol{\beta}_m)$, and $c = \dim(\boldsymbol{\beta}_m)/\mathrm{tr}(\mathbf{V}_{\mathbf{b}_m}^{-1})$ [13]. The distribution of the general precision parameter α is given by the second stage prior. It is taken as $\tau = 1/(1+\alpha)$ and $\tau \sim \mathrm{Uniform}(0,1)$ to make calculations easier. Values of τ represent the degree of our belief in prior. Leonard [15] and Leighty and Johnson [14] state that values of this precision parameter close to zero represent disbelief.

To complete the BMA procedure for HLL models, posterior distribution and posterior estimates of the log-linear parameters, given the data and model, should be obtained. For this purpose, Gibbs sampling algorithm is employed. Implementation of Gibbs sampling algorithm with these prior and likelihood settings requires full conditional distribution of each log-linear parameter given the rest of parameters, model and the data. These full conditionals are given by Demirhan and Hamurkaroglu [5]. The same procedure is followed here.

4. Analysis of dumping severity data

Variables that are included in the Dumping Severity dataset are operation (OP), hospital (HO) and dumping severity (DS). When analyzed with the assumption that all variables are nominal, even simple independence model can be found as suitable for this dataset. In addition to the independence model, various partial association models are found to be suitable [1].

Operation levels are drainage and vagotomy, 25% resection and vagotomy, 50% resection and vagotomy, and 75% resection. There are four hospitals numbered from 1 to 4. Levels of dumping severity are none, slight and moderate. Operation and dumping severity are treated as ordinal and hospital is treated as nominal variables.

Following the notation given in Section 2, $|S| = 3$, S_1 is OP, S_2 is HO, and S_3 is DS. Thus, $K_1 = K_2 = \{1,2,3,4\}$, $K_3 = \{1,2,3\}$, $K = \{(1,1,1),(1,1,2),(1,1,3),\ldots,(4,4,3)\}$, and $\mathcal{P}(S) = \{\varnothing,\{S_1\},\{S_2\},\{S_3\},\{S_1,S_2\},\{S_1,S_3\},\{S_2,S_3\},\{S_1,S_2,S_3\}\}$.

Log-linear parameter vector for the largest model is $\boldsymbol{\beta}_m = (\boldsymbol{\beta}^{c_1},\ldots,\boldsymbol{\beta}^{c_8})$. Here the largest model is not saturated due to the ordinal variables. Interactions between OP and HO, and DS and HO are represented by row effect parameters. $x^{c_2}_{m_j} = j$ for $j = 1,\ldots,4$ and $x^{c_4}_{m_i} = i$, $i = 1,2,3$. Interaction between OP and DS is represented by a homogeneous association parameter, namely $\beta^{c_5}_1$.

After these definitions, BMA is applied using Occam's Window approach. For this approach O_R and O_L are taken as 20 and 10^{-5}, respectively. To obtain $\mathbf{V}_{\mathbf{b}_m}$ and a classical estimate of the log-linear parameter vector, $\mathbf{b}_m$, the Newton-Raphson algorithm is used. Prior model probability for each model in the model space is taken as 1/18. Prior distributions of log-linear parameters are defined by using the approach of Leigthy and Johnson [14], τ of which is taken as 10^{-6}. This setting implies a noninformative prior for the log-linear parameters. Gibbs sampling is employed to obtain posterior estimate of each model parameter for a given model. Total number of iterations was 50000, 10000 of which were discarded as burn-in. To reduce the autocorrelation of the Gibbs sequence, a record has been made at end of each 200 cycles. Full conditional posterior distributions of Demirhan and Hamurkaroglu [5] are used in the Gibbs sampling algorithm.

After the application of Occam's Window, the whole model space was reduced to 8 models from 19. Therefore Occam's Razor approach was not applied. Elements of possible models and corresponding posterior model probabilities are given in Table 1.

Table 1. Elements of possible models and corresponding posterior model probabilities.

Model	Posterior model probability
$m_1 = \{\varnothing,\{S_1\},\{S_3\},\{S_1,S_3\}\}$	0.99
$m_2 = \{\varnothing,\{S_1\},\{S_2\},\{S_3\},\{S_1,S_3\}\}$	8.9 10^{-7}
$m_3 = \{\varnothing,\{S_1\},\{S_2\},\{S_3\},\{S_1,S_2\},\{S_1,S_3\}\}$	3.5 10^{-12}
$m_4 = \{\varnothing,\{S_1\},\{S_2\},\{S_3\},\{S_1,S_3\},\{S_2,S_3\}\}$	9.9 10^{-13}
$m_4 = \{\varnothing,\{S_1\},\{S_2\},\{S_3\},\{S_1,S_2\},\{S_1,S_3\},\{S_2,S_3\}\}$	4.2 10^{-18}

Obtained posterior model probability for m_1 is 0.99, therefore, it is the most appropriate model for the data, and it is not necessary to go on with the application of the Occam's razor. According to m_1, there is a homogeneous association between operation and dumping severity over the levels of these variables.

Δ of Eq. (1) is taken as $\boldsymbol{\beta}_{m_1}$ and Bayesian estimates of model parameters $(\tilde{\boldsymbol{\beta}}_{m_1})$ are obtained. $\tilde{\boldsymbol{\beta}}_{m_2},\ldots,\tilde{\boldsymbol{\beta}}_{m_5}$ are not obtained due to the very low posterior model probabilities. Results, obtained over the Gibbs sampling, are given in Table 2 for m_1. In Table 2, $\beta^{c_2}_i$ for $i = 1,2,3$ and $\beta^{c_4}_i$ for $i = 1,2$ correspond to main effects of OP, HO and DS, respectively. $\beta^{c_6}_1$ is the homogeneous association parameter between OP and DS. There is a positive and moderate homogeneous association between levels of OP and DS. This is a reasonable inference for this dataset. Levels of main effects of OP and DS are positive.

Table 2. Bayesian parameter estimates of model parameters.

Parameter	Estimate	Parameter	Estimate
β^{c_1}	0.024	$\beta_1^{c_4}$	-0.039
$\beta_1^{c_2}$	0.081	$\beta_2^{c_4}$	-0.218
$\beta_2^{c_2}$	0.070	$\beta_1^{c_6}$	0.225
$\beta_3^{c_2}$	0.152		

According to odds ratio obtained using the homogeneous association parameter $\widetilde{\beta}_1^{c_5}$, namely $\theta^{c_6} = \exp\left(\widetilde{\beta}_1^{c_6}\right) = 1.303$, the odds of one level increment in dumping severity is 1.303 times the odds of having an operation $i+1$ instead of operation i, for $i = 1,2,3$.

4. Conclusions

Agresti [1] also analyses the dumping severity data set from the classical point of view. When all the variables treated as nominal $\{\varnothing,\{s_1\},\{s_2\},\{s_3\},\{s_1,s_3\}\}$ model is obtained as the best fitting model [1]. However, in this case ordinality of operation and dumping severity variables are ignored. When operation and dumping severity are treated as ordinal and hospital is treated as nominal, the model $\{\varnothing,\{s_1\},\{s_2\},\{s_3\},\{s_1,s_2\},\{s_1,s_3\}\}$ is obtained as the best among other appropriate models [1]. Maximum likelihood estimate of the association parameter of the latter model is obtained as 0.163 by Agresti [1]. It is not suitable to compare the results of the Bayesian and classical analyses, but the inferences can be compared. We estimate the association parameter as 0.225 obtain the best fitting model as $\{\varnothing,\{s_1\},\{s_3\},\{s_1,s_3\}\}$ by BMA. We expect the more the amount of resection the more the severity of dumping and existence of an association between operation and dumping severity. The inferences obtained by the classical and BMA approaches are both compatible with our expectations. The difference between Bayesian and classical inference is the inclusion of main effect of hospital and interaction between operation and hospital. When evaluated in their own theoretical basis, inferences obtained from BMA are stronger then those obtained from classical approach due to the very high posterior model probability of the identified model. Also, having no interaction between operation and hospital, and hospital and dumping severity is reasonable. Therefore, it can be concluded that the model giving the best fit to the dumping severity data is $\{\varnothing,\{s_1\},\{s_3\},\{s_1,s_3\}\}$ and BMA is able to determine only one possible and a more parsimonious model for the dumping severity dataset.

References

[1] Agresti, A., (1984), *Analysis of Ordinal Categorical Data*, Wiley, New York.

[2] Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., (1975), *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, London.

[3] Demirhan, H., (2004), *Bayesian Estimation of The Parameters and Expected Cell Counts in Logarithmic Linear Models*, Unpublished M.Sc. thesis, Hacettepe University, Institute of Natural Sciences, Ankara.

[4] Demirhan, H., (2009), *Bayesian Approaches to the Mixed Logarithmic Linear Models*, Unpublished Ph.D. thesis, Hacettepe University, Institute of Natural Sciences, Ankara.

[5] Demirhan, H., Hamurkaroglu, C., (2006), Specification of hyper-parmeters for normal prior distributions induced on log-linear parameters, *Hacettepe Journal of Mathematics and Statistics*, 35, 1, 91-102.

[6] Demirhan, H., Hamurkaroglu, C., (2009), An Application of Bayesian Model Averaging Approach to Traffic Accidents Data Over Hierarchical Log-Linear Models, *Journal of Data Science*, 7, 497-511.

[7] Demirhan, H., Hamurkaroglu, C., (2009), Bayesian Model Averaging for Hierarchical Log-Linear Models and an Application to the Dumping Severity Data, *Abstracts Book of Applied Statistics 2009 Conference*, 42.

[8] Draper D., (1995), Assessment and propagation of model uncertainty, *Journal of the Royal Statistical Society*, Ser. B, 57, 45-97.

[9] Grizzle, J.E., Starmer, C.F., Koch, G.G., (1969), Analysis of categorical data by linear models, *Biometrics*, 25, 489-504.

[10] Hodges, J.S., (1987), Uncertainty, policy analysis and statistics, *Statistical Science*, 2, 259-291.

[11] Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., (1999), Bayesian model averaging: A tutorial (with discussion), *Statistical Science*, 14, 4, 382-417.

[12] King, R., Brooks, S.P., (2001), Prior induction in log-linear models for general contingency table analysis, *The Annals of Statistics*, 29, 3, 715-747. 7.

[13] Leamer, E.E., (1978), *Specification Searches: Ad Hoc Inference with Nonexperimetal Data*, Wiley, New York.

[14] Leighty, R.M., Johnson, W.J., (1990), A Bayesian log-linear model analysis of categorical data, *Journal of Official Statistics*, 6, 2, 133-155.

[15] Lenonard, T., (1975), Bayesian estimation methods for two-way contingency tables, *Journal of Royal Statistical Society*, Ser. B, 37, 23-37.

[16] Madigan, D., Raftery, A.E., (1994), Model selection and accounting for model uncertainty in graphical models using Occam's window, *Journal of the American Statistical Association*, 89, 1535-1546.

[17] Raftery, A.E., (1995), Bayesian modelselection in social research (with discussion), *Social Methodology*, 111-195.

[18] Raftery, A.E., (1996), Approximate Bayes factors and accounting for model uncertainty in generalised linear models, *Biometrika*, 83, 2, 251-266.

[19] Raftery, A.E., Madigan, D., Volinsky, C.T., (1996), Accounting for model uncertainty in survival analysis improves predictive performance (with discussion)}, *In Bayesian Statistics 5*, J. Bernardo, J. Berger, A. Dawid and A. Smith, (eds.), Oxford Univ. Press, 323-329.

[20] Roberts, H.V., (1996), Probablistic prediction, *Journal of the American Statistical Association*, 60, 50-62.